Statistics for Biology and Health

David G. Kleinbaum
Mitchel Klein

Survival Analysis

A Self-Learning Text

Third Edition

Springer

Statistics for Biology and Health

Series Editors
M. Gail, K. Krickeberg, J.M. Samet, A. Tsiatis, W. Wong

For further volumes:
http://www.springer.com/series/2848

David G. Kleinbaum
Mitchel Klein

Survival Analysis

A Self-Learning Text

Third Edition

Springer

David G. Kleinbaum
Department of Epidemiology
Rollins School of Public Health at
Emory University
1518 Clifton Road NE
Atlanta, GA 30322
USA
dkleinb@sph.emory.edu

Mitchel Klein
Department of Epidemiology
Rollins School of Public Health at
Emory University
1518 Clifton Road NE
Atlanta, GA 30322
USA
mklein@emory.edu

Series Editors
M. Gail
National Cancer Institute
Rockville, MD 20892
USA

K. Krickeberg
Le Châtelet
F-63270 Manglieu
France

J.M. Samet
Department of Epidemiology
School of Public Health
Johns Hopkins University
615 Wolfe Street
Baltimore, MD 21205
USA

A. Tsiatis
Department of Statistics
North Carolina State University
Raleigh, NC 27695
USA

W. Wong
Department of Statistics
Stanford University
Stanford, CA 94305
USA

SAS® and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

SPSS® is a registered trademark of SPSS Inc.

STATA® and the STATA® logo are registered trademarks of StataCorp LP.

ISSN 1431-8776
ISBN 978-1-4419-6645-2 e-ISBN 978-1-4419-6646-9
DOI 10.1007/978-1-4419-6646-9
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011938018

© Springer Science+Business Media, LLC 1996, 2005, 2012
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer Science+Business Media, LLC, 233 Spring Street, New York, NY 10013, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use in this publication of trade names, trademarks, service marks, and similar terms, even if they are not identified as such, is not to be taken as an expression of opinion as to whether or not they are subject to proprietary rights.

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

To *Rachel Robinson*

Morris Dees

Aung San Suu Kyi

John Lewis

And

countless other persons, well-known or unknown,
who have had the courage to stand up for their beliefs for the benefit of humanity.

Preface

This is the third edition of this text on survival analysis, originally published in 1996. As in the first and second editions, each chapter contains a presentation of its topic in "lecture-book" format together with objectives, an outline, key formulae, practice exercises, and a test. The "lecture-book" format has a sequence of illustrations and formulae in the left column of each page and a script in the right column. This format allows you to read the script in conjunction with the illustrations and formulae that highlight the main points, formulae, or examples being presented.

This third edition has expanded the second edition by adding one new chapter, additional sections and clarifications to several chapters, and a revised computer appendix.

The new chapter is Chapter 10, "Design Issues for Randomized Trials," which considers how to compute sample size when designing a randomized trial involving time-to-event data.

We have expanded Chapter 1 to clarify the distinction between random, independent, and noninformative censoring assumptions often made about survival data. We also added a section in Chapter 1 that introduces the Counting Process data layout that is discussed in later chapters (3, 6, and 8).

We added sections in Chapter 2 to describe how to obtain confidence intervals for the Kaplan–Meier (KM) curve and the median survival time obtained from a KM curve.

We have expanded Chapter 3 on the Cox Proportional Hazards (PH) Model by describing the use of age as the time scale instead of time-on-follow-up as the outcome variable. We also added a section that clarifies how to obtain confidence intervals for PH models that contain product terms that reflect effect modification of exposure variables of interest.

We have added sections that describe the derivation of the (partial) likelihood functions for the stratified Cox (SC) model in Chapter 5 and the extended Cox model in Chapter 6.

We have expanded Chapter 9 on competing risks to describe the Fine and Gray model for a subdistribution hazard that allows for a multivariable analysis involving a cumulative incidence curve (CIC). We also added a numerical example to illustrate the calculation of a conditional probability curve (CPC) defined from a CIC.

The Computer Appendix in the second edition of this text provided step-by-step instructions for using the computer packages STATA, SAS, and SPSS to carry out the survival analyses presented in the main text. We expanded this Appendix to include the free internet-based computer software package call R. We have also updated our description of STATA (version 10.0), SAS (version 9.2), and SPSS (version PASW 18). The application of these computer packages to survival data is described in separate self-contained sections of the Computer Appendix, with the analysis of the same datasets illustrated in each section.

In addition to the above new material, the original nine chapters have been modified slightly to correct for errata in the second edition and to add or modify exercises provided at the end of some chapters.

The authors' Web site for this textbook has the following Web-link: http://www.sph.emory.edu/dklein/surv3.htm.

This Web site includes information on how to order this second edition from the publisher and a freely downloadable zip-file containing data-files for examples used in the textbook.

Suggestions for Use

This text was originally intended for self-study, but in the 15 years since the first edition was published, it has also been effectively used as a text in a standard lecture-type classroom format. The text may also be used to supplement material covered in a course or to review previously learned material in a self-instructional course or self-planned learning activity. A more individualized learning program may be particularly suitable to a working professional who does not have the time to participate in a regularly scheduled course.

In working with any chapter, the learner is encouraged first to read the abbreviated outline and the objectives and then work through the presentation. The reader is then encouraged to read the detailed outline for a summary of the presentation, work through the practice exercises, and, finally, complete the test to check what has been learned.

Recommended Preparation

The ideal preparation for this text on survival analysis is a course on quantitative methods in epidemiology and a course in applied multiple regression. Also, knowledge of logistic regression, modeling strategies, and maximum-likelihood techniques is crucial for the material on the Cox and parametric models described in Chapters 3–9.

Recommended references on these subjects, with suggested chapter readings are:

Kleinbaum D, Kupper L, Nizam A, and Muller K, *Applied Regression Analysis and Other Multivariable Methods, Fourth Edition*, Cengage Publishers, 2007, Chapters 1–16, 22–23.

Kleinbaum D, Kupper L and Morgenstern H, *Epidemiologic Research: Principles and Quantitative Methods*, John Wiley and Sons, Publishers, New York, 1982, Chapters 20–24.

Kleinbaum D and Klein M, *Logistic Regression: A Self-Learning Text, Third Edition*, Springer Publishers, New York, 2010, Chapters 4–7, 11.

Kleinbaum D, *ActivEpi-A CD Rom Electronic Textbook on Fundamentals of Epidemiology*, Springer Publishers, New York, 2002, Chapters 13–15.

A first course on the principles of epidemiologic research would be helpful, since all chapters in this text are written from the perspective of epidemiologic research. In particular, the reader should be familiar with the basic characteristics of epidemiologic study designs, and should have some idea of the frequently encountered problem of controlling for confounding and assessing interaction/effect modification. The above reference, *ActivEpi*, provides a convenient and hopefully enjoyable way to review epidemiology.

Acknowledgments

We thank Dr. Val Gebski of the NHMRC Clinical Trials Centre, Sydney, Australia for providing continued insight on current methods of survival analysis and review of new additions to the manuscript for this edition.

Finally, David Kleinbaum and Mitch Klein thank Edna Kleinbaum and Becky Klein for their love, support, companionship, and sense of humor during the writing of this third edition.

Contents

Preface vii

Acknowledgments xi

Chapter 1 Introduction to Survival Analysis 1

Introduction 2
Abbreviated Outline 2
Objectives 3
Presentation 4
Detailed Outline 44
Practice Exercises 50
Test 52
Answers to Practice Exercises 54

Chapter 2 Kaplan-Meier Survival Curves and the Log-Rank Test 55

Introduction 56
Abbreviated Outline 56
Objectives 57
Presentation 58
Detailed Outline 83
Practice Exercises 87
Test 91
Answers to Practice Exercises 93

Chapter 3 The Cox Proportional Hazards Model and Its Characteristics 97

Introduction 98
Abbreviated Outline 98
Objectives 99
Presentation 100
Detailed Outline 145
Practice Exercises 149
Test 153
Answers to Practice Exercises 157

Chapter 4 **Evaluating the Proportional Hazards Assumption 161**

Introduction 162
Abbreviated Outline 162
Objectives 163
Presentation 164
Detailed Outline 188
Practice Exercises 191
Test 194
Answers to Practice Exercises 197

Chapter 5 **The Stratified Cox Procedure 201**

Introduction 202
Abbreviated Outline 202
Objectives 203
Presentation 204
Detailed Outline 228
Practice Exercises 231
Test 234
Answers to Practice Exercises 237

Chapter 6 **Extension of the Cox Proportional Hazards Model for Time-Dependent Variables 241**

Introduction 242
Abbreviated Outline 242
Objectives 243
Presentation 244
Detailed Outline 278
Practice Exercises 281
Test 285
Answers to Practice Exercises 287

Chapter 7 **Parametric Survival Models 289**

Introduction 290
Abbreviated Outline 290
Objectives 291
Presentation 292
Detailed Outline 345
Practice Exercises 351
Test 356
Answers to Practice Exercises 359

Chapter 8 **Recurrent Event Survival Analysis 363**

Introduction 364
Abbreviated Outline 364
Objectives 365
Presentation 366
Detailed Outline 402
Practice Exercises 408
Test 412
Answers to Practice Exercises 422

Chapter 9 **Competing Risks Survival Analysis 425**

Introduction 426
Abbreviated Outline 428
Objectives 429
Presentation 430
Detailed Outline 474
Practice Exercises 481
Test 486
Answers to Practice Exercises 492

Chapter 10 **Design Issues for Randomized Trials 497**

Introduction 498
Abbreviated Outline 498
Objectives 499
Presentation 500
Detailed Outline 518
Practice Exercises 521
Test 523
Answers to Practice Exercises 523

Computer Appendix: Survival Analysis on the Computer 525

A. Stata 527
B. SAS 570
C. SPSS 607
D. R Software 620

Test Answers 665

References 690

Index 695

1 Introduction to Survival Analysis

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_1,
© Springer Science+Business Media, LLC 2012

Introduction

This introduction to survival analysis gives a descriptive overview of the data analytic approach called **survival analysis**. This approach includes the type of problem addressed by survival analysis, the outcome variable considered, the need to take into account "censored data," what a survival function and a hazard function represent, basic data layouts for a survival analysis, the goals of survival analysis, and some examples of survival analysis.

Because this chapter is primarily descriptive in content, no prerequisite mathematical, statistical, or epidemiologic concepts are absolutely necessary. A first course on the principles of epidemiologic research would be helpful. It would also be helpful if the reader has had some experience reading mathematical notation and formulae.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

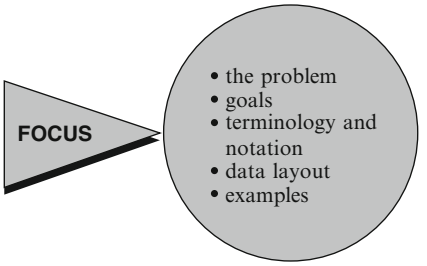
I. **What is survival analysis?** (pages 4–5)
II. **Censored data** (pages 5–8)
III. **Terminology and notation** (pages 9–15)
IV. **Goals of survival analysis** (page 16)
V. **Basic data layout for computer** (pages 16–23)
VI. **Basic data layout for understanding analysis** (pages 23–28)
VII. **Descriptive measures of survival experience** (pages 28–30)
VIII. **Example: Extended remission data** (pages 30–33)
IX. **Multivariable example** (pages 33–35)
X. **Math models in survival analysis** (pages 35–37)
XI. **Censoring assumptions** (pages 37–43)

Objectives

Upon completing the chapter, the learner should be able to:

1. Recognize or describe the type of problem addressed by a survival analysis.
2. Define what is meant by censored data.
3. Define or recognize right-censored data.
4. Give three reasons why data may be censored.
5. Define, recognize, or interpret a survivor function.
6. Define, recognize, or interpret a hazard function.
7. Describe the relationship between a survivor function and a hazard function.
8. State three goals of a survival analysis.
9. Identify or recognize the basic data layout for the computer; in particular, put a given set of survival data into this layout.
10. Identify or recognize the basic data layout, or components thereof, for understanding modeling theory; in particular, put a given set of survival data into this layout.
11. Interpret or compare examples of survivor curves or hazard functions.
12. Given a problem situation, state the goal of a survival analysis in terms of describing how explanatory variables relate to survival time.
13. Compute or interpret average survival and/or average hazard measures from a set of survival data.
14. Define or interpret the hazard ratio defined from comparing two groups of survival data.

Presentation



This presentation gives a general introduction to survival analysis, a popular data analysis approach for certain kinds of epidemiologic and other data. Here we focus on the problem addressed by survival analysis, the goals of a survival analysis, key notation and terminology, the basic data layout, and some examples.

I. What Is Survival Analysis?

We begin by describing the type of analytic problem addressed by survival analysis. Generally, survival analysis is a collection of statistical procedures for data analysis for which the outcome variable of interest is *time until an event occurs*.

Outcome variable: **Time until an event occurs**

By **time**, we mean years, months, weeks, or days from the beginning of follow-up of an individual until an event occurs; alternatively, time can refer to the **age** of an individual when an event occurs.

Event: death
disease
relapse
recovery

By **event**, we mean death, disease incidence, relapse from remission, recovery (e.g., return to work) or any designated experience of interest that may happen to an individual.

Assume 1 event

Although more than one event may be considered in the same analysis, we will assume that only one event is of designated interest. When more than one event is considered (e.g., death from any of several causes), the statistical problem can be characterized as either a recurrent event or a **competing risk** problem, which are discussed in Chaps. 8 and 9, respectively.

Time ≡ survival time

Event ≡ failure

In a survival analysis, we usually refer to the time variable as **survival time,** because it gives the time that an individual has "survived" over some follow-up period. We also typically refer to the event as a **failure,** because the event of interest usually is death, disease incidence, or some other negative individual experience. However, survival time may be "time to return to work after an elective surgical procedure," in which case failure is a positive event.

EXAMPLE

1. Leukemia patients/time in remission (weeks)
2. Disease-free cohort/time until heart disease (years)
3. Elderly (60+) population/time until death (years)
4. Parolees (recidivism study)/time until rearrest (weeks)
5. Heart transplants/time until death (months)

Five examples of survival analysis problems are briefly mentioned here. The first is a study that follows leukemia patients in remission over several weeks to see how long they stay in remission. The second example follows a disease-free cohort of individuals over several years to see who develops heart disease. A third example considers a 13-year follow-up of an elderly population (60+ years) to see how long subjects remain alive. A fourth example follows newly released parolees for several weeks to see whether they get rearrested. This type of problem is called a recidivism study. The fifth example traces how long patients survive after receiving a heart transplant.

All of the above examples are survival analysis problems because the outcome variable is time until an event occurs. In the first example, involving leukemia patients, the event of interest (i.e., failure) is "going out of remission," and the outcome is "time in weeks until a person goes out of remission." In the second example, the event is "developing heart disease," and the outcome is "time in years until a person develops heart disease." In the third example, the event is "death" and the outcome is "time in years until death." Example four, a sociological rather than a medical study, considers the event of recidivism (i.e., getting rearrested), and the outcome is "time in weeks until rearrest." Finally, the fifth example considers the event "death," with the outcome being "time until death (in months from receiving a transplant)."

We will return to some of these examples later in this presentation and in later presentations.

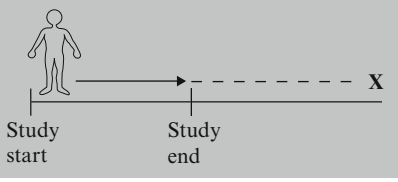
II. Censored Data

Censoring: don't know survival time exactly

Most survival analyses must consider a key analytical problem called **censoring**. In essence, censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly**.

EXAMPLE

Leukemia patients in remission:



As a simple example of censoring, consider leukemia patients followed until they go out of remission, shown here as **X**. If for a given patient, the study ends while the patient is still in remission (i.e., doesn't get the event), then that patient's survival time is considered censored. We know that, for this person, the survival time is at least as long as the period that the person has been followed, but if the person goes out of remission after the study ends, we do not know the complete survival time.

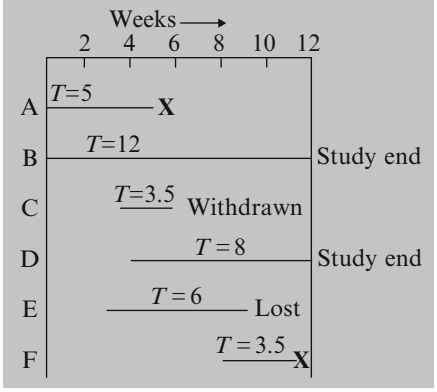
Why censor?

1. study ends – no event
2. lost to follow-up
3. withdraws

There are generally three reasons why censoring may occur:

(1) a person does not experience the event before **the study ends**;
(2) a person is **lost to follow-up** during the study period;
(3) a person **withdraws from the study** because of death (if death is not the event of interest) or some other reason (e.g., adverse drug reaction or other competing risk)

EXAMPLE



X $\Longrightarrow$ Event occurs

These situations are graphically illustrated here. The graph describes the experience of several persons followed over time. An **X** denotes a person who got the event.

Person A, for example, is followed from the start of the study until getting the event at week 5; his survival time is 5 weeks and is *not* censored.

Person B also is observed from the start of the study but is followed to the end of the 12-week study period without getting the event; the survival time here is censored because we can say only that it is *at least* 12 weeks.

Person C enters the study between the second and 3rd week and is followed until he withdraws from the study at 6 weeks; this person's survival time is censored after 3.5 weeks.

Person D enters at week 4 and is followed for the remainder of the study without getting the event; this person's censored time is 8 weeks.

Person E enters the study at week 3 and is followed until week 9, when he is lost to follow-up; his censored time is 6 weeks.

Person F enters at week 8 and is followed until getting the event at week 11.5. As with person A, there is no censoring here; the survival time is 3.5 weeks.

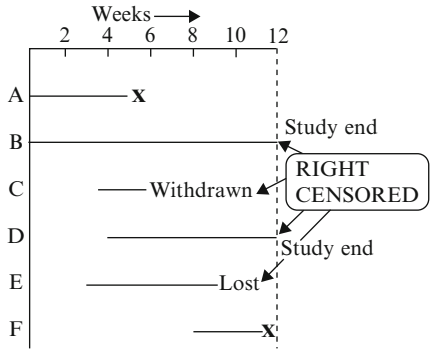
SUMMARY
Event: A, F
Censored: B, C, D, E

In **summary**, of the six persons observed, two get the event (persons A and F) and four are censored (B, C, D, and E).

Person	Survival time	Failed (1); Censored (0)
A	5	1
B	12	0
C	3.5	0
D	8	0
E	6	0
F	3.5	1

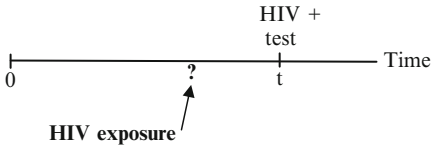
A table of the survival time data for the six persons in the graph is now presented. For each person, we have given the corresponding survival time up to the event's occurrence or up to censorship. We have indicated in the last column whether this time was censored or not (with 1 denoting failed and 0 denoting censored). For example, the data for person C is a survival time of 3.5 and a censorship indicator of 0, whereas for person F the survival time is 3.5 and the censorship indicator is 1. This table is a simplified illustration of the type of data to be analyzed in a survival analysis.

Right-censored: true survival time is equal to or greater than observed survival time



Notice in our example that for each of the four persons censored, we know that the person's true survival time becomes incomplete at the **right** side of the follow-up period, occurring when the study ends or when the person is lost to follow-up or is withdrawn. We generally refer to this kind of data as **right-censored**. For these data, the complete survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed survival time interval. Although data can also be **left-censored**, most survival data is right-censored.

Left-censored: true survival time is less than or equal to the observed survival time

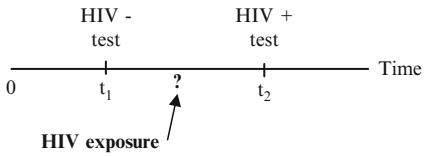


Left-censored: data can occur when a person's true survival time is less than or equal to that person's observed survival time. For example, if we are following persons until they become HIV positive, we may record a failure when a subject first tests positive for the virus. However, we may not know the exact time of first exposure to the virus, and therefore do not know exactly when the failure occurred. Thus, the survival time is censored on the left side since the true survival time, which ends at exposure, is shorter than the follow-up time, which ends when the subject's test is positive.

Event occurs between 0 and t
but
do not know the exact time.

In other words, if a person is left-censored at time t, we know they had an event between time 0 and t, but we do not know the exact time of event.

Interval-censored: true survival time is within a known time interval



Survival analysis data can also be **interval-censored**, which can occur if a subject's true (but unobserved) survival time is within a certain known specified time interval. As an example, again considering HIV surveillance, a subject may have had two HIV tests, where he/she was HIV negative at the time (say, t_1) of the first test and HIV positive at the time (t_2) of the second test. In such a case, the subject's true survival time occurred after time t_1 and before time t_2, i.e., the subject is interval-censored in the time interval (t_1, t_2).

Left censoring $\Rightarrow$ $t_1 = 0$, t_2 = upper bound

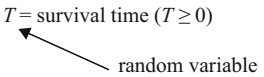
Right censoring $\Rightarrow$ t_1 = lower bound, $t_2 = \infty$

Interval-censoring actually incorporates both right-censoring and left-censoring as special cases. Left-censored data occur whenever the value of t_1 is 0 and t_2 is a known upper bound on the true survival time. In contrast, right-censored data occurs whenever the value of t_2 is infinity, and t_1 is a known lower bound on the true survival time.

Right-censored due to competing risk, e.g., death from another cause
$\Downarrow$
$t_2 = \infty$
gives upper bound for true survival time assuming that competing risk had not occurred.

If an individual is right-censored due to a competing event (e.g., death from another cause), then in this context, we consider what the true survival time would have been if the competing event had not occurred. In other words, when we state that the value of the upper bound for the true survival time is infinity for right-censored data, we are considering what would have occurred in the absence of a competing risk. Competing risks are fully discussed in Chapter 9.

III. Terminology and Notation



t = specific value for T

EXAMPLE

Survives > 5 years?
$T > t = 5$

We are now ready to introduce basic mathematical terminology and notation for survival analysis. First, we denote by a **capital *T*** the random variable for a person's survival time. Since T denotes time, its possible values include all nonnegative numbers; that is, T can be any number equal to or greater than zero.

Next, we denote by a **small letter *t*** any specific value of interest for the random variable capital T. For example, if we are interested in evaluating whether a person survives for more than 5 years after undergoing cancer therapy, **small *t*** equals 5; we then ask whether capital T exceeds 5.

d = (0, 1) random variable

$$= \begin{cases} 1 & \text{if failure} \\ 0 & \text{censored} \end{cases}$$

- study ends
- lost to follow-up
- withdraws

Finally, we denote the **small letter *d*** to define a (0,1) random variable indicating either failure or censorship. That is, $d = 1$ for failure if the event occurs during the study period, or $d = 0$ if the survival time is censored by the end of the study period. Note that if a person does not fail, that is, does not get the event during the study period, censorship is the **only** remaining possibility for that person's survival time. That is, $d = 0$ if and only if one of the following happens: a person survives until the study ends, a person is lost to follow-up, or a person withdraws during the study period.

$S(t)$ = survivor function
$h(t)$ = hazard function

We next introduce and describe two quantitative terms considered in any survival analysis. These are the **survivor function**, denoted by $S(t)$, and the **hazard function**, denoted by $h(t)$.

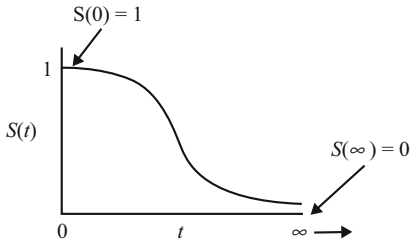
$S(t) = P(T > t)$

t	$S(t)$
1	$S(1) = P(T > 1)$
2	$S(2) = P(T > 2)$
3	$S(3) = P(T > 3)$
.	.
.	.
.	.

The survivor function $S(t)$ gives the probability that a person survives longer than some specified time ***t***: that is, $S(t)$ gives the probability that the random variable T exceeds the specified time t.

The survivor function is fundamental to a survival analysis, because obtaining survival probabilities for different values of t provides crucial summary information from survival data.

Theoretical $S(t)$:

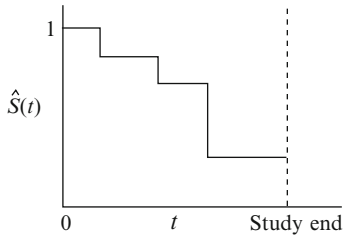


Theoretically, as t ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph, where t identifies the X-axis, all survivor functions have the following characteristics:

- they are nonincreasing; that is, they head downward as t increases;
- at time $t = 0$, $S(t) = S(0) = 1$; that is, at the start of the study, since no one has gotten the event yet, the probability of surviving past time 0 is one;
- at time $t = \infty$, $S(t) = S(\infty) = 0$; that is, theoretically, if the study period increased without limit, eventually nobody would survive, so the survivor curve must eventually fall to zero.

Note that these are **theoretical** properties of survivor curves.

$\hat{S}(t)$ in practice:

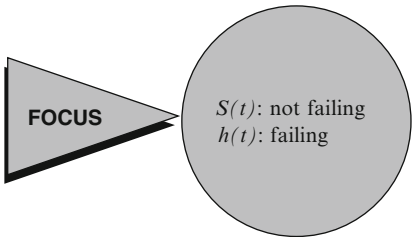


In practice, when using actual data, we usually obtain graphs that are **step functions**, as illustrated here, rather than smooth curves. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. The estimated survivor function, denoted by a caret over the S in the graph, thus may not go all the way down to zero at the end of the study.

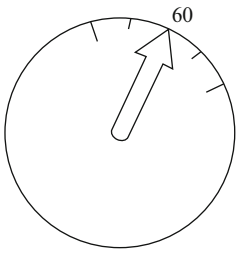
$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

The hazard function, denoted by $\boldsymbol{h(t)}$, is given by the formula: $h(t)$ equals the limit, as Δt approaches zero, of a probability statement about survival, divided by Δt, where Δt denotes a small interval of time. This mathematical formula is difficult to explain in practical terms.

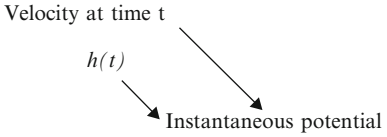
$h(t)$ = instantaneous potential



Before getting into the specifics of the formula, we give a conceptual interpretation. **The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t.** Note that, in contrast to the survivor function, which focuses on *not* failing, the hazard function focuses on failing, that is, on the event occurring. Thus, in some sense, the hazard function can be considered as giving the opposite side of the information given by the survivor function.



To get an idea of what we mean by instantaneous potential, consider the concept of velocity. If, for example, you are driving in your car and you see that your speedometer is registering 60 mph, what does this reading mean? It means that if in the next hour, you continue to drive this way, with the speedometer exactly on 60, you would cover 60 miles. This reading gives the **potential**, at the moment you have looked at your speedometer, for how many miles you will travel in the next hour. However, because you may slow down or speed up or even stop during the next hour, the 60-mph speedometer reading does not tell you the number of miles you *really* will cover in the next hour. The speedometer tells you only how fast you are going *at a given moment*; that is, the instrument gives your instantaneous potential or velocity.



Similar to the idea of velocity, a hazard function $h(t)$ gives the instantaneous potential at time t for getting an event, like death or some disease of interest, given survival up to time t. The "given" part, that is, surviving up to time t, is analogous to recognizing in the velocity example that the speedometer reading at a point in time inherently assumes that you have already traveled some distance (i.e., survived) up to the time of the reading.

Given

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

In mathematical terms, the given part of the formula for the hazard function is found in the probability statement in the numerator to the right of the limit sign. This statement is a conditional probability because it is of the form, "P of A, given B," where the P denotes probability and where the long vertical line separating A from B denotes "given." In the hazard formula, the conditional probability gives the probability that a person's survival time, T, will lie in the time interval between t and $t + \Delta t$, given that the survival time is greater than or equal to t. Because of the given sign here, the hazard function is sometimes called a **conditional failure rate**.

Conditional probabilities: $P(A|B)$

$P(t \le T < t + \Delta t \mid T \ge t)$
= P(individual fails in the interval $[t, t + \Delta t]$ | survival up to time t)

Hazard function ≡ conditional failure **rate**

$$\lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Probability per unit time

We now explain why the hazard is a **rate** rather than a probability. Note that in the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities. The numerator is the conditional probability we just discussed. The denominator is Δt, which denotes a small time interval. By this division, we obtain a probability per unit time, which is no longer a probability but a rate. In particular, the scale for this ratio is not 0 to 1, as for a probability, but rather ranges between 0 and infinity, and depends on whether time is measured in days, weeks, months, or years, etc.

Rate: 0 to ∞

$P = P(t \le T < t + \Delta t | T \ge t)$

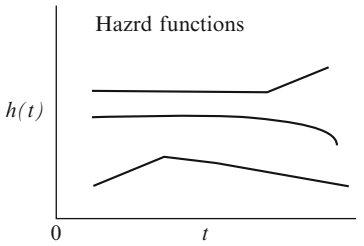
P	Δt	$P/\Delta t$ = rate
$\frac{1}{3}$	$\frac{1}{2}$ day	$\frac{1/3}{1/2} = 0.67$/day
$\frac{1}{3}$	$\frac{1}{14}$ week	$\frac{1/3}{1/14} = 4.67$/week

For example, if the probability, denoted here by P, is 1/3, and the time interval is one-half a day, then the probability divided by the time interval is 1/3 divided by 1/2, which equals 0.67 per day. As another example, suppose, for the same probability of 1/3, that the time interval is considered in weeks, so that 1/2 day equals 1/14 of a week. Then the probability divided by the time interval becomes 1/3 over 1/14, which equals 14/3, or 4.67 per week. The point is simply that the expression P divided by Δt at the right of the limit sign **does not give a probability. The value obtained will give a different number depending on the units of time used, and may even give a number larger than one**.

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}$$

Gives instantaneous potential

When we take the limit of the right-side expression as the time interval approaches zero, we are essentially getting an expression for the instantaneous probability of failing at time t per unit time. Another way of saying this is that the conditional failure rate or hazard function $h(t)$ gives the instantaneous **potential** for failing at time t per unit time, given survival up to time t.



As with a survivor function, the hazard function $h(t)$ can be graphed as t ranges over various values. The graph at the left illustrates three different hazards. In contrast to a survivor function, the graph of $h(t)$ does not have to start at 1 and go down to zero, but rather can start anywhere and go up and down in any direction over time. In particular, for a specified value of t, the hazard function $h(t)$ has the following characteristics:

- $h(t) \geq 0$
- $h(t)$ has no upper bound

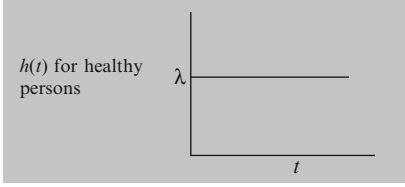
- it is always nonnegative, that is, equal to or greater than zero;
- it has no upper bound.

These two features follow from the ratio expression in the formula for $h(t)$, because both the probability in the numerator and the Δt in the denominator are nonnegative, and since Δt can range between 0 and ∞.

EXAMPLE

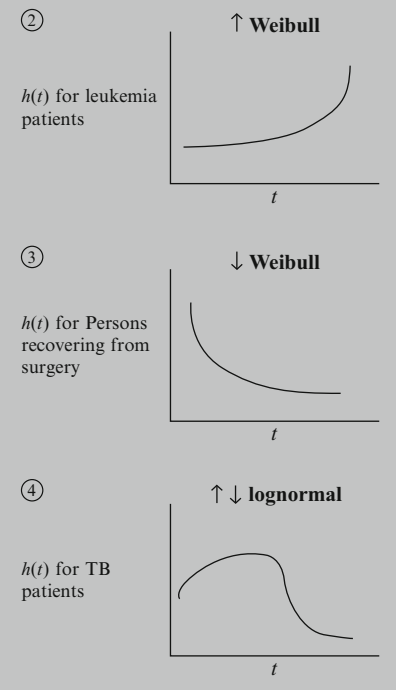
①

Constant hazard
(exponential model)



Now we show some graphs of different types of hazard functions. The first graph given shows a constant hazard for a study of healthy persons. In this graph, no matter what value of t is specified, $h(t)$ equals the same value—in this example, λ. Note that for a person who continues to be healthy throughout the study period, his/her instantaneous potential for becoming ill at any time during the period remains constant throughout the follow-up period. When the hazard function is constant, we say that the survival model is **exponential**. This term follows from the relationship between the survivor function and the hazard function. We will return to this relationship later.

EXAMPLE: (continued)



The second graph shows a hazard function that is increasing over time. An example of this kind of graph is called an **increasing Weibull** model. Such a graph might be expected for leukemia patients not responding to treatment, where the event of interest is death. As survival time increases for such a patient, and as the prognosis accordingly worsens, the patient's potential for dying of the disease also increases.

In the third graph, the hazard function is decreasing over time. An example of this kind of graph is called a **decreasing Weibull**. Such a graph might be expected when the event is death in persons who are recovering from surgery, because the potential for dying after surgery usually decreases as the time after surgery increases.

The fourth graph given shows a hazard function that is first increasing and then decreasing. An example of this type of graph is the **lognormal survival** model. We can expect such a graph for tuberculosis patients, since their potential for dying increases early in the disease and decreases later.

$S(t)$: directly describes survival
$h(t)$:
- a measure of instantaneous potential
- identify specific model form
- math model for survival analysis

Of the two functions we have considered, $S(t)$ and $h(t)$, the survivor function is more naturally appealing for analysis of survival data, simply because $S(t)$ directly describes the survival experience of a study cohort.

However, the hazard function is also of interest for the following reasons:

- it is a measure of instantaneous potential whereas a survival curve is a cumulative measure over time;
- it may be used to identify a specific model form, such as an exponential, a Weibull, or a lognormal curve that fits one's data;
- it is the vehicle by which mathematical modeling of survival data is carried out; that is, the survival model is usually written in terms of the hazard function.

Relationship of *S(t) and h(t)*:
If you know one, you can determine the other.

EXAMPLE

$h(t) = \lambda$ if and only if $S(t) = e^{-\lambda t}$

Regardless of which function $S(t)$ or $h(t)$ one prefers, **there is a clearly defined relationship between the two**. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa. For example, if the hazard function is constant, i.e., $h(t) = \lambda$, for some specific value λ, then it can be shown that the corresponding survival function is given by the following formula: $S(t)$ equals **e** to the power minus λ times t.

General formulae:

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]$$

More generally, the relationship between $S(t)$ and $h(t)$ can be expressed equivalently in either of two calculus formulae shown here.

The first of these formulae describes how the survivor function $S(t)$ can be written in terms of an integral involving the hazard function. The formula says that $S(t)$ equals the exponential of the negative integral of the hazard function between integration limits of 0 and t.

The second formula describes how the hazard function $h(t)$ can be written in terms of a derivative involving the survivor function. This formula says that $h(t)$ equals minus the derivative of $S(t)$ with respect to t divided by $S(t)$.

In any actual data analysis, a computer program can make the numerical transformation from $S(t)$ to $h(t)$, or vice versa, without the user ever having to use either formula. The point here is simply that if you know either $S(t)$ or $h(t)$, you can get the other directly.

SUMMARY

T = survival time random variable
t = specific value of T
d = (0,1) variable for failure/censorship
$S(t)$ = survivor function
$h(t)$ = hazard function

At this point, we have completed our discussion of key terminology and notation. **The key notation is *T* for the survival time variable, *t* for a specified value of *T*, and *d* for the dichotomous variable indicating event occurrence or censorship. The key terms are the survivor function *S(t)* and the hazard function *h(t)***, which are in essence opposed concepts, in that the survivor function focuses on surviving whereas the hazard function focuses on failing, given survival up to a certain time point.

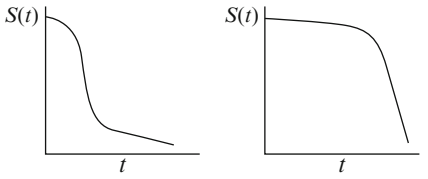
IV. Goals of Survival Analysis

We now state the basic goals of survival analysis.

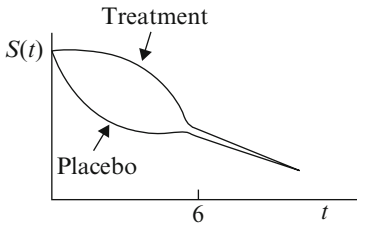
Goal 1: To estimate and interpret survivor and/or hazard functions from survival data.

Goal 2: To compare survivor and/or hazard functions.

Goal 3: To assess the relationship of explanatory variables to survival time.



Regarding the first goal, consider, for example, the two survivor functions pictured at the left, which give very different interpretations. The function farther on the left shows a quick drop in survival probabilities early in follow-up but a leveling off thereafter. The function on the right, in contrast, shows a very slow decrease in survival probabilities early in follow-up but a sharp decrease later on.



We compare survivor functions for a treatment group and a placebo group by graphing these functions on the same axis. Note that up to 6 weeks, the survivor function for the treatment group lies above that for the placebo group, but thereafter the two functions are at about the same level. This dual graph indicates that up to 6 weeks the treatment is **more effective** for survival than the placebo but has about the same effect thereafter.

Goal 3: Use math modeling, e.g., Cox proportional hazards

Goal 3 usually requires using some form of mathematical modeling, for example, the Cox proportional hazards approach, which will be the subject of subsequent chapters.

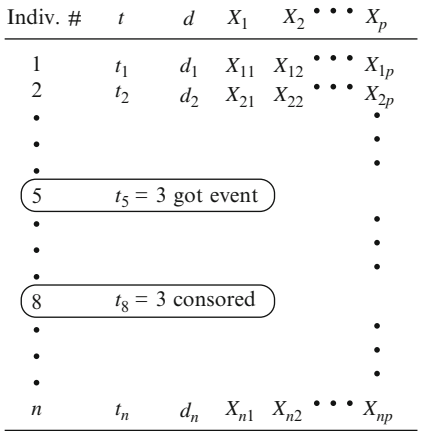
V. Basic Data Layout for Computer

We previously considered some examples of survival analysis problems and a simple data set involving six persons. We now consider the general data layout for a survival analysis. We will provide two types of data layouts, one giving the form appropriate for computer use, and the other giving the form that helps us understand how a survival analysis works.

Two types of data layouts:

- for computer use
- for understanding

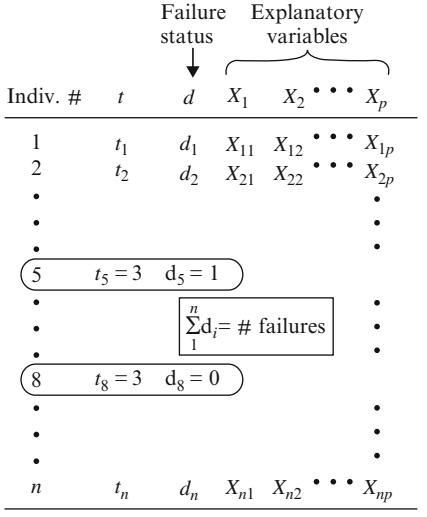
For computer:



We start by providing, in the table shown here, the basic data layout for the computer. Assume that we have a data set consisting of n persons. The first column of the table identifies each person from 1, starting at the top, to n, at the bottom.

The remaining columns after the first one provide survival time and other information for each person. The second column gives the survival time information, which is denoted t_1 for individual 1, t_2 for individual 2, and so on, up to t_n for individual n. Each of these t's gives the observed survival time regardless of whether the person got the event or is censored. For example, if person 5 got the event at 3 weeks of follow-up, then $t_5 = 3$; on the other hand, if person 8 was censored at 3 weeks, without getting the event, then $t_8 = 3$ also.

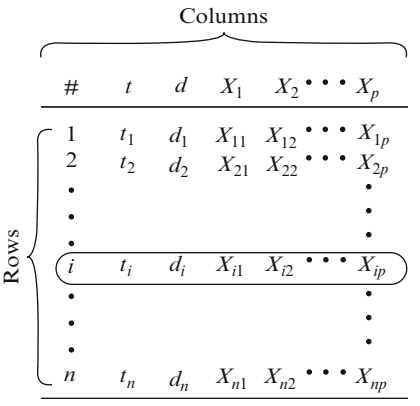
To distinguish persons who get the event from those who are censored, we turn to the third column, which gives the information for status (i.e., d) the dichotomous variable that indicates censorship status.



Thus, d_1 is 1 if person 1 gets the event or is 0 if person 1 is censored; d_2 is 1 or 0 similarly, and so on, up through d_n. In the example just considered, person 5, who failed at 3 weeks, has a d of 1; that is, d_5 equals 1. In contrast, person 8, who was censored at 3 weeks, has a d of 0; that is, d_8 equals 0.

Note that if all of the d values in this column are added up, their sum will be the total number of failures in the data set. This total will be some number equal to or less than n, because not every one may fail.

The remainder of the information in the table gives values for explanatory variables of interest. An explanatory variable, X_l is any variable like age or exposure status, E, or a product term like age $\times$ race that the investigator wishes to consider to predict survival time. These variables are listed at the top of the table as X_1, X_2, and so on, up to X_p. Below each variable are the values observed for that variable on each person in the data set.



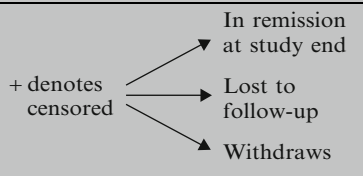
For example, in the column corresponding to X_1 are the values observed on this variable for all n persons. These values are denoted as X_{11}, X_{21}, and so on, up to X_{n1}; the first subscript indicates the person number, and the second subscript, a one in each case here, indicates the variable number. Similarly, the column corresponding to variable X_2 gives the values observed on X_2 for all n persons. This notation continues for the other X variables up through X_p.

We have thus described the basic data layout by columns. Alternatively, we can look at the table line by line, that is, by rows. For each line or row, we have the information obtained on a given individual. Thus, for individual i, the observed information is given by the values t_i, d_i, X_{i1}, X_{i2}, etc., up to X_{ip}. This is how the information is read into the computer, that is, line by line, until all persons are included for analysis.

EXAMPLE

The data: Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23



As an example of this data layout, consider the following set of data for two groups of leukemia patients: one group of 21 persons has received a certain treatment; the other group of 21 persons has received a placebo. The data come from Freireich et al., *Blood*, 1963.

As presented here, the data are not yet in tabular form for the computer, as we will see shortly. The values given for each group consist of time in weeks a patient is in remission, up to the point of the patient's either going out of remission or being censored. Here, going out of remission is a failure. A person is censored if he or she remains in remission until the end of the study, is lost to follow-up, or withdraws before the end of the study. The censored data here are denoted by a plus sign next to the survival time.

EXAMPLE: (continued)

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

	Indiv. #	t (weeks)	d (failed or censored)	X (Group)
	1	6	1	1
	2	6	1	1
	(3)	6	1	1
	4	7	1	1
	5	10	1	1
	6	13	1	1
	7	16	1	1
	8	22	1	1
GROUP	9	23	1	1
1	10	6	0	1
	11	9	0	1
	12	10	0	1
	13	11	0	1
	(14)	17	0	1
	15	19	0	1
	16	20	0	1
	17	25	0	1
	18	32	0	1
	19	32	0	1
	20	34	0	1
	21	35	0	1

Here are the data again:

Notice that the first three persons in group 1 went out of remission at 6 weeks; the next 6 persons also went out of remission, but at failure times ranging from 7 to 23. All of the remaining persons in group 1 with pluses next to their survival times are censored. For example, on line three the first person who has a plus sign next to a 6 is censored at 6 weeks. The remaining persons in group 1 are also censored, but at times ranging from 9 to 35 weeks.

Thus, of the 21 persons in group 1, nine failed during the study period, whereas the last 12 were censored. Notice also that none of the data in group 2 is censored; that is, all 21 persons in this group went out of remission during the study period.

We now put this data in tabular form for the computer, as shown at the left. The list starts with the 21 persons in group 1 (listed 1–21) and follows (on the next page) with the 21 persons in group 2 (listed 22–42). Our n for the composite group is 42.

The *second* column of the table gives the survival times in weeks for all 42 persons. The *third* column indicates failure or censorship for each person. Finally, the *fourth* column lists the values of the only explanatory variable we have considered so far, namely, group status, with 1 denoting treatment and 0 denoting placebo.

If we pick out any individual and read across the table, we obtain the line of data for that person that gets entered in the computer. For example, person #3 has a survival time of 6 weeks, and since $d = 1$, this person failed, that is, went out of remission. The X value is 1 because person #3 is in group 1. As a second example, person #14, who has an observed survival time of 17 weeks, was censored at this time because $d = 0$. The X value is again 1 because person #14 is also in group 1.

EXAMPLE: (continued)

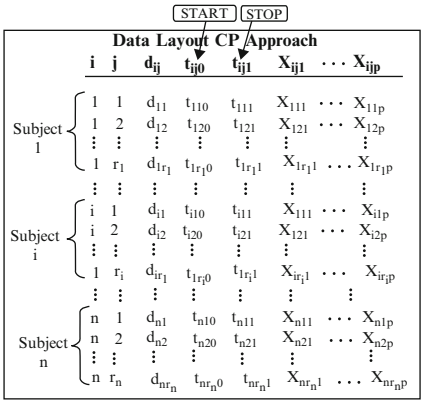
	Indiv. #	t (weeks)	d (failed or censored)	X (Group)
	22	1	1	0
	23	1	1	0
	24	2	1	0
	25	2	1	0
	26	3	1	0
	27	4	1	0
GROUP 2	28	4	1	0
	29	5	1	0
	30	5	1	0
	31	8	1	0
	32	8	1	0
	33	8	1	0
	34	8	1	0
	35	11	1	0
	36	11	1	0
	37	12	1	0
	38	12	1	0
	39	15	1	0
	40	17	1	0
	41	22	1	0
	42	23	1	0

As one more example, this time from group 2, person #32 survived 8 weeks and then failed, because $d = 1$; the X value is 0 because person #32 is in group 2.

Alternative Data Layout: **Counting Process** (Start, Stop) Format

CP Format: applies to more complicated survival analysis

- Age-at follow-up is outcome
- Time-dependent variables
- Recurrent events
- Gaps in follow-up



An alternative format for the computer is called the **Counting Process** (**CP**) format.

The CP format is useful for more complicated survival analysis situations that we discuss in later chapters, in particular when age-at-follow-up time is used as the outcome variable instead of time of follow-up (Chap. 3), when there are time-dependent variables (Chap. 6), and when there are recurrent events and/or gaps in follow-up (Chap. 8).

The general CP format is shown on the left. This format differs from the previously described "standard" data layout in two ways. First, the CP format allows **multiple lines of data for the same individual**; that is, each individual's total at-risk-follow-up time is subdivided into smaller time intervals to allow for recurrent events on the same individual. Second, there are **two time points specified for each individual**, labeled in the layout as t_{ij0} and t_{ij1}, and often referred to as **START** and **STOP times**, respectively.

The first two columns in this format are labeled **i** (for subject number) and **j** (for dataline number for the ith subject). As in the standard format, i ranges from 1 to n; also, in the CP format, j ranges from 1 to $\mathbf{r_i}$, where $\mathbf{r_i}$ denotes the number of datalines for the i-th subject.

The third column labeled $\mathbf{d_{ij}}$ denotes the failure status (1=failed, 0=censored) for the j-th dataline on the i-th subject.

The next two columns identify two time points required for each dataline, the **START time** ($\mathbf{t_{ij0}}$) and the **STOP time** ($\mathbf{t_{ij1}}$). These two columns are the primary distinguishing feature of the CP format.

Simplest CP format: 1 dataline/ subject

i	j	d_{ij}	t_{ij0}	t_{ij1}	X_{ij1}	⋯	X_{ijp}
1	1	d_{11}	0	t_1	X_{111}	⋯	X_{11p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
i	1	d_{i1}	0	t_i	X_{i11}	⋯	X_{i1p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
n	1	d_{n1}	0	t_n	X_{n11}	⋯	X_{n1p}

The simplest CP format occurs when the outcome is follow-up time since study entry and when there are no recurrent events or time-dependent covariates, as in our previously described Remission Time Dataset. In this situation, there is one dataline for each subject (i.e., r_i=1 for all i so that the only value that j takes is 1), the start time (t_{i10}) is 0 for each subject, and the stop time (t_{i11}) is the follow-up time (t) until either the event or censorship occurs.

CP Format for Group 1 of Remission Time Dataset

i	j	d_{ij}	start	stop	X(Group)
1	1	1	0	6	1
2	1	1	0	6	1
3	1	1	0	6	1
4	1	1	0	7	1
5	1	1	0	10	1
6	1	1	0	13	1
7	1	1	0	16	1
8	1	1	0	22	1
9	1	1	0	23	1
10	1	0	0	12	1
11	1	0	0	6	1
⋮	⋮	⋮	⋮	⋮	⋮
19	1	0	0	32	1
20	1	0	0	34	1
21	1	0	0	35	1

As an example, the CP format for Group 1 of the Remission Time Dataset is shown on the left. Note that the value of j is 1 throughout the table, the start times are all zero, and the stop times are the failure or censored survival times for each subject.

CP Format: First 15 Subjects-Bladder Canscer Study

i	j	d	start	stop	tx	num	size
1	1	0	0	0	0	1	1
2	1	0	0	1	0	1	3
3	1	0	0	4	0	2	1
4	1	0	0	7	0	1	1
5	1	0	0	10	0	5	1
6	1	1	0	6	0	4	1
6	2	0	6	10	0	4	1
7	1	0	0	14	0	1	1
8	1	0	0	18	0	1	1
9	1	1	0	5	0	1	3
9	2	0	5	18	0	1	3
10	1	1	0	12	0	1	1
10	2	1	12	16	0	1	1
10	3	0	16	18	0	1	1
11	1	0	0	23	0	3	3
12	1	1	0	10	0	1	3
12	2	1	10	15	0	1	3
12	3	0	15	23	0	1	3
13	1	1	0	3	0	1	1
13	2	1	3	16	0	1	1
13	3	1	16	23	0	1	1
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1
15	1	1	0	7	0	2	3
15	2	1	7	10	0	2	3
15	3	1	10	16	0	2	3
15	4	0	16	24	0	2	3

We now illustrate the CP format that allows for more than one dataline per subject as well as start times other than zero. We consider data on the first 15 subjects from a study of recurrent bladder cancer tumors (Byar, 1980; and Wei, Lin and Weissfeld, 1989). The entire dataset contained 86 patients, each followed for a variable amount of time up to 64 months. We describe how to analyze this dataset in Chapter 8 on Recurrent Event Survival Analysis. Here, we only describe how this data layout fits the CP format.

The event being analyzed is the recurrence of bladder cancer tumor after transurethral surgical excision. Each recurrence of new tumors was treated by removal at each examination.

The exposure variable of interest is drug treatment status (**tx**, 0=placebo, 1= treatment with thiotepa). Although each of 15 subjects shown here are in the placebo group (tx=0), several other subjects in the larger dataset are in the treatment group (tx=1).

The covariates listed here are initial number of tumors (**num**) and initial size of tumors (**size**) in centimeters. Both these variables have the same value for each subject (i.e., time-independent variables), although the general data layout also allows for time-dependent variables.

Notice that several subjects in this dataset, namely subjects 6, 9, 10, 12, 13, 14, and 15 have two or more datalines. Subject 6, for example, has two datalines (i.e., $r_6 = 2$), whereas subject 14 has 4 datalines (i.e., $r_{14} = 4$).

Bladder Canscer Study (cont'd)

i	j	d	start	stop	tx	num	size
6	1	1	0	6	0	4	1
6	2	0	6	10	0	4	1

The first of the two lines for subject 6 tells us that this subject had a (first) recurrent bladder cancer event (i.e., d=1) at 6 months (i.e., stop =6). This subject was then followed for another 4 months (from 6 to 10, as shown on the second dataline for this subject, where the start time is 6 and the stop time is 10. At 10 months, the subject is censored (d=0); in other words, this subject did not get a second recurrent event at 10 months, after which no further information is available.

i	j	d	start	stop	tx	num	size
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1

Subject 14 had three recurrent events, the first one at 3 months (i.e., stop =3), the second one at 9 months (i.e., stop =9), and the third one at 21 months (i.e., stop =21). This subject was followed for another 2 months (start =21 to stop =23 on dataline number j=4) without another event occurring (d=0).

CP format illustrated for other situations in later chapters.

As mentioned at the beginning of this section, the CP format is also applicable when age-at-follow-up time is used as the outcome variable instead of time of follow-up (Chapter 3), when there are time-dependent variables (Chapter 6), and when there are gaps in follow-up (Chapter 8). We will illustrate these latter situations within the later chapters just mentioned.

See Computer Appendix for computer code in CD format for SAS, STATA, and R.

In the Computer Appendix, we describe the computer code required by SAS, STATA, and R packages when the data is set up in CP format for the analysis of recurrent event survival data and when age is used as the time scale instead of time-on-study.

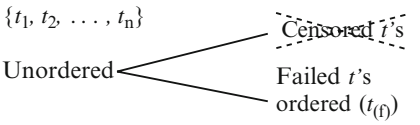
VI. Basic Data Layout for Understanding Analysis

We are now ready to look at another data layout, which is shown at the left. This layout helps provide some understanding of how a survival analysis actually works and, in particular, how survivor curves are derived.

For analysis:

Ordered failure times ($t_{(f)}$)	# of failures (m_f)	# censored in $[t_{(f}, t_{(f-1)})$ (q_f)	Risk set R ($t_{(f)}$)
$t_{(0)}= 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

The first column in this table gives ordered failure times. These are denoted by t's with subscripts within parentheses, starting with $t_{(0)}$, then $t_{(1)}$ and so on, up to $t_{(k)}$. Note that the parentheses surrounding the subscripts distinguish ordered failure times from the survival times previously given in the computer layout.



k = # of distinct times at which subjects failed ($k \leq n$)

To get ordered failure times from survival times, we must first remove from the list of unordered survival times all those times that are censored; we are thus working only with those times at which people failed. We then order the remaining failure times from smallest to largest, and count ties only once. The value k gives the number of distinct times at which subjects failed.

EXAMPLE

Remission Data: Group 1
(n = 21, 9 failures, k = 7)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Remission Data: Group 2
(n = 21, 21 failures, k = 12)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 1$	2	0	21 persons survive ≥ 1 wks
$t_{(2)} = 2$	2	0	19 persons survive ≥ 2 wks
$t_{(3)} = 3$	1	0	17 persons survive ≥ 3 wks
$t_{(4)} = 4$	2	0	16 persons survive ≥ 4 wks
$t_{(5)} = 5$	2	0	14 persons survive ≥ 5 wks
$t_{(6)} = 8$	4	0	12 persons survive ≥ 8 wks
$t_{(7)} = 11$	2	0	8 persons survive ≥ 11 wks
$t_{(8)} = 12$	2	0	6 persons survive ≥ 12 wks
$t_{(9)} = 15$	1	0	4 persons survive ≥ 15 wks
$t_{(10)} = 17$	1	0	3 persons survive ≥ 17 wks
$t_{(11)} = 22$	1	0	2 persons survive ≥ 22 wks
$t_{(12)} = 23$	1	0	1 person survive ≥ 23 wks
Totals	21	0	

ties

For example, using the remission data for group 1, we find that 9 of the 21 persons failed, including 3 persons at 6 weeks and 1 person each at 7, 10, 13, 16, 22, and 23 weeks. These nine failures have $k = 7$ distinct survival times, because three persons had survival time 6 and we only count one of these 6's as distinct. The first ordered failure time for this group, denoted as $t_{(1)}$, is 6; the second ordered failure time $t_{(2)}$, is 7, and so on up to the seventh ordered failure time of 23.

Turning to group 2, we find that although all 21 persons in this group failed, there are several ties. For example, two persons had a survival time of 1 week; two more had a survival time of 2 weeks; and so on. In all, we find that there were $k = 12$ distinct survival times out of the 21 failures. These times are listed in the first column for group 2.

Note that for both groups we inserted a row of data giving information at time 0. We will explain this insertion when we get to the third column in the table.

The *second column* in the data layout gives frequency counts, denoted by m_f, of those persons who failed at each distinct failure time. When there are no ties at a certain failure time, then $m_f = 1$. Notice that in group 1, there were three ties at 6 weeks but no ties thereafter. In group 2, there were ties at 1, 2, 4, 5, 8, 11, and 12 weeks. In any case, the sum of all the m_f's in this column gives the total number of failures in the group tabulated. This sum is 9 for group 1 and 21 for group 2.

EXAMPLE: (continued)

q_j = censored in $[t_{(j)}, t_{(j+1)}]$

Remission Data: Group 1

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$ (ties)	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Remission Data: Group 1

#	t(weeks)	d	X(group)
1	6	1	1
2	6	1	1
3	6	1	1
4	7	1	1
5	10	1	1
6	13	1	1
7	16	1	1
8	22	1	1
9	23	1	1
10	6	0	1
11	9	0	1
12	10	0	1
13	11	0	1
14	17	0	1
15	19	0	1
16	20	0	1
17	25	0	1
18	32	0	1
19	32	0	1
20	34	0	1
21	35	0	1

The *third column* gives frequency counts, denoted by q_f, of those persons censored in the time interval starting with failure time $t_{(f)}$ up to the next failure time denoted $t_{(f+1)}$. Technically, because of the way we have defined this interval in the table, we include those persons censored at the beginning of the interval.

For example, the remission data, for group 1 includes 5 nonzero q's: $q_1 = 1$, $q_2 = 1$, $q_3 = 2$, $q_5 = 3$, $q_7 = 5$. Adding these values gives us the total number of censored observations for group 1, which is 12. Moreover, if we add the total number of q's (12) to the total number of m's (9), we get the total number of subjects in group 1, which is 21.

We now focus on group 1 to look a little closer at the q's. At the left, we list the unordered group 1 information followed (on the next page) by the ordered failure time information. We will go back and forth between these two tables (and pages) as we discuss the q's. Notice that in the table here, one person, listed as #10, was censored at week 6. Consequently, in the table at the top of the next page, we have $q_1 = 1$, which is listed on the second line corresponding to the ordered failure time $t_{(1)}$, which equals 6.

The next q is a little trickier, it is derived from the person who was listed as #11 in the table here and was censored at week 9. Correspondingly, in the table at the top of the next page, we have $q_2 = 1$ because this one person was censored within the time interval that starts at the second ordered failure time, 7 weeks, and ends just before the third ordered failure time, 10 weeks. We have *not* counted person #12 (who was censored at week 10) here because this person's censored time is exactly at the end of the interval. We count this person in the following interval.

EXAMPLE: (continued)

Group 1 using ordered failure times

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

We now consider, from the table of unordered failure times, person #12 who was censored at 10 weeks, and person #13, who was censored at 11 weeks. Turning to the table of ordered failure times, we see that these two times are within the third ordered time interval, which starts and includes the 10-week point and ends just before the 13th week. As for the remaining q's, we will let you figure them out for practice.

One last point about the q information. We inserted a row at the top of the data for each group corresponding to time 0. This insertion allows for the possibility that persons may be censored after the start of the study but before the first failure. In other words, it is possible that q_0 may be nonzero. For the two groups in this example, however, no one was censored before the first failure time.

EXAMPLE

Risk Set: $R(t_f)$ is the set of individual for whom

Remission Data: Group 1

$t_{(f)}$	$m_{(f)}$	$q_{(f)}$	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

The last column in the table gives the "**risk set**." The risk set is not a numerical value or count but rather a collection of individuals. By definition, the risk set $R(t_{(f)})$ is the collection of individuals who have survived at least to time $t_{(f)}$; that is, each person in $R(t_{(f)})$ has a survival time that is $t_{(f)}$ or longer, regardless of whether the person has failed or is censored.

For example, we see that at the start of the study everyone in group 1 survived at least 0 weeks, so the risk set at time 0 consists of the entire group of 21 persons. The risk set at 6 weeks for group 1 also consists of all 21 persons, because all 21 persons survived at least as long as 6 weeks. These 21 persons include the 3 persons who failed at 6 weeks, because they survived and were still at risk just up to this point.

EXAMPLE: (continued)

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
$t_{(0)} = 0$	0	0	21 persons survive ≥ 0 wks
$t_{(1)} = 6$	3	1	21 persons survive ≥ 6 wks
$t_{(2)} = 7$	1	1	17 persons survive ≥ 7 wks
$t_{(3)} = 10$	1	2	15 persons survive ≥ 10 wks
$t_{(4)} = 13$	1	0	12 persons survive ≥ 13 wks
$t_{(5)} = 16$	1	3	11 persons survive ≥ 16 wks
$t_{(6)} = 22$	1	0	7 persons survive ≥ 22 wks
$t_{(7)} = 23$	1	5	6 persons survive ≥ 23 wks
Totals	9	12	

Now let's look at the risk set at 7 weeks. This set consists of 17 persons in group 1 that survived at least 7 weeks. We omit everyone in the X-ed area. Of the original 21 persons, we therefore have excluded the three persons who failed at 6 weeks and the one person who was censored at 6 weeks. These four persons did not survive at least 7 weeks. Although the censored person may have survived longer than 7 weeks, we must exclude him or her from the risk set at 7 weeks because we have information on this person only up to 6 weeks.

To derive the other risk sets, we must exclude all persons who either failed or were censored before the start of the time interval being considered. For example, to obtain the risk set at 13 weeks for group 1, we must exclude the five persons who failed before, but not including, 13 weeks and the four persons who were censored before, but not including, 13 weeks. Subtracting these 9 persons from 21, leaves 12 persons in group 1 still at risk for getting the event at 13 weeks. Thus, the risk set consists of these 12 persons.

How we work with censored data: Use all information up to time of censorship; don't throw away information.

The importance of the table of ordered failure times is that we can work with censored observations in analyzing survival data. Even though censored observations are incomplete, in that we don't know a person's survival time exactly, we can still make use of the information we have on a censored person up to the time we lose track of him or her. Rather than simply throw away the information on a censored person, we use all the information we have on such a person up until time of censorship. (Nevertheless, most survival analysis techniques require a key assumption that censoring is independent, i.e., censored subjects are not at increased risk for failure. See Chap. 9 on competing risks for further details.)

EXAMPLE

$t_{(f)}$	m_f	q_f		$R(t_{(f)})$
6	3	1	✓	21 persons
7	1	1	✓	17 persons
10	1	2	✓	15 persons
13	1	0	✓	12 persons
16	1	③	✓	11 persons
22	1	0		7 persons
23	1	5		6 persons

For example, for the three persons in group 1 who were censored between the 16th and 22nd weeks, there are at least 16 weeks of survival information on each that we don't want to lose. These three persons are contained in all risk sets up to the 16th week; that is, they are each at risk for getting the event up to 16 weeks. Any survival probabilities determined before, and including, 16 weeks should make use of data on these three persons as well as data on other persons at risk during the first 16 weeks.

Having introduced the basic terminology and data layouts to this point, we now consider some data analysis issues and some additional applications.

VII. Descriptive Measures of Survival Experience

EXAMPLE

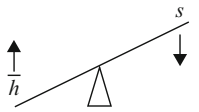
Remission times (in weeks) for two groups of leukemia patients

Group 1 (Treatment) $n = 21$	Group 2 (Placebo) $n = 21$
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23
$\overline{T}_1$ (ignoring + 's) = 17.1	$\overline{T}_2 = 8.6$
$\overline{h}_1 = \frac{9}{359} = .025$	$\overline{h}_2 = \frac{21}{182} = .115$

$$\text{Average hazard rate } (\overline{h}) = \frac{\#\text{failures}}{\sum_{i=1}^{n} t_i}$$

We first return to the remission data, again shown in untabulated form. Inspecting the survival times given for each group, we can see that most of the treatment group's times are longer than most of the placebo group's times. If we ignore the plus signs denoting censorship and simply average all 21 survival times for each group we get an average, denoted by $\boldsymbol{T}$ "bar," of **17.1** weeks survival for the treatment group and **8.6** weeks for the placebo group. Because several of the treatment group's times are censored, this means that group 1's true average is even larger than what we have calculated. Thus, it appears from the data (without our doing any mathematical analysis) that, regarding survival, the treatment is more effective than the placebo.

As an alternative to the simple averages that we have computed for each group, another descriptive measure of each group is the **average hazard rate**, denoted as $\boldsymbol{h}$ "bar." This rate is defined by dividing the total number of failures by the sum of the observed survival times. For group 1, h "bar" is 9/359, which equals **.025**. For group 2, h "bar" is 21/182, which equals. **115**.



As previously described, the hazard rate indicates failure potential rather than survival probability. Thus, the higher the average hazard rate, the lower is the group's probability of surviving.

In our example, the average hazard for the treatment group is smaller than the average hazard for the placebo group.

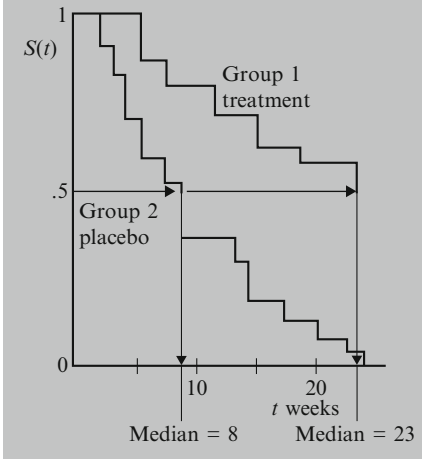
Placebo hazard > treatment hazard: suggests that treatment is more effective than placebo

Thus, using average hazard rates, we again see that the treatment group appears to be doing better overall than the placebo group; that is, the treatment group is less prone to fail than the placebo group.

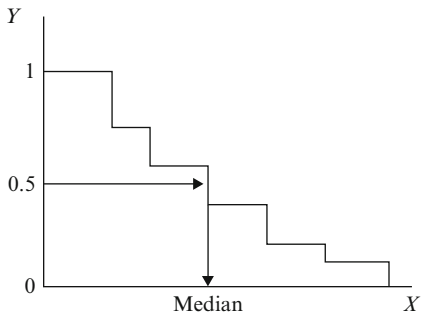
Descriptive measures ($\overline{T}$ and $\overline{h}$) give **overall** comparison; they do not give comparison over time.

The descriptive measures we have used so far—the ordinary average and the hazard rate average—provide overall comparisons of the treatment group with the placebo group. These measures don't compare the two groups at different points in time of follow-up. Such a comparison is provided by a graph of survivor curves.

EXAMPLE



Here we present the **estimated survivor curves** for the treatment and placebo groups. The method used to get these curves is called the Kaplan–Meier method, which is described in Chap. 2. When estimated, these curves are actually step functions that allow us to compare the treatment and placebo groups over time. The graph shows that the survivor function for the treatment group consistently lies above that for the placebo group; this difference indicates that the treatment appears effective at all points of follow-up. Notice, however, that the two functions are somewhat closer together in the first few weeks of follow-up, but thereafter are quite spread apart. This widening gap suggests that the treatment is more effective later during follow-up than it is early on.



Median (treatment) = 23 weeks
Median (placebo) = 8 weeks

Also notice from the graph that one can obtain estimates of the median survival time, the time at which the survival probability is.5 for each group. Graphically, the median is obtained by proceeding horizontally from the 0.5 point on the *Y*-axis until the survivor curve is reached, as marked by an arrow, and then proceeding vertically downward until the *X*-axis is crossed at the median survival time.

For the treatment group, the median is 23 weeks; for the placebo group, the median is 8 weeks. Comparison of the two medians reinforces our previous observation that the treatment is more effective overall than the placebo.

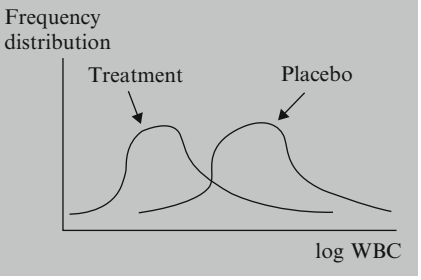
VIII. Example: Extended Remission Data

Group 1		Group 2	
t (weeks)	log WBC	*t* (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

Before proceeding to another data set, we consider the remission example data (Freireich et al., *Blood*, 1963) in an **extended form**. The table at the left gives the remission survival times for the two groups with additional information about white blood cell count for each person studied. In particular, each person's log white blood cell count is given next to that person's survival time. The epidemiologic reason for adding log WBC to the data set is that this variable is usually considered an important predictor of survival in leukemia patients; the higher the WBC, the worse the prognosis. Thus, any comparison of the effects of two treatment groups needs to consider the possible **confounding effect** of such a variable.

EXAMPLE: CONFOUNDING

Treatment group: $\overline{\log \text{WBC}} = 1.8$
Placebo group: $\overline{\log \text{WBC}} = 4.1$
Indicates **confounding** of treatment effect by log WBC



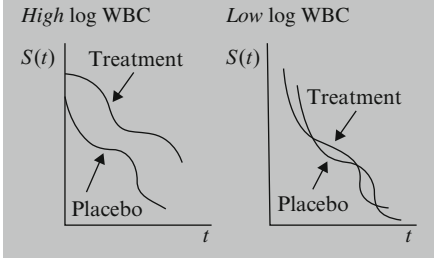
Although a full exposition of the nature of confounding is not intended here, we provide a simple scenario to give you the basic idea. Suppose all of the subjects in the treatment group had very low log WBC, with an average, for example, of 1.8, whereas all of the subjects in the placebo group had very high log WBC, with an average of 4.1. We would have to conclude that the results we've seen so far that compare treatment with placebo groups may be misleading.

The additional information on log WBC would suggest that the treatment group is surviving longer simply because of their low WBC and not because of the efficacy of the treatment itself. In this case, we would say that **the treatment effect is confounded by the effect of log WBC**.

Need to adjust for imbalance in the distribution of log WBC

More typically, the distribution of log WBC may be quite different in the treatment group than in the control group. We have illustrated one extreme in the graph at the left. Even though such an extreme is not likely, and is not true for the data given here, the point is that some attempt needs to be made to adjust for whatever imbalance there is in the distribution of log WBC. However, if high log WBC count was a consequence of the treatment, then white blood cell count should not be controlled for in the analysis.

EXAMPLE: INTERACTION



Treatment by log WBC interaction

Another issue to consider regarding the effect of log WBC is **interaction**. What we mean by interaction is that the effect of the treatment may be different, depending on the level of log WBC. For example, suppose that for persons with high log WBC, survival probabilities for the treatment are consistently higher over time than for the placebo. This circumstance is illustrated by the first graph at the left. In contrast, the second graph, which considers only persons with low log WBC, shows no difference in treatment and placebo effect over time. In such a situation, we would say that **there is strong treatment by log WBC interaction**, and we would have to qualify the effect of the treatment as depending on the level of log WBC.

Need to consider:

- interaction;
- confounding.

The example of interaction we just gave is but one way interaction can occur; on the other hand, interaction may not occur at all. As with confounding, it is beyond our scope to provide a thorough discussion of interaction. In any case, the assessment of interaction is something to consider in one's analysis in addition to confounding that involves explanatory variables.

The problem:
Compare two groups after adjusting for confounding and interaction.

Thus, with our extended data example, the basic **problem** can be described as follows: to compare the survival experience of the two groups after adjusting for the possible confounding and/or interaction effects of log WBC.

EXAMPLE

	Individual #	t (weeks)	d	X_1 (Group)	X_2 (log WBC)
Group 1	1	6	1	1	2.31
	2	6	1	1	4.06
	3	6	1	1	3.28
	4	7	1	1	4.43
	5	10	1	1	2.96
	6	13	1	1	2.88
	7	16	1	1	3.60
	8	22	1	1	2.32
	9	23	1	1	2.57
	10	6	0	1	3.20
	11	9	0	1	2.80
	12	10	0	1	2.70
	13	11	0	1	2.60
	14	17	0	1	2.16
	15	19	0	1	2.05
	16	20	0	1	2.01
	17	25	0	1	1.78
	18	32	0	1	2.20
	19	32	0	1	2.53
	20	34	0	1	1.47
	21	35	0	1	1.45
Group 2	22	1	1	0	2.80
	23	1	1	0	5.00
	24	2	1	0	4.91
	25	2	1	0	4.48
	26	3	1	0	4.01
	27	4	1	0	4.36
	28	4	1	0	2.42
	29	5	1	0	3.49
	30	5	1	0	3.97
	31	8	1	0	3.52
	32	8	1	0	3.05
	33	8	1	0	2.32
	34	8	1	0	3.26
	35	11	1	0	3.49
	36	11	1	0	2.12
	37	12	1	0	1.50
	38	12	1	0	3.06
	39	15	1	0	2.30
	40	17	1	0	2.95
	41	22	1	0	2.73
	42	23	1	0	1.97

The problem statement tells us that we are now considering two explanatory variables in our extended example, whereas we previously considered a single variable, group status. The data layout for the computer needs to reflect the addition of the second variable, log WBC. The extended table in computer layout form is given at the left. Notice that we have labeled the two explanatory variables X_1 (for group status) and X_2 (for log WBC). The variable X_1 is our primary study or exposure variable of interest here, and the variable X_2 is an extraneous variable that we are interested in accounting for because of either confounding or interaction.

Analysis alternatives:

- stratify on log WBC;
- use math modeling, e.g., proportional hazards model.

As implied by our extended example, which considers the possible confounding or interaction effect of log WBC, we need to consider methods for adjusting for log WBC and/or assessing its effect in addition to assessing the effect of treatment group. The two most popular alternatives for analysis are the following:

- to stratify on log WBC and compare survival curves for different strata; or
- to use mathematical modeling procedures such as the proportional hazards or other survival models; such methods will be described in subsequent chapters.

IX. Multivariable Example

- Describes general multivariable survival problem.
- Gives analogy to regression problems.

We now consider one other example. Our purpose here is to describe a more general type of multivariable survival analysis problem. The reader may see the analogy of this example to multiple regression or even logistic regression data problems.

EXAMPLE

13-year follow-up of fixed cohort from Evans County, Georgia

$n = 170$ white males (60+)

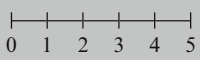
T = years until death

Event = death

Explanatory variables:
- exposure variable
- confounders
- interaction variables

Exposure:

Social Network Index (**SNI**)



We consider a data set developed from a 13-year follow up study of a fixed cohort of persons in Evans County Georgia, during the period 1967–1980 (Schoenbach et al., *Amer. J. Epid.*, 1986). From this data set, we focus on a portion containing $n = 170$ white males who are age 60 or older at the start of follow-up in 1967.

For this data set, the outcome variable is T, time in years until death from start of follow-up, so the event of interest is **death**. Several explanatory variables are measured, one of which is considered the primary exposure variable; the other variables are considered as potential confounders and/or interaction variables.

The primary exposure variable is a measure called Social Network Index (SNI). This is an ordinal variable derived from questionnaire measurement and is designed to assess the extent to which a study subject has social contacts of various types. With the questionnaire, a scale is used with values ranging from 0 (absence of any social network) to 5 (excellent social network).

EXAMPLE: (continued)

Study goal: to determine whether **SNI** is protective against death,
i.e., **SNI** ↗ ⇒ S(t) ↗.

Explanatory variables:

SNI — Exposure variable

AGE, **SBP**, **CHR**, **QUET**, **SOCL** — Potential confounders/ interaction variables

Note: $\text{QUET} = \dfrac{\text{Weight}}{(\text{height})} \times 100$

The problem:

To describe the relationship between **SNI** and time to death, after controlling for **AGE, SBP, CHR, QUET**, and **SOCL**.

Goals:

- Measure of effect (adjusted)
- Survivor curves for different SNI categories (adjusted)
- Decide on variables to be adjusted; determine method of adjustment

The study's goal is to determine whether one's social network, as measured by SNI, is protective against death. If this study hypothesis is correct, then the higher the social network score, the longer will be one's survival time.

In evaluating this problem, several explanatory variables, in addition to SNI, are measured at the start of follow-up. These include AGE, systolic blood pressure (SBP), an indicator of the presence or absence of some chronic disease (CHR), body size as measured by Quetelet's index (QUET = weight over height squared times 100), and social class (SOCL).

These five additional variables are of interest because they are thought to have their own special or collective influence on how long a person will survive. Consequently, these variables are viewed as potential confounders and/ or interaction variables in evaluating the effect of social network on time to death.

We can now clearly state the problem being addressed by this study: To describe the relationship between SNI and time to death, controlling for AGE, SBP, CHR, QUET, and SOCL.

Our goals in using survival analysis to solve this problem are as follows:

- to obtain some measure of effect that will describe the relationship between SNI and time until death, after adjusting for the other variables we have identified;
- to develop survival curves that describe the probability of survival over time for different categories of social networks; in particular, we wish to compare the survival of persons with excellent networks to the survival of persons with poor networks. Such survival curves need to be adjusted for the effects of other variables.
- to achieve these goals, two intermediary goals are to decide which of the additional variables being considered need to be adjusted and to determine an appropriate method of adjustment.

The computer data layout for this problem is given below. The first column lists the 170 individuals in the data set. The second column lists the survival times, and the third column lists failure or censored status. The remainder of the columns list the 6 explanatory variables of interest, starting with the exposure variable SNI and continuing with the variables to be accounted for in the analysis.

Computer layout: 13-year follow-up study (1967–1980) of a fixed cohort of $n = 170$ white males (60+) from Evans County, Georgia

#	t	d	SNI	AGE	SBP	CHR	QUET	SOCL
1	t_1	d_1	SNI_1	AGE_1	SBP_1	CHR_1	$QUET_1$	$SOCL_1$
2	t_2	d_2	SNI_2	AGE_2	SBP_2	CHR_2	$QUET_2$	$SOCL_2$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
170	t_{170}	d_{170}	SNI_{170}	AGE_{170}	SBP_{170}	CHR_{170}	$QUET_{170}$	$SOCL_{170}$

X. Math Models in Survival Analysis

General framework

Controlling for $\boldsymbol{C}_1, \boldsymbol{C}_2, \ldots \boldsymbol{C}_p$.

SNI study:

$E = \text{SNI} \Rightarrow D = \text{survival time}$

Controlling for **AGE, SBP, CHR, QUET**, and **SOCL**

It is beyond the scope of this presentation to provide specific details of the survival analysis of these data. Nevertheless, the problem addressed by these data is closely analogous to the typical multivariable problem addressed by linear and logistic regression modeling. Regardless of which modeling approach is chosen, the typical problem concerns describing the relationship between an exposure variable (e.g., E) and an outcome variable (e.g., D) after controlling for the possible confounding and interaction effects of additional variables (e.g., C_1, C_2, and so on up to C_p). In our survival analysis example, E is the social network variable SNI, D is the survival time variable, and there are $p = 5$ C variables, namely, AGE, SBP, CHR, QUET, and SOCL.

	Model	Outcome
follow-up time info not used	Survival analysis	Time to event (with censoring)
	Linear regression	Continuous (SBP)
	Logistic regression	Dichotomous (CHD yes/no)

Nevertheless, an important distinction among modeling methods is the type of outcome variable being used. In survival analysis, the outcome variable is "time to an event," and there may be censored data. In linear regression modeling, the outcome variable is generally a continuous variable, like blood pressure. In logistic modeling, the outcome variable is a dichotomous variable, like CHD status, yes or no. And with linear or logistic modeling, we usually do not have information on follow-up time available.

As with linear and logistic modeling, one statistical goal of a survival analysis is to obtain some measure of effect that describes the exposure–outcome relationship adjusted for relevant extraneous variables.

Measure of effect:

Linear regression:
regression coefficient β

In linear regression modeling, the measure of effect is usually some regression coefficient β.

Logistic regression:
odds ratio e^{β}

In logistic modeling, the measure of effect is an odds ratio expressed in terms of an exponential of one or more regression coefficients in the model, for example, e to the β.

Survival analysis:
hazard ratio e^{β}

In survival analysis, the measure of effect typically obtained is called a **hazard ratio**; as with the logistic model, this hazard ratio is expressed in terms of an exponential of one or more regression coefficients in the model.

EXAMPLE

SNI study: hazard ratio (HR) describes relationship between SNI and T, after controlling for covariates.

Thus, from the example of survival analysis modeling of the social network data, one may obtain a hazard ratio that describes the relationship between SNI and survival time (T), after controlling for the appropriate covariates.

Interpretation of HR (like OR):

HR = 1 ⇒ no relationship

HR = 10 ⇒ exposed hazard 10 times unexposed

HR = 1/10 ⇒ exposed hazard 1/10 times unexposed

The hazard ratio, although a different measure from an odds ratio, nevertheless has a similar interpretation of the strength of the effect. A hazard ratio of 1, like an odds ratio of 1, means that there is no effect; that is, 1 is the null value for the exposure–outcome relationship. A hazard ratio of 10, on the other hand, is interpreted like an odds ratio of 10; that is, the exposed group has ten times the hazard of the unexposed group. Similarly, a hazard ratio of 1/10 implies that the exposed group has one-tenth the hazard of the unexposed group.

XI. Censoring Assumptions

Three assumptions about censoring:

Independent (vs. non-independent) censoring
Random (vs. non-random) censoring
Non-informative (vs. informative) censoring

There are three assumptions about censoring often considered for survival data: **independent censoring**, **random censoring**, and **non-informative censoring**. Although these assumptions have similarities, they are also somewhat different and are often confused in the textbook and published literature as being interchangeable synonyms.

Mathematic definitions have been provided elsewhere.

Mathematical definitions of independent (vs. nonindependent), random (vs. nonrandom), and non-informative (vs. informative) censoring have been given elsewhere (Kalbfleisch and Prentice, 1980; Klein and Moeschberger, 2003). Here, however, we prefer to provide more intuitive definitions and examples.

Independent (vs. non-independent) censoring
- most useful
- affects validity

The assumption of independent censoring is the most useful of the three types for drawing correct inferences that compare the survival experience of two or more groups (e.g., treatment vs. placebo). In particular, the presence of non-independent censoring typically affects the validity of one's estimated effect. Random censoring is a stronger assumption and more restrictive than independent censoring.

Random (vs. non-random) censoring
- more restrictive than independent,
  i.e., random ⇒ indep,
  whereas indep ⇏ random.

Random Censoring:

	Failure rate
Censored	Not censored
$h_{Ce}(t) = h_{NCe}(t)$	

To be more specific, **random censoring** essentially means that *subjects who are censored at time t should be representative of all the study subjects who remained at risk at time t with respect to their survival experience.* In other words, the failure rate for subjects who are censored is assumed to be equal to the failure rate for subjects who remained in the risk set who are not censored.

Independent censoring:

	Failure rate	
Subgrp	Censored	Not censored
A	$h_{A,Ce}(t) = h_{A,NCe}(t)$	
B	$h_{B,Ce}(t) = h_{B,NCe}(t)$	

Independent censoring essentially means that *within any subgroup of interest, the subjects who are censored at time t should be representative of all the subjects in that subgroup who remained at risk at time t with respect to their survival experience.* In other words, censoring is independent provided that it is random within any subgroup of interest. We illustrate these ideas with an example.

EXAMPLE

Group A

Time	# at risk	# events	# survived
0–3 yrs	100	20	80

3-yr risk = 20/100 = 0.20
3-yr survival = 80/100 =0.80

Suppose that we are interested in estimating the 3-year survival (for some disease) among those in Group A. We follow 100 individuals intially disease free for 3 years. Over the 3-year period, 20 contract disease. We estimate the 3-year risk of disease for those in Group A to be 0.20 and the 3-year survival to be 0.80 (since 80 of 100 survived).

Time	# at risk	# events	# survived
0–3	100	20	80
	40 leave study		
3–5	40	5	35
	5-year survival?		

Now suppose we wish to continue the study for another 2 years in order to estimate the 5-year survival for Group A. We want to continue the following for the 80 individuals who participated in the study and survived for the first 3 years. However, half or 40 of those 80 individuals refused to continue in the study and were therefore lost to follow-up (censored). Of the other 40 individuals who remained in the study, 5 individuals contract the disease. With this information, what is the estimate of the 5-year survival for Group A and under what assumptions?

EXAMPLE: (continued)

What happened to 40 individuals who were censored at 3 years? *Don't know*

Assuming indep and random censoring:

40 at risk at time 5
similar to
40 censored at time 3

i.e.,
expect 5 events from 40 censored at time 3
since 5 events from 40 at risk

Estimated # of cases over 5 years:

20 + 5 + 5
first 3 years next 2 years censored cases

= 30 estimated cases from original 100 over 5 years

Estimated 5-year survival = 70/100 = 0.70

If we know what happened to the 40 individuals who were censored at the 3-year mark, then we could sum the total number of events and the total number of individuals who survived (out of the original 100 at risk). **Under an assumption of independent and random censoring, we assume that the 40 individuals who were censored were similar to the 40 who remained at risk with respect to their survival experience.** Since 5 of the 40 who remained in the study after 3 years contracted disease over the next 2 years, we estimate that 5 of the 40 who were censored also contracted the disease over the same time period, even though their disease experience was unobserved.

So over the course of the 5 years: 20 contracted disease in the first 3 years, 5 were observed to get disease after 3 years, and 5 of the censored individuals were estimated to have contracted disease. This yields 20 + 5 + 5 = 30 who are estimated to have contract disease leaving 70 of the original 100 who have survived over the 5-year period. The estimated 5-year survival among Group A is 0.70 under the assumptions of random and independent censoring.

The idea:
Assume survival experience of subjects censored at t is *as expected* if randomly selected from subjects who are at risk at t.

The idea behind independent and random censoring is that it is *as if* the subjects censored at time t were randomly selected to be censored from the group of subjects who were in the risk set at time t. Even though the censored subjects were not randomly selected, their survival experience would be expected to be the same as if they had been randomly selected from the risk set at time t.

So far, there is no distinction between independent and random censoring.

Reason: Only considering one group

In this example, there is no distinction made between independent censoring and random censoring. The reason is because we are only considering one group of individuals (i.e., there are no predictor variables considered). The distinction comes if we consider more than one group for comparison. We illustrate this distinction by continuing the example.

EXAMPLE: (continued)

Group B

Time	# at risk	# events	# survived
0–3	100	40	60
	10 leave study		
3–5	50	10	40

Failure risk from 3 to 5 yrs $= 10/50 =$ **0.20**.

We extend the previous example to include 100 subjects from Group B who are disease free at the start of follow-up. The goal is to estimate their 5-year survival and compare it to the 5-year survival for Group A. Suppose over the first 3-year period, 40 of the 100 individuals contract disease. Then, of the 60 who survive the first 3 years, 10 refuse to continue in the study and are therefore censored. For the 50 who remain in the study, 10 individuals contract disease by the 5^{th} year (10 of 50 = 20%).

Assuming independent censoring:
expect $\mathbf{0.20} \times 10 = 2$ cases from 10 censored at time 3

Estimated # of cases over 5 years:

$$\underset{\text{first 3 years}}{40} + \underset{\text{next 2 years}}{40} + \underset{\text{censored cases}}{2}$$

= 52 estimated cases from original 100 over 5 years

Estimated 5-year survival = 48/100 = 0.48

Under independent censoring, we estimate that 20% or 2 of the 10 censored subjects contract disease by the 5^{th} year.

So, over the course of the 5 years among the original 100 in Group B: 40 contracted disease in the first 3 years, 10 were observed to get disease after 3 years, and 2 of the censored individuals were estimated to the contracted disease. This yields 40 + 10 + 2 = 52 who are estimated to have contracted disease, leaving 48 of the original 100 who survived over the 5-year period. The estimated 5-year survival among Group B is 0.48 under the assumptions of independent censoring.

Groups A and B combined

Time	# at risk			# events			# survived		
	A	B	total	A	B	total	A	B	total
0–3	100	100	**200**	20	40	**60**	80	60	**140**
	40 from **A** and **10** from **B** leave study								
3–5	40	50	90	5	10	15	35	40	75

$$\begin{cases} p_A(\text{censored}) = 40/80 = 0.50 \text{ or } 50\% \\ p_B(\text{censored}) = 10/60 = 0.17 \text{ or } 17\% \\ p_A(\text{censored}) \gg p_B(\text{censored}) \end{cases}$$

	Group A	Group B
5-yr survival	0.70	0.48

$\Downarrow$

Censoring not random

Over all, combining both groups, there were 200 originally at risk of whom 60 contracted disease within the first 3 years (20 from Group A and 40 from Group A) leaving 140 who have survived for the first 3 years (80 from Group A and 60 from Group B). At the 3-year mark, 50 subjects were censored (40 in Group A and 10 in Group B).

A much **higher proportion of censoring occurred in Group A** (i.e., 40/80 = 0.50) **than in Group B** (i.e., 10/60 = 0.17).

Moreover, subjects in **Group A** had a higher survival probability than those in **Group B**.

Therefore, the censoring was not random.

EXAMPLE (continued)

Random censoring
within **Group A** *and* within **Group B**
⇓
Independent censoring
(i.e., random censoring conditional on covariates)

Nevertheless,
(overall) random censoring not met

However, conditional on each level of covariates (conditional on group status in this example), the censoring was random. Therefore, the censoring was independent because **independent censoring is random censoring conditional on each level of covariates**.

Nevertheless, the random censoring assumption was not met overall because the censored individuals were not representative of all who remained in the risk set at time t with respect to the rate of failure.

ALTERNATIVE EXAMPLE

Time	# at risk			# events			# survived		
	A	B	total	A	B	total	A	B	total
0–3	100	100	200	20	40	60	80	60	140
40 from **A** and **30** from **B** leave study									
3–5	40	30	70	5	10	15	35	20	55

$p_{\mathbf{A}}(\text{censored}) = 40/80 = 0.50$ or 50%
$p_{\mathbf{B}}(\text{censored}) = 30/60 = 0.50$ or 50%

$p_{\mathbf{A}}(\text{censored}) = p_{\mathbf{B}}(\text{censored})$
⇓
Random censoring (overall)

If instead in the previous example, suppose **40** of 80 from Group A and **30** of 60 from Group B were censored at the 3-year mark, as shown in the table on the left.

Then the censoring would have been random because an equal proportion of Group A and Group B would have been censored from the risk set and those censored would be representative of those who remained at risk.

Non-informative censoring depends on
- distribution of time-to-event
- distribution of time-to-censorship

We next consider the assumption of **non-informative censoring**. Whether censoring is non-informative or informative depends on two distributions: (1) the distribution of the time-to-event random variable and (2) the distribution of time-to-censorship random variable.

Time-to-event random variable (T):
Distribution of survival times assuming:
- no loss-to-follow-up
- study continues until all subjects get event

We can conceptualize the distribution for the time-to-event random variable by considering the distribution of survival times if there was no loss to follow-up and the study did not end until all subjects got the event.

Time-to-censorship random variable (C):
Distribution of censoring times assuming:
- study ends before all subjects get event
- censored subjects do not get event prior to the end of study

Similarly, we can conceptualize a time-to-censorship random variable by considering the distribution of censoring times for those subjects who would not have gotten the event if the study ended before all subjects got the event.

Non-informative censoring:

$$T \text{ distribution} \overset{\text{no information}}{\nLeftrightarrow} C \text{ distribution}$$

Note: must still need to know which subjects are censored or not censored.

Non-informative
Independent
Random

- Often all justifiable together
- Not all equivalent

Non-informative censoring occurs if the distribution of survival times (T) provides no information about the distribution of censorship times (C), and vice versa. Otherwise, the censoring is **informative**. Note, however, that the data must still identify which subjects are or are not censored.

The assumption of non-informative censoring is often justifiable when censoring is independent and/or random; nevertheless, these three assumptions are not equivalent.

EXAMPLE: Independent and random but informative censoring

Subject A gets event
⇓
Subject B (randomly selected) gets event, e.g., family member of Subject A leaves study

Assume: censored subjects represent subjects at risk at any time

Then

- independent and random censoring
- **informative censoring** since T ⇒ C. (i.e., T distribution specifies C distribution)

To illustrate how independent censoring could be different from non-informative censoring, we describe an artificial example where the censoring is informative but also random and independent.

Suppose every time an individual gets an event, another individual in the study is randomly selected to leave the study, e.g., after an event, a family member decides to leave the study. If those censored were representative of those who remained in the risk set, then the censoring would be random and independent. However, the censoring would also be informative as **the censoring mechanism would be related to the time-to-event distribution** (since events cause censorships). In fact, if this was the only mechanism in which individuals were censored, the distribution of survival times would completely specify the distribution of censoring times (highly informative).

EXAMPLE: Not independent censoring

- Drug side effect causes censoring
- Censored subjects not representative of subjects still at risk
- Censored subjects more vulnerable than subjects still at risk

⇓

Assuming independent censoring would overestimate survival

To see how bias can occur if the censoring is not independent, consider a drug study in which some individuals are censored from the study due to the occurrence of some side effects. It may be that the unobserved survival experience among those who are censored due to a drug side effect is not representative of those who remained in the study. If those with a side effect are more vulnerable to the health outcome, then we would likely overestimate their survival with an assumption of independent censoring.

Independent censoring most relevant: affects validity

Many of the analytic techniques discussed in the chapters that follow Kaplan–Meier survival estimation, the log rank test, and the Cox model, rely on an assumption of independent censoring for valid inference in the presence of right-censored data.

Chapters

✓ 1. Introduction

2. Kaplan–Meier Survival Curves and the Log–Rank Test

This presentation is now complete. We suggest that you review the material covered here by reading the detailed outline that follows. Then do the practice exercises and test.

In Chap. 2 we describe how to estimate and graph survival curves using the Kaplan–Meier (KM) method. We also describe how to test whether two or more survival curves are estimating a common curve. The most popular such test is called the log–rank test.

Detailed Outline

I. What is survival analysis? (pages 4–5)

A. Type of problem addressed: outcome variable is **time until an event occurs**.

B. Assume one event of interest; more than one type of event implies **a competing risk** problem.

C. Terminology: time = survival time; event = failure.

D. Examples of survival analysis:

i. leukemia patients/time in remission

ii. disease-free cohort/time until heart disease

iii. elderly population/time until death

iv. parolees/time until rearrest (recidivism)

v. heart transplants/time until death

II. Censored data (pages 5–8)

A. Definition: don't know survival time exactly.

B. Typical reasons: study ends, loss to follow-up, withdrawal from study.

C. Example that illustrates (right-) censoring.

D. Right-censoring: true survival time is equal to or greater than observed survival times.

E. Left-censoring: true survival time is less than or equal to observed survival time

F. Interval-censoring: true survival time is within a known time interval (t1, t2)

G. Interval-censoring incorporates right- and left-censoring as special cases, i.e.,

right-censoring $\Rightarrow t_1$ =lower bound, $t_2=\infty$;

left-censoring $\Rightarrow t_1=0$, t_2=upper bound.

III. Terminology and notation (pages 9–15)

A. Notation: T = survival time random variable:

t = specific value for T

d = (0–1) variable for failure/censorship status

B. Terminology: $S(t)$ = survivor function

$h(t)$ = hazard function

C. Properties of survivor function:

- theoretically, graph is smooth curve, decreasing from $S(t) = 1$ at time $t = 0$ to $S(t) = 0$ at $t = \infty$;
- in practice, graph is step function that may not go all the way to zero at end of study if not everyone studied gets the event.

D. Hazard function formula:

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t | T \ge t)}{\Delta t}$$

E. Hazard function properties:
- $h(t)$ gives instantaneous potential for event to occur given survival up to time t;
- instantaneous potential idea is illustrated by velocity;
- hazard function also called "conditional failure rate";
- $h(t) \ge 0$; has no upper bound; not a probability; depends on time units.

F. Examples of hazard curves:
i. exponential
ii. increasing Weibull
iii. decreasing Weibull
iv. log normal

G. Uses of hazard function:
- gives insight about conditional failure rates;
- identifies specific model form;
- math model for survival analysis is usually written in terms of hazard function.

H. Relationship of $S(t)$ to $h(t)$: if you know one, you can determine the other:
- example: $h(t) = \lambda$ if and only if $S(t) = e^{-\lambda t}$
- general formulae:

$$S(t) = \exp\left[-\int_0^t h(u)du\right]$$

$$h(t) = -\left[\frac{d\,S(t)/dt}{S(t)}\right]$$

IV. Goals of survival analysis (page 16)

A. Estimate and interpret survivor and/or hazard functions.

B. Compare survivor and/or hazard functions.

C. Assess the relationship of explanatory variables to survival time.

V. Basic data layout for computer (pages 16–23)

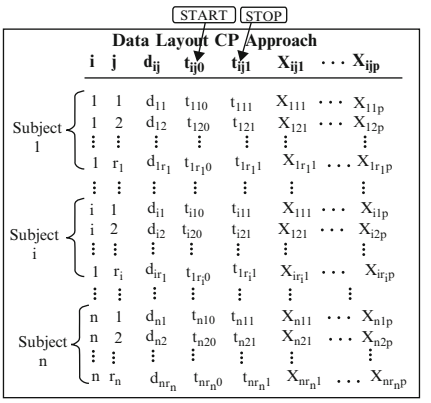
A. General layout:

#	t	d	X_1	$X_2 \ldots X_p$
1	t_1	d_1	X_{11}	$X_{12} \ldots X_{1p}$
2	t_2	d_2	X_{21}	$X_{22} \ldots X_{2p}$
.	.	.		.
.	.	.		.
.	.	.		.
i	t_i	d_i	X_{i1}	$X_{i2} \ldots X_{ip}$
.	.	.		.
.	.	.		.
.	.	.		.
n	t_n	d_n	X_{n1}	$X_{n2} \ldots X_{np}$

B. Example: Remission time data

C. Alternative Data Layout for Computer: Couuting Process (Start, Stop) Format

- Useful for more complicated survival analysis:
  i. Age-at-follow-up as time scale (Chapter 3)
  ii. Time-dependent variables (Chapter 6)
  iii. Recurrent events (Chapter 7)
- CP data layout



- Simplest CP format: 1 dataline subject

i	j	d_{ij}	t_{ij0}	t_{ij1}	X_{ij1}	⋯	X_{ijp}
1	1	d_{11}	0	t_1	X_{111}	⋯	X_{11p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
i	1	d_{i1}	0	t_i	X_{i11}	⋯	X_{i1p}
⋮	⋮	⋮	⋮	⋮	⋮		⋮
n	1	d_{n1}	0	t_n	X_{n11}	⋯	X_{n1p}

- Example from Remission Time Dataset
- Example from Study of Recurrent Bladder Cancer Tumors (Byar, 1980; Wei, Lin, and Weisfeld, 1989)
- Computer Appendix gives programming code

VI. Basic data layout for understanding analysis (pages 23–28)

A. General layout:

Ordered failure times ($t_{(f)}$)	# of failures (m_f)	# censored in $[t_{(f)}, t_{(f+1)})$ (q_f)	Risk set $R(t_{(f)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

Note: k = # of distinct times at which subjects failed; n = # of subjects ($k \leq n$); $R(t_{(f)})$, the risk set, is the set of individuals whose survival times are at least $t_{(f)}$ or larger.

B. Example: Remission time data
Group 1 ($n = 21$, 9 failures, $k = 7$);
Group 2 ($n = 21$, 21 failures, $k = 12$)

C. How to work with censored data:
Use all information up to the time of censorship; don't throw away information.

VII. Descriptive measures of survival experience (pages 28–30)

A. Average survival time (ignoring censorship status):

$$\overline{T} = \frac{\sum_{i=1}^{n} t_i}{n}$$

$\overline{T}$ underestimates the true average survival time, because censored times are included in the formula.

B. Average hazard rate:

$$\bar{h} = \frac{\#\text{ failures}}{\sum_{i=1}^{n} t_i}$$

C. Descriptive measures $\overline{T}$ and $\overline{h}$ give overall comparison; estimated survivor curves give comparison over time.

D. Estimated survivor curves are step function graphs.

E. Median survival time: graphically, proceed horizontally from 0.5 on the *Y*-axis until reaching graph, then vertically downward until reaching the *X*-axis.

VIII. Example: Extended remission data (pages 30–33)

A. Extended data adds log WBC to previous remission data.

B. Need to consider **confounding** and **interaction**.

C. Extended data problem: compare survival experience of two groups, after adjusting for confounding and interaction effects of log WBC.

D. Analysis alternatives:

i. stratify on log WBC and compare survival curves for different strata;

ii. use math modeling, e.g., proportional hazards model.

IX. Multivariable example (pages 33–35)

A. The problem: to describe the relationship between social network index (**SNI**) and time until death, controlling for **AGE**, systolic blood pressure (**SBP**), presence or absence of chronic disease (**CHR**), Quetelet's index (**QUET** – a measure of body size), and social class (**SOCL**).

B. Goals:

- to obtain an adjusted measure of effect;
- to obtain adjusted survivor curves for different SNI categories;
- to decide on variables to be adjusted.

C. The data: 13-year follow-up study (1967–1980) of a fixed cohort of $n = 170$ white males (60+) from Evans County, Georgia.

X. Math models in survival analysis (pages 35–37)

A. Survival analysis problem is analogous to typical multivariable problem addressed by linear and/or logistic regression modeling: describe relationship of exposure to outcome, after accounting for possible confounding and interaction.

B. Outcome variable (time to event) for survival analysis is different from linear (continuous) or logistic (dichotomous) modeling.

C. Measure of effect typically used in survival analysis: hazard ratio (**HR**).

D. Interpretation of HR: like OR. SNI study: **HR** describes relationship between SNI and *T*, after controlling for covariates.

XI. Censoring assumptions (pages 37–43)

A. Three different assumptions about censoring:

i. Independent (vs. Non-independent) censoring

a. most useful- concerns validity of estimated effect

ii. Random (vs. Non-random) censoring

a. more restrictive than independent censoring

iii. Non-informative (vs. Informative) censoring

a. typically affects efficiency of estimated effect

B. Examples.

Practice Exercises

True or False (Circle T or F):

T F 1. In a survival analysis, the outcome variable is dichotomous.

T F 2. In a survival analysis, the event is usually described by a (0, 1) variable.

T F 3. If the study ends before an individual has gotten the event, then his or her survival time is censored.

T F 4. If, for a given individual, the event occurs **before** the person is lost to follow-up or withdraws from the study, then this person's survival time is censored.

T F 5. $S(t) = P(T > t)$ is called the hazard function.

T F 6. The hazard function is a probability.

T F 7. Theoretically, the graph of a survivor function is a smooth curve that decreases from $S(t) = 1$ at $t = 0$ to $S(t) = 0$ at $t = \infty$.

T F 8. The survivor function at time t gives the instantaneous potential per unit time for a failure to occur, given survival up to time t.

T F 9. The formula for a hazard function involves a conditional probability as one of its components.

T F 10. The hazard function theoretically has no upper bound.

T F 11. Mathematical models for survival analysis are frequently written in terms of a hazard function.

T F 12. One goal of a survival analysis is to compare survivor and/or hazard functions.

T F 13. Ordered failure times are censored data.

T F 14. Censored data are used in the analysis of survival data up to the time interval of censorship.

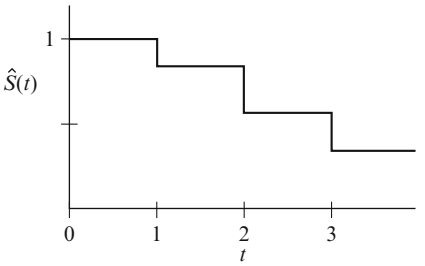
T F 15. A typical goal of a survival analysis involving several explanatory variables is to obtain an adjusted measure of effect.

16. Given the following survival time data (in weeks) for n = 15 subjects,
1, 1, 1+, 1+, 1+, 2, 2, 2, 2+, 2+, 3, 3, 3+, 4+, 5+
where + denotes censored data, complete the following table:

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
0	0	0	15 persons survive $\geq$ 0 weeks
1			
2			
3			

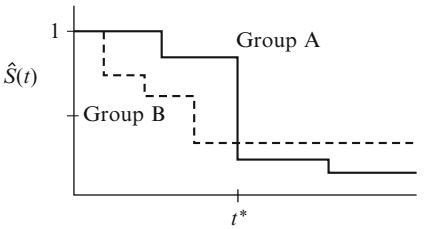
Also, compute the average survival time ($\overline{T}$) and the average hazard rate ($\overline{h}$) using the raw data (ignoring + signs for $\overline{T}$).

17. Suppose that the estimated survivor curve for the above table is given by the following graph:



What is the median survival time for this cohort?

Questions 18–20 consider the comparison of the following two survivor curves:



18. Which group has a better survival prognosis **before** time t^*?
19. Which group has a better survival prognosis **after** time t^*?
20. Which group has a longer median survival time?

Test

True or False (Circle T or F):

T F 1. Survival analysis is a collection of statistical procedures for data analysis for which the outcome variable is **time until an event occurs**.

T F 2. In survival analysis, the term "event" is synonymous with "failure."

T F 3. If a given individual is lost to follow-up or withdraws from the study before the end of the study without the event occurring, then the survival time for this individual is said to be "censored."

T F 4. In practice, the survivor function is usually graphed as a smooth curve.

T F 5. The survivor function ranges between 0 and ∞.

T F 6. The concept of instantaneous potential is illustrated by velocity.

T F 7. A hazard rate of one per day is equivalent to seven per week.

T F 8. If you know the form of a hazard function, then you can determine the corresponding survivor curve, and vice versa.

T F 9. One use of a hazard function is to gain insight about conditional failure rates.

T F 10. If the survival curve for group 1 lies completely above the survival curve for group 2, then the median survival time for group 2 is longer than that for group 1.

T F 11. The risk set at 6 weeks is the set of individuals whose survival times are less than or equal to 6 weeks.

T F 12. If the risk set at 6 weeks consists of 22 persons, and 4 persons fail and 3 persons are censored by the 7th week, then the risk set at 7 weeks consists of 18 persons.

T F 13. The measure of effect used in survival analysis is an odds ratio.

T F 14. If a hazard ratio comparing group 1 relative to group 2 equals 10, then the potential for failure is ten times higher in group 1 than in group 2.

T F 15. The outcome variable used in a survival analysis is different from that used in linear or logistic modeling.

16. State two properties of a hazard function.

17. State three reasons why hazard functions are used.

18. State three goals of a survival analysis.

19. The following data are a sample from the 1967–1980 Evans County study. Survival times (in years) are given for two study groups, each with 25 participants. Group 1 has no history of chronic disease (CHR = 0), and group 2 has a positive history of chronic disease (CHR = 1):

Group 1 (CHR = 0): 12.3+, 5.4, 8.2, 12.2+, 11.7, 10.0, 5.7, 9.8, 2.6, 11.0, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3.0

Group 2 (CHR = 1): 5.8, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 1.8, 3.1, 11.4, 2.4, 1.4, 5.9, 1.6, 2.8, 4.9, 3.5, 6.5, 9.9, 3.6, 5.2, 8.8, 7.8, 4.7, 3.9

For group 1, complete the following table involving ordered failure times:

	$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
Group 1:	0.0	0	0	25 persons survived ≥ 0 years
	1.8	1	0	25 persons survived ≥ 1.8 years
	2.2			
	2.5			
	2.6			
	3.0			
	3.5			
	3.8			
	5.3			
	5.4			
	5.7			
	6.6			
	8.2			
	8.7			
	9.2			
	9.8			
	10.0			
	10.2			
	10.7			
	11.0			
	11.1			
	11.7			

20. For the data of Problem 19, the average survival time ($\overline{T}$) and the average hazard rate ($\overline{h}$) for each group are given as follows:

	$\overline{T}$	$\overline{h}$
Group 1:	7.5	.1165
Group 2:	5.3	.1894

a. Based on the above information, which group has a better survival prognosis? Explain briefly.

b. How would a comparison of survivor curves provide additional information to what is provided in the above table?

Answers to Practice Exercises

1. F: the outcome is continuous; time until an event occurs.
2. T
3. T
4. F: the person fails, i.e., is not censored.
5. F: $S(t)$ is the survivor function.
6. F: the hazard is a rate, not a probability.
7. T
8. F: the hazard function gives instantaneous potential.
9. T
10. T
11. T
12. T
13. F: ordered failure times are data for persons who are failures.
14. T
15. T
16.

$t_{(f)}$	m_f	q_f	$R(t_{(f)})$
0	0	0	15 persons survive ≥ 0 weeks
1	2	3	15 persons survive ≥ 1 weeks
2	3	2	10 persons survive ≥ 2 weeks
3	2	3	5 persons survive ≥ 3 weeks

$$\overline{T} = \frac{33}{15} = 2.2,; \quad \overline{h} = \frac{7}{33} = 0.22$$

17. Median = 3 weeks
18. Group A
19. Group B
20. Group A

2 Kaplan-Meier Survival Curves and the Log-Rank Test

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_2,
© Springer Science+Business Media, LLC 2012

Introduction

We begin with a brief review of the purposes of survival analysis, basic notation and terminology, and the basic data layout for the computer.

We then describe how to estimate and graph survival curves using the **Kaplan-Meier (KM)** method. The estimated survival probabilities are computed using a **product limit formula**.

Next, we describe how to compare two or more survival curves using the **log–rank test** of the null hypothesis of a common survival curve. For two groups, the log–rank statistic is based on the summed observed minus expected score for a given group and its variance estimate. For several groups, a computer should always be used because the log–rank formula is more complicated mathematically. The test statistic is approximately chi-square in large samples with $G - 1$ degrees of freedom, where G denotes the number of groups being compared.

Several alternatives to the log–rank test will be briefly described. These tests are variations of the log rank test that weigh each observation differently. They are also large sample chi-square tests with $G - 1$ degrees of freedom.

Finally, we describe how to compute confidence intervals for the KM curve and for the median survival time.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

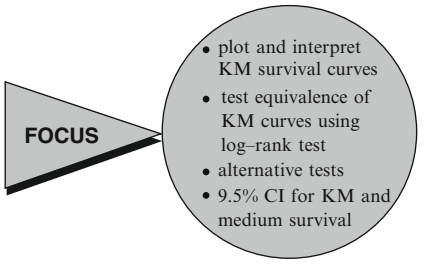
I. **Review** (pages 58–60)
II. **An example of Kaplan-Meier curves** (pages 61–65)
III. **General features of KM curves** (pages 66–67)
IV. **The log-rank test for two groups** (pages 67–71)
V. **The log-rank test for several groups** (pages 71–73)
VI. **Alternatives to the log rank test** (pages 73–78)
VII. **Confidence intervals for KM curves** (pages 78–79)
VIII. **Confidence intervals for the median survival time** (page 80)
IX. **Summary** (page 81)

Objectives

Upon completing the chapter, the learner should be able to:

1. Compute Kaplan-Meier (KM) probabilities of survival, given survival time and failure status information on a sample of subjects.
2. Interpret a graph of KM curves that compare two or more groups.
3. Draw conclusions as to whether or not two or more survival curves are the same based on computer results that provide a log–rank test and/or an alternative test.
4. Decide whether the log–rank test or one of the alternatives to this test is more appropriate for a given set of survival data.
5. Compute a 95% confidence interval for a KM survival probability.
6. Compute a 95% confidence interval for the median survival time obtained from a KM curve.

Presentation



This presentation describes how to plot and interpret survival data using Kaplan-Meier (KM) survival curves and how to test whether or not two or more KM curves are equivalent using the log–rank test. We also describe alternative tests to the log–rank test. Furthermore, we provide formulae for computing 95% confidence intervals for a KM curve and for the median survival time.

I. Review

Event: death
disease
relapse

We begin by reviewing the basics of survival analysis. Generally, survival analysis is a collection of statistical procedures for the analysis of data in which the outcome variable of interest is **time until an event occurs**. By **event**, we mean death, disease incidence, relapse from remission, or any designated experience of interest that may happen to an individual.

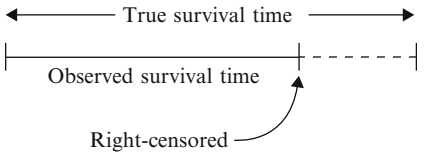
Time = survival time

Event = failure

When doing a survival analysis, we usually refer to the time variable as **survival time**. We also typically refer to the event as a **failure**.

Censoring: Don't know survival time exactly

Most survival analyses consider a key data analytical problem called **censoring**. In essence, censoring occurs when we have some information about individual survival time, but **we don't know the survival time exactly**.



Most survival time data is right-censored, because the true survival time interval, which we don't really know, has been cut off (i.e., censored) at the right side of the observed time interval, giving us an observed survival time that is shorter than the true survival time. We want to use the observed survival time to draw implications about the true survival time.

NOTATION

T = survival time
random variable
t = specific value for T

As notation, we denote by a **capital *T*** the random variable for a person's survival time. Next, we denote by a **small letter *t*** any specific value of interest for the variable T.

d = (0, 1) random variable

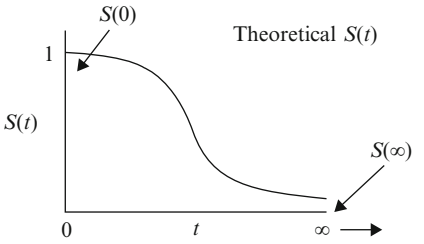
$= \begin{cases} 1 & \text{if failure} \\ 0 & \text{if censored} \end{cases}$

We let d denote a (0,1) random variable indicating either censorship or failure. A person who does not fail, that is, does not get the event during the study period, must have been censored either before or at the end of the study.

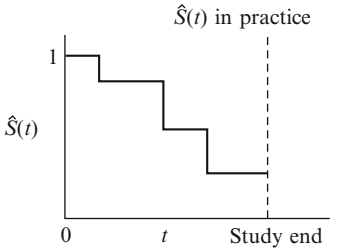
$S(t)$ = survivor function

$= \Pr(T > t)$

The survivor function, denoted by ***S(t)***, gives the probability that the random variable ***T*** exceeds the specified time t.



Theoretically, as t ranges from 0 up to infinity, the survivor function is graphed as a decreasing smooth curve, which begins at $S(t) = 1$ at $t = 0$ and heads downward toward zero as t increases toward infinity.

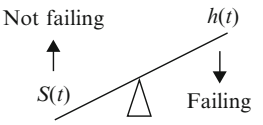


In practice, using data, we usually obtain estimated survivor curves that are **step functions**, as illustrated here, rather than smooth curves.

$h(t)$ = hazard function

= instantaneous potential given survival up to time t

The hazard function, denoted by ***h*(*t*)**, gives the **instantaneous potential** per unit time for the event to occur given that the individual has survived up to time t.



In contrast to the survivor function, which focuses on **not** failing, the hazard function focuses on failing; in other words, the higher the average hazard, the worse the impact on survival. The hazard is a **rate**, rather than a probability. Thus, the values of the hazard function range between zero and infinity.

$h(t)$ is a rate: 0 to ∞

Regardless of which function $S(t)$ or $h(t)$ one prefers, **there is a clearly defined relationship between the two**. In fact, if one knows the form of $S(t)$, one can derive the corresponding $h(t)$, and vice versa.

General Data Layout:

Indiv. #	t	d	X_1	$X_2 \ldots X_p$
1	t_1	d_1	X_{11}	$X_{12} \ldots X_{1p}$
2	t_2	d_2	X_{21}	$X_{22} \ldots X_{2p}$
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
n	t_n	d_n	X_{n1}	$X_{n2} \ldots X_{np}$

The general data layout for a survival analysis is given by the table shown here. The first column of the table identifies the study subjects. The second column gives the observed survival time information. The third column gives the information for d, the dichotomous variable that indicates censorship status. The remainder of the information in the table gives values for explanatory variables of interest.

Alternative (ordered) data layout:

Ordered failure times, $t_{(f)}$	# of failures m_f	# censored in $[t_{(f)}, t_{(f+1)})$, q_f	Risk set, $R(t_{(f)})$
$t_{(0)} = 0$	$m_0 = 0$	q_0	$R(t_{(0)})$
$t_{(1)}$	m_1	q_1	$R(t_{(1)})$
$t_{(2)}$	m_2	q_2	$R(t_{(2)})$
.	.	.	.
.	.	.	.
.	.	.	.
$t_{(k)}$	m_k	q_k	$R(t_{(k)})$

An alternative data layout is shown here. This layout is the basis upon which **Kaplan-Meier** survival curves are derived. The first column in the table gives ordered survival times from smallest to largest. The second column gives frequency counts of failures at each distinct failure time. The third column gives frequency counts, denoted by q_f, of those persons censored in the time interval starting with failure time $t_{(f)}$ up to but not including the next failure time, denoted by $t_{(f+1)}$. The last column gives the **risk set**, which denotes the collection of individuals who have survived at least to time $t_{(f)}$.

Table of ordered failures:

- Uses all information up to time of censorship;
- $S(t)$ is derived from $R(t)$.

To estimate the survival probability at a given time, we make use of the risk set at that time to include the information we have on a censored person up to the time of censorship, rather than simply throw away all the information on a censored person.

Survival probability:
Use **Kaplan-Meier (KM)** method.

The actual computation of such a survival probability can be carried out using the Kaplan-Meier (KM) method. We introduce the KM method in the next section by way of an example.

II. An Example of Kaplan-Meier Curves

EXAMPLE

The data: remission times (weeks) for two groups of leukemia patients

Group 1 ($n = 21$) treatment	Group 2 ($n = 21$) placebo
6, 6, 6, 7, 10, 13, 16, 22, 23, 6+, 9+, 10+, 11+, 17+, 19+, 20+, 25+, 32+, 32+, 34+, 35+,	1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, 23

Note: + denotes censored

	# failed	# censored	Total
Group 1	9	12	21
Group 2	21	0	21

Descriptive statistics:

$$\bar{T}_1(\text{ignoring} + \text{'s}) = 17.1,\ \bar{T}_2 = 8.6$$

$$\bar{h}_1 = .025,\ \bar{h}_2 = .115,\ \frac{\bar{h}_2}{\bar{h}_1} = 4.6$$

The data for this example derive from a study of the remission times in weeks for two groups of leukemia patients, with 21 patients in each group. Group 1 is the treatment group and group 2 is the placebo group. The basic question of interest concerns comparing the survival experience of the two groups.

Of the 21 persons in group 1, 9 failed during the study period and 12 were censored. In contrast, none of the data in group 2 are censored; that is, all 21 persons in the placebo group went out of remission during the study period.

In Chapter 1, we observed for this data set that group 1 appears to have better survival prognosis than group 2, suggesting that the treatment is effective. This conclusion was supported by descriptive statistics for the average survival time and average hazard rate shown. Note, however, that descriptive statistics provide overall comparisons but do not compare the two groups at different times of follow-up.

EXAMPLE: (continued)

Ordered failure times:

Group 1 (treatment)

$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
6	21	3	1
7	17	1	1
10	15	1	2
13	12	1	0
16	11	1	3
22	7	1	0
23	6	1	5
>23	—	—	—

Group 2 (placebo)

$t_{(f)}$	n_f	m_f	q_f
0	21	0	0
1	21	2	0
2	19	2	0
3	17	1	0
4	16	2	0
5	14	2	0
8	12	4	0
11	8	2	0
12	6	2	0
15	4	1	0
17	3	1	0
22	2	1	0
23	1	1	0

Group 2: no censored subjects
Group 2 (placebo)

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	1
1	21	2	0	$19/21 = .90$
2	19	2	0	$17/21 = .81$
3	17	1	0	$16/21 = .76$
4	16	2	0	$14/21 = .67$
5	14	2	0	$12/21 = .57$
8	12	4	0	$8/21 = .38$
11	8	2	0	$6/21 = .29$
12	6	2	0	$4/21 = .19$
15	4	1	0	$3/21 = .14$
17	3	1	0	$2/21 = .10$
22	2	1	0	$1/21 = .05$
23	1	1	0	$0/21 = .00$

A table of ordered failure times is shown here for each group. These tables provide the basic information for the computation of KM curves.

Each table begins with a survival time of zero, even though no subject actually failed at the start of follow-up. The reason for the zero is to allow for the possibility that some subjects might have been censored before the earliest failure time.

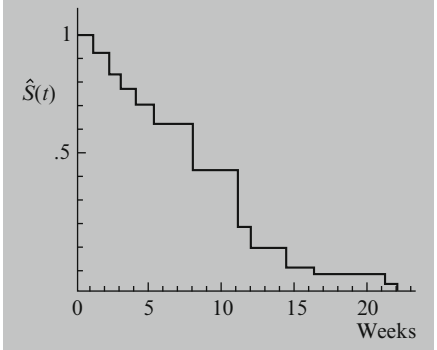
Also, each table contains a column denoted as n_f that gives the number of subjects in the risk set at the start of the interval. Given that the risk set is defined as the collection of individuals who have survived at least to time $t_{(f)}$, it is assumed that n_f includes those persons failing at time $t_{(f)}$. In other words, n_f counts those subjects at risk for failing instantaneously prior to time $t_{(f)}$.

We now describe how to compute the KM curve for the table for group 2. The computations for group 2 are quite straightforward because there are no censored subjects for this group.

The table of ordered failure times for group 2 is presented here again with the addition of another column that contains survival probability estimates. These estimates are the KM survival probabilities for this group. We will discuss the computations of these probabilities shortly.

EXAMPLE: (continued)

KM Curve for Group 2 (Placebo)



$S(t) = \Pr(T > t)$

Group 2 (placebo)

$t_{(f)}$	n_f	m_f	q_f	$\check{S}(t_{(f)})$
0	21	0	0	1
1	21	2	0	19/21 = .90
2	19	2	0	17/21 = .81
3	17	1	0	16/21 = .76
4	16	2	0	14/21 = .67
5	14	2	0	12/21 = .57
8	12	4	0	8/21 = .38
11	8	2	0	6/21 = .29
12	6	2	0	4/21 = .19
15	4	1	0	3/21 = .14
17	3	1	0	2/21 = .10
22	2	1	0	1/21 = .05
23	1	1	0	0/21 = .00

$$\hat{S}(t_{(f)}) = \frac{\#\text{ surviving past } t_{(f)}}{21}$$

No censorship in group 2
Alternative formula: KM approach

A plot of the KM survival probabilities corresponding to each ordered failure time is shown here for group 2. Empirical plots such as this one are typically plotted as a step function that starts with a horizontal line at a survival probability of 1 and then steps down to the other survival probabilities as we move from one ordered failure time to another.

We now describe how the survival probabilities for the group 2 data are computed. Recall that a survival probability gives the probability that a study subject survives past a specified time.

Thus, considering the group 2 data, the probability of surviving past zero is unity, as it will always be for any data set.

Next, the probability of surviving past the first ordered failure time of 1 week is given by 19/21 or (.90) because 2 people failed at 2 week, so that 19 people from the original 21 remain as survivors past 2 week.

Similarly, the next probability concerns subjects surviving past 2 weeks, which is 17/21 (or. 81) because 2 subjects failed at 1 week and 2 subjects failed at 2 weeks leaving 17 out of the original 21 subjects surviving past 2 weeks.

The remaining survival probabilities in the table are computed in the same manner, that is, we count the number of subjects surviving past the specified time being considered and divide this number by 21, the number of subjects at the start of follow-up.

Recall that no subject in group 2 was censored, so the q column for group 2 consists entirely of zeros. If some of the q's had been nonzero, an alternative formula for computing survival probabilities would be needed. This alternative formula is called the Kaplan-Meier (KM) approach and can be illustrated using the group 2 data even though all values of q are zero.

EXAMPLE

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} = \frac{14}{21} = .67$$

$\Pr(T > t_{(f)} \mid T \geq t_{(f)})$

$$\hat{S}(4) = 1 \times \textcircled{\frac{19}{21}} \times \frac{17}{19} \times \textcircled{\frac{16}{17}} \times \frac{14}{16} = \frac{14}{21} = .67$$

$\frac{19}{21} = \Pr(T > 1 \mid T \geq 1)$

$\frac{16}{17} = \Pr(T > 3 \mid T \geq 3)$

17 = # in risk set at week 3

$$\hat{S}(4) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \textcircled{\frac{14}{16}}$$

$$\hat{S}(8) = 1 \times \frac{19}{21} \times \frac{17}{19} \times \frac{16}{17} \times \frac{14}{16} \times \frac{12}{14} \times \textcircled{\frac{8}{12}}$$

KM formula = product limit formula

Group 1 (treatment)

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	(1)
6	21	3	1	$1 \times \frac{18}{21}$
		.		
		.		
		.		

For example, an alternative way to calculate the survival probability of exceeding 4 weeks for the group 2 data can be written using the KM formula shown here. This formula involves the product of conditional probability terms. That is, each term in the product is the probability of exceeding a specific ordered failure time $t_{(f)}$ given that a subject survives up to that failure time.

Thus, in the KM formula for survival past 4 weeks, the term 19/21 gives the probability of surviving past the first ordered failure time, 1 week, given survival up to the first week. Note that all 21 persons in group 2 survived up to 1 week, but that 2 failed at 1 week, leaving 19 persons surviving past 1 week.

Similarly, the term 16/17 gives the probability of surviving past the third ordered failure time at week 3, given survival up to week 3. There were 17 persons who survived up to week 3 and 1 of these then failed, leaving 16 survivors past week 3. Note that the 17 persons in the denominator represents the number in the risk set at week 3.

Notice that the product terms in the KM formula for surviving past 4 weeks stop at the 4th week with the component 14/16. Similarly, the KM formula for surviving past 8 weeks stops at the eighth week.

More generally, any KM formula for a survival probability is limited to product terms up to the survival week being specified. That is why the KM formula is often referred to as a "product-limit" formula.

Next, we consider the KM formula for the data from group 1, where there are several censored observations.

The estimated survival probabilities obtained using the KM formula are shown here for group 1.

The first survival estimate on the list is $\hat{S}(0) - 1$, as it will always be, because this gives the probability of surviving past time zero.

EXAMPLE: (continued)

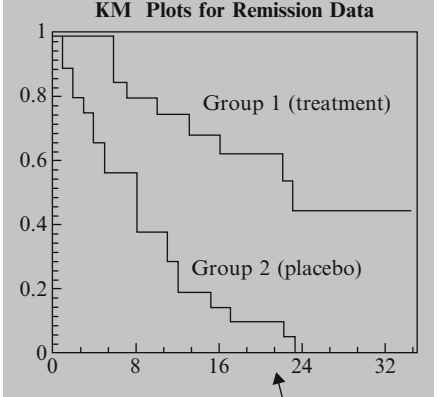
Group 1 (treatment)

$t_{(f)}$	n_f	m_f	q_f	$\hat{S}(t_{(f)})$
0	21	0	0	(1)
6	21	3	1	$1 \times \left(\frac{18}{21}\right) = .8571$
7	17	1	1	$.8571 \times \left(\frac{16}{17}\right) = .8067$
10	15	1	2	$.8067 \times \frac{14}{15} = .7529$
13	12	1	0	$.7529 \times \frac{11}{12} = .6902$
16	11	1	3	$.6902 \times \frac{10}{11} = .6275$
22	7	1	0	$.6275 \times \frac{6}{7} = .5378$
23	6	1	5	$.5378 \times \frac{5}{6} = .4482$

Fraction at $t_{(f)}$: $\Pr(T > t_{(f)} \mid T \geq t_{(f)})$

Not available at $t_{(f)}$: failed prior to $t_{(f)}$ or censored prior to $t_{(f)}$

group 1 only



Obtain KM plots from computer package, e.g., SAS, Stata, SPSS R

The other survival estimates are calculated by multiplying the estimate for the immediately preceding failure time by a fraction. For example, the fraction is 18/21 for surviving past week 6, because 21 subjects remain up to week 6 and 3 of these subjects fail to survive past week 6. The fraction is 16/17 for surviving past week 7, because 17 people remain up to week 7 and 1 of these fails to survive past week 7. The other fractions are calculated similarly.

For a specified failure time $t_{(f)}$, the fraction may be generally expressed as the conditional probability of surviving past time $t_{(f)}$, given availability (i.e., in the risk set) at time $t_{(f)}$. This is exactly the same formula that we previously used to calculate each product term in the product limit formula used for the group 2 data.

Note that a subject might not be available at time $t_{(f)}$ for one of two reasons: (1) either the subject has failed prior to $t_{(f)}$, or (2) the subject has been censored prior to $t_{(f)}$. Group 1 has censored observations, whereas group 2 does not. Thus, for group 1, censored observations have to be taken into account when determining the number available at $t_{(f)}$.

Plots of the KM curves for groups 1 and 2 are shown here on the same graph. Notice that the KM curve for group 1 is consistently higher than the KM curve for group 2. These figures indicate that group 1, which is the treatment group, has better survival prognosis than group 2, the placebo group. Moreover, as the number of weeks increases, the two curves appear to get farther apart, suggesting that the beneficial effects of the treatment over the placebo are greater the longer one stays in remission.

The KM plots shown above can be easily obtained from most computer packages that perform survival analysis, including SAS, Stata, SPSS, and R. All the user needs to do is provide a KM computer program with the basic data layout and then provide appropriate commands to obtain plots.

III. General Features of KM Curves

General KM formula:

$$\hat{S}(t_{(f)}) = \hat{S}(t_{(f-1)}) \times \hat{Pr}(T > t_{(f)} | T \geq t_{(f)})$$

The general formula for a KM survival probability at failure time $t_{(f)}$ is shown here. This formula gives the probability of surviving past the previous failure time $t_{(f-1)}$, multiplied by the conditional probability of surviving past time $t_{(f)}$, given survival to *at least* time $t_{(f)}$.

KM formula = product limit formula

$$\hat{S}(t_{(f-1)}) = \prod_{i=1}^{f-1} \hat{Pr}(T > t_{(i)} | T \geq t_{(i)})$$

The above KM formula can also be expressed as a product limit if we substitute for the survival probability $\hat{S}(t_{(f-1)})$, the product of all fractions that estimate the conditional probabilities for failure times $t_{(f-1)}$ and earlier.

EXAMPLE

$$\hat{S}(10) = .8067 \times \frac{14}{15} = .7529$$

$$= \boxed{\frac{18}{21} \times \frac{16}{17}} \times \frac{14}{15}$$

For example, the probability of surviving past 10 weeks is given in the table for group 1 (page 65) by .8067 times 14/15, which equals .7529. But the .8067 can be alternatively written as the product of the fractions 18/21 and 16/17. Thus, the product limit formula for surviving past 10 weeks is given by the triple product shown here.

$$\hat{S}(16) = .6902 \times \frac{10}{11}$$

$$= \boxed{\frac{18}{21} \times \frac{16}{17} \times \frac{14}{15} \times \frac{11}{12}} \times \frac{10}{11}$$

Similarly, the probability of surviving past 16 weeks can be written either as .6902 × 10/11, or equivalently as the five-way product of fractions shown here.

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{Pr}[T > t_{(i)} | T \geq t_{(i)}]$$
$$= \hat{S}(t_{(f-1)}) \times \hat{Pr}(T > t_{(f)} | T \geq t_{(f)})$$

The general expression for the product limit formula for the KM survival estimate is shown here together with the general KM formula given earlier. Both expressions are equivalent.

Math proof:

Pr(A and B) = Pr(A) × Pr(B | A) always

A simple mathematical proof of the KM formula can be described in probability terms. One of the basic rules of probability is that the probability of a joint event, say A and B, is equal to the probability of one event, say A, times the conditional probability of the other event, B, given A.

$A = "T \geq t_{(f)}" \rightarrow A \text{ and } B = B$
$B = "T > t_{(f)}"$
$\Pr(A \text{ and } B) = \Pr(B) = S(t_{(f)})$

If we let A be the event that a subject survives to at least time $t_{(f)}$ and we let B be the event that a subject survives past time $t_{(f)}$, then the joint event A and B simplifies to the event B, which is inclusive of A. It follows that the probability of A and B equals the probability of surviving past time $t_{(f)}$.

No failures during $t_{(f-1)} < T < t_{(f)}$
$\Pr(A) = \Pr(T > t_{(f-1)}) = S(t_{(f-1)})$

Also, because $t_{(f)}$ is the next failure time after $t_{(f-1)}$, there can be no failures after time $t_{(f-1)}$ and before time $t_{(f)}$. Therefore, the probability of A is equivalent to the probability of surviving past the $(f-1)$th ordered failure time.

$\Pr(B|A) = \Pr(T > t_{(f)} | T \geq t_{(f)})$

Furthermore, the conditional probability of B given A is equivalent to the conditional probability in the KM formula.

Thus, from Pr(A and B) formula,

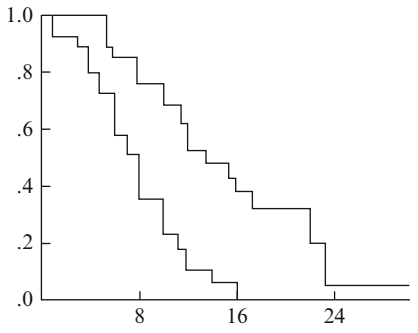
$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B \mid A)$$
$$S(t_{(f)}) = S(t_{(f-1)}) \times \Pr(T > t_{(f)} | T \geq t_{(f)})$$

Thus, using the basic rules of probability, the KM formula can be derived.

IV. The Log-Rank Test for Two Groups

We now describe how to evaluate whether or not KM curves for two or more groups are statistically equivalent. In this section we consider two groups only. The most popular testing method is called the log–rank test.

Are KM curves statistically equivalent?



When we state that two KM curves are "statistically equivalent," we mean that, based on a testing procedure that compares the two curves in some "overall sense," we do not have evidence to indicate that the true (population) survival curves are different.

- Chi-square test
- Overall comparison of KM curves
- Observed versus expected counts
- Categories defined by ordered failure times

The log–rank test is a large-sample chi-square test that uses as its test criterion a statistic that provides an overall comparison of the KM curves being compared. This (log–rank) statistic, like many other statistics used in other kinds of chi-square tests, makes use of observed versus expected cell counts over categories of outcomes. The categories for the log–rank statistic are defined by each of the ordered failure times for the entire set of data being analyzed.

EXAMPLE

Remission data: $n = 42$

# failures			# in risk set	
$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}
1	0	2	21	21
2	0	2	21	19
3	0	1	21	17
(4)	0	2	21	16
5	0	2	21	14
6	3	0	21	12
7	1	0	17	12
8	0	4	16	12
(10)	1	0	15	8
11	0	2	13	8
12	0	2	12	6
13	1	0	12	4
15	0	1	11	4
16	1	0	11	3
17	0	1	10	3
22	1	1	7	2
23	1	1	6	1

As an example of the information required for the log–rank test, we again consider the comparison of the treatment (group 1) and placebo (group 2) subjects in the remission data on 42 leukemia patients.

Here, for each ordered failure time, $t_{(f)}$, in the entire set of data, we show the numbers of subjects (m_{if}) failing at that time, separately by group (i), followed by the numbers of subjects (n_{if}) in the risk set at that time, also separately by group.

Thus, for example, at week 4, no subjects failed in group 1, whereas two subjects failed in group 2. Also, at week 4, the risk set for group 1 contains 21 persons, whereas the risk set for group 2 contains 16 persons.

Similarly, at week 10, 1 subject failed in group 1, and no subjects failed at group 2; the risk sets for each group contain 15 and 8 subjects, respectively.

Expected cell counts:

$$e_{1f} = \left(\frac{n_{1f}}{n_{1f} + n_{2f}}\right) \times \left(m_{1f} + m_{2f}\right)$$

↑ Proportion in risk set ↑ # of failures over both groups

$$e_{2f} = \left(\frac{n_{2f}}{n_{1f} + n_{2f}}\right) \times \left(m_{1f} + m_{2f}\right)$$

We now expand the previous table to include expected cell counts and observed minus expected values for each group at each ordered failure time. The formula for the expected cell counts is shown here for each group. For group 1, this formula computes the expected number at time f (i.e., e_{1f}) as the proportion of the total subjects in both groups who are at risk at time f, that is, $n_{1f}/(n_{1f} + n_{2f})$, multiplied by the total number of failures at that time for both groups (i.e., $m_{1f} + m_{2f}$). For group 2, e_{2f} is computed similarly.

EXAMPLE

Expanded Table (Remission Data)

		# failures		# in risk set		# expected		Observed-expected	
f	$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	e_{1f}	e_{2f}	$m_{1f}-e_{1f}$	$m_{2f}-e_{2f}$
1	1	0	2	21	21	(21/42) × 2	(21/42) × 2	−1.00	1.00
2	2	0	2	21	19	(21/40) × 2	(19/40) × 2	−1.05	1.05
3	3	0	1	21	17	(21/38) × 1	(17/38) × 1	−0.55	0.55
4	4	0	2	21	16	(21/37) × 2	(16/37) × 2	−1.14	1.14
5	5	0	2	21	14	(21/35) × 2	(14/35) × 2	−1.20	1.20
6	6	3	0	21	12	(21/33) × 3	(12/33) × 3	1.09	−1.09
7	7	1	0	17	12	(17/29) × 1	(12/29) × 1	0.41	−0.41
8	8	0	4	16	12	(16/28) × 4	(12/28) × 4	−2.29	2.29
9	10	1	0	15	8	(15/23) × 1	(8/23) × 1	0.35	−0.35
10	11	0	2	13	8	(13/21) × 2	(8/21) × 2	−1.24	1.24
11	12	0	2	12	6	(12/18) × 2	(6/18) × 2	−1.33	1.33
12	13	1	0	12	4	(12/16) × 1	(4/16) × 1	0.25	−0.25
13	15	0	1	11	4	(11/15) × 1	(4/15) × 1	−0.73	0.73
14	16	1	0	11	3	(11/14) × 1	(3/14) × 1	0.21	−0.21
15	17	0	1	10	3	(10/13) × 1	(3/13) × 1	−0.77	0.77
16	22	1	1	7	2	(7/9) × 2	(2/9) × 2	−0.56	0.56
17	23	1	1	6	1	(6/7) × 2	(1/7) × 2	−0.71	0.71
Totals		9	(21)			19.26	(10.74)	−10.26	(−10.26)

# of failure times

$$O_i - E_i = \sum_{f=1}^{17} \left(m_{if} - e_{if}\right), \qquad i = 1, 2$$

When two groups are being compared, the log–rank test statistic is formed using the sum of the observed minus expected counts over all failure times for one of the two groups. In this example, this sum is −10.26 for group 1 and 10.26 for group 2. We will use the group 2 value to carry out the test, but as we can see, except for the minus sign, the difference is the same for the two groups.

EXAMPLE

$O_1 - E_1 = -10.26$
$O_2 - E_2 = 10.26$

Two groups:

$O_2 - E_2 =$ summed observed minus expected score for group 2

$$\text{Log–rank statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)}$$

For the two-group case, the log–rank statistic, shown here at the left, is computed by dividing the square of the summed observed minus expected score for one of the groups — say, group 2 — by the variance of the summed observed minus expected score.

$$\mathrm{Var}(O_i - E_i) = \sum_j \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$
$$i = 1,\ 2$$

The expression for the estimated variance is shown here. For two groups, the variance formula is the same for each group. This variance formula involves the number in the risk set in each group (n_{if}) and the number of failures in each group (m_{if}) at time f. The summation is over all distinct failure times.

H_0: no difference between survival curves

The null hypothesis being tested is that there is no overall difference between the two survival curves. Under this null hypothesis, the log–-rank statistic is approximately chi-square with one degree of freedom. Thus, a P-value for the log–rank test is determined from tables of the chi-square distribution.

Log-rank statistic $\sim\chi^2$ with 1 df under H_0

Computer programs:
Stata's "sts test":

- descriptive statistics for KM curves
- log–rank statistic
- Alternative statistics to log–rank statistic

Several computer programs are available for calculating the log–rank statistic. For example the **Stata** package has a command called **"sts test"** that computes descriptive information about Kaplan-Meier curves, the log–rank statistic, and alternative statistics to the log–rank statistic, to be described later. Other packages, like **SAS** and **SPSS**, have procedures that provide results similar to those of **Stata**. A comparison of Stata, SAS, SPSS and R procedures and output is provided in the Computer Appendix at the back of this text.

EXAMPLE

Using Stata: Remission Data

Group	Events observed	Events expected
1	9	19.25
2	21	10.75
Total	30	30.00

Log rank = chi2(2) = 16.79
P-Value = Pr > chi2 = 0.000

For the remission data, the edited printout from using the **Stata** "sts test" procedure is shown here. The log–rank statistic is 16.79 and the corresponding P-value is zero to three decimal places. This P-value indicates that the null hypothesis should be rejected. We can therefore conclude that the treatment and placebo groups have significantly different KM survival curves.

EXAMPLE

$O_2 - E_2 = 10.26$
$\text{Var}(O_2 - E_2) = 6.2685$

$$\text{Log - rank statistic} = \frac{(O_2 - E_2)^2}{\widehat{\text{Var}}(O_2 - E_2)} = \frac{(10.26)^2}{6.2685} = 16.793$$

Although the use of a computer is the easiest way to calculate the log–rank statistic, we provide here some of the details of the calculation. We have already seen from earlier computations that the value of $O_2 - E_2$ is 10.26. The estimated variance of $O_2 - E_2$ is computed from the variance formula above to be 6.2685. The log–rank statistic then is obtained by squaring 10.26 and dividing by 6.285, which yields 16.793, as shown on the computer printout.

Approximate formula:

$$X^2 \approx \sum_{i}^{\#\text{ of groups}} \frac{(O_i - E_i)^2}{E_i}$$

An approximation to the log–rank statistic, shown here, can be calculated using observed and expected values for each group without having to compute the variance formula. The approximate formula is of the classic chi-square form that sums over each group being compared the square of the observed minus expected value divided by the expected value.

EXAMPLE

$$X^2 = \frac{(-10.26)^2}{19.26} + \frac{(10.26)^2}{10.74} = 15.276$$

Log-rank statistic = 16.793

The calculation of the approximate formula is shown here for the remission data. The expected values are 19.26 and 10.74 for groups 1 and 2, respectively. The chi-square value obtained is 15.276, which is slightly smaller than the log–rank statistic of 16.793.

V. The Log-Rank Test for Several Groups

H_0: All survival curves are the same.

Log-rank statistics for > 2 groups involves variances and covariances of $O_i - E_i$.

Matrix formula: See Appendix at end of this chapter.

The log–rank test can also be used to compare three or more survival curves. The null hypothesis for this more general situation is that all survival curves are the same.

Although the same tabular layout can be used to carry out the calculations when there are more than two groups, the test statistic is more complicated mathematically, involving both variances and covariances of summed observed minus expected scores for each group. A convenient mathematical formula can be given in matrix terms. We present the matrix formula for the interested reader in an Appendix at the end of this chapter.

Use computer program for calculations.

We will not describe further details about the calculation of the log–rank statistic, because a computer program can easily carry out the computations from the basic data file. Instead, we illustrate the use of this test with data involving more than two groups.

G (≥ 2) groups:
log–rank statistic $\sim \chi^2$ with $G - 1$ df

If the number of groups being compared is G (≥ 2), then the log–rank statistic has approximately a large sample chi-square distribution with $G - 1$ degrees of freedom. Therefore, the decision about significance is made using chi-square tables with the appropriate degrees of freedom.

Approximation formula:

$$X^2 \approx \sum_{i}^{\#\text{ofgroups}} \frac{(O_i - E_i)^2}{E_i}$$

Not required because computer program calculates the exact log–-rank statistic

The approximate formula previously described involving only observed and expected values without variance or covariance calculations can also be used when there are more than two groups being compared. However, practically speaking, the use of this approximate formula is not required as long as a computer program is available to calculate the exact log–rank statistic.

We now provide an example to illustrate the use of the log–rank statistic to compare more than two groups.

EXAMPLE

vets.dat: survival time in days,

n 137

Veteran's Administration Lung Cancer Trial

Column 1: Treatment (standard = 1, test = 2)
Column 2: Cell type 1 (large = 1, other = 0)
Column 3: Cell type 2 (adeno = 1, other = 0)
Column 4: Cell type 3 (small = 1, other = 0)
Column 5: Cell type 4 (squamous = 1, other = 0)
Column 6: Survival time (days)
Column 7: (Performance Status)
(0 = worst . . . 100 = best)
Column 8: Disease duration (months)
Column 9: Age
Column 10: Prior therapy (none = 0, some = 1)
Column 11: Status (0 = censored, 1 = died)

The data set "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (*The Statistical Analysis of Survival Time Data*, John Wiley, pp. 223–224, 1980). A complete list of the variables is shown here. Failure status is defined by the status variable (column 11).

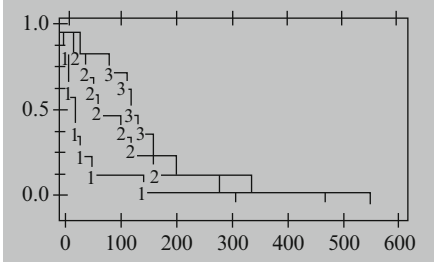
Among the variables listed, we now focus on the performance status variable (column 7). This variable is a continuous variable, so before we can obtain KM curves and the log–rank test, we need to categorize this variable.

EXAMPLE: (continued)

Performance Status Categories

Group #	Categories	Size
1	0–59	52
2	60–74	50
3	75–100	35

KM curves for performance status groups



Group	Events observed	Events expected
1	50	26.30
2	47	55.17
3	31	46.53
Total	128	128.00

Log-rank = chi2(2) = 29.18
P-value = Pr > chi2 = 0.0000
$G = 3$ groups; $df = G - 1 = 2$

Log-rank test is highly significant.

Conclude significant difference among three survival curves.

If, for the performance status variable, we choose the categories 0–59, 60–74, and 75–100, we obtain three groups of sizes 52, 50, and 35, respectively.

The KM curves for each of three groups are shown here. Notice that these curves appear to be quite different. A test of significance of this difference is provided by the log–rank statistic.

An edited printout of descriptive information about the three KM curves together with the log–rank test results are shown here. These results were obtained using the Stata package.

Because three groups are being compared here, $G = 3$ and the degrees of freedom for the log–rank test is thus $G - 1$, or 2. The log–rank statistic is computed to be 29.181, which has a P-value of zero to three decimal places. Thus, the conclusion from the log–rank test is that there is a highly significant difference among the three survival curves for the performance status groups.

VI. Alternatives to the Log Rank Test

Alternative tests supported by Stata

Wilcoxen
Tarone-Ware
Peto
Flemington-Harrington

There are several alternatives to the log rank test offered by Stata, SAS, SPSS, and R designed to test the hypothesis that two or more survival curves are equivalent. In this section we describe the **Wilcoxon**, the **Tarone-Ware**, the **Peto**, and the **Flemington-Harrington** test. All of these tests are variations of the log rank test and are easily implemented in Stata.

Log rank uses

$$O_i - E_i = \sum_f \left(m_{if} - e_{if}\right)$$

i = group #
$f = f$th failure time

In describing the differences among these tests, recall that the log rank test uses the summed observed minus expected score $O - E$ in each group to form the test statistic. This simple sum gives the same weight – namely, unity – to each failure time when combining observed minus expected failures in each group.

Weighting the test statistic for two groups

Test statistic:

$$\frac{\left(\sum_f w(t_{(f)})\left(m_{if} - e_{if}\right)\right)^2}{\text{var}\left(\sum_j w(t_{(f)})\left(m_{if} - e_{if}\right)\right)}$$

i = 1, 2
$f = f$th failure time
$w(t_{(f)})$ = weight at fth failure time

The Wilcoxon, Tarone-Ware, Peto, and Flemington-Harrington test statistics are variations of the log rank test statistic and are derived by applying different weights at the f-th failure time (as shown on the left for two groups).

Wilcoxon Test

- $w(t_f) = n_f$ (number at risk)
- Earlier failures receive more weight
- Appropriate if treatment effect is strongest in earliest phases of administration

The Wilcoxon test (called the Breslow test in SPSS) weights the observed minus expected score at time t_f by the number at risk n_f, over all groups at time t_f. Thus, the Wilcoxon test places more emphasis on the information at the beginning of the survival curve where the number at risk is large allowing early failures to receive more weight than later failures. This type of weighting may be used to assess whether the effect of a treatment on survival is strongest in the earlier phases of administration and tends to be less effective over time.

Weights Used for Various Test Statistics

Test Statistic	$w(t_{(f)})$
Log rank	1
Wilcoxon	n_f
Tarone-Ware	$\sqrt{n_f}$
Peto	$\tilde{s}(t_{(f)})$
Flemington-Harrington	$\hat{S}(t_{(f-1)})^p \times[1 - \hat{S}(t_{(f-1)})]^q$

The Tarone-Ware test statistic also applies more weight to the early failure times by weighting the observed minus expected score at time $t_{(f)}$ by the square root of the number at risk $\sqrt{n_f}$. The Peto test weights the f-th failure time by the survival estimate $\tilde{s}(t_{(f)})$ calculated over all groups combined. This survival estimate $\tilde{s}(t_{(f)})$ is similar but not exactly equal to the Kaplan-Meier survival estimate. The Flemington-Harrington test uses the Kaplan-Meier survival estimate $\hat{S}(t)$ over all groups to calculate its weights for the f-th failure time, $\hat{S}(t_{(f-1)})^p[1 - \hat{S}(t_{(f-1)})]^q$. The weights for each of these test statistics are summarized on the left.

Flemington-Harrington Test

$$w(t) = \hat{S}\left(t_{(f-1)}\right)^{p}\left[1 - \hat{S}\left(t_{(f-1)}\right)\right]^{q}$$

if p = 1 and q = 0, $w(t) = \hat{S}(t_{(f-1)})$
if p = 0 and q = 1,
  $w(t) = 1 - \hat{S}(t_{(f-1)})$
if p = 0 and q = 0,
  w(t) = 1 (log rank test)

The Flemington-Harrington test allows the most flexibility in terms of the choice of weights because the user provides the values of p and q. For example, if p = 1 and q = 0 then $w(t) = \hat{S}(t_{(f-1)})$ which gives more weight for the earlier survival times when $\hat{S}(t_{(f-1)})$ is close to one. However, if p = 0 and q = 1 then $w(t) = 1 - \hat{S}(t_{(f-1)})$ in which case the later survival times receive more weight. If p = 0 and q = 0 then w(t) = 1, and the Flemington-Harrington test reduces to the log rank test.

Comparisons of Test Results: Remission Data, Testing Treatment (RX)

Test	Chi-square (1 df)	P-value
Log rank	16.79	0.0000
Wilcoxon	13.46	0.0002
Tarone-Ware	15.12	0.0001
Peto	14.08	0.0002
FH (p = 3, q = 1)	8.99	0.0027
FH (p = 1, q = 3)	12.26	0.005

On the left is a comparison of test results for the effect of treatment (vs. placebo) using the remission data. The log rank chi-square statistic (also displayed previously in this chapter) is the highest among these tests at 16.79. The Flemington–Harrington (FH) test with p = 3 and q = 1 yielded the lowest chi-square value at 8.99, although with this weighting it is not immediately obvious which part of the survival curve is getting the most weight. However, all the test results are highly significant yielding a similar conclusion to reject the null hypothesis.

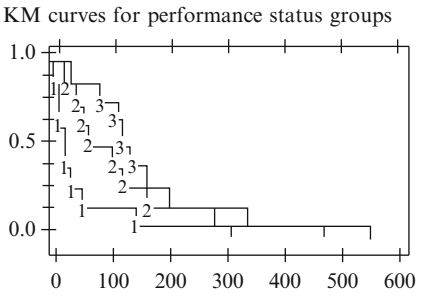
Vets Data, 3-Level Performance Status

Test	Chi-square (2 df)	P-value
Log rank	29.18	0.0000
Wilcoxon	46.10	0.0000

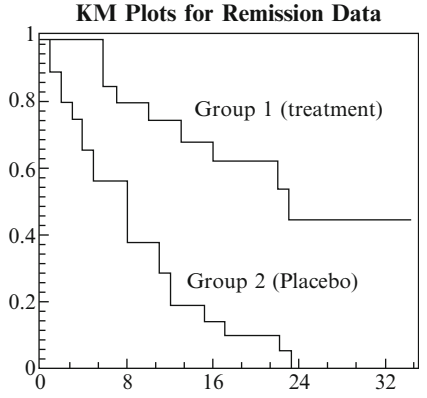
Remission Data, 2-Level Treatment

Test	Chi-square (1 df)	P-value
Log rank	16.79	0.0000
Wilcoxon	13.46	0.0002

On the left are comparisons of the log rank and Wilcoxon tests for the 3-level performance status variable from the vets dataset discussed in the previous section. The Wilcoxon test yields a higher chi-square value (46.10) than the log rank test (29.18). In contrast, the log rank test for the effect of treatment (RX) from the remissions data yields a higher chi-square value (16.79) than the Wilcoxon test (13.46). However, both the Wilcoxon and log rank tests are highly significant for both performance status and for treatment variables.



A comparison of survival curves gives insight into why the Wilcoxon test yields a higher chi-square value than the log rank test for the 3-level performance status variable. The 3 curves being compared are farthest apart in the early part of followup before becoming closer later. By contrast, a comparison of the 2 curves for treatment shows the curves diverging over time.



Choosing a Test

- Results of different weightings usually lead to similar conclusions
- The best choice is test with most power
- Power depends on how the null is violated
- There may be a clinical reason to choose a particular weighting
- Choice of weighting should be a priori

In general, the various weightings should provide similar results and will usually lead to the same decision as to whether the null hypothesis is rejected. The choice of which weighting of the test statistic to use (e.g., log rank or Wilcoxon) depends on which test is believed to provide the greatest statistical power, which in turn depends on how it is believed the null hypothesis is violated.

If there is a clinical reason to believe the effect of an exposure is more pronounced toward the beginning (or end) of the survival function, then it makes sense to use a weighted test statistic. However, **one should make an a priori decision on which statistical test to use** rather than fish for a desired p-value. Fishing for a desired result may lead to bias.

Stratified log rank test

- $O - E$ scores calculated within strata
- $O - E$ scores then summed across strata
- Allows control of stratified variable

```
Stratified log-rank test
->1wbc3 = 1
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        0      2.91
1          |        4      1.09
-----------+--------------------
Total      |        4      4.00
           |
->1wbc3 = 2
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        5      7.36
1          |        5      2.64
-----------+--------------------
Total      |       10     10.00
->1wbc3 = 3
           |  Events    Events
rx         | observed  expected
-----------+--------------------
0          |        4      6.11
1          |       12      9.89
-----------+--------------------
Total      |       16     16.00
->Total
           |             Events
           |  Events    expected
rx         | observed      (*)
-----------+--------------------
0          |        9     16.38
1          |       21     13.62
-----------+--------------------
Total      |       30     30.00
```

The stratified log rank test is another variation of the log rank test. With this test the summed observed minus expected scores $O - E$ are calculated within strata of each group and then summed across strata. **The stratified log rank test provides a method of testing the equivalence of survival curves controlling for the stratified variable**. An example of the stratified log rank test is presented next using the remission data.

On the left is Stata output from performing a stratified log rank test for the effect of treatment (RX) stratified by a 3-level variable (LWBC3) indicating low, medium, or high log white blood cell count (coded 1, 2, and 3, respectively).

Within each stratum of LWBC3, the expected number of events is calculated for the treated group (RX = 0) and for the placebo group (RX = 1). The total expected number of events for the treated group is found by summing the expected number of events over the three strata: 2.91 + 7.36 + 6.11 = 16.38. Similarly the total expected number of events for the placebo group is calculated: 1.09 + 2.64 + 9.89 = 13.62. This compares to 9 observed cases from the treated group and 21 observed cases from the placebo group yielding a chi-square value of 10.14 with 1 degree of freedom (for 2 levels of treatment) and a corresponding p-value of 0.0014.

Recall that when we did not control for log white blood cell count, the log rank test for the effect of treatment yielded a chi-square value of 16.79 and a corresponding p-value rounded to 0.0000.

Log rank unstratified

$$O_i - E_i = \sum_j \left(m_{if} - e_{if}\right)$$

i = group #, f = fth failure time

Log rank stratified

$$O_i - E_i = \sum_s \sum_f \left(m_{ifs} - e_{ifs}\right)$$

i = group #, f = jth failure time, s = stratum #

Stratified or unstratified (G groups)
Under H_0:

log rank statistic $\sim\chi^2$ with G − 1 df

The only difference between the unstratified and stratified approaches is that for the unstratified approach, the observed minus expected number of events for each failure time is summed over all failure times for each group (i). With the stratified approach, the observed minus expected number of events is summed over all failure times for each group within each stratum and then summed over all strata. Either way, the null distribution is chi-square with G − 1 degrees of freedom, where G represents the number of groups being compared (not the number of strata).

Can stratify with other tests
Wilcoxon, Tarone-Ware, Peto, Flemington-Harrington

The stratified approach can also be applied to any of the weighted variations of the log rank test (e.g., Wilcoxon). A limitation of the stratified approach is the reduced sample size within each stratum. This is particularly problematic with the remission dataset, which has a small sample size to begin with.

Limitation
Sample-size may be small within strata

Alternatively
Test associations using modeling

- Can simultaneously control covariates
- Shown in next chapter

We have shown how the stratified log rank test can be used to test the effect of treatment while controlling for log white blood cell count. In the next chapter we show how modeling can be used to test an association of a predictor variable while simultaneously controlling for other covariates.

VII. Confidence intervals for KM curves

95% CI for the KM curve:

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\hat{Var}[\hat{S}_{KM}(t)]}$$

where **Greenwood's formula** for $\mathrm{Var}[\hat{S}_{KM}(t)]$ is given by

$$\hat{Var}[\hat{S}_{KM}(t)] = \left(\hat{S}_{KM}(t)\right)^2 \sum_{f:t_{(f)} \le t} \left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

$t_{(f)}$ = f-ordered failure time
m_f = number of failures at $t_{(f)}$,
n_f = number in the risk set at $t_{(f)}$,

We now describe how to calculate (95%) confidence intervals (CIs) for the estimated Kaplan–Meier (KM) curve.

The 95% CI formula for estimated KM probability at any time point over follow-up has the general large sample form shown on the left, where $\hat{S}_{KM}(t)$denotes the KM survival estimate at time t and $\mathrm{Var}[\hat{S}_{KM}(t)]$ denotes variance of $\hat{S}_{KM}(t)$. The most common approach used to calculate this variance uses the **Greenwood's formula**, also shown on the left.

$$\frac{m_f}{n_f(n_f - m_f)} = \frac{1}{n_f}\left(\frac{m_f}{n_f - m_f}\right) \quad \text{(weight; Conditional risk)}$$

$$= \frac{1}{n_f} P[T > t_{(f)} \mid T \geq t_{(f)}]$$

t	n	m	$\frac{m}{n(n-m)}$	$\sum \frac{m}{n(n-m)}$	S(t)	Var[S(t)]
0	21	0	0	0	1	0
6	**21**	**3**	**0.0079**	**0.0079**	**0.857**	**0.0058**
7	17	1	0.0037	0.0116	0.807	0.0076
10	**15**	**1**	**0.0048**	**0.0164**	**0.753**	**0.0093**
13	12	1	0.0076	0.0240	0.690	0.0114
16	11	1	0.0091	0.0330	0.628	0.0130
22	7	1	0.0238	0.0569	0.538	0.0164
23	6	1	0.0333	0.0902	0.448	0.0181

6 weeks: $0.857 \pm 1.96\sqrt{(.857)^2\,(0.0079)}$

$= 0.857 \pm 0.149$ **= (.708, 1.006)** → 1.0

10 weeks: $\frac{m_3}{n_3(n_3 - m_3)} = \frac{1}{15(14)} = .0048$

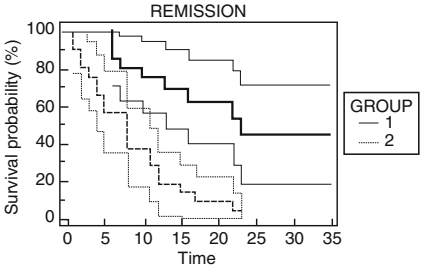
$$\sum_{f:t_{(f)} \leq t=10} \left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

$= 0.0079 + 0.0037 + 0.0048 = 0.0164.$

$\hat{Var}[\hat{S}_{KM}(10)] = (0.753)^2(.0164) = 0.0093$

95% CI : $0.753 \pm 1.96\sqrt{0.0093} = (\mathbf{0.564}, \mathbf{0.942})$

Same CI at t =11 and 12, since no events occurred at those times.



The summation component of Greenwood's formula essentially gives at each failure time $t_{(f)}$, a weighted (by $1/n_f$) average of the conditional risk of failing at those failure times prior to $t_{(f)}$. Thus, the variance formula gives the square of the KM coordinate at each event time weighted by the cumulative estimate of the risk at time t.

We illustrate how Greenwood's variance is calculated for the treatment group (Group 1) of the remission times data described earlier. The layout for this computation is shown n the left.

At 6 weeks, the estimate of the survival function is 0.857. There were three events at 6 weeks and 21 patients at risk. Therefore, $m_f/n_f(n_f - m_f) = 3/(21\times18) = 0.0079$. As this is the only component of the sum, the variance is then $0.0079 \times 0.857^2 = 0.0058$. The corresponding 95% confidence interval is shown on the left, where the upper level should be modified to 1.

At 10 weeks, the estimate of the survival function is 0.753. There was 1 event at 10 months and 15 patients at risk. Therefore, $m_f/n_f(n_f\text{-}m_f) = 1/(15\times14) = 0.0048$.

There were two other risk components prior to this time, 0.0079 at time 6 and 0.0037 at time 7 and their sum is 0.0164. The variance at 10 weeks is then 0.0164×0.753^2, which equals 0.0093.

The 95% CI for the proportion of patients at 10 weeks is shown on the left to have the limits (0.564, 0.942). Note that since the variance is only defined at event times, the 95% confidence interval remains the same at 11 and 12 weeks also.

On the left we show the KM curves and their corresponding 95% confidence intervals for Group 1 (treatment) and Group 2 (placebo) for the remission time data.

VIII. Confidence intervals for the median survival time

Remission data example:
Group 1 median = 8 weeks
95% CI ?

Formula for 95% CI derived from:

$$\frac{\left(\hat{S}_{\mathrm{KM}}(\mathrm{M}) - 0.5\right)^2}{\hat{\mathrm{Var}}[\hat{S}_{\mathrm{KM}}(\mathrm{M})]} \sim \chi^2_1 \text{ where}$$

M = true (unknown) median survival time, i.e., $S_{KM}(M) = 0.5$

$\hat{S}_{KM}(M)$ = estimated survival probability from KM curve at the true median survival time

$\hat{\mathrm{Var}}[\hat{S}_{KM}(M)]$ uses Greenwood's formula.

95% CI for median survival:

$$\boxed{\left(\hat{S}_{KM}(t) - 0.5\right)^2 \; < \; 3.84 \hat{Var}[\hat{S}_{KM}(t)]}$$

$t_{(i)}$	$\hat{S}(t)$	$(\hat{S}(t)-0.5)^2$	3.84 $\mathrm{Var}(\hat{S}(t))$	Inequality satisfied?
0	1	0.250	-	
1	0.90	0.160	0.016	N
2	0.81	0.096	0.028	N
3	0.76	0.068	0.033	N
4	**0.67**	**0.029**	**0.041**	**Y**
5	**0.57**	**0.005**	**0.045**	**Y**
8	**0.38**	**0.014**	**0.044**	**Y**
11	0.29	0.44	0.038	N
12	0.19	0.096	0.028	N
15	0.14	0.130	0.022	N
17	0.10	0.160	0.016	N
22	0.05	0.203	0.008	N
23	0.00	0.050	0	N

0.096 > 0.028 so the inequality is not satisfied
0.014 > 0.044 and the inequality is satisfied

Caution (ref B&C, 1982): upper limit should be adjusted to reflect censoring, e.g., SAS's LIFETEST adjusts above 95% CI from (**4,8**) to (**4, 11**).

Returning again to the remission time dataset, we now consider the calculation of the 95% CI for the median of the remission times for the placebo group. Recall that the median survival for this group is 8 weeks.

Brookmeyer and Crowley (1982) proposed a simple way to calculate the CI for the median survival time based on the fact that the square of a *standardized function* of the survival curve around the true (unknown) median value (M) is asymptotically χ^2distributed. This relationship is shown mathematically on the left.

Using the above result about the standardized survival curve, a general formula for the 95% CI for the median survival is provided by the inequality expression shown on the left. **The times for which this inequality holds are plausible values of the true median, while the boundaries represent upper and lower times for the 95% CI for the median**. The lower boundary may be 0 and the upper boundary may not always exist.

For the remission time data, the calculation of the CI around the median of 8 weeks is given in the table shown on the left. Since the inequality in the CI formula is satisfied in the range t=4 weeks to 8 weeks, the resulting 95% CI is (**4,8**).

Brookmeyer and Crowley (B&C) caution that the upper limit for these estimates should be adjusted to reflect censoring. They recommend reporting semiopen intervals that extend one event beyond the event that satisfies the inequality; some packages (e.g., SAS, R) incorporate this recommendation.

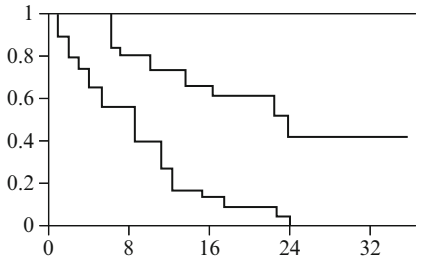
In this example, the 95% CI for the median survival time obtained from SAS output is given by the interval (4, 11), with the upper limit now extended beyond 8 weeks but not including the event at month 11.

IX. Summary

KM curves:

We now briefly summarize this presentation. First, we described how to estimate and graph survival curves using the Kaplan-Meier (KM) method.

KM curves:



$t_{(f)}$: fth ordered failure time

$$\hat{S}(t_{(f)}) = \prod_{i=1}^{f} \hat{P}r[T > t_{(i)} | T \geq t_{(i)}]$$
$$= \hat{S}(t_{(f-1)}) \times \hat{P}r(T > t_{(f)} | T \geq t_{(f)})$$

To compute KM curves, we must form a data layout that orders the failure times from smallest to largest. For each ordered failure time, the estimated survival probability is computed using the **product limit formula** shown here. Alternatively, this estimate can be computed as the product of the survival estimate for the previous failure time multiplied by the conditional probability of surviving past the current failure time.

Log-rank test:

H_0: common survival curve for all groups

$$\text{Log - rank statistic} = \frac{(O_2 - E_2)^2}{\text{Var}(O_2 - E_2)}$$

log–rank statistic $\sim \chi^2$ with $G - 1$ df under H_0

G = # of groups

When survival curves are being compared, the log–rank test gives a statistical test of the null hypothesis of a common survival curve. For two groups, the log–rank statistic is based on the summed observed minus expected scores for a given group and its variance estimate. For several groups, a computer should always be used since the log–rank formula is more complicated mathematically. The test statistic is approximately chi-square in large samples with $G - 1$ degrees of freedom, where G denotes the number of groups being compared.

Greenwood's Variance formula:

$$\hat{Var}[\hat{S}_{KM}(t)] = (\hat{S}_{KM}(t))^2 \times \sum_{f:t_{(f)} \leq t} \left[\frac{m_f}{n_f (n_f - m_f)} \right]$$

Large sample confidence intervals for KM curves can be computed based on Greenwood's formula, shown at the left, for the variance of an estimated KM survival probability.

95% CI for KM:

$$\hat{S}_{KM}(t) \pm 1.96\sqrt{\hat{Var}\left[\hat{S}_{KM}(t)\right]}$$

The expression for the 95% confidence interval is shown below Greenwood's formula.

95% CI for median survival:

$$\left(\hat{S}_{KM}(t) - 0.5\right)^2 3.84\,\hat{Var}\left[\hat{S}_{KM}(t)\right]$$

A large sample confidence interval formula for the median of the KM curve can also be computed using the inequality formula shown here on the left. The upper and lower boundaries of t for which this inequality holds provide the 95% confidence limits.

Chapters

1. Introduction
✓2. Kaplan-Meier Survival Curves and the Log-Rank Test

This presentation is now complete. You can review this presentation using the detailed outline that follows and then try the practice exercises and test.

Next:

3. The Cox Proportional Hazards Model and Its Characteristics

Chapter 3 introduces the Cox proportional hazards (PH) model, which is the most popular mathematical modeling approach for estimating survival curves when considering several explanatory variables simultaneously.

Detailed Outline

I. Review (pages 58–60)

A. The outcome variable is (survival) time until an event (failure) occurs.

B. Key problem: **censored data**, i.e., don't know survival time exactly.

C. Notation:
T = survival time random variable
t = specific value of T
d = (0, 1) variable for failure/censorship status
S(t) = survivor function
h(t) = hazard function

D. Properties of survivor function:
   i. theoretically, graph is smooth curve, decreasing from $S(t) = 1$ at time $t = 0$ to $S(t) = 0$ at $t = \infty$;
   ii. in practice, graph is step function.

E. Properties of $h(t)$:
   i. instantaneous potential for failing given survival up to time;
   ii. $h(t)$ is a rate; ranges from 0 to ∞.

F. Relationship of $S(t)$ to $h(t)$: if you know one you can determine the other.

G. Goals of survival analysis: estimation of survivor and hazard functions; comparisons and relationships of explanatory variables to survival.

H. Data layouts
   i. for the computer;
   ii. for understanding the analysis: involves **risk sets**.

II. An Example of Kaplan-Meier Curves (pages 61–65)

A. Data are from study of remission times in weeks for two groups of leukemia patients (21 in each group).

B. Group 1 (treatment group) has several censored observations, whereas group 2 has no censored observations.

C. Table of ordered failure times is provided for each group.

D. For group 2 (all noncensored), survival probabilities are estimated directly and plotted. Formula used is

$$\hat{S}(t_{(f)}) = \frac{\#\text{ surviving past } t_{(f)}}{21}.$$

E. Alternative approach for group 2 is given by a **product limit** formula.

F. For group 1, survival probabilities calculated by multiplying estimate for immediately preceding failure time by a conditional probability of surviving past current failure time, i.e.,

$$\hat{S}_{(f)} = \hat{S}_{(f-1)}\hat{P}r\left[T > t_{(f)} \middle| T \geq t_{(f)}\right].$$

III. General Features of KM Curves (pages 66–67)

A. Two alternative general formulae:

$$S_{(f)} = \prod_{i=1}^{f} \Pr\left[T > t_{(i)} \middle| T \geq t_{(i)}\right] \quad \text{(product limit formula)}$$

$$S_{(f)} = S_{(f-1)} \Pr\left[T > t_{(f)} \middle| T \geq t_{(f)}\right]$$

B. Second formula derived from probability rule:

$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B|A)$$

IV. The Log-Rank Test for Two Groups (pages 67–71)

A. Large sample chi-square test; provides overall comparison of KM curves.

B. Uses observed versus expected counts over categories of outcomes, where categories are defined by ordered failure times for entire set of data.

C. Example provided using remission data involving two groups:

i. expanded table described to show how expected and observed minus expected cell counts are computed.

ii. for ith group at time f, where $i = 1$ or 2:
observed counts $= m_{if}$,
expected counts $= e_{if}$, where
expected counts = (proportion in risk set) × (# failures over both groups),

$$\text{i.e., } e_{if} = \left(\frac{n_{if}}{n_{1f} + n_{2f}}\right)(m_{1f} + m_{2f}).$$

D. Log-rank statistic for two groups:

$$\boxed{\frac{(O_i - E_i)^2}{\text{Var}(O_i - E_i)}},$$

where $i = 1.\ 2$,

$$O_i - E_i = \sum_f \left(m_{if} - e_{if}\right), \text{ and}$$

$$\text{Var}(O_i - E_i)$$

$$= \sum_f \frac{n_{1f} n_{2f} \left(m_{1f} + m_{2f}\right)\left(n_{1f} + n_{2f} - m_{1f} - m_{2f}\right)}{\left(n_{1f} + n_{2f}\right)^2 \left(n_{1f} + n_{2f} - 1\right)}$$

$$i = 1,\ 2$$

E. H_0: no difference between survival curves.

F. Log-rank statistic $\sim \chi^2$ with 1 *df* under H_0.

G. Approximate formula:

$$X^2 = \sum_{i=1}^{G} \frac{(O_i - E_i)^2}{E_i}, \text{ where } G = 2 = \#\text{ of groups}$$

H. Remission data example: Log-rank statistic = 16.793, whereas $X^2 = 15.276$.

V. The Log-Rank Test for Several Groups (pages 71–73)

A. Involves variances and covariances; matrix formula in Appendix.

B. Use computer for calculations.

C. Under H_0, log–rank statistic $\sim\chi^2$ with $G - 1$ *df*, where $G = \#$ of groups.

D. Example provided using vets.dat with interval variable "performance status"; this variable is categorized into $G = 3$ groups, so *df* for log-–rank test is $G - 1 = 2$, log–rank statistic is 29.181 ($P = 0.0$).

VI. Alternatives to the Log-Rank Test (pages 73–78)

A. Alternative tests supported by Stata: Wilcoxen, Tarone-Ware, Peto, and Flemington-Harrington.

B. Alternative tests differ by applying different weights at the j-th failure time.

C. The choice of alternative depends on the reason for the belief that the effect is more pronounced towards the beginning (or end) of the survival function.

D. The stratified-log–rank test is a variation of the log–rank test that controls for one or more stratified variables.

VII. Confidence Intervals for KM Curves (pages 78–79)

A. General form of 95% CI:

$$\hat{S}_{KM}(\mathrm{t}) \pm 1.96\sqrt{\hat{Var}[\hat{S}_{KM}(\mathrm{t})]}$$

B. $\hat{\mathrm{Var}}[\hat{\mathrm{S}}_{\mathrm{KM}}(\mathrm{t})]$uses Greenwood's formula:

$$\hat{Var}[\hat{S}_{KM}(\mathrm{t})] = \left(\hat{S}_{KM}(\mathrm{t})\right)^2 \sum_{f:t_{(f)}\le t}\left[\frac{m_f}{n_f(n_f - m_f)}\right]$$

C. Example using Remission Time Data

VIII. Confidence Intervals for the Median Survival Time (page 80)

A. General form of 95% CI where values of t satisfying the following inequality provide confidence limits:

$$\left(\hat{S}_{KM}(t) - 0.5\right)^2 < 3.84\hat{Var}[\hat{S}_{KM}(t)]$$

B. Example using Remission Time Data

IX. Summary (page 81)

Practice Exercises

1. The following data are a sample from the 1967-1980 Evans County study. Survival times (in years) are given for two study groups, each with 25 participants. Group 1 has no history of chronic disease (CHR = 0), and group 2 has a positive history of chronic disease (CHR = 1):

   Group 1 (CHR = 0): 12.3+, 5.4, 8.2, 12.2+, 11.7, 10.0, 5.7, 9.8, 2.6, 11.0, 9.2, 12.1+, 6.6, 2.2, 1.8, 10.2, 10.7, 11.1, 5.3, 3.5, 9.2, 2.5, 8.7, 3.8, 3.0

   Group 2 (CHR = 1): 5.8, 2.9, 8.4, 8.3, 9.1, 4.2, 4.1, 1.8, 3.1, 11.4, 2.4, 1.4, 5.9, 1.6, 2.8, 4.9, 3.5, 6.5, 9.9, 3.6, 5.2, 8.8, 7.8, 4.7, 3.9

   a. Fill in the missing information in the following table of ordered failure times for groups 1 and 2:

Group 1					Group 2				
$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)})$	$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)})$
0.0	25	0	0	1.00	0.0	25	0	0	1.00
1.8	25	1	0	.96	1.4	25	1	0	.96
2.2	24	1	0	.92	1.6	24	1	0	.92
2.5	23	1	0	.88	1.8	23	1	0	.88
2.6	22	1	0	.84	2.4	22	1	0	.84
3.0	21	1	0	.80	2.8	21	1	0	.80
3.5	20				2.9	20	1	0	.76
3.8	19	1	0	.72	3.1	19	1	0	.72
5.3	18	1	0	.68	3.5	18	1	0	.68
5.4	17	1	0	.64	3.6	17	1	0	.64
5.7	16	1	0	.60	3.9				
6.6	15	1	0	.56	4.1				
8.2	14	1	0	.52	4.2				
8.7	13	1	0	.48	4.7	13	1	0	.48
9.2					4.9	12	1	0	.44
9.8	10	1	0	.36	5.2	11	1	0	.40
10.0	9	1	0	.32	5.8	10	1	0	.36
10.2	8	1	0	.28	5.9	9	1	0	.32
10.7	7	1	0	.24	6.5	8	1	0	.28
11.0	6	1	0	.20	7.8	7	1	0	.24
11.1	5	1	0	.16	8.3	6	1	0	.20
11.7	4				8.4	5	1	0	.16
					8.8	4	1	0	.12
					9.1				
					9.9				
					11.4	1	1	0	.00

   b. Based on your results in part a, plot the KM curves for groups 1 and 2 on the same graph. Comment on how these curves compare with each other.

c. Fill in the following expanded table of ordered failure times to allow for the computation of expected and observed minus expected values at each ordered failure time. Note that your new table here should combine both groups of ordered failure times into one listing and should have the following format:

$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	e_{1f}	e_{2f}	$m_{1f}-e1_f$	$m_{2f}-e2_f$
1.4	0	1	25	25	.500	.500	−.500	.500
1.6	0	1	25	24	.510	.490	−.510	.510
1.8	1	1	25	23	1.042	.958	−.042	.042
2.2	1	0	24	22	.522	.478	.478	−.478
2.4	0	1	23	22	.511	.489	−.511	.511
2.5	1	0	23	21	.523	.477	.477	−.477
2.6	1	0	22	21	.516	.484	.484	−.484
2.8	0	1	21	21	.500	.500	−.500	.500
2.9	0	1	21	20	.512	.488	−.512	.512
3.0	1	0	21	19	.525	.475	.475	−.475
3.1								
3.5								
3.6								
3.8								
3.9	0	1	18	16	.529	.471	−.529	.529
4.1	0	1	18	15	.545	.455	−.545	.545
4.2	0	1	18	14	.563	.437	−.563	.563
4.7	0	1	18	13	.581	.419	−.581	.581
4.9	0	1	18	12	.600	.400	−.600	.600
5.2	0	1	18	11	.621	.379	−.621	.621
5.3	1	0	18	10	.643	.357	.357	−.357
5.4	1	0	17	10	.630	.370	.370	−.370
5.7	1	0	16	10	.615	.385	.385	−.385
5.8	0	1	15	10	.600	.400	−.600	.600
5.9	0	1	15	9	.625	.375	−.625	.625
6.5	0	1	15	8	.652	.348	−.652	.652
6.6	1	0	15	7	.682	.318	.318	−.318
7.8	0	1	14	7	.667	.333	−.667	.667
8.2	1	0	14	6	.700	.300	.300	−.300
8.3	0	1	13	6	.684	.316	−.684	.684
8.4	0	1	13	5	.722	.278	−.722	.722
8.7	1	0	13	4	.765	.235	.335	−.335
8.8	0	1	12	4	.750	.250	−.750	.750
9.1	0	1	12	3	.800	.200	−.800	.800
9.2								
9.8								
9.9								
10.0	1	0	9	1	.900	.100	.100	−.100
10.2	1	0	8	1	.888	.112	.112	−.112
10.7	1	0	7	1	.875	.125	.125	−.125
11.0	1	0	6	1	.857	.143	.143	−.143
11.1	1	0	5	1	.833	.167	.167	−.167
11.4	0	1	4	1	.800	.200	−.800	.800
11.7	1	0	4	0	1.000	.000	.000	.000
Totals	22	25			30.79	16.21		

d. Use the results in part c to compute the log–rank statistic. Use this statistic to carry out the log–rank test for these data. What is your null hypothesis and how is the test statistic distributed under this null hypothesis? What are your conclusions from the test?

2. The following data set called "anderson.dat" consists of remission survival times on 42 leukemia patients, half of whom get a certain new treatment therapy and the other half of whom get a standard treatment therapy. The exposure variable of interest is treatment status ($Rx = 0$ if new treatment, $Rx = 1$ if standard treatment). Two other variables for control as potential confounders are log white blood cell count (i.e., logwbc) and sex. Failure status is defined by the relapse variable (0 if censored, 1 if failure). The data set is listed as follows:

Subj	Survt	Relapse	Sex	log WBC	*Rx*
1	35	0	1	1.45	0
2	34	0	1	1.47	0
3	32	0	1	2.20	0
4	32	0	1	2.53	0
5	25	0	1	1.78	0
6	23	1	1	2.57	0
7	22	1	1	2.32	0
8	20	0	1	2.01	0
9	19	0	0	2.05	0
10	17	0	0	2.16	0
11	16	1	1	3.60	0
12	13	1	0	2.88	0
13	11	0	0	2.60	0
14	10	0	0	2.70	0
15	10	1	0	2.96	0
16	9	0	0	2.80	0
17	7	1	0	4.43	0
18	6	0	0	3.20	0
19	6	1	0	2.31	0
20	6	1	1	4.06	0
21	6	1	0	3.28	0
22	23	1	1	1.97	1
23	22	1	0	2.73	1
24	17	1	0	2.95	1
25	15	1	0	2.30	1
26	12	1	0	1.50	1
27	12	1	0	3.06	1
28	11	1	0	3.49	1

(*Continued on next page*)

Subj	Survt	Relapse	Sex	log WBC	*Rx*
29	11	1	0	2.12	1
30	8	1	0	3.52	1
31	8	1	0	3.05	1
32	8	1	0	2.32	1
33	8	1	1	3.26	1
34	5	1	1	3.49	1
35	5	1	0	3.97	1
36	4	1	1	4.36	1
37	4	1	1	2.42	1
38	3	1	1	4.01	1
39	2	1	1	4.91	1
40	2	1	1	4.48	1
41	1	1	1	2.80	1
42	1	1	1	5.00	1

a. Suppose we wish to describe KM curves for the variable logwbc. Because logwbc is continuous, we need to categorize this variable before we compute KM curves. Suppose we categorize logwbc into three categories - low, medium, and high - as follows:

   low (0–2.30), $n = 11$;
   medium (2.31–3.00), $n = 14$;
   high (>3.00), $n = 17$.

   Based on this categorization, compute and graph KM curves for each of the three categories of logwbc. (You may use a computer program to assist you or you can form three tables of ordered failure times and compute KM probabilities directly.)

b. Compare the three KM plots you obtained in part a. How are they different?

c. Below is an edited printout of the log–rank test comparing the three groups.

Group	Events observed	Events expected
1	4	13.06
2	10	10.72
3	16	6.21
Total	30	30.00

   Log-rank = chi2(2) = 26.39
   P-value = Pr > chi2 = 0.0000

   What do you conclude about whether or not the three survival curves are the same?

Test

To answer the questions below, you will need to use a computer program (from SAS, Stata, SPSS, R or any other package you are familiar with) that computes and plots KM curves and computes the log–rank test. Freely downloadable files can be obtained from weblink http://www.sph.emory.edu/dkleinb/surv3.htm.

1. For the vets.dat data set described in the presentation:
   a. Obtain KM plots for the two categories of the variable cell type 1 (1 = large, 0 = other). Comment on how the two curves compare with each other. Carry out the log–rank, and draw conclusions from the test(s).
   b. Obtain KM plots for the four categories of cell type - large, adeno, small, and squamous. Note that you will need to recode the data to define a single variable which numerically distinguishes the four categories (e.g., 1 = large, 2 = adeno, etc.). As in part a, compare the four KM curves. Also, carry out the log–rank for the equality of the four curves and draw conclusions.
2. The following questions consider a data set from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," *Med. J. Aust*., 1991). These data comprise the times in days spent by heroin addicts from entry to departure from one of two methadone clinics. There are two further covariates, namely, prison record and methadone dose, believed to affect the survival times. The data set name is addicts.dat. A listing of the variables is given below:

   Column 1: Subject ID
   Column 2: Clinic (1 or 2)
   Column 3: Survival status (0 = censored, 1 = departed from clinic)
   Column 4: Survival time in days
   Column 5: Prison record (0 = none, 1 = any)
   Column 6: Methadone dose (mg/day)

a. Compute and plot the KM plots for the two categories of the "clinic" variable and comment on the extent to which they differ.

b. A printout of the log–rank and Wilcoxon tests (using Stata) is provided below. What are your conclusions from this printout?

Group	Events observed	Events expected
1	122	90.91
2	28	59.09
Total	150	150.00

Log-rank = chi2(1) = 27.89
P-value = Pr > chi2 = 0.0000
Wilcoxon = chi2(1) = 11.63
P-value = Pr > chi2 = 0.0007

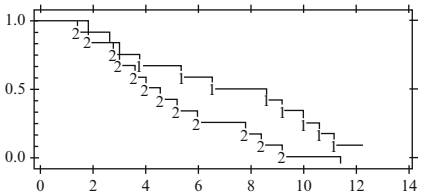
c. Compute and evaluate KM curves and the log–rank test for comparing suitably chosen categories of the variable "Methadone dose." Explain how you determined the categories for this variable.

Answers to Practice Exercises

1. a.

Group 1					Group 2				
$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)})$	$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)})$
0.0	25	0	0	1.00	0.0	25	0	0	1.00
1.8	25	1	0	.96	1.4	25	1	0	.96
2.2	24	1	0	.92	1.6	24	1	0	.92
2.5	23	1	0	.88	1.8	23	1	0	.88
2.6	22	1	0	.84	2.4	22	1	0	.84
3.0	21	1	0	.80	2.8	21	1	0	.80
3.5	20	1	0	.76	2.9	20	1	0	.76
3.8	19	1	0	.72	3.1	19	1	0	.72
5.3	18	1	0	.68	3.5	18	1	0	.68
5.4	17	1	0	.64	3.6	17	1	0	.64
5.7	16	1	0	.60	3.9	16	1	0	.60
6.6	15	1	0	.56	4.1	15	1	0	.56
8.2	14	1	0	.52	4.2	14	1	0	.52
8.7	13	1	0	.48	4.7	13	1	0	.48
9.2	12	2	0	.40	4.9	12	1	0	.44
9.8	10	1	0	.36	5.2	11	1	0	.40
10.0	9	1	0	.32	5.8	10	1	0	.36
10.2	8	1	0	.28	5.9	9	1	0	.32
10.7	7	1	0	.24	6.5	8	1	0	.28
11.0	6	1	0	.20	7.8	7	1	0	.24
11.1	5	1	0	.16	8.3	6	1	0	.20
11.7	4	1	3	.12	8.4	5	1	0	.16
					8.8	4	1	0	.12
					9.1	3	1	0	.08
					9.9	2	1	0	.04
					11.4	1	1	0	.00

b. KM curves for CHR data:



Group 1 appears to have consistently better survival prognosis than group 2. However, the KM curves are very close during the first 4 years, but are quite separate after 4 years, although they appear to come close again around 12 years.

c. Using the expanded table format, the following information is obtained:

$t_{(f)}$	m_{1f}	m_{2f}	n_{1f}	n_{2f}	e_{1f}	e_{2f}	$m_{1f}-e_{1f}$	$m_{2f}-e_{2f}$
1.4	0	1	25	25	.500	.500	−.500	.500
1.6	0	1	25	24	.510	.490	−.510	.510
1.8	1	1	25	23	1.042	.958	−.042	.042
2.2	1	0	24	22	.522	.478	.478	−.478
2.4.	0	1	23	22	.511	.489	−.511	.511
2.5.	1	0	23	21	.523	.477	.477	−.477
2.6	1	0	22	21	.516	.484	.484	−.484
2.8	0	1	21	21	.500	.500	−.500	.500
2.9	0	1	21	20	.512	.488	−.512	.512
3.0	1	0	21	19	.525	.475	.475	−.475
3.1	0	1	20	19	.513	.487	−.513	.513
3.5	1	1	20	18	1.053	.947	−.053	.053
3.6	0	1	19	17	.528	.472	−.528	.528
3.8	1	0	19	16	.543	.457	.457	−.457
3.9	0	1	18	16	.529	.471	−.529	.529
4.1	0	1	18	15	.545	.455	−.545	.545
4.2	0	1	18	14	.563	.437	−.563	.563
4.7	0	1	18	13	.581	.419	−.581	.581
4.9	0	1	18	12	.600	.400	−.600	.600
5.2	0	1	18	11	.621	.379	−.621	.621
5.3	1	0	18	10	.643	.357	.357	−.357
5.4	1	0	17	10	.630	.370	.370	−.370
5.7	1	0	16	10	.615	.385	.385	−.385
5.8	0	1	15	10	.600	.400	−.600	.600
5.9	0	1	15	9	.625	.375	−.625	.625
6.5	0	1	15	8	.652	.348	−.652	.652
6.6	1	0	15	7	.682	.318	.318	−.318
7.8	0	1	14	7	.667	.333	−.667	.667
8.2	1	0	14	6	.700	.300	.300	−.300
8.3	0	1	13	6	.684	.316	−.684	.684
8.4	0	1	13	5	.722	.278	−.722	.722
8.7	1	0	13	4	.765	.235	.335	−.335
8.8	0	1	12	4	.750	.250	−.750	.750
9.1	0	1	12	3	.800	.200	−.800	.800
9.2	2	0	12	2	1.714	.286	.286	−.286
9.8	1	0	10	2	.833	.167	.167	−.167
9.9	0	1	9	2	.818	.182	−.818	.818
10.0	1	0	9	1	.900	.100	.100	−.100
10.2	1	0	8	1	.888	.112	.112	−.112
10.7	1	0	7	1	.875	.125	.125	−.125
11.0	1	0	6	1	.857	.143	.143	−.143
11.1	1	0	5	1	.833	.167	.167	−.167
11.4	0	1	4	1	.800	.200	−.800	.800
11.7	1	0	4	0	1.000	.000	.000	.000
Totals	22	25			30.79	16.21	−8.790	8.790

d. The log–rank statistic can be computed from the totals of the expanded table using the formulae:

$$\text{log-rank statistic} = \frac{(O_i - E_i)^2}{\widehat{\text{Var}}(O_i - E_i)}$$

$$\text{Var}(O_i - E_i) = \sum_i \frac{n_{1f} n_{2f} (m_{1f} + m_{2f})(n_{1f} + n_{2f} - m_{1f} - m_{2f})}{(n_{1f} + n_{2f})^2 (n_{1f} + n_{2f} - 1)}$$

The variance turns out to be 9.658, so that the log–-rank statistic is $(8.79)^2/9.658 = 7.993$.

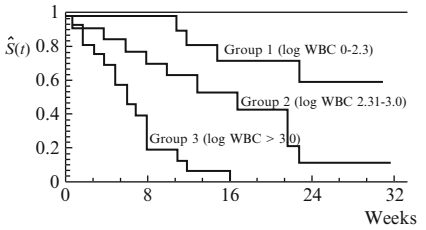
Using Stata, the results for the log–rank test are given as follows:

Group	Events observed	Events expected
1	22	30.79
2	25	16.21
Total	47	47.00

Log-rank = chi2(1) = 7.99
P-value = Pr > chi2 = 0.0047

The log–rank test gives highly significant results. This indicates that there is a significant difference in survival between the two groups.

2. a. For the Anderson dataset, the KM plots for the three categories of log WBC are shown below:



b. The KM curves are quite different with group 1 having consistently better survival prognosis than group 2, and group 2 having consistently better survival prognosis than group 3. Note also that the difference between group 1 and 2 is about the same over time, whereas group 2 appears to diverge from group 3 as time increases.

c. The log–rank statistic (26.391) is highly significant with P-values equal to zero to three decimal places. These results indicate that there is some overall difference between the three curves.

Appendix: Matrix Formula for the Log-Rank Statistic for Several Groups

For $i = 1, 2, \ldots, G$ and $f = 1, 2, \ldots, k$, where $G =$ # of groups and $k =$ # of distinct failure times,

$n_{if} =$ # at risk in ith group at fth ordered failure time

$m_{if} =$ observed # of failures in ith group at fth ordered failure time

$e_{if} =$ expected # of failures in ith group at fth ordered failure time

$$= \left(\frac{n_{if}}{n_{1f} + n_{2f}}\right)\left(m_{1f} + m_{2f}\right)$$

$$n_f = \sum_{i=1}^{G} n_{if}$$

$$m_f = \sum_{i=1}^{G} m_{if}$$

$$O_i - E_i = \sum_{f=1}^{k} \left(m_{if} - e_{if}\right)$$

$$\mathrm{Var}(O_i - E_i) = \sum_{f=1}^{k} \left(\frac{n_{if}\left(n_f - n_{if}\right)m_{if}\left(n_f - m_f\right)}{n_f^2\left(n_f - 1\right)}\right)$$

$$\mathrm{Cov}(O_i - E_i,\ O_l - E_l) = \sum_{f=1}^{k} \left(\frac{-n_{if}\, n_{lf}\, m_f\left(n_f - m_f\right)}{n_f^2\left(n_f - 1\right)}\right)$$

$$\mathbf{d} = (O_1 - E_1,\ O_2 - E_2,\ \ldots,\ O_{G-1} - E_{G-1})'$$

$$\mathbf{V} = ((v_{il}))$$

where $v_{ii} = \mathrm{Var}\,(O_i - E_i)$ and $v_{il} = \mathrm{Cov}\,(O_i - E_i, O_l - E_l)$ for $i = 1, 2, \ldots, G - 1$; $l = 1, 2, \ldots, G - 1$.

Then, the log–rank statistic is given by the matrix product formula:

$$\boxed{\text{Log-rank statistic} \;=\; \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}}$$

which has approximately a chi-square distribution with $G - 1$ degrees of freedom under the null hypothesis that all G groups have a common survival curve.

3 The Cox Proportional Hazards Model and Its Charact-eristics

D.G. Kleinbaum and M.Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_3,
© Springer Science+Business Media, LLC 2012

Introduction

We begin by discussing some computer results using the Cox PH model, without actually specifying the model; the purpose here is to show the similarity between the Cox model and standard linear regression or logistic regression.

We then introduce the Cox model and describe why it is so popular. In addition, we describe its basic properties, including the meaning of the proportional hazards assumption and the Cox likelihood. We also describe how and why we might consider using "age as the time scale" instead of "time-on follow-up" as the outcome variable.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

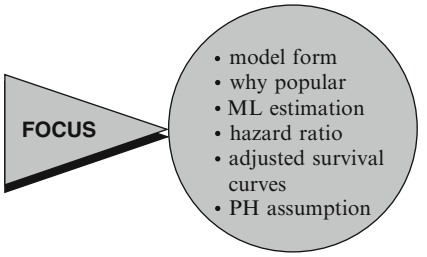
I. **A computer example using the Cox PH model** (pages 100–108)
II. **The formula for the Cox PH model** (pages 108–110)
III. **Why the Cox PH model is popular** (pages 110–112)
IV. **ML estimation of the Cox PH model** (pages 112–114)
V. **Computing the hazard ratio** (pages 114–117)
VI. **Interval estimation: interaction** (pages 117–119)
VII. **Adjusted survival curves using the Cox PH model** (pages 120–123)
VIII. **The meaning of the PH assumption** (pages 123–127)
IX. **The Cox likelihood** (pages 127–131)
X. **Using age as the time scale** (pages 131–142)
XI. **Summary** (pages 143–144)

Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize the general form of the Cox PH model.
2. State the specific form of a Cox PH model appropriate for the analysis, given a survival analysis scenario involving one or more explanatory variables.
3. State or recognize the form and properties of the baseline hazard function in the Cox PH model.
4. Give three reasons for the popularity of the Cox PH model.
5. State the formula for a designated hazard ratio of interest given a scenario describing a survival analysis using a Cox PH model, when
   a. there are confounders but no interaction terms in the model;
   b. there are both confounders and interaction terms in the model.
6. State or recognize the meaning of the PH assumption.
7. Determine and explain whether the PH assumption is satisfied when the graphs of the hazard functions for two groups cross each other over time.
8. State or recognize what is an adjusted survival curve.
9. Compare and/or interpret two or more adjusted survival curves.
10. Given a computer printout involving one or more fitted Cox PH models,
    a. compute or identify any hazard ratio(s) of interest;
    b. carry out and interpret a designated test of hypothesis;
    c. carry out, identify or interpret a confidence interval for a designated hazard ratio;
    d. evaluate interaction and confounding involving one or more covariates.
11. Give an example of how the Cox PH likelihood is formed.
12. Given left truncated survival data, describe how and when you would consider using "age as the time scale" instead of "time-on follow-up" as the outcome variable.
13. Given left-truncated survival data, state the hazard function formula that uses "age as the time scale" as the outcome variable.
14. Illustrate the difference between an "open cohort" and a "closed cohort".

Presentation



This presentation describes the Cox proportional hazards (PH) model, a popular mathematical model used for analyzing survival data. Here, we focus on the model form, why the model is popular, maximum likelihood (ML) estimation of the model parameters, the formula for the hazard ratio, how to obtain adjusted survival curves, and the meaning of the PH assumption.

I. A Computer Example Using the Cox PH Model

EXAMPLE

Leukemia Remission Data

Group 1($n = 21$)		Group 2($n = 21$)	
t (weeks)	log WBC	t (weeks)	log WBC
6	2.31	1	2.80
6	4.06	1	5.00
6	3.28	2	4.91
7	4.43	2	4.48
10	2.96	3	4.01
13	2.88	4	4.36
16	3.60	4	2.42
22	2.32	5	3.49
23	2.57	5	3.97
6+	3.20	8	3.52
9+	2.80	8	3.05
10+	2.70	8	2.32
11+	2.60	8	3.26
17+	2.16	11	3.49
19+	2.05	11	2.12
20+	2.01	12	1.50
25+	1.78	12	3.06
32+	2.20	15	2.30
32+	2.53	17	2.95
34+	1.47	22	2.73
35+	1.45	23	1.97

+ denotes censored observation

We introduce the Cox PH model using computer output from the analysis of remission time data (Freireich et al., *Blood*, 1963), which we previously discussed in Chapters 1 and 2. The data set is listed here at the left.

These data involve two groups of leukemia patients, with 21 patients in each group. Group 1 is the treatment group, and group 2 is the placebo group. The data set also contains the variable log WBC, which is a well-known prognostic indicator of survival for leukemia patients.

For this example, the basic question of interest concerns comparing the survival experience of the two groups adjusting for the possible confounding and/or interaction effects of log WBC.

EXAMPLE: (continued)

T = weeks until going out of remission
X_1 = group status = E
X_2 = log WBC (confounding?)

Interaction?
$X_3 = X_1 \times X_2$ = group status × log WBC

Computer results for three Cox PH models using the Stata package

Other computer packages provide similar information.

Computer Appendix: uses Stata, SAS, and SPSS on the same dataset.

We are thus considering a problem involving two explanatory variables as predictors of survival time T, where T denotes "weeks until going out of remission." We label the explanatory variables X_1 (for group status) and X_2 (for log WBC). The variable X_1 is the primary study or exposure variable of interest. The variable X_2 is an extraneous variable that we are including as a possible confounder or effect modifier.

Note that if we want to evaluate the possible interaction effect of log WBC on group status, we would also need to consider a third variable, that is, the product of X_1 and X_2.

For this dataset, the computer results from fitting three different Cox proportional hazards models are presented below. The computer package used is Stata. This is one of several packages that have procedures for carrying out a survival analysis using the Cox model. The information printed out by different packages will not have exactly the same format, but they will provide similar information. A comparison of output using Stata, SAS, SPSS, and R procedures on the same dataset is provided in the computer appendix at the back of this text.

Edited Output From Stata:

Model 1:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.509	0.410	3.68	0.000	4.523	2.027	10.094
No. of subjects = 42		Log likelihood = −86.380			Prob > chi2 = 0.0001		

Model 2:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486
No. of subjects = 42		Log likelihood = −72.280			Prob > chi2 = 0.0000		

Model 3:

	Coef.	Std. Err.	z	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Rx	2.355	1.681	1.40	0.161	10.537	0.391	284.201
log WBC	1.803	0.447	4.04	0.000	6.067	2.528	14.561
Rx x log WBC	−0.342	0.520	−0.66	0.510	0.710	0.256	1.967
No. of subjects = 42		Log likelihood = −72.066			Prob > chi2 = 0.0000		

EDITED OUTPUT FROM STATA

Model 1:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

No. of subjects = 42 Log likelihood = −86.380

Hazard ratios

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

No. of subjects = 42 Log likelihood = −72.066

We now describe how to use the computer printout to evaluate the possible effect of treatment status on remission time adjusted for the potential confounding and interaction effects of the covariate log WBC. For now, we focus only on five columns of information provided in the printout, as presented at the left for all three models.

For each model, the first column identifies the **variables** that have been included in the model. The second column gives estimates of **regression coefficients** corresponding to each variable in the model. The third column gives **standard errors** of the estimated regression coefficients. The fourth column gives **p-values** for testing the significance of each coefficient. The fifth column, labeled as **Haz. Ratio**, gives e^{Coef} for each variable in each model.

Models 1 and 2: $e^{coef} = HR$
Model 3: HR formula more complicated

As we discuss later in this chapter, e^{Coef} gives an estimated **hazard ratio (HR)** for the effect of each variable adjusted for the other variables in a model (e.g. Models 1 and 2) without product terms. With product terms such as Rx × log WBC in Model 3, the hazard ratio formula is more complicated, as we also discuss later.

Except for the Haz. Ratio column, these computer results are typical of output found in standard linear regression printouts. As the printout suggests, we can analyze the results from a Cox model in a manner similar to the way we would analyze a linear regression model.

EXAMPLE: (continued)

Same dataset for each model
$n = 42$ subjects
T = time (weeks) until out of remission
Model 1: *Rx* only
Model 2: *Rx* and log WBC
Model 3: *Rx*, log WBC, and *Rx* × log WBC

We now distinguish among the output for the three models shown here. All three models are using the same remission time data on 42 subjects. The outcome variable for each model is the same: time in weeks until a subject goes out of remission. However, the independent variables are different for each model. Model 1 contains only the treatment status variable, indicating whether a subject is in the treatment or placebo group. Model 2 contains two variables, treatment status and log WBC. And model 3 contains an interaction term defined as the product of treatment status and log WBC.

EDITED OUTPUT: ML ESTIMATION

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

No. of subjects = 42 Log likelihood = −72.066

We now focus on the output for model 3. The method of estimation used to obtain the coefficients for this model, as well as the other two models, is maximum likelihood (ML) estimation. Note that a p-value of 0.510 is obtained for the coefficient of the product term for the interaction of treatment with log WBC. This p-value indicates that there is no significant interaction effect, so that we can drop the product term from the model and consider the other two models instead.

EXAMPLE: (continued)

$P = 0.510 : \dfrac{-0.342}{-0.520} = -0.66 = Z$ Wald statistic

LR statistic: uses Log likelihood = −72.066

−2 ln L (log likelihood statistic) = −2 × (−72.066) = 144.132

The p-value of 0.510 that we have just described is obtained by dividing the coefficient −0.342 of the product term by its standard error of 0.520, which gives −0.66, and then assuming that this quantity is approximately a standard normal or *Z* variable. This *Z* statistic is known as a **Wald statistic,** which is one of two test statistics typically used with ML estimates. The other test statistic, called the **likelihood ratio,** or *LR* statistic, makes use of the log likelihood statistic. The log likelihood statistic is obtained by multiplying the "Log likelihood" in the Stata output by −2 to get −2 ln L.

Edited Output

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280
−2 ln L = −2× (-72.280) = 144.550

We now look at the output for model 2, which contains two variables. The treatment status variable (*Rx*) represents the exposure variable of primary interest. The log WBC variable is being considered as a confounder. Our goal is to describe the effect of treatment status adjusted for log WBC. Note that for Model 2, −2 ln L equals 144.550.

EXAMPLE: (continued)

LR (interaction in model 3)
$= -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}})$

In general:
$LR = -2 \ln L_{\text{R}} - (-2 \ln L_{\text{F}})$

To use the likelihood ratio (LR) statistic to test the significance of the interaction term, we need to compute the difference between the log likelihood statistic of the reduced model which does not contain the interaction term (model 2) and the log likelihood statistic of the full model containing the interaction term (model 3). In general, the LR statistic can be written in the form $-2 \ln L_R$ minus $-2 \ln L_F$, where *R* denotes the reduced model and *F* denotes the full model.

EXAMPLE: (continued)

LR (interaction in model 3)
$= -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}})$
$= (-2 \times -72.280) - (-2 \times -72.066)$
$= 144.550 - 144.132 = 0.428$

(LR is χ^2 with 1 d.f. under H_0: no interaction.)
$0.40 < P < 0.50$, **not significant**
Wald test $P = 0.510$

To obtain the LR statistic in this example, we compute 144.550 minus 144.132 to obtain 0.428. Under the null hypothesis of no interaction effect, the test statistic has a chi-square distribution with p degrees of freedom, where p denotes the number of predictors being assessed. The p-value for this test is between 0.40 and 0.50, which indicates no significant interaction. Although the p-values for the Wald test (0.510) and the LR test are not exactly the same, both p-values lead to the same conclusion.

LR $\neq$ Wald
When in doubt, use the LR test.

In general, the LR and Wald statistics may not give exactly the same answer. Statisticians have shown that of the two test procedures, the LR statistic has better statistical properties, so when in doubt, you should use the LR test.

OUTPUT

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

We now focus on how to assess the effect of treatment status adjusting for log WBC using the model 2 output, again shown here.

Three statistical objectives.

1. **test for significance of effect**
2. **point estimate of effect**
3. **confidence interval for effect**

There are three statistical objectives typically considered. One is to **test for the significance** of the treatment status variable, adjusted for log WBC. Another is to obtain a **point estimate of the effect** of treatment status, adjusted for log WBC. And a third is to obtain a **confidence interval for this effect**. We can accomplish these three objectives using the output provided, without having to explicitly describe the formula for the Cox model being used.

EXAMPLE: (continued)

Test for treatment effect:
Wald statistic: $P = 0.002$ (highly significant)
LR statistic: compare
$-2 \log L$ from model 2 with
$-2 \log L$ from model without Rx variable
Printout not provided here

Conclusion: treatment effect is significant, after adjusting for log WBC

To test for the significance of the treatment effect, the p-value provided in the table for the Wald statistic is 0.002, which is highly significant. Alternatively, a likelihood ratio (LR) test could be performed by comparing the log likelihood statistic (144.559) for model 2 with the log likelihood statistic for a model which does not contain the treatment variable. This latter model, which should contain only the log WBC variable, is not provided here, so we will not report on it other than to note that the LR test is also very significant. Thus, these test results show that using model 2, the treatment effect is significant, after adjusting for log WBC.

EXAMPLE: (continued)

Point estimate:

$$\widehat{HR} = 3.648 = e^{1.294}$$

Coefficient of treatment variable

A point estimate of the effect of the treatment is provided in the *HR* column by the value 3.648. This value gives the estimated hazard ratio (HR) for the effect of the treatment; in particular, we see that the hazard for the placebo group is 3.6 times the hazard for the treatment group. Note that the value 3.648 is calculated as e to the coefficient of the treatment variable; that is, *e* to the 1.294 equals 3.648.

To describe the confidence interval for the effect of treatment status, we consider the output for the extended table for model 2 given earlier.

Output

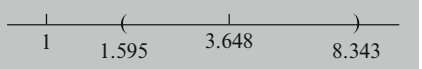
Model 2:

	Coef.	Std. Err.	z	P>|z|	Haz. Ratio	[95% Conf. Interval]	
Rx	1.294	0.422	3.07	0.002	3.648	1.595	8.343
Log WBC	1.604	0.329	4.87	0.000	4.975	2.609	9.486

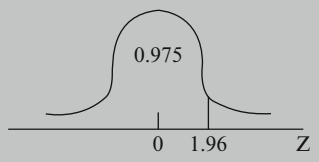
No. of subjects = 42 Log likelihood = −72.280 Prob > chi2 = 0.0000

EXAMPLE: (continued)

95% confidence interval for the *HR*: (1.595, 8.343)



95% CI for β_1: $1.294 \pm (1.96)\,(0.422)$



95% CI for $HR = e^{\beta_1}$

$$\exp\left[\hat{\beta}_1 \pm 1.96 s_{\hat{\beta}_1}\right] = e^{1.294 \pm 1.96(0.422)}$$

From the table, we see that a 95% confidence interval for the treatment effect is given by the range of values 1.595–8.343. This is a confidence interval for the hazard ratio (HR), which surrounds the point estimate of 3.648 previously described. Notice that this confidence interval is fairly wide, indicating that the point estimate is somewhat unreliable. As expected from the low p-value of 0.002, the confidence interval for HR does not contain the null value of 1.

The calculation of the confidence interval for HR is carried out as follows:

1. Compute a 95% confidence interval for the regression coefficient of the *Rx* variable (β_1). The large sample formula is 1.294 plus or minus 1.96 times the standard error 0.422, where 1.96 is the 97.5 percentile of the standard normal or *Z* distribution.
2. Exponentiate the two limits obtained for the confidence interval for the regression coefficient of *Rx*.

Stata: provides CI directly

Other packages: provide $\hat{\beta}$ and $s_{\hat{\beta}}$

The Stata output provides the required confidence interval directly, so that the user does not have to carry out the computations required by the large sample formula. Other computer packages may not provide the confidence interval directly, but, rather, may provide only the estimated regression coefficients and their standard errors.

EDITED OUTPUT

Model 1:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

No. of subjects = 42 Log likelihood = −86.380

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

To this point, we have made use of information from outputs for models 2 and 3, but have not yet considered the model 1 output, which is shown again here. Note that model 1 contains only the treatment status variable, whereas model 2, shown below, contains log WBC in addition to treatment status. Model 1 is sometimes called the "crude" model because it ignores the effect of potential covariates of interest, like log WBC.

EXAMPLE: (continued)

HR for model 1 (4.523) is higher than *HR* for model 2 (3.648)

Confounding: crude versus adjusted $\widehat{HR}$ are meaningfully different.

Model 1 can be used in comparison with model 2 to evaluate the potential confounding effect of the variable log WBC. In particular, notice that the value in the *HR* column for the treatment status variable is 4.523 for model 1, but only 3.648 for model 2. Thus, the crude model yields an estimated hazard ratio that is somewhat higher than the corresponding estimate obtained when we adjust for log WBC. If we decide that the crude and adjusted estimates are meaningfully different, we then say that there is confounding due to log WBC.

Confounding due to log WBC ⇒ must control for log WBC, i.e., prefer model 2 to model 1.

Once we decide that confounding is present, we then **must** control for the confounder, in this case, log WBC, in order to obtain a valid estimate of the effect. Thus, we prefer model 2, which controls for log WBC, to model 1, which does not.

If no confounding, then consider precision: e.g., if 95% CI is narrower for model 2 than model 1, we prefer model 2.

Note that if we had decided that there is no "meaningful" confounding, then we would not need to control for log WBC to get a valid answer. Nevertheless, we might wish to control for log WBC anyhow, to obtain a more precise estimate of the hazard ratio. That is, if the confidence interval for the HR is narrower when using model 2 than when using model 1, we would prefer model 2 to model 1 for **precision** gain.

EDITED OUTPUT: Confidence Intervals

	[95% Conf. Interval]	
Rx model 1	2.027	10.094
	width = **8.067**	
	width = **6.748**	
Rx model 2	1.595	8.343
log WBC	2.609	9.486

The confidence intervals for *Rx* in each model are shown here at the left. The interval for *Rx* in model 1 has width equal to 10.094 minus 2.027, or **8.067**; for model 2, the width is 8.343 minus 1.595, or **6.748**. Therefore, model 2 gives a more precise estimate of the hazard ratio than does model 1.

EXAMPLE: (continued)

Model 2 is best model.

$\widehat{HR} = 3.648$ statistically significant

95% CI for *HR*: (1.6, 8.3)

Our analysis of the output for the three models has led us to conclude that model 2 is the best of the three models and that, using model 2, we get a statistically significant hazard ratio of 3.648 for the effect of the treatment, with a 95% confidence interval ranging between 1.6 and 8.3.

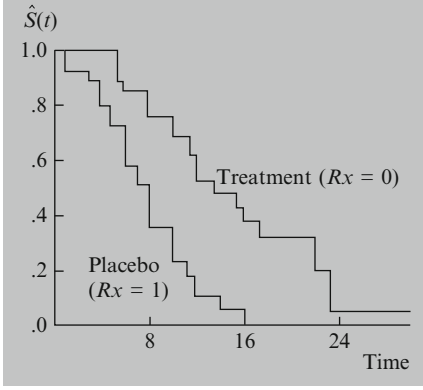
Cox model formulae not specified

Analysis strategy and methods for Cox model analogous to those for logistic and classical linear models.

Note that we were able to carry out this analysis without actually specifying the formulae for the Cox PH models being fit. Also, the strategy and methods used with the output provided have been completely analogous to the strategy and methods one uses when fitting logistic regression models (see Kleinbaum and Klein, *Logistic Regression*, Chapters 6 and 7, 2010), and very similar to carrying out a classical linear regression analysis (see Kleinbaum et al., *Applied Regression Analysis*, 4th ed., Chapter 16, 2008).

EXAMPLE: (continued)

Survival Curves Adjusted for log WBC (Model 2)



In addition to the above analysis of this data, we can also obtain survival curves for each treatment group, **adjusted** for the effects of log WBC and based on the model 2 output. Such curves, sketched here at the left, give additional information to that provided by estimates and tests about the hazard ratio. In particular, these curves describe how the treatment groups compare over the time period of the study.

For these data, the survival curves show that the treatment group consistently has higher survival probabilities than the placebo group after adjusting for log WBC. Moreover, the difference between the two groups appears to widen over time.

Adjusted survival curves	KM curves
Adjusted for covariates	No covariates
Use fitted Cox model	No Cox model fitted

Note that adjusted survival curves are mathematically different from Kaplan–Meier (KM) curves. KM curves do not adjust for covariates and, therefore, are not computed using results from a fitted Cox PH model.

Nevertheless, for these data, the plotted KM curves (which were described in Chapter 2) are similar in appearance to the adjusted survival curves.

Remainder:

- Cox model formula
- basic characteristics of Cox model
- meaning of PH assumption

In the remainder of this presentation, we describe the Cox PH formula and its basic characteristics, including the meaning of the PH assumption and the Cox likelihood.

II. The Formula for the Cox PH Model

$$h(t,\mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i}$$

$\mathbf{X} = (X_1, X_2, \ldots X_p)$
explanatory/predictor variables

$h_0(t)$	$\times$	$e^{\sum_{i=1}^{p}\beta_i X_i}$
Baseline hazard		Exponential
Involves t but not X's		Involves X's but not t (X's are time-independent)

The Cox PH model is usually written in terms of the hazard model formula shown here at the left. This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by the bold $\mathbf{X}$. That is, the bold $\mathbf{X}$ represents a collection (sometimes called a "vector") of predictor variables that is being modeled to predict an individual's hazard.

The Cox model formula says that the hazard at time t is the product of two quantities. The first of these, $h_0(t)$, is called the **baseline hazard** function. The second quantity is the exponential expression e to the linear sum of $\beta_i X_i$, where the sum is over the p explanatory X variables.

An important feature of this formula, which concerns the proportional hazards (*PH*) assumption, is that the baseline hazard is a function of t, but does not involve the X's. In contrast, the exponential expression shown here, involves the X's, but does not involve t. The X's here are called **time-independent** X's.

X's involving *t*: time-dependent

Requires extended Cox model (no PH)

It is possible, nevertheless, to consider *X*'s which do involve *t*. Such *X*'s are called **time-dependent** variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the **extended Cox model**.

Time-dependent variables: Chapter 6

The use of time-dependent variables is discussed in Chapter 6. For the remainder of this presentation, we will consider time-independent *X*'s only.

Time-independent variable: Values for a given individual do not change over time; e.g., SEX and SMK

Assumed not to change once measured

A time-independent variable is defined to be any variable whose value for a given individual does not change over time. Examples are SEX and smoking status (SMK). Note, however, that a person's smoking status may actually change over time, but for purposes of the analysis, the SMK variable is assumed not to change once it is measured, so that only one value per individual is used.

AGE and WGT values do not change much, or effect on survival depends on one measurement.

Also note that although variables like AGE and weight (WGT) change over time, it may be appropriate to treat such variables as time-independent in the analysis if their values do not change much over time or if the effect of such variables on survival risk depends essentially on the value at only one measurement.

$$X_1 = X_2 = \cdots = X_p = 0$$

$$h(t, \mathbf{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i}$$

$$= h_0(t)\, e^0$$

$$= h_0(t)$$

Baseline hazard

The Cox model formula has the property that if all the *X*'s are equal to zero, the formula reduces to the baseline hazard function. That is, the exponential part of the formula becomes *e* to the zero, which is 1. This property of the Cox model is the reason why $h_0(t)$ is called the baseline function.

No *X*'s in model: $h(t,\mathbf{X}) = h_0(t)$.

Or, from a slightly different perspective, the Cox model reduces to the baseline hazard when no *X*'s are in the model. Thus, $h_0(t)$ may be considered as a starting or "baseline" version of the hazard function, prior to considering any of the *X*'s.

$h_0(t)$ is unspecified.

Cox model: **semiparametric**

Another important property of the Cox model is that the baseline hazard, $h_0(t)$, is an unspecified function. It is this property that makes the Cox model a **semiparametric** model.

Example: Parametric Model

Weibull:
$h(t, \mathbf{X}) = \lambda p t^{p-1}$
where $\lambda = \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$
and $h_0(t) = p t^{p-1}$

In contrast, a **parametric** model is one whose functional form is completely specified, except for the values of the unknown parameters. For example, the Weibull hazard model is a parametric model and has the form shown here, where the unknown parameters are λ, p, and the β_i's. Note that for the Weibull model, $h_0(t)$ is given by $\lambda p t^{p-1}$ (see Chapter 7).

Semiparametric property
⇓
Popularity of the Cox model

One of the reasons why the Cox model is so popular is that it is semiparametric. We discuss this and other reasons in the next section (III) concerning why the Cox model is so widely used.

III. Why the Cox PH Model Is Popular

Cox PH model is "robust": Will closely approximate correct parametric model

A key reason for the popularity of the Cox model is that, even though the baseline hazard is not specified, reasonably good estimates of regression coefficients, hazard ratios of interest, and adjusted survival curves can be obtained for a wide variety of data situations. Another way of saying this is that the Cox PH model is a "robust" model, so that the results from using the Cox model will closely approximate the results for the correct parametric model.

If correct model is:

Weibull	⇒	Cox model will approximate Weibull
Exponential	⇒	Cox model will approximate exponential

For example, if the correct parametric model is Weibull, then use of the Cox model typically will give results comparable to those obtained using a Weibull model. Or, if the correct model is exponential, then the Cox model results will closely approximate the results from fitting an exponential model.

Prefer parametric model if sure of correct model, e.g., use goodness-of-fit test (Lee, 1982).

We would prefer to use a parametric model if we were sure of the correct model. Although there are various methods for assessing goodness of fit of a parametric model (for example, see Lee, *Statistical Methods for Survival Data Analysis*, 1982), we may not be completely certain that a given parametric model is appropriate.

When in doubt, the Cox model is a "safe" choice.

Thus, when in doubt, as is typically the case, the Cox model will give reliable enough results so that it is a "safe" choice of model, and the user does not need to worry about whether the wrong parametric model is chosen.

In addition to the general "robustness" of the Cox model, the specific form of the model is attractive for several reasons.

$$h(t,\mathbf{X}) = \underbrace{h_0(t)}_{\text{Baseline hazard}} \times \underbrace{e^{\sum_{i=1}^{p}\beta_i X_i}}_{\begin{array}{c}\text{Exponential}\\ \Downarrow \\ 0 \le h(t,\mathbf{X}) < \infty \text{ always}\end{array}}$$

As described previously, the specific form of the Cox model gives the hazard function as a product of a baseline hazard involving t and an exponential expression involving the X's without t. The exponential part of this product is appealing because it ensures that the fitted model will always give estimated hazards that are non-negative.

$$h_0(t) \times \underbrace{\sum_{i=1}^{p}\beta_i X_i}_{\text{Linear}} \Downarrow \text{Might be} < 0$$

We want such nonnegative estimates because, by definition, the values of any hazard function must range between zero and plus infinity, that is, a hazard is always nonnegative. If, instead of an exponential expression, the X part of the model were, for example, linear in the X's, we might obtain negative hazard estimates, which are not allowed.

Even though $h_0(t)$ is unspecified, we can estimate the β's.

Measure of effect: hazard ratio (*HR*) involves only β's, without estimating $h_0(t)$.

Another appealing property of the Cox model is that, even though the baseline hazard part of the model is unspecified, it is still possible to estimate the β's in the exponential part of the model. As we will show later, all we need are estimates of the β's to assess the effect of explanatory variables of interest. The measure of effect, which is called a hazard ratio, is calculated without having to estimate the baseline hazard function.

Can estimate $h(t,\mathbf{X})$ and $S(t, \mathbf{X})$ for Cox model using a minimum of assumptions.

Note that the hazard function $h(t,\mathbf{X})$ and its corresponding survival curves $\mathbf{S}(t,\mathbf{X})$ can be estimated for the Cox model even though the baseline hazard function is not specified. Thus, with the Cox model, using a minimum of assumptions, we can obtain the primary information desired from a survival analysis, namely, a hazard ratio and a survival curve.

Cox model preferred to **logistic** model.

⇓	⇓
Uses survival times and censoring	Uses (0,1) outcome; ignores survival times and censoring

One last point about the popularity of the Cox model is that it is preferred over the logistic model when survival time information is available and there is censoring. That is, the Cox model uses more information, the survival times, than the logistic model, which considers a (0, 1) outcome and ignores survival times and censoring.

IV. ML Estimation of the Cox PH Model

$$h(t, \mathbf{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i}$$

We now describe how estimates are obtained for the parameters of the Cox model. The parameters are the β's in the general Cox model formula shown here. The corresponding estimates of these parameters are called maximum likelihood (ML) estimates and are denoted as $\hat{\beta}_i$.

ML estimates: $\hat{\beta}_i$

	Coef.	Std.Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

No. of subjects = 42 Log likelihood = −72.280

As an example of ML estimates, we consider once again the computer output for one of the models (model 2) fitted previously from remission data on 42 leukemia patients.

Estimated model:

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t)\, e^{1.294\, Rx\, +\, 1.604\, \log\, \mathrm{WBC}}$$

The Cox model for this example involves two parameters, one being the coefficient of the treatment variable (denoted here as *Rx*) and the other being the coefficient of the log WBC variable. The expression for this model is shown at the left, which contains the estimated coefficients 1.294 for *Rx* and 1.604 for log white blood cell count.

ML estimates: maximize likelihood function L

L = joint probability of observed data = $L(\beta)$

As with logistic regression, the ML estimates of the Cox model parameters are derived by maximizing a likelihood function, usually denoted as L. The likelihood function is a mathematical expression which describes the joint probability of obtaining the data actually observed on the subjects in the study as a function of the unknown parameters (the β's) in the model being considered. L is sometimes written notationally as $L(\beta)$ where β denotes the collection of unknown parameters.

The expression for the likelihood is developed at the end of the chapter. However, we give a brief overview below.

L is a partial likelihood:

- considers probabilities only for subjects who fail
- does not consider probabilities for subjects who are censored

The formula for the Cox model likelihood function is actually called a "partial" likelihood function rather than a (complete) likelihood function. The term "partial" likelihood is used because the likelihood formula considers probabilities only for those subjects who fail, and does not explicitly consider probabilities for those subjects who are censored. Thus the likelihood for the Cox model does not consider probabilities for all subjects, and so it is called a "partial" likelihood.

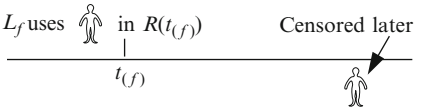
Number of failure times

$$L = L_1 \times L_2 \times L_3 \times \cdots \times L_k = \prod_{j=1}^{k} L_j$$

where
L_f = portion of L for the jth failure time given the risk set $R\,(t_{(f)})$

In particular, the partial likelihood can be written as the product of several likelihoods, one for each of, say, k failure times. Thus, at the f-th failure time, L_f denotes the likelihood of failing at this time, given survival up to this time. Note that the set of individuals at risk at the jth failure time is called the "risk set," $R(t_{(f)})$, and this set will change – actually get smaller in size – as the failure time increases.

Information on censored subjects used prior to censorship.



Thus, although the partial likelihood focuses on subjects who fail, survival time information prior to censorship is used for those subjects who are censored. That is, a person who is censored *after* the f-th failure time is part of the risk set used to compute L_f even though this person is censored later.

Steps for obtaining ML estimates:

- form L from model
- maximize ln L by solving

$$\frac{\partial \ln L}{\partial \beta_i} = 0$$
$$i = 1, \ldots, p(\#\text{ of parameters})$$

Once the likelihood function is formed for a given model, the next step for the computer is to maximize this function. This is generally done by maximizing the natural log of L, which is computationally easier.

Solution by iteration:

- guess at solution
- modify guess in successive steps
- stop when solution is obtained

The maximization process is carried out by taking partial derivatives of log of L with respect to each parameter in the model, and then solving a system of equations as shown here. This solution is carried out using **iteration**. That is, the solution is obtained in a stepwise manner, which starts with a guessed value for the solution, and then successively modifies the guessed value until a solution is finally obtained.

Statistical inferences for hazard ratios: (See Section I, pages 100–107)

Test hypotheses	Confidence intervals
Wald test LR test	Large sample 95% CI

$\widehat{HR} = e^{\hat{\beta}}$ for a (0, 1) exposure variable (no interaction)

Once the ML estimates are obtained, we are usually interested in carrying out statistical inferences about hazard ratios defined in terms of these estimates. We illustrated previously how to test hypotheses and form confidence intervals for the hazard ratio in Section I above. There, we described how to compute a Wald test and a likelihood ratio (LR) test. We also illustrated how to calculate a large sample 95% confidence interval for a hazard ratio. The estimated hazard ratio (HR) was computed by exponentiating the coefficient of a (0,1) exposure variable of interest. Note that the model contained no interaction terms involving exposure.

V. Computing the Hazard Ratio

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})}$$

where

$$\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)$$

and

$$\mathbf{X} = (X_1, X_2, \cdots, X_p)$$

denote the set of X's for two individuals

In general, a hazard ratio (HR) is defined as the hazard for one individual divided by the hazard for a different individual. The two individuals being compared can be distinguished by their values for the set of predictors, that is, the X's.

We can write the hazard ratio as the estimate of $h(t,\mathbf{X}^*)$ divided by the estimate of $h(t,\mathbf{X})$, where $\mathbf{X}^*$ denotes the set of predictors for one individual, and $\mathbf{X}$ denotes the set of predictors for the other individual.

To interpret $\widehat{HR}$, want $\widehat{HR} > 1$, i.e., $\hat{h}(t, \mathbf{X}^*) > \hat{h}(t, \mathbf{X})$.

Typical coding: $\mathbf{X}^*$: group with larger h
$\mathbf{X}$: group with smaller h

Note that, as with an odds ratio, it is easier to interpret an HR that exceeds the null value of 1 than an HR that is less than 1. Thus, the X's are typically coded so that group with the larger hazard corresponds to $\mathbf{X}^*$, and the group with the smaller hazard corresponds to $\mathbf{X}$. As an example, for the remission data described previously, the placebo group is coded as $X_1^* = 1$, and the treatment group is coded as $X_1 = 0$.

EXAMPLE: Remission Data

$\mathbf{X}^* = (X_1^*, X_2^*, \cdots, X_p^*)$, where $X_1^* = 1$ denotes **placebo** group.

$\mathbf{X} = (X_1, X_2, \ldots, X_p)$, where $X_1 = 0$ denotes **treatment** group.

$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\cancel{\hat{h}_0(t)}\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\cancel{\hat{h}_0(t)} e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}}$$

We now obtain an expression for the *HR* formula in terms of the regression coefficients by substituting the Cox model formula into the numerator and denominator of the hazard ratio expression. This substitution is shown here. Notice that the only difference in the numerator and denominator are the *X**'s versus the *X*'s. Notice also that the baseline hazards will cancel out.

$$\widehat{HR} = \frac{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i^*}}{\hat{h}_0(t)\, e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}} = e^{\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)}$$

Using algebra involving exponentials, the hazard ratio formula simplifies to the exponential expression shown here. Thus, the hazard ratio is computed by exponentiating the sum of each β_i "hat" times the difference between X_i^* and X_i.

$$\boxed{\widehat{HR} = \exp\left[\sum_{i=1}^{p} \beta_i \left(X_i^* - X_i\right)\right]}$$

An alternative way to write this formula, using exponential notation, is shown here. We will now illustrate the use of this general formula through a few examples.

EXAMPLE

$\mathbf{X} = (X_1, X_2, \ldots, X_p) = (X_1)$, where X_1 denotes (0, 1) exposure status ($p = 1$)
$X_1^* = 1, X_1 = 0$

$$\widehat{HR} = \exp\left[\hat{\beta}_1 \left(X_1^* - X_1\right)\right]$$
$$= \exp\left[\hat{\beta}_1 (1 - 0)\right] = e^{\hat{\beta}_1}$$

Suppose, for example, there is only one X variable of interest, X_1, which denotes (0,1) exposure status, so that $p = 1$. Then, the hazard ratio comparing exposed to unexposed persons is obtained by letting $X_1^* = 1$ and $X_1 = 0$ in the hazard ratio formula. The estimated hazard ratio then becomes e to the quantity β_1 "hat" times 1 minus 0, which simplifies to e to the β_1 "hat."

Model 1:

	Coef.	Std. Err.	P > |z|	Haz. Ratio
Rx	1.509	0.410	0.000	4.523

Recall the remission data printout for Model 1, which contains only the *Rx* variable, again shown here. Then the estimated hazard ratio is obtained by exponentiating the coefficient 1.509, which gives the value 4.523 shown in the *HR* column of the output.

EXAMPLE 2

Model 2:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	1.294	0.422	0.002	3.648
log WBC	1.604	0.329	0.000	4.975

$\mathbf{X}^* = (1, \log \text{WBC})$, $\mathbf{X} = (0, \log \text{WBC})$
HR for effect of *Rx* adjusted for log WBC:

As a second example, consider the output for Model 2, which contains two variables, the *Rx* variable and log WBC. Then to obtain the hazard ratio for the effect of the *Rx* variable adjusted for the log WBC variable, we let the vectors $\mathbf{X}^*$ and $\mathbf{X}$ be defined as $\mathbf{X}^* = (1, \log \text{WBC})$ and $\mathbf{X} = (0, \log \text{WBC})$. Here we assume that log WBC is the same for $\mathbf{X}^*$ and $\mathbf{X}$ though unspecified.

EXAMPLE 2: (continued)

$$\widehat{HR} = \exp\left[\hat{\beta}_1(X_1^* - X_1) + \hat{\beta}_1(X_2^* - X_2)\right]$$
$$= \exp[1.294(1-0) + 1.604(\log\ \text{WBC} - \log\ \text{WBC})]$$
$$= \exp[1.294(1) + 1.604(0)] = \boxed{e^{1.294}}$$

The estimated hazard ratio is then obtained by exponentiating the sum of two quantities, one involving the coefficient 1.294 of the *Rx* variable, and the other involving the coefficient 1.604 of the log WBC variable. Since the log WBC value is fixed, however, this portion of the exponential is zero, so that the resulting estimate is simply e to the 1.294.

General rule: If X_1 is a (0,1) exposure variable, then

$\widehat{HR} = e^{\hat{\beta}_1}$ (= effect of exposure adjusted for other X's)

provided no other X's are product terms involving exposure.

This second example illustrates the general rule that the hazard ratio for the effect of a (0,1) exposure variable which adjusts for other variables is obtained by exponentiating the estimated coefficient of the exposure variable. This rule has the proviso that the model does not contain any product terms involving exposure.

EXAMPLE 3

Model 3:

	Coef.	Std. Err.	p > |z|	Haz. Ratio
Rx	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
Rx × log WBC	−0.342	0.520	0.510	0.710

Want *HR* for effect of *Rx* adjusted for log WBC.

Placebo subject:

$$\mathbf{X}^* = (X_1^* = 1,\ X_2^* = \log\ \text{WBC},\ X_3^* = 1 \times \log\ \text{WBC})$$

Treated subject:

$$\boldsymbol{X} = (X_1 = 0, X_2 = \log \text{WBC}, X_3 = 0 \times \log \text{WBC})$$

$$\widehat{HR} = \exp\left[\sum_{i=1}^{3} \hat{\beta}_i(X_i^* - X_i)\right]$$
$$\widehat{HR} = \exp[2.355(1-0) + 1.803(\log\ \text{WBC} - \log\ \text{WBC}) + (-0.342)(1 \times \log\ \text{WBC} - 0 \times \log\ \text{WBC})]$$
$$\boxed{= \exp[2.355 - 0.342\ \log\ \text{WBC}]}$$

We now give a third example which illustrates how to compute a hazard ratio when the model does contain product terms. We consider the printout for Model 3 of the remission data shown here.

To obtain the hazard ratio for the effect of *Rx* adjusted for log WBC using Model 3, we consider $\mathbf{X}^*$ and $\mathbf{X}$ vectors which have three components, one for each variable in the model. The $\mathbf{X}^*$ vector, which denotes a placebo subject, has components $X_1^* = 1$, $X_2^* = \log \text{WBC}$ and $1 \times \log \text{WBC}$. The $\mathbf{X}$ vector, which denotes a treated subject, has components $X_1 = 0$, $X_2 = \log \text{WBC}$ and $X_3 = 0 \times \log \text{WBC}$. Note again that, as with the previous example, the value for log WBC is treated as fixed, though unspecified.

Using the general formula for the hazard ratio, we must now compute the exponential of the sum of three quantities, corresponding to the three variables in the model. Substituting the values from the printout and the values of the vectors $\mathbf{X}^*$ and $\mathbf{X}$ into this formula, we obtain the exponential expression shown here. Using algebra, this expression simplifies to the exponential of 2.355 minus 0.342 times log WBC.

EXAMPLE: (continued)

log WBC = 2:

$$\widehat{HR} = \exp[2.355 - 0.342(2)]$$
$$= e^{1.671} = 5.32$$

log WBC = 4:

$$\widehat{HR} = \exp[2.355 - 0.342(4)]$$
$$= e^{0.987} = 2.68$$

In order to get a numerical value for the hazard ratio, we must specify a value for log WBC. For instance, if log WBC = 2, the estimated hazard ratio becomes 5.32, whereas if log WBC = 4, the estimated hazard ratio becomes 2.68. Thus, we get different hazard ratio values for different values of log WBC, which should make sense since log WBC is an effect modifier in Model 3.

General rule for (0, 1) exposure variables when there are product terms:

$$\boxed{\widehat{HR} = \exp\left[\hat{\beta} + \sum \hat{\delta}_j W_j\right]}$$

where

$\hat{\beta}$ = coefficient of E

$\hat{\delta}_j$ = coefficient of $E \times W_j$

($\widehat{HR}$ does not contain coefficients of non-product terms)

The example we have just described using Model 3 illustrates a general rule which states that the hazard ratio for the effect of a (0,1) exposure variable in a model which contains product terms involving this exposure with other X's can be written as shown here. Note that β "hat" denotes the coefficient of the exposure variable and the δ "hats" are coefficients of product terms in the model of the form $E \times W_j$. Also note that this formula does not contain coefficients of nonproduct terms other than those involving E.

EXAMPLE

Model 3:

$\hat{\beta}$ = coefficient of Rx (E)

$\hat{\delta}_1$ = coefficient of $Rx \times$ log WBC (W_1 = log WBC)

$$\widehat{HR}\,(\text{Model 3}) = \exp\left[\hat{\beta} + \hat{\delta}_1 \log \text{WBC}\right]$$
$$= \exp[2.355 - 0.342 \log \text{WBC}]$$

For Model 3, β "hat" is the coefficient of the Rx variable, and there is only one δ "hat" in the sum, which is the coefficient of the product term $Rx \times$ log WBC. Thus, there is only one W, namely W_1 = log WBC. The hazard ratio formula for the effect of exposure is then given by exponentiating β "hat" plus δ "hat" times log WBC. Substituting the estimates from the printout into this formula yields the expression obtained previously, namely the exponential of 2.355 minus 0.342 times log WBC.

VI. Interval Estimation: Interaction

We have previously illustrated in Model 2 of the Remission Time Data how to obtain a 95% interval estimate of the HR when there is only one regression coefficient of interest, e.g., the HR is of the form $\exp[\beta_1]$.

Model 2:

$h(t, X) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$

$HR = \exp[\beta_1]$

Large sample 95% confidence interval:

$$\exp\left[\hat{\beta}_1 \pm 1.96\sqrt{\hat{Var}\hat{\beta}_1}\right]$$

where

$$s_{\hat{\beta}_1} = \sqrt{\hat{Var}\hat{\beta}_1}$$

The procedure typically used to obtain a large sample 95% confidence interval (CI) for the parameter is to compute *the exponential of the estimate of the parameter plus or minus a percentage point of the normal distribution times the estimated standard error of the estimate*. Note that the square root of the estimated variance is the standard error.

No interaction: simple formula

Interaction: complex formula

This computation is relatively simple when there are no interaction effects in the model. However, when there is interaction, the computational formula for the estimated standard error is more complex.

Model 3:

$$h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log WBC + \beta_3 (Rx \times \log WBC)]$$

$$HR = \exp[\beta_1 + \beta_3 \log WBC]$$

Suppose we focus on Model 3, shown here on the left, and, again, we assume that Rx is a (0, 1) exposure variable of interest. Then the formula for the HR for the effect of Rx controlling for the variable log WBC is given on the left underneath the model formula.

Interaction: variance calculation difficult

No interaction: variance directly from printout

The difficult part in computing the CI for a HR involving interaction effects is the calculation for the estimated variance. When there is no interaction, so that the parameter of interest is a single regression coefficient, this variance is obtained directly from the listing of estimated coefficients and corresponding standard errors.

$$\hat{HR} = exp[\hat{\ell}],$$
where $\ell = \beta_1 + \beta_3 \log WBC$

For Model 3, we can alternatively write the estimated HR formula as $\exp[\hat{\ell}]$, where ℓ is the linear function $\beta_1 + \beta_3 \log WBC$ and $\hat{\ell}$ is the estimate of this linear function using the ML estimates.

$$\boxed{\begin{array}{l}95\%\text{ CI for HR} = \exp[\hat{\ell}] \\ \exp[\hat{\ell} \pm 1.96\sqrt{\text{Var}\hat{\ell})}]\end{array}}$$

To obtain a 95% CI for $\exp[\ell]$ we must exponentiate the CI for ℓ. The formula is shown on the left.

General Formula:
can consider any ℓ, e.g.,

$$\ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots \delta_k W_k,$$

where $X_1 = (0, 1)$ exposure variable and β_1 = coeff of X_1,
δ_j = coeff of $X_1 \times W_j$, j=1,. . ., k

This CI formula, though motivated by our example using Model 3, is actually the general formula for the 95% CI for any HR of interest from a Cox PH model. In general, for a model with a (0, 1) exposure variable X_1 and interaction terms $X_1 \times W_1$, . . ., $X_1 \times W_k$, the linear function may take any form of interest, as shown in the left.

$Var(\hat{\ell}) = Var(\hat{\beta}_1 + \hat{\delta}_1 W_1 + \cdots \hat{\delta}_k W_k)$
where the estimates $\hat{\beta}_1, \hat{\delta}_i, \ldots, \hat{\delta}_k$ are correlated, so one must use
$Var(\hat{\beta}_1), Cov(\hat{\beta}_1, \hat{\delta}_i)$ *and* $Cov(\hat{\delta}_i, \hat{\delta}_j)$

When the HR involves interaction effects, the estimated variance considers a linear sum of estimated regression coefficients. Because the coefficients in the linear sum are estimated from the same data set, these coefficients are correlated with one another. Consequently, the calculation of the estimated variance must consider both the variances and the covariances of the estimated coefficients, which makes computations somewhat cumbersome.

Computer packages SAS and STATA compute $\hat{\text{Var}}\,\hat{\ell}$ as part of the program options (see Computer Appendix).

Nevertheless, most computer packages that have procedures for fitting survival analysis models like the Cox model provide for computing the estimated variance of linear functions like $\hat{\ell}$ as part of the program options. See the Computer Appendix for details on the use of the "contrast" option in SAS and the "lincom" option in STATA.

General formula for $\hat{\text{Var}}\,\hat{\ell}$:

$$\hat{\text{Var}}(\hat{\ell}) = \hat{\text{Var}}(\hat{\beta}_1) + \sum_j W_j^2 \hat{\text{Var}}(\hat{\delta}_1) + 2\sum_j W_j \hat{Cov}(\hat{\beta}_1, \hat{\delta}_j) + 2\sum_j \sum_k W_j W_k \hat{Cov}(\hat{\delta}_j, \hat{\delta}_k)$$

For the interested reader, we provide here the general formula for the estimated variance of the linear function $\hat{\ell}$.

- Variances and covariances provided in the computer output
- User specifies W's values of interest.

In applying this formula, the user obtains the estimated variances and covariances from the variancecovariance output. The user must specify values of interest for the effect modifiers defined by the Ws in the model.

Model 3:
$\ell = \beta_1 + \beta_3 \log \text{WBC}$

$$\text{Var}(\hat{\ell}) = \text{Var}(\hat{\beta}_1) + (\log WBC)^2 \text{Var}(\hat{\beta}_3) + 2(\log WBC)^2 Cov(\hat{\beta}_1, \hat{\beta}_3)$$

Applying this variance formula to Model 3, we obtain the variance expression shown on the left. Since log WBC is the only effect modifier here, the user would need to specify log WBC values of interest, e.g., log WBC =2 and log WBC=4.

95% CI for Rx in Model 3 (SAS edited output):

log WBC	$\widehat{HR}$	S.E.	Conf	Limits
2	5.3151	3.8410	1.2894	21.9101
4	2.6809	1.6520	0.8013	8.9700

Using the "contrast" option in SAS's PHREG procedure, we show on the left computed 95% CI's for the Rx variable for two choices of log WBC. When log WBC is 2, the estimated HR is 5.32 with a 95% CI given by the limits (1.29, 21.91), whereas when log WBC is 4, the estimated HR is 2.68 with a 95% CI given by the limits (0.80, 8.97).

CI results suggest log WBC by Rx interaction but conflict with nonsignificant interaction test result. Note: small study size (n=42)

These results suggest the interaction of log WBC with Rx, and they conflict with the previously reported nonsignificance of the test for interaction in Model 3, which might primarily be attributed to the small sample size (n=42) of this study.

VII. Adjusted Survival Curves Using the Cox PH model

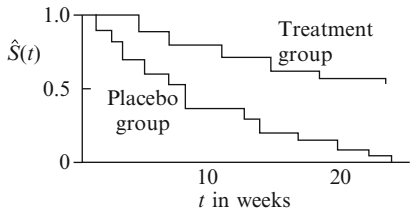
The two primary quantities desired from a survival analvsis point of viem are estimated hazard ratios and estimated survival curves. Having just described how to compute hazard ratios, we now turn to estimation of survival curves using the Cox model.

Two primary quantities:

1. estimated hazard ratios
2. estimated survival curves

No model: use KM cunzes

Recall that if no model is used to fit survival data, a survival curve can be estimated using a Kaplan–Meier method. Such KM curves are plotted as step functions as shown here for the remission data example.



Cox model: adjusted survival curves (also step fuuctions).

When a Cox model is used to fit survival data, survival curves can be obtained that adjust for the explanatory variables used as predictors. These are called **adjusted survival curves**, and, like KM curves, these are also plotted as step functions.

Cox model hazard function:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

Cox model survival function:

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{\sum_{i=1}^{p} \beta_i X_i}}$$

The hazard function formula for the Cox PH model, shown here again, can be converted to a corresponding survival function formula as shown below. This survival function formula is the basis for determining adjusted survival curves. Note that this formula says that the survival function at time t for a subject with vector $\mathbf{X}$ as predictors is given by a baseline survival function $S_0(t)$ raised to a power equal to the exponential of the sum of β_i times X_i.

Estimated survival function:

$$\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{e^{\sum_{i=1}^{p} \hat{\beta}_i X_i}}$$

The expression for the estimated survival function can then be written with the usual "hat" notation as shown here.

$\hat{S}_0(t)$ and $\hat{\beta}_i$ are provided by the computer program. The X_i must be specified by the investigator.

The estimates of $\hat{S}_0(t)$ and $\hat{\beta}_i$ are provided by the computer program that fits the Cox model. The X's, however, must first be specified by the investigator before the computer program can compute the estimated survival curve.

EXAMPLE: Model 2 Remission Data

$\hat{h}(t,\mathbf{X}) = \hat{h}_0(t)e^{1.294\,Rx\,+\,1.604\,\log\,\mathrm{WBC}}$
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp(1.294\,Rx\,+\,1.604\,\log\,\mathrm{WBC})}$

For example, if we consider model 2 for the remission data, the fitted model written in terms of both the hazard function and corresponding survival function is given here.

Specify values for $\mathbf{X} = (Rx, \log \mathrm{WBC})$

We can obtain a specific survival curve by specifying values for the vector $\mathbf{X}$, whose component variables are Rx and log WBC.

$Rx = 1$, log WBC = 2.93:

$$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_1\overline{Rx}+\hat{\beta}_2\overline{\log\ \mathrm{WBC}}\right)}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(1.294(0.5)+1.604(2.93))}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(5.35)} = \boxed{\left[\hat{S}_0(t)\right]^{210.6}}$$

For instance, if $Rx = 1$ and log WBC = 2.93, the estimated survival curve is obtained by substituting these values in the formula as shown here, and carrying out the algebra to obtain the expression circled. Note that the value 2.93 is the overall mean log WBC for the entire dataset of 42 subjects.

$Rx = 0$, log WBC = 2.93:

$$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp(1.294(0)+1.604(2.93))}$$
$$= \left[\hat{S}_0(t)\right]^{\exp(4.70)} = \boxed{\left[\hat{S}_0(t)\right]^{109.9}}$$

Also, if $Rx = 0$ and log WBC = 2.93, the estimated survival curve is obtained as shown here.

Each of the circled expressions gives **adjusted** survival curves, where the adjustment is for the values specified for the X's. Note that for each expression, a survival probability can be obtained for any value of t.

Adjusted Survival Curves

$Rx = 1$, log WBC = 2.93:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{400.9}$
$Rx = 0$, log WBC = 2.93:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{109.9}$

The two formulae just obtained, again shown here, allow us to compare survival curves for different treatment groups adjusted for the covariate log WBC. Both curves describe estimated survival probabilities over time assuming the same value of log WBC, in this case, the value 2.93.

Typically, use $X = \overline{X}$ or X_{median}

Computer uses $\overline{X}$

Typically, when computing adjusted survival curves, the value chosen for a covariate being adjusted is an average value like an arithmetic mean or a median. In fact, most computer programs for the Cox model automatically use the mean value over all subjects for each covariate being adjusted.

EXAMPLE: (continued)

Remission data ($n = 42$):
$\overline{\log\ \mathrm{WBC}} = 2.93$

In our example, the mean log WBC for all 42 subjects in the remission data set is 2.93. That is why we chose this value for log WBC in the formulae for the adjusted survival curve.

General formulae for adjusted survival curves comparing two groups:

Exposed subjects:

$$\hat{S}(t, \mathbf{X}_1) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1(1) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i\right]}$$

Unexposed subjects:

$$\hat{S}(t, \mathbf{X}_0) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1(0) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i\right]}$$

More generally, if we want to compare survival curves for two levels of an exposure variable, and we want to adjust for several covariates, we can write the formula for each curve as shown here. Note that we are assuming that the exposure variable is variable X_1, whose estimated coefficient is β_1 "hat," and the value of X_1 is 1 for exposed and 0 for unexposed subjects.

General formula for adjusted survival curve for all covariates in the model:

$$\hat{S}(t, \overline{\mathbf{X}}) = \left[\hat{S}_0(t)\right]^{\exp\left[\sum \hat{\beta}_i \bar{X}_i\right]}$$

Also, if we want to obtain an adjusted survival curve which adjusts for all covariates in the model, the general formula which uses the mean value for each covariate is given as shown here. This formula will give a single adjusted survival curve rather than different curves for each exposure group.

EXAMPLE

Single survival curve for Cox model containing *Rx* and log WBC:

$\overline{Rx} = 0.50$

$\overline{\log\ \text{WBC}} = 2.93$

$$\begin{aligned}\hat{S}(t, \mathbf{X}) &= \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_1 \overline{Rx} + \hat{\beta}_2 \overline{\log\ \text{WBC}}\right)} \\ &= \left[\hat{S}_0(t)\right]^{\exp(1.294(0.5) + 1.604(2.93))} \\ &= \left[\hat{S}_0(t)\right]^{\exp(5.35)} = \boxed{\left[\hat{S}_0(t)\right]^{210.6}}\end{aligned}$$

To illustrate this formula, suppose we again consider the remission data, and we wish to obtain a single survival curve that adjusts for both *Rx* and log WBC in the fitted Cox model containing these two variables. Using the mean value of each covariate, we find that the mean value for *Rx* is 0.5 and the mean value for log WBC is 2.93, as before.

To obtain the single survival curve that adjusts for *Rx* and log WBC, we then substitute the mean values in the formula for the adjusted survival curve for the model fitted. The formula and the resulting expression for the adjusted survival curve are shown here. (Note that for the remission data, where it is of interest to compare two exposure groups, the use of a single survival curve is not appropriate.)

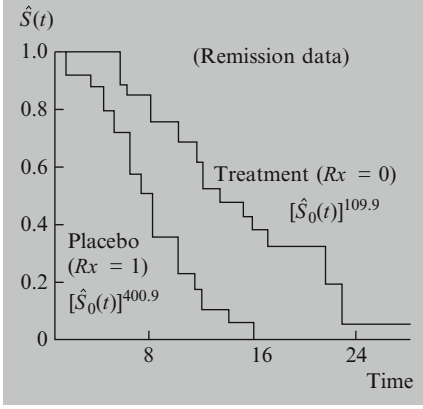
Compute survival probability by specifying value for t in $\hat{S}(t, \overline{\mathbf{X}}) = [\hat{S}_0(t)]^{210.6}$

From this expression for the survival curve, a survival probability can be computed for any value of t that is specified. When graphing this survival curve using a computer package, the values of t that are chosen are the failure times of all persons in the study who got the event. This process is automatically carried out by the computer without having the user specify each failure time.

Computer uses t's which are failure times.

EXAMPLE

Adjusted Survival Curves for Treatment and Placebo Groups



The graph of adjusted survival curves obtained from fitting a Cox model is usually plotted as a step function. For example, we show here the step functions for the two adjusted survival curves obtained by specifying either 1 or 0 for treatment status and letting log WBC be the mean value 2.93.

Next section: PH assumption

- explain meaning
- when PH **not** satisfied

We now turn to the concept of the proportional hazard (PH) assumption. In the next section, we explain the meaning of this assumption and we give an example of when this assumption is not satisfied.

Later presentations:

- how to evaluate PH
- analysis when PH not met

In later presentations, we expand on this subject, describing how to evaluate statistically whether the assumption is met and how to carry out the analysis when the assumption is not met.

VIII. The Meaning of the PH Assumption

PH: *HR* is constant over time, i.e., $\hat{h}(t,\mathbf{X}^*) = \text{constant} \times \hat{h}(t,\mathbf{X})$

$$\widehat{HR} = \frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \frac{\hat{h}_0(t)\exp\left[\sum \hat{\beta}_i X_i^*\right]}{\hat{h}_0(t)\exp\left[\sum \hat{\beta}_i X_i\right]} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

where $\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_p^*\right)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote the set of X's for two individuals.

The PH assumption requires that the *HR* is constant over time, or equivalently, that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

To understand the PH assumption, we need to reconsider the formula for the *HR* that compares two different specifications $\mathbf{X}^*$ and $\mathbf{X}$ for the explanatory variables used in the Cox model. We derived this formula previously in Section V, and we show this derivation again here. Notice that the baseline hazard function $\hat{h}_0(t)$ appears in both the numerator and denominator of the hazard ratio and cancels out of the formula.

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

does not involve t.

The final expression for the hazard ratio therefore involves the estimated coefficients β_i "hat" and the values of $\mathbf{X}^*$ and $\mathbf{X}$ for each variable. However, because the baseline hazard has canceled out, the final expression does not involve time t.

Let (Constant)

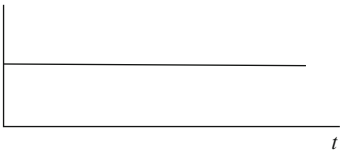
$$\hat{\theta} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

then

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}$$

Thus, once the model is fitted and the values for $\mathbf{X}^*$ and $\mathbf{X}$ are specified, the value of the exponential expression for the estimated hazard ratio is a constant, which does not depend on time. If we denote this constant by θ "hat," then we can write the hazard ratio as shown here. This is a mathematical expression which states the proportional hazards assumption.

$\widehat{HR}$ ($\mathbf{X}^*$ versus $\mathbf{X}$)



Graphically, this expression says that the estimated hazard ratio comparing any two individuals plots as a constant over time.

$$\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X})$$

Proportionality constant (not dependent on time)

Another way to write the proportional hazards assumption mathematically expresses the hazard function for individual $\mathbf{X}^*$ as θ "hat" times the hazard function for individual $\mathbf{X}$, as shown here. This expression says that the hazard function for one individual is proportional to the hazard function for another individual, where the proportionality constant is θ "hat," which does not depend on time.

EXAMPLE: Remission Data

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t)e^{1.294\,Rx + 1.604\ \log\ \mathrm{WBC}}$$

$$\widehat{HR} = \frac{\hat{h}(t, Rx = 1, \log \mathrm{WBC} = 2.93)}{\hat{h}(t, Rx = 0, \log \mathrm{WBC} = 2.93)} = \exp[1.294] = 3.65 \text{ Constant}$$

Placebo

$$\hat{h}(t, Rx = 1, \log \mathrm{WBC} = 2.93) = 3.65\, \hat{h}(t, Rx = 0, \log \mathrm{WBC} = 2.93)$$

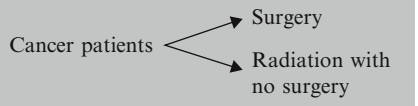
Treatment

3.65 = proportionality constant

To illustrate the proportional hazard assumption, we again consider the Cox model for the remission data involving the two variables Rx and log WBC. For this model, the estimated hazard ratio that compares placebo ($Rx = 1$) with treated ($Rx = 0$) subjects controlling for log WBC is given by e to the 1.294, which is 3.65, a constant.

Thus, the hazard for placebo group ($Rx = 1$) is 3.65 times the hazard for the treatment group ($Rx = 0$), and the value, 3.65, is the same regardless of time. In other words, using the above model, the hazard for the placebo group is proportional to the hazard for the treatment group, and the proportionality constant is 3.65.

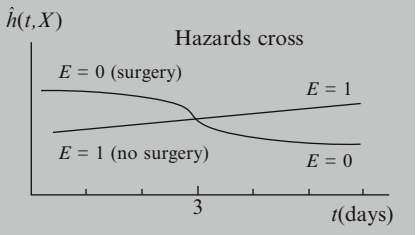
EXAMPLE: PH Not Satisfied



$$E = \begin{cases} 0 & \text{if surgery} \\ 1 & \text{if no surgery} \end{cases}$$

$$h(t,\mathbf{X}) = h_0(t)e^{\beta E}$$

Is the above Cox PH model appropriate?
Note:
Serious surgery $\Rightarrow$ High risk for death early



2 days: $\dfrac{\hat{h}(t=2,\ E=1)}{\hat{h}(t=2,\ E=0)} < 1$

but

5 days: $\dfrac{\hat{h}(t=5,\ E=1)}{\hat{h}(t=5,\ E=0)} > 1$

To further illustrate the concept of proportional hazards, we now provide an example of a situation for which the proportional hazards assumption is *not* satisfied.

For our example, we consider a study in which cancer patients are randomized to either surgery or radiation therapy without surgery. Thus, we have a (0,1) exposure variable denoting surgery status, with 0 if a patient receives surgery and 1 if not. Suppose further that this exposure variable is the only variable of interest, so that a Cox PH model for the analysis of this data, as shown here, will contain only the one variable E, denoting exposure.

Now the question we consider here is whether the above Cox model containing the variable E is an appropriate model to use for this situation. To answer this question we note that when a patient undergoes serious surgery, as when removing a cancerous tumor, there is usually a high risk for complications from surgery or perhaps even death early in the recovery process, and once the patient gets past this early critical period, the benefits of surgery, if any, can then be observed.

Thus, in a study that compares surgery to no surgery, we might expect to see hazard functions for each group that appear as shown here. Notice that these two functions cross at about 3 days, and that prior to 3 days, the hazard for the surgery group is higher than the hazard for the no surgery group, whereas after 3 days, the hazard for the surgery group is lower than the hazard for the no surgery group.

Looking at the above graph more closely, we can see that at 2 days, when $t = 2$, the hazard ratio of non-surgery ($E = 1$) to surgery ($E = 0$) patients yields a value less than 1. In contrast, at $t = 5$ days, the hazard ratio of nonsurgery to surgery yields a value greater than 1.

EXAMPLE: (continued)

Given the above description, ***HR* is not constant over time**.

Cox PH model inappropriate because PH model assumes constant *HR*:

$h(t,\mathbf{X}) = h_0(t)e^{\beta E}$

$\widehat{HR} = \dfrac{\hat{h}(t,\ E=1)}{\hat{h}(t,\ E=0)} = e^{\hat{\beta}}$

Thus, if the above description of the hazard functions for each group is accurate, the hazard ratios are not constant over time. That is, the hazard ratio is some number less than 1 before 3 days and greater than 1 after 3 days.

It is therefore inappropriate to use a Cox PH model for this situation, because the PH model assumes a constant hazard ratio across time, whereas our situation yields a hazard ratio that varies with time.

In fact, if we use a Cox PH model, shown here again, the estimated hazard ratio comparing exposed to unexposed patients at any time is given by the constant value e to the β "hat," which does not vary over time.

General rule:
If the hazards cross, then a Cox PH model is not appropriate.

This example illustrates the general rule that if the hazards cross, then the PH assumption cannot be met, so that a Cox PH model is inappropriate.

Analysis when Cox PH model not appropriate? See Chapters 5 and 6.

It is natural to ask at this point, if the Cox PH model is inappropriate, how should we carry out the analysis? The answer to this question is discussed in Chapters 5 and 6. However, we will give a brier reply with regard to the surgery study example just described.

EXAMPLE: (continued)

Surgery study analysis options:

- stratify by exposure (use KM curves)
- start analysis at 3 days; use Cox PH model
- fit PH model for < 3 days and for > 3 days; get $\widehat{HR}$ (< 3 days) and $\widehat{HR}$ (> 3 days)
- include time-dependent variable (e.g., E $\times$ t); use extended Cox model

Actually for the surgery study there are several oplions available for the analysis. These include:

- analyze by stratifying on the exposure variable; that is, do not fit any model, and, instead oblain Kaplan-Meier curves for each exposure group separately;
- start the analysis at three days, and use a Cox PH model on three-day survivors;
- fit Cox model for less than three days and a different Cox model lor greater than three days to get two difterent hazard ratio estimates, one for each of these two lime periods;
- fit a modified Cox model that includes a time-dependent variable which measures the interaction of exposure with time. This model is called an **extended Cox model**.

Different options may lead to different conclusions.

Further discussion of these options is given in subsequent chapters. We point out here, that different options may lead to different conclusions, so that the investigator may have to weigh the relative merits of each option in light of the data actually obtained before deciding on any particular option as best.

Hazards cross $\Rightarrow$ PH not met

but

? $\Rightarrow$ PH met

See Chapter 4: Evaluating PH Assumption

One final comment before concluding this section: although we have shown that when the hazards cross, the PH assumption is not met, we have not shown how to decide when the PH assumption is met. This is the subject of Chapter 4 entitled, "Evaluating the PH Assumption."

IX. The Cox Likelihood

Likelihood

- Typically based on outcome distribution
- Outcome distribution not specified for Cox model
- Cox likelihood based on order of events rather than their distribution
  - Called partial likelihood

Typically, the formulation of a likelihood function is based on the distribution of the outcome. However, one of the key features of the Cox model is that there is not an assumed distribution for the outcome variable (i.e., the time to event). Therefore, in contrast to a parametric model, a full likelihood based on the outcome distribution cannot be formulated for the Cox PH model. Instead, the construction of the **Cox likelihood is based on the observed order of events** rather than the joint distribution of events. Thus the Cox likelihood is called a "partial" likelihood.

Illustration

Scenario:

- Gary, Larry, Barry have lottery tickets
- Winning tickets chosen at times t_1, t_2, ...
- Each person ultimately chosen
- Can be chosen only once

To illustrate the idea underlying the formulation of the Cox model, consider the following scenario. Suppose Gary, Larry, and Barry are each given a lottery ticket. Winning tickets are chosen at times t_j ($j = 1,2, \ldots$). Assume each person is ultimately chosen and once a person is chosen he cannot be chosen again (i.e., he is out of the risk set). What is the probability that the order each person is chosen is first Barry, then Gary, and finally Larry?

Question:
What is the probability that the order chosen is as follows?

1. Barry
2. Gary
3. Larry

Answer:

$$\text{Probability} = \frac{1}{3} \times \frac{1}{2} \times \frac{1}{1} = \frac{1}{6}$$

Barry Gary Larry

The probability that Barry's ticket is chosen before Gary's and Larry's is one out of three. Once Barry's ticket is chosen, it cannot be chosen again. The probability that Gary's ticket is then chosen before Larry's is one out of two. Once Barry's and Gary's tickets are chosen, they cannot be chosen again which means that Larry's ticket must be chosen last. This yields a probability of 1/6 for this given order of events (see left).

Scenario:

Barry – 4 tickets
Gary – 1 ticket
Larry – 2 tickets

Now consider a modification of the previous scenario. Suppose Barry has 4 tickets, Gary has 1 ticket, and Larry has 2 tickets; now what is the probability that the order each person is chosen is first Barry, then Gary, and finally Larry?

Question:
What is the probability that the order chosen is as follows?

1. Barry
2. Gary
3. Larry

Barry, Gary, and Larry have 7 tickets in all and Barry owns 4 of them so Barry's probability of being chosen first is 4 out of 7. After Barry is chosen, Gary has 1 of the 3 remaining tickets and after Barry and Gary are chosen, Larry owns the remaining 2 tickets. This yields a probability of 4/21 for this order (see left).

Answer:

$$\text{Probability} = \frac{4}{7} \times \frac{1}{3} \times \frac{2}{2} = \frac{4}{21}$$

For this scenario

Subject's number of tickets affects probability

For Cox model

Subject's pattern of covariates affects likelihood of ordered events

For this scenario, the probability of a particular order is affected by the number of tickets held by each subject. For a Cox model, the likelihood of the observed order of events is affected by the pattern of covariates of each subject.

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

SURVT = Survival time (in years)
STATUS = 1 for event, 0 for censorship
SMOKE = 1 for a smoker, 0 for a nonsmoker

To illustrate this connection, consider the dataset shown on the left. The data indicate that Barry got the event at TIME = 2 years. Gary got the event at 3 years, Harry was censored at 5 years, and Larry got the event at 8 years. Furthermore, Barry and Larry were smokers whereas Gary and Harry were nonsmokers.

Cox PH model

$h(t) = h_0(t)e^{\beta_1 SMOKE}$

ID	Hazard
Barry	$h_0(t)e^{\beta_1}$
Gary	$h_0(t)e^0$
Harry	$h_0(t)e^0$
Larry	$h_0(t)e^{\beta_1}$

Individual hazards (Cox likelihood) analogous to number of tickets (lottery scenario) For example, smokers analogous to persons with extra lottery tickets

Consider the Cox proportional hazards model with one predictor, SMOKE. Under this model the hazards for Barry, Gary, Harry, and Larry can be expressed as shown on the left. The individual hazards are determined by whether the subject was a smoker or nonsmoker.

The individual level hazards play an analogous role toward the construction of the Cox likelihood as the number of tickets held by each subject plays for the calculation of the probabilities in the lottery scenario discussed earlier in this section. The subjects who smoke are analogous to persons given extra lottery tickets, thereby affecting the probability of a particular order of events.

Cox Likelihood

$$L = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

On the left is the Cox likelihood for these data. Notice the likelihood is a product of three terms, which correspond to the three event times. Barry got the event first at TIME = 2 years. At that time, all four subjects were at risk for the event. The first product (L_1) has the sum of the four subjects' hazards in the denominator and Barry's hazard in the numerator. Gary got the event next at 3 years when Gary, Harry, and Larry were still in the risk set. Consequently, the second product (L_2) has the sum of the three hazards for the subjects still at risk in the denominator and Gary's hazard in the numerator. Harry was censored at 5 years, which occurred between the second and third event. Therefore, when Larry got the final event at 8 years, nobody else was at risk for the event. As a result, the third product (L_3) just has Larry's hazard in the denominator and the numerator.

Likelihood is product of 3 terms

$$L = L_1 \times L_2 \times L_3$$

$$L_1 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right]$$

$$L_2 = \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right]$$

$$L_3 = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

t_1, time = 2, four at risk (L_1)
t_2, time = 3, three at risk (L_2)
t_3, time = 8, one at risk (L_3)

For each term:

Numerator – single hazard
Denominator – sum of hazards

Baseline hazard, $h_0(t)$ cancels

$$L = \left[\frac{e^{\beta_1}}{e^{\beta_1} + e^0 + e^0 + e^{\beta_1}}\right] \times \left[\frac{e^0}{e^0 + e^0 + e^{\beta_1}}\right] \times \left[\frac{e^{\beta_1}}{e^{\beta_1}}\right]$$

Thus, L does not depend on $h_0(t)$

Data A

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

Data B

ID	TIME	STATUS	SMOKE
Barry	1	1	1
Gary	7	1	0
Harry	8	0	0
Larry	63	1	1

Comparing datasets

- TIME variable differs
- Order of events the same
- Cox PH likelihood the same

To summarize, the likelihood in our example consists of a product of three terms (L_1, L_2, and L_3) corresponding to the ordered failure times (t_1, t_2, and t_3). The denominator for the term corresponding to time t_j (j = 1, 2, 3) is the sum of the hazards for those subjects still at risk at time t_j, and the numerator is the hazard for the subject who got the event at t_j.

A key property of the Cox likelihood is that the baseline hazard cancels out in each term. Thus, the form of the baseline hazard need not be specified in a Cox model, as it plays no role in the estimation of the regression parameters. By factoring $h_0(t)$ in the denominator and then canceling it out of each term, the likelihood for Barry, Gary, and Larry can be rewritten as shown on the left.

As we mentioned earlier, the Cox likelihood is determined by the order of events and censorships and not by the distribution of the outcome variable. To illustrate this point, compare datasets A and B on the left, and consider the likelihood for a Cox PH model with smoking status as the only predictor. Although the values for the variable TIME differ in the two datasets, the Cox likelihood will be the same using either dataset because the order of the outcome (TIME) remains unchanged.

General Approach

- k failure times
- Likelihood a product of K terms
- Construction of each term similar to Barry, Gary, and Larry

$$L = L_1 \times L_2 \times L_3 \times \ldots \times L_k$$
$$= \prod_{f=1}^{k} L_f$$

We have used a small dataset (four observations with three failure times) for ease of illustration. However, the approach can be generalized. Consider a dataset with k failure times and let L_f denote the contribution to the likelihood corresponding to the f-th failure time. Then the Cox likelihood can be formulated as a product of each of the k terms as shown on the left. Each of the terms L_f is constructed in a similar manner as with the data for Gary, Larry, and Barry.

Obtaining maximum likelihood estimates

Solve system of equations

$$\frac{\partial \ln L}{\partial \beta_i} = 0, \quad i = 1, 2, 3, \ldots, p$$

p = # of parameters

Once the likelihood is formulated, the question becomes: *which values of the regression parameters would maximize L*? The process of maximizing the likelihood is typically carried out by setting the partial derivative of the natural log of L to zero and then solving the system of equations (called the score equations).

X. Using Age as the Time Scale

Outcome variable:
time until an event occurs

where "time" is measured as

time-on-study (years, months, etc., of follow-up from study entry)

or

age at follow-up

Recall that when we introduced the topic of survival analysis in Chapter 1, we wrote that the "time" variable used as the outcome variable could be measured as **time-on-study** (i.e., follow-up time since study entry) in years, months, weeks, or days from the beginning of follow-up. We also wrote that, alternatively, we might use **age as the time scale**, so that time is measured as **age at follow-up** until either an event or censorship occurs. In this section, we focus on the use of age as the time scale, and describe when such use is appropriate, provide the form of the Cox PH model in this situation, and illustrate its use.

Time 0:
starting time of the true survival time

Possible choices for time 0:

- Study entry
- Beginning of treatment
- Disease onset
- Disease diagnosis
- Surgery
- Point in calendar time
- Birth
- Conception

A key decision in any survival analysis is where to define the starting point for determining individual's "true" survival time, which we call **time 0**. Depending on the study, choices for time 0 might be: the time the subject enters the study, the time the subject begins treatment, the time of disease onset, the time of diagnosis, a point in calendar time, the time of a seminal event (e.g., surgery), birth, or conception. If we define time 0 at birth, then an individual's survival time is represented by their age.

Time 0 not necessarily equal to t_0, where

t_0 = time when subject's survival time is first observed

e.g., if survival time is measured by age and subject enters study at age 45

$\Downarrow$

t_0 = age 45 but time 0 < age 45 since time 0 = age-at-birth

Time 0 is not necessarily the time point where a subject's survival time is first observed (which we call time **t_0**). For example, if survival time is measured by age at follow-up and a subject enters the study at age 45, then t_0=45 years for this subject. In this example, the subject's survival time has been **left-truncated** at $t_0 = 45$, which we now define.

Left truncation:

- subject not observed before t_0
- if subject has event before t_0, then not included in the study
- if subject has event after t_0, then included in the study and assumed not at risk for event until t_0

Left truncation at time t_0 is defined as follows:

The subject is not observed from time 0 to t_0. If the subject has the event before time t_0, then that subject is not included in the study. If the subject has the event after time t_0, the subject is included in the study but with the caveat that the subject was not at risk to be an *observed* event until time t_0.

Two types of left truncation:

Type 1: subject has event before t_0 and not included in the study,

e.g.,

E causes death before study entry

$\Downarrow$

Bias: effect of E underestimated

We note that there are two types of left truncation at t_0. The first type of left truncation occurs if the subject has the event before t_0 and thus is not included in the study. If, for example, the exposure (E) under study causes individuals to die before they could enter the study, this could lead to a (selective) **survival bias** that would underestimate the effect of exposure.

Type 2: $t_0 > 0$ and $t > t_0$

where t = observed survival time

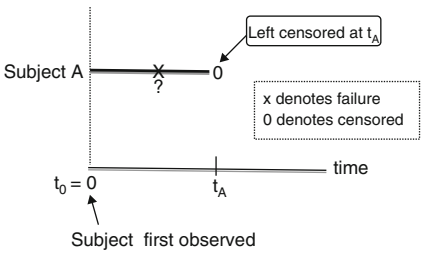
The second type of left truncation occurs if the subject survives beyond time t_0 (i.e., $t > t_0$). This is required in order for the subject to have his/her survival time observed.

Study entry $\Rightarrow$ subject survives until t_0

Type 1: subject not included in the study

Type 2: subject included in the study

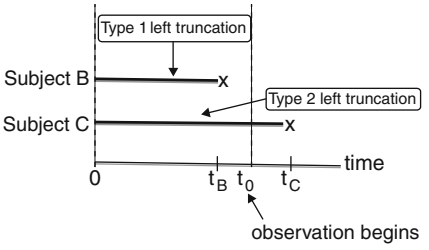
Thus, a condition of the subject's entry into the study is that they survive until time t_0. If they do not meet that condition, then their left truncation is of the first type and thus not included in the study. If they do survive past time t_0, then their left truncation is of the second type.



Left truncation (of both types) at time t is commonly confused with left censorship at time t. If a subject is **left censored at time t**, then that subject is (i) included in the study, (ii) known to be event free at time 0, (iii) known to be at risk for the event after time 0, and (iv) known to have had the event before time t but with the exact time of event being unknown.

Example: left censored data
Subject A: $t_0 = 0$
$t_A = 6$
true $t = ? < 6$

For example, for subject A in the above graph, suppose death is the outcome and a patient is first treated for some illness at the time of their first visit to the clinic (t_0=0). Further, suppose that the patient does not show up at the next scheduled clinic visit 6 months later because that patient had died in the interim. If the specific month of death is unable to be ascertained, then that patient is included in the study and **left censored** at $t_A = 6$ months.



In contrast, the diagram on the left illustrates the two types of left truncation.

In this diagram, subject B provides an example of left truncation of the first type that would occur if an individual died between disease onset (time 0) and disease diagnosis (time t_0) and thus was never included in the study. In this example, having the disease was a necessary condition for study inclusion, whereas subject B died before it was known that he/she had the disease.

Example: Type 1 left truncation
Subject B: Time $0 < t_0$
not included in the study
Example: Type 2 left truncation
Subject C: Time $0 < t_0$ but first observed at t_0

Subject C illustrates Type 2 left truncation, since he/she developed the disease (time 0) prior to being diagnosed with the disease at time t_0 and was observed after t_0.

Being observed at time t means:
If event at $t \Rightarrow$ recorded event at t

One clarifying point is that when we say a subject is observed at time t, we do not necessarily mean that the subject is observed in an active prospective manner. Rather, what we mean by a subject being observed at time t is as follows: *if that subject had an event at time t, then the subject would be recorded in the study as having had an event at time t.*

2 approaches for measuring survival time:

Time-on-study
vs.
Age-at-follow-up

We now compare two approaches of measuring survival time. One approach is to measure survival time as **time-on-study** and the other is to measure survival time as **age-at-follow-up** until either an event or censorship. The choice of approach determines the risk set at the time of each event. We illustrate this idea with hypothetical data.

- Choice determines the risk set.

Hypothetical Survival Data

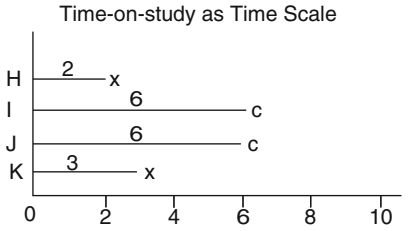
Subject	t	d	a_0	a
H	2	1	65	67
I	6	0	65	71
J	6	0	74	80
K	3	1	75	78

Consider the data shown on the left on four different subjects, for each of which we have identified time-on-follow-up (**t**), whether failed or censored (**d**), age at study entry ($\mathbf{a}_0$), and age at the end of follow-up (**a**). Note that t is simply $a - \mathbf{a}_0$, the difference between age at follow-up time and age at study entry.

Time-on-study Layout

f	$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
1	2	4	0	0	H,I,J,K
2	3	3	2	2	I,J,K

Using time-on-study (i.e., from entry into the study) as the time scale, the data layout based on ordered follow-up times is shown on the left, below which is shown a graphical representation that follows each subject from the time of study entry. There are only two failures and these occur at follow-up times 2 (subject H) and 3 (subject K).



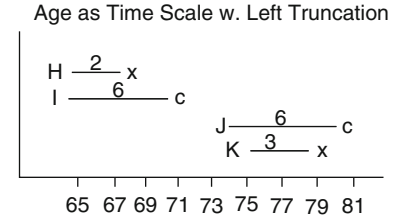
The risk set at the first failure time ($t_{(1)} = 2$) consists of all four subjects, and the risk set at the second failure time ($t_{(2)} = 3$) contains subjects I, J, and K. Subjects I and J are censored after time $t_{(2)} = 3$, i.e., the value for q_2 is 2 at this time. Notice that the risk set at time $t_{(2)} = 3$ is contained in the risk set at time $t_{(1)} = 2$, which is generally true when the outcome variable being considered is time-on-follow-up. The data layout represented here, in which the size of the risk set always decreases over time, is called a **closed cohort**.

$R(t_{(1)} = 2) = \{H, I, J, K\}$
$R(t_{(2)} = 3) = \{I, J, K\}$
I and J censored after $t_{(2)} \Rightarrow q_2 = 2$
{I, J, K} contained in {H, I, J, K}

Age as Time Scale Layout

f	$a_{(f)}$	n_f	m_f	q_f	$R(a_f)$
1	67	2	1	1	H,I
2	78	2	1	1	J,K

Now let's consider the data layout that would result if we used age as the time scale, which is shown on the left. Below this layout is a graphical representation that follows each subject from age at study entry.



Using age as the time scale, the first failure time is at age $a_{(1)} = 67$ (for subject H), and there are two subjects (H and I) in the risk set at this time; subject I is in the risk set at $a_{(1)} = 67$ because (s)he entered the study at age 65 and was still at risk when subject H failed. However, because subjects J and K did not enter the study until ages 74 and 75, these subjects are not in the risk set at $a_{(1)} = 67$.

First failure: $R(a_{(1)} = 67) = \{H, I\}$

I still at risk at $a_{(1)} = 67$ but
J and K not in study at $a_{(1)} = 67$

Second failure: $R(a_{(2)} = 78) = \{J, K\}$

H and J no longer at risk at $a_{(2)} = 78$
I censored between $a_{(1)} = 67$ and $a_{(2)} = 78 \Rightarrow q_1 = 1$
J censored after $a_{(2)} = 78 \Rightarrow q_2 = 1$

{J, K} not contained in {H, I}

The second failure time is at age $a_{(2)} = 78$ (for subject K). The only two subjects in the risk set at $a_{(2)} = 78$ are subjects J and K, since subject H failed at age 67 and subject I was censored at age 71. The values in the q column are 1 at failure age 67 (for subject I) and 1 at failure age 72 (for subject J). In contrast to the previous data layout, the risk set at the later failure age (containing J, K) is not a subset of the risk set at the first failure age (containing H, I), but, rather, is a mutually exclusive subset. This data layout, in which the size of the risk set may increase or decrease over time, is called an **open cohort**.

Time-on-Study vs. Age as Time Scale

- Closed cohort vs. Open cohort
- How we decide which to use?

We thus see that using time-on-study as the time scale can give a different view of the survival data (i.e., a closed cohort) than found when using age as the time scale (i.e., an open cohort). So which time scale should be used and how do we make such a decision in general?

Key issue:
Did all subjects first become at risk at their study entry?

To answer this question, a key issue is to determine whether all subjects in the study first begin to be at risk for the outcome at the time they enter the study.

Clinical trial:

- Subjects start to be followed for the outcome after random allocation

Suppose the study is a **clinical trial** to compare, say, treatment and placebo groups, and subjects start to be followed shortly after random allocation into one of these two groups.

- Reasonable to assume subjects start to be at risk upon study entry

⇓

Time-on-Study typically used as the outcome

(Covariates may also be controlled)

Then, it may be reasonable to assume that study subjects begin to be at risk for the outcome upon entry into the study. In such a situation, using time-on-study as the time scale is typically appropriate. Further, covariates of interest may be controlled for by stratification and/or being entered into a regression model (e.g., Cox PH model) as predictors in addition to the treatment status variable.

Observational study:

- Subjects already at risk prior to study entry
- Unknown time or age when first at risk

Suppose, instead of the above scenario, the **study is observational** (i.e., not a clinical trial) and subjects are already at risk for the outcome prior to their study entry. Also, suppose the time or age at which subjects first became at risk is unknown.

- Example: Subjects with high blood pressure enter study, but unknown date or age when first diagnosed (prior to study entry).

For example, the subjects may all have high blood pressure when the study begins and are then followed until a coronary event occurs (or censorship); such subjects already had high blood pressure when recruited for the study, but the date or their age when their high blood pressure condition was first diagnosed is assumed unknown.

- Reasonable to assume that

$$T = t_r + t$$

where
T = true survival time
t_r = time at risk prior to study entry
t = observed time-on-study

In this situation, it seems reasonable that the time at risk prior to study entry ($\mathbf{t_r}$), which is unknown, contributes to the true survival time ($\mathbf{T}$) for the individual, although only the observed time-on-study ($\mathbf{t}$) is actually available to analyze. The individual's true (i.e., total) survival time is therefore underestimated by the time-on-study information (obtained from study entry), i.e., the true survival time is **left-truncated**.

Left-truncated survival data
⇓
Time-on-study questionable

So, for the situation where we have left-truncated survival data, the use of time-on-study follow-up times that ignores unknown delayed entry time may be questioned.

Subject	**t**	**d**	$\mathbf{a_0}$	**a**
H	2	1	65	67
I	6	0	65	71
J	6	0	74	80
K	3	1	75	78

Recall that although both subjects I and J were censored at follow-up time 6 (say, in weeks) from study entry, subject I entered the study at age 65, whereas subject J entered the study at age 74. Because subject J is 9 years older than subject I upon study entry, and recognizing that age is a well-known risk factor for most diseases, e.g., coronary disease, we would expect subject J to have higher potential for failing (i.e., higher hazard rate) than subject I at study entry. However, if we just use time-on-study follow-up times in our analysis, we are not taking into account the increased failure potential for subject J over subject I at study entry.

Subject J is 9 years older than Subject I
⇓
h(t| subject J) > h(t| subject I).
But, using time-on-study approach does not account for this difference.

One modified approach:
Use time-on-study, but control for a_0, e.g.,

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp[\sum \beta_i X_i + \gamma a_0]$$

⇓
OK provided model correctly specified but not always appropriate.

One way to account for the age difference at entry would simply to control for age at entry (i.e., $\mathbf{a_0}$) as a covariate in one's survival analysis by adding the variable $\mathbf{a_0}$ to a Cox PH model. This approach is reasonable provided the model is specified correctly (e.g., proportional hazards assumption is met for age).

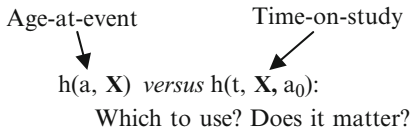
Alternatively, may consider using age as the time scale.

Alternatively, considering subjects I and J, who have entered at the same time but are 9 years different in age, we might consider using age as the time scale to represent a subject's potential for failure, which we now describe.

$h(a, \mathbf{X}) = h_0(a) \exp[\sum \beta_i X_i]$

X denotes set of covariates,
e.g., **X** = (Rx, BMI, SMK)
$h_0(a)$ = baseline hazard

In a Cox PH model that uses age as the time scale (as shown on the left), the outcome variable will be age-at-event (a) rather than time-on-study (t). **X** denotes the set of covariates in the model, e.g., **X** = (Rx, BMI, SMK). The baseline hazard function $h_0(a)$ is an unspecified function of **a** (rather than **t**).



At this point, we might again ask, when, if at all, would using a model based on h(a, **X**) be preferable to simply using a model of the form h(t, **X**, a_0) where t denotes time-on follow-up, and a_0 denotes age at entry?

It depends!
And, it might not matter!
(often same results, if model well-specified)

The answer is that "it depends". Moreover, in many situations, it might not matter, since use of either model form will often lead to essentially the same results, provided the model is well-specified in each case.

Prefer h(a, **X**) provided

- age is stronger determinant of outcome than time-on-study
- $h_0(a)$ is unspecified, so that age is not modeled as a covariate i.e., avoids mispecifying the model as linear in a_0 when a_0^2 also needed

On one hand, using h(a, **X**) may be preferable if age is a much stronger determinant of the outcome than time-on-study, i.e., age at event may have a larger effect on the hazard than time-on-study (Korn et al. 1997). Also, because age is taken into account in an unspecified baseline hazard $h_0(a)$, a more effective control of age may result that avoids the possibility of mispecifying the way the age at entry ($\mathbf{a_0}$) might be entered into a time-on-study model, e.g., using only a linear term when a quadratic term such as is also required for model adequacy.

Prefer h(t, **X**, a_0) provided

- time-on-study is stronger determinant of outcome than age
- age at entry ($\mathbf{a_0}$) is effectively controlled in the model using a linear and/ or possibly higher/ order term (or age is controlled by stratification)

On the other hand, h(t, **X**, a_0) may be preferable if time-on-study is a stronger determinant of the outcome than age at the event, as in a randomized clinical trial. Also, a time-on-study model would seem appropriate if age at entry ($\mathbf{a_0}$) is "effectively controlled" (e.g., using a quadratic term if necessary) or is stratified in the model.

Alternative Cox PH models for age-truncated survival data:

Let

t = follow-up time,

a = attained age at event or censorship

$\mathbf{a}_0$ = age at enrollment into study (Note: $t = a - a_0$)

$\mathbf{X} = (X_1, X_2, \ldots, X_k)$, vector of predictors, not including a_0

β_i = regression coeff. corresponding to X_i.

γ = regression coeff. if a_0 included in model

We now specify several alternative forms that a Cox PH model might take to account for risk-truncated survival data. As previously introduced, our notation uses **t** to denote time-on-study follow-up time, **a** to denote attained age at the event or censorship, $\mathbf{a_0}$ to denote age at study entry, **X** to denote the vector of predictor variables, not including age, β_i to denote the vector of Cox model coefficients corresponding to **X**, and γ_1 to denote the coefficient of a_0 if model includes a_0.

time on study

Model 0:
$h(t,\mathbf{X}) = h_0(t)\exp[\Sigma\beta_i X_i]$,
unadjusted for a_0

Model 1:
$h(t,\mathbf{X},a_0) = h_0(t)\exp[\Sigma\beta_i X_i + \gamma_1 a_0]$,
adjusted for a_0 as linear covariate

Model 2:
$h(t,\mathbf{X},a_0) = h_0(t)\exp[\Sigma\beta_i X_i + \gamma_1 a_0 + \gamma_2 a_0^2]$
adjusted for a_0 with quadratic covariate

Model 3:
$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\Sigma\beta_i X_i]$,
stratified by a_0 or birth cohort, g= 1, …, s

On the left, we provide seven different Cox PH models that might be considered to analyze risk-truncated survival data.

Models 0–3 use an analysis based on time-on-study follow-up, whereas Models 4–6 consider age as the time scale.

age as time scale

Model 4:
$h(a,\mathbf{X}) = h_0(a)\exp[\Sigma\beta_i X_i]$,
unadjusted for left truncation at a_0

Model 5:
$h(a,\mathbf{X}) = h_0(a|a_0)\exp[\Sigma\beta_i X_i]$,
adjusted for left truncation at a_0

Model 6:
$h_g(a,\mathbf{X}) = h_{0g}(a|a_0)\exp[\Sigma\beta_i X_i]$,
adjusted for left truncation at a_0 and stratified by birth cohort, g= 1, …, s

Of all these models, Model 0 is the least appropriate since this model uses time-on-study as the outcome and does not adjust for age at entry ($\mathbf{a_0}$) in any way.

Models 1–3 control for a_0 differently
- Model 1: linear effect of $\mathbf{a_0}$
- Model 2: quadratic effect of $\mathbf{a_0}$
- Model 3: stratifies on $\mathbf{a_0}$ or on birth cohorts defined from $\mathbf{a_0}$ (uses Stratified Cox PH model)

Models 1–3 are time-on-study models that control for age at entry ($\mathbf{a_0}$), but do so differently. Model 1 controls for $\mathbf{a_0}$ as a covariate and assumes a **linear effect of a_0**. Model 2, in contrast, assumes that $\mathbf{a_0}$ has **both linear and quadratic effects**. Model 3 **stratifies on either $\mathbf{a_0}$ or on birth cohort** defined from $\mathbf{a_0}$. Model 3 is called a Stratified Cox (SC) PH model, which we describe in detail in Chapter 5.

Models 1–3 reasonable

- if all study subjects begin risk at study entry
- if models provide effective control of $\mathbf{a_0}$

Models 1–3 are all reasonable if we assume that study subjects begin to be at risk upon study entry, as in a randomized clinical trial. Moreover, even for an observational study design in which subjects have different ages at entry, these models may appear justifiable if they provide effective control of $\mathbf{a_0}$.

Model 3
$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\sum \beta_i X_i]$

- alternative method of control
- may account for advances in medical management if stratified on birth cohort
- stratifying on either $\mathbf{a_0}$ or on birth cohort likely to give similar results unless enrollment over long time period

Model 3 controls for entry age by stratifying either on age at entry ($\mathbf{a_0}$) or on birth cohort based on $\mathbf{a_0}$. Model 3 provides an alternative way to control for age without explicitly putting $\mathbf{a_0}$ as a covariate in the model (as was done in Models 1 and 2). If we stratify by birth cohort instead of by $\mathbf{a_0}$, we can account for possible advances in medical management in later birth cohorts. Nevertheless, stratifying by age at entry or stratifying by birth cohort would likely give similar results unless enrollment happens over a long period of time. In the latter case, we recommend stratifying on birth cohort.

Models 4–6:

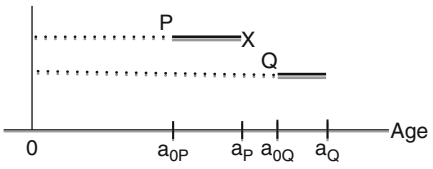
- outcome is age-at-event
- differ in baseline hazard

Models 4–6 use age-at-event or censorship rather than time-on-study as the outcome variable. These models differ in the way the baseline hazard function is specified.

Model 4: $h(a, \mathbf{X}) = h_0(a) \exp[\sum \beta_i X_i]$

- does not adjust for left-truncation at a_0
- assumes risk starts at birth
- data layout describes **closed cohort**

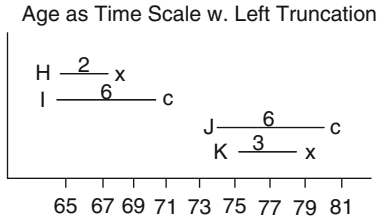
Model 4 uses $h_0(a)$ to indicate that, although age is the outcome, the model does not adjust for left truncation at the entry age (a_0). In effect, this baseline hazard assumes that each subject's *observed risk period* started at birth.



R(a) = {P, Q} using Model 4 even though Q enrolled after P failed

In other words, Model 4 allows keeping in the risk set R($\mathbf{a_P}$) any subject (e.g., subject Q in the figure at left) who was not under study at age $\mathbf{a_P}$ but who enrolled later (at age $\mathbf{a_{0Q}}$). Here, subject Q is in the risk set R($\mathbf{a_P}$) because we assume he is at risk from birth (Age=0) when subject P fails at $\mathbf{a_P}$. The data layout with ordered failure ages is thus a **closed cohort** that starts with all subjects in the risk set at birth.

Previous example:

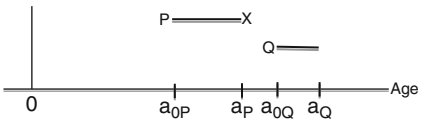


R(a = 67) = {H, I, J, K} using Model 4 since all four subjects at risk from birth x until H fails at age a = 67

If Model 4 were applied to our previous example involving four subjects, subjects J and K would be incorrectly included in the risk set R (**a**=67) when subject H failed, even though both these subjects were enrolled after age 67. This model ***inappropriately*** assumes that all subjects were at risk from birth; it does not adjust for age-truncation.

Model 5: $h(a, \mathbf{X}) = h_0(a|a_0)\exp[\sum \beta_i X_i]$

- adjusts for left-truncation at a_0
- data layout describes **open cohort**



R(a) = {P} and R(a*) = {Q} using Model 5 because Q enrolled after P failed

Model 5, on the other hand, accounts for left truncation by age at entry. The baseline hazard $h_0(a \mid a_0)$ is used to indicate that the data layout with ordered failure ages is an **open cohort**. For this model, the risk set R(**a**) at time **a** contains only those subjects who are under study at age **a**.

Previous example with H, I, J, K: R(a=67) = {H, I} and R(a=78) = {J, K} using Model 5 since J and K had not enrolled when H failed at 67 and {H, I} were not used in study when K failed at 78.

If Model 5 was applied to our previous example, subjects J and K, who had not enrolled when subject H failed at 67, would not be in the risk set R(**a**=67). Also, subjects H and I, who were no longer in the study when subject K failed at age 78, would not be in the risk set R(**a**=78).

	Model 3	Model 6
Stratifies on birth cohort?	Yes	Yes
Adjusts for age-truncation?	No	Yes

Model 6 is similar to Model 3 stratified on birth cohort. However, Model 6 adjusts for age truncation, whereas Model 3 does not. As with Model 3, Model 6 is intended to account for possible advances in medical management in later birth cohorts. Model 6 would not be necessary if we are considering a study in which everyone is enrolled within a short period of time.

Summary about Models 0–6:
Models 0 and 4:

- Both inappropriate
- Model 0 does not adjust for age
- Model 4 incorrectly assumes that all subjects are at risk from birth

Models 1–3, 5, 6

- All adjust for age-at-entry (a_0)
- Question: Do they differ in practice?

Pencina et al. (2007):

- Compare estimated regression coefficients for Models 1–6
- Consider Model 5 (age-truncated age scale) most appropriate conceptually
- Consider Models 1 and 2 (covariate adjusted for a_0) "attempts to approximate Model 5"
- Used numerical simulations and practical examples from Framingham

Conclusions:

- correct adjustment for the age at entry is crucial
- Model 1 inferior (and biased)
- Little practical or meaningful difference between Models 2 through 6

Cox PH Regression Coefficients (± se) for two CHD risk factors among men- Framingham Heart Disease Study (Pencina et al, 2007)

Time-on-study	Model			Age-time-scale		
	1	2	3	**4**	5	6
	linear	quad	strat	unadj	age-trunc	strat
Diabetic *versus* non-diabetic (n=2439)						
	0.48* ± 0.21	0.49* ± 0.21	0.48* ± 0.21	**0.23*** ± 0.20	0.47* ± 0.21	0.45* ± 0.21
Education: post– HS *versus* HS or less (n = 2177)						
	–0.43* ± 0.15	–0.40* ± 0.15	–0.43* ± 0.15	**0.18** ± 0.15	–0.43* ± 0.16	–0.38* ± 0.15

*The coefficient is significantly different from zero at the 0.05 level.

In summary, of the seven models we have presented, Models 0 and 4 are inappropriate because Model 0 does not account for age at all and Model 4 ignores age truncation by incorrectly assuming that all study subjects were observed for the outcome from birth.

The other five models (i.e., 1–3, 5, 6) all adjust for age at study entry in some way. A logical question at this point is whether in practice, it makes a difference which model is used to analyze age-truncated survival data?

The above question was actually addressed by Pencina et al. (Statist. Med., 2007) by comparing Models 1–6 above in terms of the estimated regression coefficients they produce. These authors consider Model 5, the age-truncated age scale model, to be "possibly the most appropriate refinement" to account for age-truncation. They also view time-on-study Models 1 and 2, which use linear and/or quadratic terms to adjust for entry age as a covariate as "attempts to approximate" Model 5.

Nevertheless, by considering numerical simulations as well as four practical examples from the Framingham Heart Study, Pencina et al. conclude that **correct adjustment for the age at entry is crucial** in reducing bias of the estimated coefficients. The unadjusted age-scale model (Model 1) is inferior to any of the five other models considered, regardless of their choice of time scale. Moreover, if correct adjustment for age at entry is made when considering Models 2–6, their analyses suggest that **there exists little if any practical or meaningful difference in the estimated regression coefficients depending on the choice of time scale**.

To illustrate, we show on the left results from Pencina et al. corresponding to Models 1–6 applied to 12-year follow-up Framingham Heart Study data. The outcome considered here is coronary heart disease (**CHD**) in men.

These results focus on two risk factors measured at baseline: diabetes mellitus status and education status, the latter categorized into two groups defined by post-high-school education (yes/no). The estimated regression coefficients (separately) relating these two risk factors to **CHD** outcome are presented in the table.

Summary of Framingham results from Pencina et al.:

- Model 4 inferior to other models
- Results for Models 1–3, 5, and 6 are similar
- Directions of estimated coefficients were as anticipated conceptually, e.g., + diabetes and smoking − education
- Quadratic terms (Model 2) were significant, suggesting that Model 1 is mispecified but
- Did not materially influence magnitude or significance of exposure variables (e.g., diabetes, smoking, education)

As expected, the table shows a substantial difference in the coefficient of the risk group variable estimated by the unadjusted age-scale model (Model 4) and the five other models. Moreover, the results for Models 1–3, 5, and 6 are all quite similar.

Pencina also point out that the directions of coefficients for these five models are in the directions anticipated conceptually, e.g., diabetes coefficients are positive, whereas education coefficients are negative.

The quadratic baseline age term (Model 2) was significant for both CHD risk factors. This suggests potential misspecification in the modeling of the relationship between CHD and age introduced by Model 1, which treats entry age as linear. However, its inclusion in the time-on-study model did not materially influence the magnitude or significance of the estimated exposure variable (diabetes or education status) coefficient.

Data Layout for Age-as-Time Scale: **CP format for age-truncated survival data (Model 5)**

Subj#	**d**	$\mathbf{a_0}$	**a**	$\mathbf{X_1}$	...	$\mathbf{X_p}$
1	d_1	a_{01}	a_1	X_{11}	...	X_{1p}
2	d_2	a_{02}	a_2	X_{21}	...	X_{2p}
3	d_3	a_{03}	a_3	X_{31}	...	X_{3p}
⋮	⋮	⋮	⋮	⋮	...	⋮
n	d_n	a_{0n}	a_n	X_{n1}	...	X_{np}

When using age as the time scale and accounting for age truncation (i.e., using Model 5 above), the data layout requires the counting process (**CP**) in start–stop format previously introduced in Section VI of Chapter 1 with a_0 as the start variable and a as the stop variable. However, since we are not considering recurrent events data here, the CP format for age-truncated survival data has a simpler form, involving only one line of data for each study subject, as shown on the left. The computer code needed to program the analysis is described in the Computer Appendix for STATA, SAS, SPSS, or R packages.

Model 4 layout: Set $a_0 = 0$ for all subjects or use "standard" layout (w/0 a_0 column)

Subj#	**d**	**a**	$\mathbf{X_1}$	...	$\mathbf{X_p}$
1	d_1	a_1	X_{11}	...	X_{1p}
2	d_2	a_2	X_{21}	...	X_{2p}
3	d_3	a_3	X_{31}	...	X_{3p}
⋮	⋮	⋮	⋮	...	⋮
n	d_n	a_n	X_{n1}	...	X_{np}

Note that the CP format corresponding to Model 4, which assumes the starting time is birth, would modify the Model 4 layout by letting $a_0 = 0$ in the a_0 column for all subjects. Nevertheless, this layout would be equivalent to the "standard" layout that omits **a_0 column** and simply treats the **a column** data as time-on-study information. Again, since Model 4 appears to be inferior to the other models, we caution the reader not to use this format unless the risk period was observed since birth.

XI. Summary

In this section we briefly summarize the content covered in this presentation.

1. Review: $S(t)$, $h(t)$, data layout, etc.
2. Computer example of Cox model:
   - estimate *HR*
   - test hypothesis about *HR*
   - obtain confidence intervals
3. Cox model formula:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p}\beta_i X_i}$$

4. Why popular: Cox PH model is "robust"

5. ML estimation: maximize a partial likelihood $L = L(\beta)$ = joint probability of observed data

6. Hazard ratio formula:

$$\widehat{HR} = \exp\left[\sum_{i=1}^{p}\hat{\beta}_i(X_i^* - X_i)\right]$$

7. Interval estimation-interaction: HR=exp[ℓ],

$$\text{where } \ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_k W_K$$
$$\beta_1 = \text{coeff. of } X_1, \text{ and}$$
$$\delta_j = \text{coeff. of } X \times W_j,\ j = 1, \ldots, k$$

95% CI for HR=exp[ℓ]:

$$\exp[\hat{\ell} \pm 1.96\sqrt{\widehat{\text{Var}}\hat{\ell})}]$$

Most computer packages, e.g., SAS, STATA, compute $\widehat{\text{Var}}\hat{\ell}$ as part of the program options (see Computer Appendix).

- We began with a computer example that uses the Cox PH model. We showed how to use the output to estimate the *HR*, and how to test hypotheses and obtain confidence intervals about the hazard ratio.
- We then provided the formula for the hazard function for the Cox PH model and described basic features of this model. The most important feature is that the model contains two components, namely, a baseline hazard function of time and an exponential function involving *X*'s but not time.
- We discussed reasons why the Cox model is popular, the primary reason being that the model is "robust" for many different survival analysis situations.
- We then discussed ML estimation of the parameters in the Cox model, and pointed out that the ML procedure maximizes a "partial" likelihood that focuses on probabilities at failure times only.
- Next, we gave a general formula for estimating a hazard ratio that compared two specifications of the *X*'s, defined as $\mathbf{X}^*$ and $\mathbf{X}$. We illustrated the use of this formula when comparing two exposure groups adjusted for other variables.
- We then described how to obtain a 95% CI for the HR when the hazard model contains interaction terms of the form $X_1 \times W_j$, where X_1 is a (0,1) exposure variable and W_j is an effect modifier of exposure. The formula is shown at the left. In this formula, the $\widehat{\text{Var}}\,\hat{\ell}$ is difficult to calculate without the use of a computer program.

  Fortunately, most computer packages have procedures for calculating this formula as part of the program options, e.g., SAS's "contrast" option and STATA's "lincom" option.

8. Adjusted survival curves: 0 or 1
   Comparing E groups:
   $$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[\hat{\beta}_1 E + \sum_{i\neq 1}\hat{\beta}_i \bar{X}_i\right]}$$
   Single curve:
   $$\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[\sum \hat{\beta}_i \bar{X}_i\right]}$$

9. PH assumption:
   $$\frac{\hat{h}(t,\mathbf{X}^*)}{\hat{h}(t,\mathbf{X})} = \hat{\theta} \text{ (a constant over } t)$$
   i.e., $\hat{h}(t,\mathbf{X}^*) = \hat{\theta}\hat{h}(t,\mathbf{X})$
   Hazards cross $\Rightarrow$ PH not met

10. Derivation of Cox PH Likelihood

11. Using "age-as-the-time scale" instead of "time-on-follow-up"
    Reason: account for left truncation of age
    Cox PH model that adjusts for age truncation:
    $$h(a,X) = h_0(a|a_0)\exp[\sum \beta_i X_i]$$
    where $\mathbf{a}$ = age at event or censorship
    $\mathbf{a_0}$ = age at study entry
    Data Layout: CP (start–stop) format

- We then defined an adjusted survival curve and presented formulas for adjusted curves comparing two groups adjusted for other variables in the model and a formula for a single adjusted curve that adjusts for all X's in the model. Computer packages for these formulae use the mean value of each X being adjusted in the computation of the adjusted curve.
- We described the PH assumption as meaning that the hazard ratio is constant over time, or equivalently that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time. We also showed that for study situations in which the hazards cross, the PH assumption is not met.
- We then showed how the Cox likelihood is derived using ordered failure times.
- Finally, we considered the use of "age as the time scale" instead of "time-on-follow-up" as the outcome variable, described why such use is appropriate to account for left truncation of age, provided the form of the Cox PH model in this situation, illustrate its use, and described the data layout required using a "stop-start" Counting Process (CP) format.

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Survival Curves and the Log–Rank Test
3. The Cox Proportional Hazards Model and Its Characteristics
4. Evaluating the Proportional Hazards Assumption
5. The Stratified Cox Procedure
6. Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

This presentation is now complete. We recommend that the reader review the detailed outline that follows and then do the practice exercises and test.

The next Chapter (4) describes how to evaluate the PH assumption. Chapters 5 and 6 describe methods for carrying out the analysis when the PH assumption is not met.

Detailed Outline

I. A computer example using the Cox PH model (pages 100–108)

A. Printout shown for three models involving leukemia remission data.

B. Three explanatory variables of interest: treatment status, log WBC, and product term; outcome is time until subject goes out of remission.

C. Discussion of how to evaluate which model is best.

D. Similarity to classical regression and logistic regression.

II. The formula for the Cox PH model (pages 108–110)

A. $\boxed{h(t, \mathbf{X}) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]}$

B. $h_0(t)$ is called the **baseline hazard function**.

C. $\mathbf{X}$ denotes a collection of p explanatory variables $X_1, X_2, \ldots, X_p$.

D. The model is semiparametric because $h_0(t)$ is unspecified.

E. Examples of the Cox model using the leukemia remission data.

F. Survival curves can be derived from the Cox PH model.

III. Why the Cox PH model is popular (pages 110–112)

A. Can get an estimate of effect (the hazard ratio) without needing to know $h_0(t)$.

B. Can estimate $h_0(t)$, $h(t, \mathbf{X})$, and survivor functions, even though $h_0(t)$ is not specified.

C. The e part of the formula is used to ensure that the fitted hazard is nonnegative.

D. The Cox model is "robust": it usually fits the data well no matter which parametric model is appropriate.

IV. ML estimation of the Cox PH model (pages 112–114)

A. Likelihood function is maximized.

B. L is called a partial likelihood, because it uses survival time information only on failures, and does not use censored information explicitly.

C. L makes use of the risk set at each time that a subject fails.

D. Inferences are made using standard large sample ML techniques, e.g., Wald or likelihood ratio tests and large sample confidence intervals based on asymptotic normality assumptions.

V. Computing the hazard ratio (pages 114–117)

A. Formula for hazard ratio comparing two individuals, and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

$$\boxed{\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right]}$$

B. Examples are given using a (0, 1) exposure variable, potential confounders, and potential effect modifiers.

C. Typical coding identifies $\mathbf{X}^*$ as the group with the larger hazard and $\mathbf{X}$ as the group with the smaller hazard, e.g., $X_1^* = 1$ for unexposed group and $X_1 = 0$ for exposed group.

VI. Interval estimation: interaction (pages 117–119)

A. Example- Model 3 from Remission Time Data

i. $h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 logWBC + \beta_3(Rx \times logWBC)]$

ii. $HR = \exp[\beta_1 + \beta_3 logWBC]$

B. General form of HR:
$HR = \exp[\ell]$, where
$\ell = \beta_1 + \delta_1 W_1 + \delta_2 W_2 + \ldots + \delta_k W_k$,
$X_1 = (0, 1)$ exposure variable, β_1 = coeff of X_1, and δ_j = coeff of $X_1 \times W_j$, j=1,..., k

C. General form of 95% CI for $HR = \exp[\ell]$:

$$\exp\left[\hat{\ell} \pm 1.96\sqrt{\hat{Var}\,\hat{\ell})}\right], \text{ where}$$

$$\text{Var}\left(\hat{\ell}\right) = \text{Var}\left(\hat{\beta}_1 + \hat{\delta}_1 W_1 + \ldots + \hat{\delta}_k W_k\right)$$

D. Computation of var complicated.

i. Computer programs, e.g., in SAS, STATA, can do this for the user.

ii. Otherwise, user must carry out complicated calculation using formula for var:

$$\hat{Var}\left(\hat{\ell}\right) = \hat{Var}\left(\hat{\beta}_1\right) + \sum_j W_j^2 \hat{Var}\left(\hat{\delta}_j\right) + 2\sum_j W_j \hat{Cov}\left(\hat{\beta}_1, \hat{\delta}_j\right) + 2\sum_j \sum_k W_j W_k \hat{Cov}\left(\hat{\delta}_j, \hat{\delta}_k\right)$$

iii. Variances and covariances provided in computer output

iv. User specifies W's values of interest.

v. Model 3 formula for $\hat{Var}(\hat{\ell})$:

$$\hat{Var}\left(\hat{\ell}\right) = \hat{Var}\left(\hat{\beta}_1\right) + (\log WBC)^2 \hat{Var}\left(\hat{\beta}_3\right) + 2(\log WBC)^2 Cov\left(\hat{\beta}_1, \hat{\beta}_3\right)$$

E. Example of 95% CI: Model 3

VII. Adjusted survival curves using the Cox PH model (pages 120–123)

A. Survival curve formula can be obtained from hazard ratio formula:

$$S(t, \mathbf{X}) = [S_0(t)]^{\exp[\sum \beta_i X_i]}$$

where $S_0(t)$ is the baseline survival function that corresponds to the baseline hazard function $h_0(t)$.

B. To graph $S(t, \mathbf{X})$, must specify values for $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

C. To obtain "adjusted" survival curves, usually use overall mean values for the X's being adjusted.

D. Examples of "adjusted" $S(t, \mathbf{X})$ using leukemia remission data.

VIII. The meaning of the PH assumption (pages 123–127)

A. Hazard ratio formula shows that hazard ratio is independent of time:

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \theta$$

B. Baseline hazard function not involved in the HR formula.

C. Hazard ratio for two $\mathbf{X}$'s are proportional:

$$h(t, \mathbf{X}^*) = \theta\, h(t, \mathbf{X})$$

D. An example when the PH assumption is not satisfied: hazards cross

IX. The Cox likelihood (pages 127–131)

A. Lottery Example

B. Likelihood based on order of events

X. Using age as the time scale (pages 131–142)

A. Definition of Left Truncation

i. Type I: subject has event before t_0 and not included in study

ii. Type II: $t_0 > 0$ and $t > t_0$ where t_0 time when first observed t = observed survival time

B. Left Truncation versus Left Censoring

C. Time-on-study versus Age-as-time scale: Closed cohort versus Open cohort

D. When to use Age-as-time-scale

i. It depends

a. Type of study

b. Well-defined model involving a_0 where a_0 denotes age at entry

E. Alternative Models
   i. Time-on-study
      a. outcome is t = time since first observed
      b. consider control for a_0, e.g.,

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp\left[\sum \beta_i X_i + \gamma_1 a_0\right] \text{ or}$$

$$h(t, \mathbf{X}, a_0) = h_0(t) \exp\left[\sum \beta_i X_i + \gamma_1 a_0 + \gamma_2 a_0^2\right]$$

   ii. Age-as-time-scale
      a. outcome is a = age at event or censorship
      b. adjusting for age truncation, e.g.,

$$h(a, \mathbf{X}) = h_0(a \mid a_0) \exp\left[\sum \beta_i X_i\right] \text{ or}$$

$$h_g(a, \mathbf{X}) = h_{0g}(a \mid a_0) \exp\left[\sum \beta_i X_i\right]$$

F. Example from Pencina et al (2007)
   i. age-as-time-scale model
      a. need to adjust for age truncation subjects
      b. incorrect results if subjects assumed to be observed from birth
   ii. time-on-study model: when a_0 is controlled, results similar to age-as-time-scale model.
   iii. overall recommendation: correct adjustment for the age at entry is crucial

XI. Summary (pages 143–144)

Practice Exercises

1. In a 10-year follow-up study conducted in Evans County, Georgia, involving persons 60 years or older, one research question concerned evaluating the relationship of social support to mortality status. A Cox proportional hazards model was fit to describe the relationship of a measure of social network to time until death. The social network index was denoted as SNI, and took on integer values between 0 (poor social network) to 5 (excellent social network). Variables to be considered for control in the analysis as either potential confounders or potential effect modifiers were AGE (treated continuously), RACE (0,1), and SEX (0,1).

   a. State an initial PH model that can be used to assess the relationship of interest, which considers the potential confounding and interaction effects of the AGE, RACE, and SEX (assume no higher than two-factor products involving SNI with AGE, RACE, and SEX).

   b. For your model in part 1a, give an expression for the hazard ratio that compares a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates being controlled.

   c. Describe how you would test for interaction using your model in part 1a. In particular, state the null hypothesis, the general form of your test statistic, with its distribution and degrees of freedom under the null hypothesis.

   d. Assuming a revised model containing no interaction terms, give an expression for a 95% interval estimate for the adjusted hazard ratio comparing a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates in your model.

   e. For the no-interaction model described in part 1d, give an expression (i.e., formula) for the estimated survival curve for a person with SNI = 4, adjusted for AGE, RACE, and SEX, where the adjustment uses the overall mean value for each of the three covariates.

   f. Using the no-interaction model described in part 1d, if the estimated survival curves for persons with SNI = 4 and SNI = 2 adjusted for (mean) AGE, RACE, and SEX are plotted over time, will these two estimated survival curves cross? Explain briefly.

   g. For the (interaction) model described in Part 1a, what is the formula for the 95% CI for the HR that compares a person with SNI = 4 to a person with SNI = 2 and the same values of the covariates being controlled?

2. For this question, we consider the survival data for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their book (*The Statistical Analysis of Survival Time Data, Wiley*, 1980). The variables in this dataset are listed as follows:

	Variable#	Variable name	Coding
	1	Treatment	Standard = 1, test = 2
Four indicator variables for cell type	2	Cell type 1	Large = 1, other = 0
	3	Cell type 2	Adeno = 1, other = 0
	4	Cell type 3	Small = 1, other = 0
	5	Cell type 4	Squamous = 1, other = 0
	6	Survival time	(Days) integer counts
	7	Performance status	0 = worst, . . . , 100 = best
	8	Disease duration	(Months) integer counts
	9	Age	(Years) integer counts
	10	Prior therapy	None = 0, some = 10
	11	Status	0 = censored, 1 = died

For these data, a Cox PH model was fitted yielding the following edited computer results:
Response: survival time

Variable name	Coef.	Std. Err.	$p > \|z\|$	Haz. Ratio	[95% Conf.	interval]
1 Treatment	0.290	0.207	0.162	1.336	0.890	2.006
3 Adeno cell	0.789	0.303	0.009	2.200	1.216	3.982
4 Small cell	0.457	0.266	0.086	1.579	0.937	2.661
5 Squamous cell	−0.400	0.283	0.157	0.671	0.385	1.167
7 Perf. status	−0.033	0.006	0.000	0.968	0.958	0.978
8 Disease dur.	0.000	0.009	0.992	1.000	0.982	1.018
9 Age	−0.009	0.009	0.358	0.991	0.974	1.010
10 Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054

Log likelihood = −475.180

a. State the Cox PH model used to obtain the above computer results.

b. Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with large cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.

c. Using the printout above, what is the hazard ratio that compares persons with adeno cell type with persons with squamous cell type? Explain your answer using the general hazard ratio formula for the Cox PH model.

d. Based on the computer results, is there an effect of treatment on survival time? Explain briefly.

e. Give an expression for the estimated survival curve for a person who was given the test treatment and who had a squamous cell type, where the variables to be adjusted are performance status, disease duration, age, and prior therapy.

f. Suppose a revised Cox model is used which contains, in addition to the variables already included, the product terms: treatment × performance status; treatment × disease duration; treatment × age; and treatment × prior therapy. For this revised model, give an expression for the hazard ratio for the effect of treatment, adjusted for the other variables in the model.

3. The data for this question contain survival times of 65 multiple myeloma patients (references Krall et al., "A Step-up Procedure for Selecting Variables Associated with Survival Data," *Biometrics*, vol. 31, pp. 49–57, 1975). A partial list of the variables in the dataset is given below:

   Variable 1: observation number
   Variable 2: survival time (in months) from time of diagnosis
   Variable 3: survival status (0 = alive, 1 = dead)
   Variable 4: platelets at diagnosis (0 = abnormal, 1 = normal)
   Variable 5: age at diagnosis (years)
   Variable 6: sex (1 = male, 2 = female)

Below, we provide edited computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results.

Model 1:

Variable	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Platelets	0.470	2.854	.869	1.600	0.006	429.689
Age	0.000	0.037	.998	1.000	0.930	1.075
Sex	0.183	0.725	.801	1.200	0.290	4.969
Platelets × age	−0.008	0.041	.850	0.992	0.915	1.075
Platelets × sex	−0.503	0.804	.532	0.605	0.125	2.924
	Log likelihood = −153.040					

Model 2:

Platelets	−0.725	0.401	.071	0.484	0.221	1.063
Age	−0.005	0.016	.740	0.995	0.965	1.026
Sex	−0.221	0.311	.478	0.802	0.436	1.476
	Log likelihood = −153.253					

Model 3:

Platelets	−0.706	0.401	.078	0.493	0.225	1.083
Age	−0.003	0.015	.828	0.997	0.967	1.027
	Log likelihood = −153.509					

Model 4:

Platelets	−0.705	0.397	.076	0.494	0.227	1.075
Sex	−0.204	0.307	.506	0.815	0.447	1.489
	Log likelihood = −153.308					

Model 5:

Platelets	−0.694	0.397	.080	0.500	0.230	1.088
	Log likelihood = −153.533					

a. For model 1, give an expression for the hazard ratio for the effect of the platelet variable adjusted for age and sex.

b. Using your answer to part 3a, compute the estimated hazard ratio for a 40-year-old male. Also compute the estimated hazard ratio for a 50-year-old female.

c. Carry out an appropriate test of hypothesis to evaluate whether there is any significant interaction in model 1. What is your conclusion?

d. Considering models 2–5, evaluate whether age and sex need to be controlled as confounders?

e. Which of the five models do you think is the best model and why?

f. Based on your answer to part 3e, summarize the results that describe the effect of the platelet variable on survival adjusted for age and sex.

g. Why might you consider using age-as-the-time-scale instead of time-on-follow-up as the outcome to analyze these data?

Test

1. Consider a hypothetical 2-year study to investigate the effect of a passive smoking intervention program on the incidence of upper respiratory infection (URI) in newborn infants. The study design involves the random allocation of one of three intervention packages (A, B, C) to all healthy newborn infants in Orange County, North Carolina, during 1985. These infants are followed for 2 years to determine whether or not URI develops. The variables of interest for using a survival analysis on these data are:

   T = time (in weeks) until URI is detected or time until censored
   s = censorship status (= 1 if URI is detected, = 0 if censored)
   PS = passive smoking index of family during the week of birth of the infant
   DC = daycare status (= 1 if outside daycare, = 0 if only daycare is in home)
   BF = breastfeeding status (= 1 if infant is breastfed, = 0 if infant is not breastfed)
   T_1 = first dummy variable for intervention status (= 1 if A, = 0 if B, = −1 if C)
   T_2 = second dummy variable for intervention status (= 1 if B, = 0 if A, = −1 if C).

   a. State the Cox PH model that would describe the relationship between intervention package and survival time, controlling for *PS*, *DC*, and *BF* as confounders and effect modifiers. In defining your model, use only two factor product terms involving exposure (i.e., intervention) variables multiplied by control variables in your model.

   b. Assuming that the Cox PH model is appropriate, give a formula for the hazard ratio that compares a person in intervention group A with a person in intervention group C, adjusting for *PS*, *DC*, and *BF*, and assuming interaction effects.

   c. Assuming that the PH model in part 1a is appropriate, describe how you would carry out a chunk test for interaction; i.e., state the null hypothesis, describe the test statistic and give the distribution of the test statistic and its degrees of freedom under the null hypothesis.

d. Assuming no interaction effects, how would you test whether packages A, B, and C are equally effective, after controlling for *PS*, *DC*, and *BF* in a Cox PH model without interaction terms (i.e., state the two models being compared, the null hypothesis, the test statistic, and the distribution of the test statistic under the null hypothesis).

e. For the no-interaction model considered in parts 1c and 1d, give an expression for the estimated survival curves for the effect of intervention A adjusted for *PS*, *DC*, and *BF*. Also, give similar (but different) expressions for the adjusted survival curves for interventions B and C.

2. The data for this question consists of a sample of 50 persons from the 1967–1980 Evans County Study. There are two basic independent variables of interest: AGE and chronic disease status (CHR), where CHR is coded as 0 = none, 1 = chronic disease. A product term of the form AGE × CHR is also considered. The dependent variable is time until death, and the event is death. The primary question of interest concerns whether CHR, considered as the exposure variable, is related to survival time, controlling for AGE. The edited output of computer results for this question is given as follows:

Model 1:

Variable	Coef.	Std. Err.	Chi-sq	$p > \|z\|$
CHR	0.8595	0.3116	7.61	.0058
	Log likelihood = −142.87			
Model 2:				
CHR	0.8051	0.3252	6.13	.0133
AGE	0.0856	0.0193	19.63	.0000
	Log likelihood = −132.45			
Model 3:				
CHR	1.0009	2.2556	0.20	.6572
AGE	0.0874	0.0276	10.01	.0016
CHR × AGE	−0.0030	0.0345	0.01	.9301
	Log likelihood = −132.35			

a. State the Cox PH model that allows for main effects of CHR and AGE as well as the interaction effect of CHR with AGE.

b. Carry out the test for significant interaction; i.e., state the null hypothesis, the test statistic, and its distribution under the null hypothesis. What are your conclusions about interaction?

c. Assuming no interaction, should AGE be controlled? Explain your answer on the basis of confounding and/or precision considerations.

d. If, when considering plots of various hazard functions over time, the hazard function for persons with CHR = 1 crosses the hazard function for persons with CHR = 0, what does this indicate about the use of any of the three models provided in the printout?

e. Using model 2, give an expression for the estimated survival curve for persons with CHR = 1, adjusted for AGE. Also, give an expression for the estimated survival curve for persons with CHR = 0, adjusted for AGE.

f. What is your overall conclusion about the effect of CHR on survival time based on the computer results provided from this study?

3. The data for this question contain remission times of 42 multiple leukemia patients in a clinical trial of a new treatment. The variables in the dataset are given below:

Variable 1: survival time (in weeks)
Variable 2: status (1 = in remission, 0 = relapse)
Variable 3: sex (1 = female, 0 = male)
Variable 4: log WBC
Variable 5: *Rx* status (1 = placebo, 0 = treatment)

Below, we provide computer results for several different Cox models that were fit to this dataset. A number of questions will be asked about these results starting below.

Model 1:

Variable	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Rx	0.894	1.815	.622	2.446	0.070	85.812
Sex	−1.012	0.752	.178	0.363	0.083	1.585
log WBC	1.693	0.441	.000	5.437	2.292	12.897
Rx × Sex	1.952	0.907	.031	7.046	1.191	41.702
Rx × log WBC	−0.151	0.531	.776	0.860	0.304	2.433
	Log likelihood = −69.515					

Model 2:

Rx	0.405	0.561	.470	1.500	0.499	4.507
Sex	−1.070	0.725	.140	0.343	0.083	1.422
log WBC	1.610	0.332	.000	5.004	2.610	9.592
Rx × Sex	2.013	0.883	.023	7.483	1.325	42.261
	Log likelihood = −69.555					

Model 3:

Rx	0.587	0.542	.279	1.798	0.621	5.202
Sex	−1.073	0.701	.126	0.342	0.087	1.353
Rx × Sex	1.906	0.815	.019	6.726	1.362	33.213
	Log likelihood = −83.475					

Model 4:

Rx	1.391	0.457	.002	4.018	1.642	9.834
Sex	0.263	0.449	.558	1.301	0.539	3.139
log WBC	1.594	0.330	.000	4.922	2.578	9.397
	Log likelihood = −72.109					

a. Use the above computer results to carry out a chunk test to evaluate whether the two interaction terms in model 1 are significant. What are your conclusions?

b. Evaluate whether you would prefer model 1 or model 2. Explain your answer.

c. Using model 2, give an expression for the hazard ratio for the effect of the *Rx* variable adjusted for SEX and log WBC.

d. Using your answer in part 3c, compute the hazard ratio for the effect of *Rx* for males and for females separately.

e. By considering the potential confounding of log WBC, determine which of models 2 and 3 you prefer. Explain.

f. Of the models provided which model do you consider to be best? Explain.

Answers to Practice Exercises

1. a. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1\,\text{SNI} + \beta_2\,\text{AGE} + \beta_3\,\text{RACE} + \beta_4\,\text{SEX} + \beta_5\,\text{SNI} \times \text{AGE} + \beta_6\,\text{SNI} \times \text{RACE} + \beta_7\,\text{SNI} \times \text{SEX}]$

   b. $HR = \exp[2\beta_1 + 2(\text{AGE})\beta_5 + 2(\text{RACE})\beta_6 + 2(\text{SEX})\beta_7]$

   c. H_0: $\beta_5 = \beta_6 = \beta_7 = 0$. Likelihood ratio test statistic: $-2\ln L_R - (-2\ln L_F)$, which is approximately X^2_3 under H_0, where R denotes the reduced model (containing no product terms) under H_0, and F denotes the full model (given in part la above).

   d. 95% CI for adjusted HR:

   $$\exp\left[2\hat{\beta}_1 \pm 1.96 \times 2\sqrt{\hat{\text{Var}}\left(\hat{\beta}_1\right)}\right]$$

   e. $\hat{S}(t,\mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[4\hat{\beta}_1 + (\overline{\text{AGE}})\hat{\beta}_2 + (\overline{\text{RACE}})\hat{\beta}_3 + (\overline{\text{SEX}})\hat{\beta}_4\right]}$

   f. The two survival curves will **not** cross, because both are computed using the same proportional hazards model, which has the property that the hazard functions, as well as their corresponding estimated survivor functions, will not cross.

   g. 95% CI for HR = $\exp[\ell]$: $\exp[\hat{\ell} \pm 1.96\sqrt{\hat{\text{Var}}\ \hat{\ell}}]$ where $\ell = 2\beta_1 + 2(\text{AGE})\beta_5 + 2(\text{RACE})\beta_6 + 2(\text{SEX})\beta_7$

2. a. $h\,(t,\mathbf{X}) = h_0\,(t)\exp\,[\beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \ldots + \beta_{10} X_{10}]$

   b. Adeno cell type: $\mathbf{X}^*$ = (treatment, 1, 0, 0, perfstat, disdur, age, prther)
   Large cell type: $\mathbf{X}$ = (treatment, 0, 0, 0, perfstat, disdur, age, prther)

   $$\begin{aligned} HR = \frac{h(t,\,\mathbf{X}^*)}{h(t,\,\mathbf{X})} &= \exp\,\left[\sum_{i=1}^{p}\beta_i\left(X_i^* - X_i\right)\right] \\ &= \exp\,[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) \\ &\quad + \hat{\beta}_5(0-0) + 0 + \ldots + 0] \\ &= .\exp\,\left[\hat{\beta}_3\right] = \exp[0.789] = 2.20 \end{aligned}$$

   c. Adeno cell type: $\mathbf{X}^*$ = (treatment, 1, 0, 0, perfstat, disdur, age, prther)
   Squamous cell type: $\mathbf{X}$ = (treatment, 0, 0, 1, perfstat, disdur, age, prther)

   $$\begin{aligned} HR = \frac{h(t,\,\mathbf{X}^*)}{h(t,\,\mathbf{X})} &= \exp\,\left[\sum_{i=1}^{p}\beta_i\left(X_i^* - X_i\right)\right] \\ &= \exp\,[0 + \hat{\beta}_3(1-0) + \hat{\beta}_4(0-0) \\ &\quad + \hat{\beta}_5(0-1) + 0 + \ldots + 0] \\ &= \exp\,\left[\hat{\beta}_3 - \hat{\beta}_5\right] = \exp\,[0.789 \\ &\quad - (-0.400)] = \exp\,[1.189] = 3.28 \end{aligned}$$

d. There does not appear to be an effect of treatment on survival time, adjusted for the other variables in the model. The hazard ratio is 1.3, which is close to the null value of one, the p-value of 0.162 for the Wald test for treatment is not significant, and the 95% confidence interval for the treatment effect correspondingly includes the null value.

e. $\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[2\hat{\beta}_1 + \hat{\beta}_5 + \overline{(\text{perfstat})}\hat{\beta}_7 + \overline{(\text{disdur})}\hat{\beta}_8 + \overline{(\text{age})}\hat{\beta}_9 + \overline{(\text{prther})}\hat{\beta}_{10}\right]}$

f. $$HR = \frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp[\beta_1 + (\text{perfstat})\beta_{11} + (\text{disdur})\beta_{12} + (\text{age})\beta_{13} + (\text{prther})\beta_{14}]$$

where β_1 is the coefficient of the treatment variable and β_{11}, β_{12}, β_{13}, and β_{14} are the coefficients of product terms involving treatment with the four variables indicated.

3. a. $\widehat{HR} = \exp[0.470 + (-0.008)\text{age} + (-0.503)\text{sex}]$

b. 40-year-old male:

$$\widehat{HR} = \exp[0.470 + (-0.008)40 + (-0.503)1] = 0.70$$

50-year-old Female:

$$\widehat{HR} = \exp[0.470 + (-0.008)50 + (-0.503)2] = 0.39$$

c. The LR (chunk) test for the significance of both interaction terms simultaneously yields the following likelihood ratio statistic which compares models 1 and 2:

$$LR = [(-2 \times -153.253) - (-2 \times -153.040)]$$
$$= 306.506 - 306.080 = 0.426$$

This statistic is approximately chi-square with 2 degrees of freedom under the null hypothesis of no interaction. This LR statistic is highly nonsignificant. Thus, we conclude that there is no significant interaction in the model (1).

d. The gold-standard hazard ratio is 0.484, which is obtained for model 2. Note that model 2 contains no interaction terms and controls for both covariates of interest. When either age or sex or both are dropped from the model, the hazard ratio (for platelets) does not change appreciably. Therefore, it appears that neither age nor sex need to be controlled for confounding.

e. Models 2–5 are all more or less equivalent, since they all give essentially the same hazards ratio and confidence interval for the effect of the platelet variable. A political choice for best model would be the gold-standard model (2), because the critical reviewer can see both age and sex being controlled in model 2.

f.
- The point estimate of the hazard ratio for normal versus abnormal platelet count is $0.484 = 1/2.07$, so that the hazard for an abnormal count is twice that for a normal count.
- There is a borderline significant effect of platelet count on survival adjusted for age and sex ($P = .071$).
- The 95% CI for the hazard ratio is given by $0.221 < HR < 1.063$, which is quite wide and therefore shows a very imprecise estimate.

4. Subjects may already at risk for the outcome prior to their study entry (at diagnosis). If so, then the time at risk prior to study entry contributes to the true survival time (say, T) for the individual, although only the observed time-on-study (t), is actually available to be analyzed. The individual's survival time is therefore underestimated by the time-on-study information (obtained from study entry), i.e., the true survival time is **left-truncated**. However, if age-as-the-time-scale as the outcome is considered, then it is possible to adjust for this left truncation by using age at entry in a hazard model of the form

$$h(a, \mathbf{X}) = h_0(a|a_0) \exp\left[\sum \beta_i X_i\right]$$

where a denotes age at follow-up, a_0 denotes age at study entry, and $h_0(a|a_0)$ is a baseline hazard that adjusts for age truncation at a_0.

4 Evaluating the Proportional Hazards Assumption

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_4,
© Springer Science+Business Media, LLC 2012

Introduction

We begin with a brief review of the characteristics of the Cox proportional hazards (PH) model. We then give an overview of three methods for checking the PH assumption: graphical, goodness-of-fit (GOF), and time-dependent variable approaches.

We then focus on each of the above approaches, starting with graphical methods. The most popular graphical approach involves the use of "log–log" survival curves. A second graphical approach involves the comparison of "observed" with "expected" survival curves.

The GOF approach uses a test statistic or equivalent p-value to assess the significance of the PH assumption. We illustrate this test and describe some of its advantages and drawbacks.

Finally, we discuss the use of time-dependent variables in an extended Cox model as a third method for checking the PH assumption. A more detailed description of the use of time-dependent variables is provided in Chapter 6.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

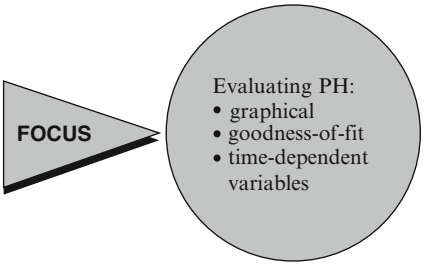
I. **Background** (pages 164–165)
II. **Checking the proportional hazards assumption: overview** (pages 165–167)
III. **Graphical approach 1: log–log plots** (pages 167–175)
IV. **Graphical approach 2: observed versus expected plots** (pages 175–180)
V. **The goodness-of-fit (GOF) testing approach** (pages 181–183)
VI. **Assessing the PH assumption using time-dependent covariates** (pages 183–187)

Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize three general approaches for evaluating the PH assumption.
2. Summarize how log–log survival curves may be used to assess the PH assumption.
3. Summarize how observed versus expected plots may be used to assess the PH assumption.
4. Summarize how GOF tests may be used to assess the PH assumption.
5. Summarize how time-dependent variables may be used to assess the PH assumption.
6. Describe — given survival data or computer output from a survival analysis that uses a Cox PH model — how to assess the PH assumption for one or more variables in the model using:
   a. a graphical approach
   b. the GOF approach
   c. an extended Cox model with time-dependent covariates
7. State the formula for an extended Cox model that provides a method for checking the PH assumption for one or more of the time-independent variables in the model, given survival analysis data or computer output that uses a Cox PH model.

Presentation



This presentation describes three approaches for evaluating the proportional hazards (PH) assumption of the Cox model — a graphical procedure, a goodness-of-fit testing procedure, and a procedure that involves the use of time-dependent variables.

I. Background

Cox PH model:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

$\mathbf{X} = (X_1, X_2, \ldots, X_p)$ explanatory/predictor variables

$h_0(t)$	$\times$	$e^{\sum_{i=1}^{p} \beta_i X_i}$
Baseline hazard Involves t but not X's		Exponential Involves X's but not t (X's are time-independent)

X's involving t: **time-dependent**
Requires **extended Cox model** (no PH)

Chapter 6

Recall from the previous chapter that the general form of the Cox PH model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by the bold $\mathbf{X}$.

The Cox model formula says that the hazard at time t is the product of two quantities. The first of these, $h_0(t)$, is called the **baseline hazard** function. The second quantity is the exponential expression e to the linear sum of $\beta_i X_i$, where the sum is over the p explanatory X variables.

An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the X's, whereas the exponential expression involves the X's, but does not involve t. The X's here are called **time-independent** X's.

It is possible, nevertheless, to consider X's that do involve t. Such X's are called **time-dependent** variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption, and is called the **extended Cox model**. We will discuss this extended Cox model in Chapter 6 of this series.

Hazard ratio formula:

$$\widehat{HR} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i \left(X_i^* - X_i\right)\right]$$

where $\mathbf{X}^* = \left(X_1^*, X_2^*, \cdots, X_p^*\right)$
and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$
denote the two sets of X's.

From the Cox PH model, we can obtain a general formula, shown here, for estimating a hazard ratio that compares two specifications of the X's, defined as $\mathbf{X}^*$ and $\mathbf{X}$.

Adjusted survival curves comparing E groups:

$$\hat{S}(t, \mathbf{X}) = \left[\hat{S}_0(t)\right]^{\exp\left[\beta_1 E + \sum_{i \neq 1} \beta_i \overline{X}_i\right]}$$

We can also obtain from the Cox model an expression for an adjusted survival curve. Here we show a general formula for obtaining adjusted survival curves comparing two groups adjusted for other variables in the model. Below this, we give a formula for a single adjusted survival curve that adjusts for all X's in the model. Computer packages for these formulae use the mean value of each X being adjusted in the computation of the adjusted curve.

Single curve:

$$S(t, \overline{\mathbf{X}}) = \left[\hat{S}_0(t)\right]^{\exp\left[\sum_{i=1}^{p} \hat{\beta}_i \overline{X}_i\right]}$$

PH assumption:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}, \text{ constant over } t$$

i.e., $\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X})$

The Cox PH model assumes that the hazard ratio comparing any two specifications of predictors is constant over time. Equivalently, this means that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

Hazards cross: $\Rightarrow$ PH not met

Hazards don't cross $\not\Rightarrow$ PH met

The PH assumption is not met if the graph of the hazards cross for two or more categories of a predictor of interest. However, even if the hazard functions do not cross, it is possible that the PH assumption is not met. Thus, rather than checking for crossing hazards, we must use other approaches to evaluate the reasonableness of the PH assumption.

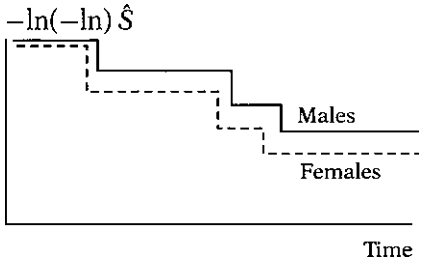
II. Checking the Proportional Hazards Assumption: Overview

There are three general approaches for assessing the PH assumption, again listed here. We now briefly overview each approach, starting with graphical techniques.

Three approaches:

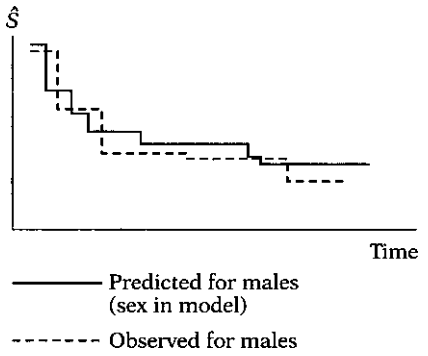
- graphical
- goodness-of-fit test
- time-dependent variables

Graphical techniques:
–ln(–ln) *S* curves parallel?



There are two types of graphical techniques available. The most popular of these involves comparing **estimated –ln(–ln) survivor curves** over different (combinations of) categories of variables being investigated. We will describe such curves in detail in the next section. Parallel curves, say comparing males with females, indicate that the PH assumption is satisfied, as shown in this illustration for the variable Sex.

Observed vs. predicted: Close?



An alternative graphical approach is to compare **observed with predicted** survivor curves. The observed curves are derived for categories of the variable being assessed, say, Sex, without putting this variable in a PH model. The predicted curves are derived with this variable included in a PH model. If observed and predicted curves are close, then the PH assumption is reasonable.

Goodness-of-fit (GOF) tests:

- Large sample *Z* or chi-square statistics
- Gives p-value for evaluating PH assumption for each variable in the model.

p-value large ⇒ PH satisfied
(e.g. $P > 0.10$)
p-value small ⇒ PH not satisfied
(e.g. $P < 0.05$)

A second approach for assessing the PH assumption involves goodness-of-fit (GOF) tests. This approach provides large sample *Z* or chi-square statistics which can be computed for each variable in the model, adjusted for the other variables in the model. A p-value derived from a standard normal statistic is also given for each variable. This p-value is used for evaluating the PH assumption for that variable. A nonsignificant (i.e., large) p-value, say greater than 0.10, suggests that the PH assumption is reasonable, whereas a small p-value, say less than 0.05, suggests that the variable being tested does not satisfy this assumption.

Time-dependent covariates:

Extended Cox model:
Add product term involving some function of time.

When time-dependent variables are used to assess the PH assumption for a time-independent variable, the Cox model is extended to contain **product** (i.e., interaction) **terms** involving the time-independent variable being assessed and some function of time.

EXAMPLE

$h(t, X) = h_0(t) \exp[\beta \text{Sex} + \delta(\text{Sex} \times t)]$
$\delta \neq 0$. PH assumption violated

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable "Sex $\times$ t" in addition to Sex. If the coefficient of the product term turns out to be significant, we can conclude that the PH assumption is violated for Sex.

GOF provides test statistic
Graphical: subjective
Time-dependent: computationally cumbersome
GOF: global, may not detect specific departures from PH

The GOF approach provides a single test statistic for each variable being assessed. This approach is not as subjective as the graphical approach nor as cumbersome computationally as the time-dependent variable approach. Nevertheless, a GOF test may be too "global" in that it may not detect specific departures from the PH assumption that may be observed from the other two approaches.

III. Graphical Approach 1: Log–Log Plots

- log–log survival curves
- observed versus expected survival curves

The two graphical approaches for checking the PH assumption are comparing log–log survival curves and comparing observed versus expected survival curves. We first explain what a −ln −ln survival curve is and how it is used.

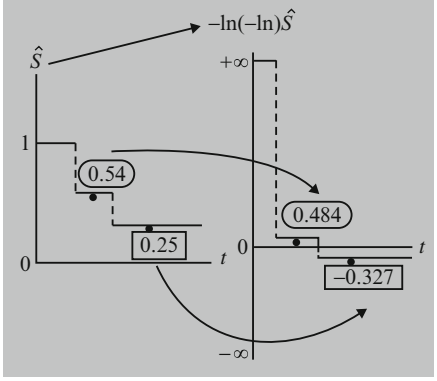
log−log $\hat{S}$ = transformation of $\hat{S}$
= −ln−ln $\hat{S}$)

- ln $\hat{S}$ is negative ⇒ − (ln $\hat{S}$) is positive.
- can't take log of ln $\hat{S}$, but can take log of (−ln $\hat{S}$).
- −ln(−ln $\hat{S}$) may be positive or negative.

A log–log survival curve is simply a transformation of an estimated survival curve that results from taking the natural log of an estimated survival probability *twice*. Mathematically, we write a log–log curve as −ln(−ln $\hat{S}$). Note that the log of a probability such as $\hat{S}$ is always a negative number. Because we can only take logs of positive numbers, we need to negate the first log before taking the second log. The value for −ln(−ln $\hat{S}$) may be positive or negative, either of which is acceptable, because we are not taking a third log.[1]

[1] An equivalent way to write to −ln(−ln$\hat{S}$) is $-\ln\left(\int_0^t h(u)du\right)$, where $\int_0^t h(u)du$ is called the "cumulative hazard" function. This result follows from the formula $S(t) = \exp\left[-\int_0^t h(u)du\right]$, which relates the survivor function to the hazard function (see p. 15 in Chapter 1).

EXAMPLE



As an example, in the graph at left, the estimated survival probability of 0.54 is transformed to a log–log value of 0.484. Similarly, the point 0.25 on the survival curve is transformed to a −ln −ln value of −0.327.

Note that because the survival curve is usually plotted as a step function, so will the log–log curve be plotted as a step function.

EXAMPLE

$\hat{S} = 0.54$; want $-\ln(-\ln 0.54)$
$-\ln(-\ln 0.54) = -\ln(0.616)$
since $\ln(0.54) = -0.616$
$-\ln(0.616) = 0.484$
since $\ln(0.616) = -0.484$
Thus, $-\ln(-\ln 0.54) = 0.484$

To illustrate the computation of a log–log value, suppose we start with an estimated survival probability of 0.54. Then the log–log transformation of this value is $-\ln(-\ln 0.54)$, which is $-\ln(0.616)$, because $\ln(0.54)$ equals -0.616. Now, continuing further, $-\ln(0.616)$ equals 0.484, because $\ln(0.616)$ equals -0.484. Thus, the transformation $-\ln(-\ln 0.54)$ equals 0.484.

ANOTHER EXAMPLE

$\hat{S} = 0.25$; want $-\ln(-\ln 0.25)$
$-\ln(-\ln 0.25) = -\ln(1.386) = -0.327$
Thus, $-\ln(-\ln 0.25) = -0.327$

As another example, if the estimated survival probability is 0.25, then $-\ln(-\ln 0.25)$ equals $-\ln(1.386)$, which equals -0.327.

y-axis scale:

$$\begin{array}{c|c|c|} 1 & & +\infty \\ 0 & \hat{S} & \\ & & -\infty \end{array} \quad -\ln(1-\ln)\hat{S}$$

Note that the scale of the *y*-axis of an estimated survival curve ranges between 0 and 1, whereas the corresponding scale for a $-\ln(-\ln)$ curve ranges between $-\infty$ and $+\infty$.

log–log $\hat{S}$ for the Cox PH model:

We now show why the PH assumption can be assessed by evaluating whether or not log–log curves are parallel. To do this, we must first describe the log–log formula for the Cox PH model.

Cox PH hazard function:

$$h(t, \mathbf{X}) = h_0(t)e^{\sum_{i=1}^{p} \beta_i X_i}$$

↓ From math

Cox PH survival function:

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{\sum_{i=1}^{p} \beta_i X_i}}$$

↗ Baseline survival function.

We start with the formula for the survival curve that corresponds to the hazard function for the Cox PH model. Recall that there is a mathematical relationship between any hazard function and its corresponding survival function. We therefore can obtain the formula shown here for the survival curve for the Cox PH model. In this formula, the expression $S_0(t)$ denotes the baseline survival function that corresponds to the baseline hazard function $h_0(t)$.

log–log⇒ takes logs twice

The log–log formula requires us to take logs of this survival function twice. The first time we take logs we get the expression shown here.

log #1:

$$\ln\ S(t, \mathbf{X}) = e^{\sum_{i=1}^{p} \beta_i X_i} \times\ \ln\ S_0(t)$$

$$0 \leq S(t, \mathbf{X}) \leq 1$$

Now since $S(t, \mathbf{X})$ denotes a survival probability, its value for any t and any specification of the vector $\mathbf{X}$ will be some number between 0 and 1. It follows that the natural log of any number between 0 and 1 is a negative number, so that the log of $S(t, \mathbf{X})$ as well as the log of $S_0(t)$ are both negative numbers. This is why we have to put a minus sign in front of this expression before we can take logs a second time, because there is no such thing as the log of a negative number.

ln(probability) = negative value, so $\ln S(t,\mathbf{X})$ and $\ln S_0(t)$ are negative.

But $-\ln\ S(t,\mathbf{X})$ is positive, which allows us to take logs again.

log #2:

$$\ln[-\ln S(t, \mathbf{X})]$$

$$= \ln\left[-e^{\sum_{i=1}^{p} \beta_i X_i} \times\ \ln\ S_0(t)\right]$$

$$= \ln\left[e^{\sum_{i=1}^{p} \beta_i X_i}\right] + \ln[-\ln\ S_0(t)]$$

$$= \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln\ S_0(t)]$$

$$= \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln\ S_0(t)]$$

Thus, when taking the second log, we must obtain the log of $-\ln\ S(t, \mathbf{X})$, as shown here. After using some algebra, this expression can be rewritten as the sum of two terms, one of which is the **linear sum of the** $\beta_i X_i$ and the other is the **log of the negative log of the baseline survival function**.

or

$$-\ln[-\ln\ S(t, \mathbf{X})]$$

$$= -\sum_{i=1}^{p} \beta_i X_i - \ln[-\ln\ S_0(t)]$$

This second log may be either positive or negative, and we aren't taking any more logs, so we actually don't have to take a second negative. However, for consistency's sake, a common practice is to put a minus sign in front of the second log to obtain the $-\ln\ -\ln$ expression shown here. Nevertheless, some software packages do not use a second minus sign.

Two individuals:
$\mathbf{X}_1 = (X_{11}, X_{12}, \ldots, X_{1p})$
$\mathbf{X}_2 = (X_{21}, X_{22}, \ldots, X_{2p})$

Now suppose we consider two different specifications of the $\mathbf{X}$ vector, corresponding to two different individuals, $\mathbf{X}_1$ and $\mathbf{X}_2$.

$$\begin{cases} \ln[-\ln\ S(t, \mathbf{X}_1)] \\ \quad = \sum_{i=1}^{p} \beta_i X_{1i} + \ln[-\ln\ S_0(t)] \\ \ln[-\ln\ S(t, \mathbf{X}_2)] \\ \quad = \sum_{i=1}^{p} \beta_i X_{2i} + \ln[-\ln\ S_0(t)] \end{cases}$$

Then the corresponding log–log curves for these individuals are given as shown here, where we have simply substituted $\mathbf{X}_1$ and $\mathbf{X}_2$ for $\mathbf{X}$ in the previous expression for the log–log curve for any individual $\mathbf{X}$.

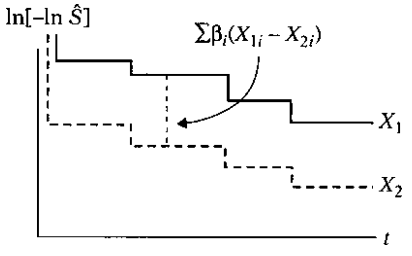
$$\ln[-\ln\ S(t, \mathbf{X}_1)] - (\ln[-\ln\ S(t, \mathbf{X}_2)]) = \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})$$

does not involve t

Subtracting the second log–log curve from the first yields the expression shown here. This expression is a linear sum of the differences in corresponding predictor values for the two individuals. Note that the baseline survival function has dropped out, so that the difference in log–log curves involves an expression that does not involve time t.

$$-\ln[-\ln\ S(t, \mathbf{X}_1)] = \ln[-\ln\ S(t, \mathbf{X}_2)] + \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})$$

Alternatively, using algebra, we can write the above equation by expressing the log–log survival curve for individual $\mathbf{X}_1$ as the log–log curve for individual $\mathbf{X}_2$ plus a linear sum term that is independent of t.



The above formula says that if we use a Cox PH model and we plot the estimated log–log survival curves for individuals on the same graph, the two plots would be approximately parallel. The distance between the two curves is the linear expression involving the differences in predictor values, which does not involve time. Note, in general, if the vertical distance between two curves is constant, then the curves are parallel.

Graphical approach using log–log plots: PH model is appropriate if "empirical" plots of log–log survival curves are parallel.

The parallelism of log–log survival plots for the Cox PH model provides us with a graphical approach for assessing the PH assumption. That is, if a PH model is appropriate for a given set of predictors, one should expect that empirical plots of log–log survival curves for different individuals will be approximately parallel.

Empirical plots: use $-\ln[-\ln \hat{S}]$ where

1. $\hat{S}$ is a KM curve
2. $\hat{S}$ is an adjusted survival curve for predictors satisfying the PH assumption; predictor being assessed not included in model

By empirical plots, we mean plotting log–log survival curves based on Kaplan–Meier (KM) estimates that do not assume an underlying Cox model. Alternatively, one could plot log–log survival curves which have been adjusted for predictors already assumed to satisfy the PH assumption but have not included the predictor being assessed in a PH model.

EXAMPLE

Clinical trial of leukemia patients:
T = weeks until patient goes out of remission

Predictors (X's):
Rx (= 1 if placebo, 0 if treatment)
log WBC

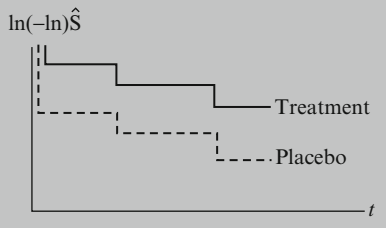
Cox PH model:

$h(t, \mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$
Assessing PH assumption: compare log–log survival curves for categories of Rx and log WBC

As an example, suppose we consider the comparison of treatment and placebo groups in a clinical trial of leukemia patients, where survival time is time, in weeks, until a patient goes out of remission. Two predictors of interest in this study are treatment group status (1 = placebo, 0 = treatment), denoted as Rx, and log white blood cell count (log WBC), where the latter variable is being considered as a confounder.

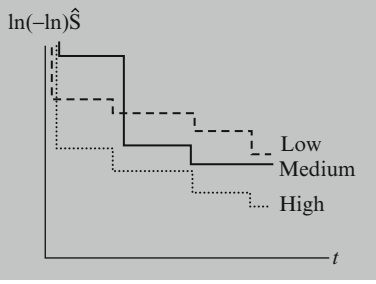
A Cox PH model involving both these predictors would have the form shown at the left. To assess whether the PH assumption is satisfied for either or both of these variables, we would need to compare log–log survival curves involving categories of these variables.

One-at-a-time strategy: Rx variable



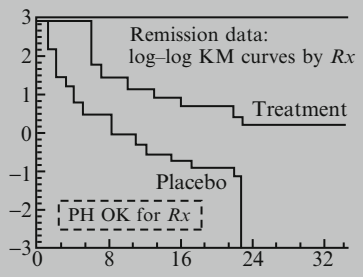
One strategy to take here is to consider the variables one at a time. For the Rx variable, this amounts to plotting log–log KM curves for treatment and placebo groups and assessing parallelism. If the two curves are approximately parallel, as shown here, we would conclude that the PH assumption is satisfied for the variable Rx. If the two curves intersect or are not parallel in some other way, we would conclude that the PH assumption is not satisfied for this variable.

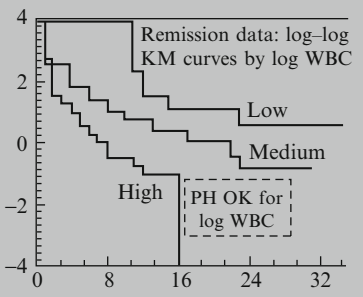
One-at-a-time strategy: log WBC

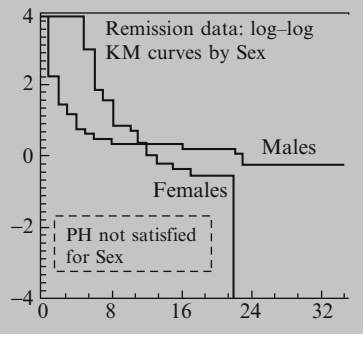


For the log WBC variable, we need to categorize this variable into categories – say, low, medium, and high – and then compare plots of log–log KM curves for each of the three categories. In this illustration, the three log–log Kaplan–Meier curves are clearly nonparallel, indicating that the PH assumption is not met for log WBC.

EXAMPLE: Computer Results







The above examples are sketches of some of the possibilities that could occur from comparisons of log–log curves. For the actual data set containing 42 leukemia patients, computer results are shown here for each variable separately. Similar output using Stata, SAS, SPSS, and R packages is provided in the Computer Appendix.

We first show the log–log KM curves by treatment, *Rx*. Notice that the two log–log curves are roughly parallel, indicating that the *Rx* variable satisfies the PH assumption when being considered by itself.

Here we show the log–log KM curves by log WBC, where we have divided this variable into low (below 2.3), medium (between 2.3 and 3), and high (above 3) values. Notice that there is some indication of nonparallelism below 8 days, but that overall the three curves are roughly parallel. Thus, these plots suggest that the PH assumption is more or less satisfied for the variable log WBC, when considered alone.

As a third example, we consider the log–log KM plots categorized by Sex from the remission data. Notice that the two curves clearly intersect, and are therefore noticeably nonparallel. Thus, the variable, Sex, when considered by itself, does not appear to satisfy the PH assumption and therefore should not be incorporated directly into a Cox PH model containing the other two variables, *Rx* and log WBC.

Problems with log–log survival curve approach:

How parallel is parallel?
Recommend:

- subjective decision
- conservative strategy: assume PH is OK unless strong evidence of nonparallelism

The above examples suggest that there are some problems associated with this graphical approach for assessing the PH assumption. The main problem concerns how to decide "how parallel is parallel?" This decision can be quite subjective for a given data set, particularly if the study size is relatively small. We recommend that one should use a conservative strategy for this decision by assuming the PH assumption is satisfied unless there is **strong** evidence of nonparallelism of the log–log curves.

How parallel is parallel?
Recommend:

- many categories ⇒ data "thins out"
- different categorizations may give different graphical pictures

Another problem concerns how to categorize a continuous variable like log WBC. If many categories are chosen, the data "thins out" in each category, making it difficult to compare different curves. [Also, one categorization into, say, three groups may give a different graphical picture from a different categorization into three groups.]

Recommend:

- small # of categories (2 or 3)
- meaningful choice
- reasonable balance (e.g., terciles)

In categorizing continuous variables, we recommend that the number of categories be kept reasonably small (e.g., two or three) if possible, and that the choice of categories be as meaningful as possible and also provide reasonable balance of numbers (e.g., as when using terciles).

How to evaluate several variables simultaneously?

In addition to the two problems just described, another problem with using log–log survival plots concerns how to evaluate the PH assumption for several variables simultaneously.

Strategy:

- categorize variables separately
- form combinations of categories
- compare log–log curves on same graph

One strategy for simultaneous comparisons is to categorize all variables separately, form combinations of categories, and then compare log–log curves for all combinations on the same graph.

Drawback:

- data "thins out"
- difficult to identify variables responsible for nonparallelism

A drawback of this strategy is that the data will again tend to "thin out" as the number of combinations gets even moderately large. Also, even if there are sufficient numbers for each combined category, it is often difficult to determine which variables are responsible for any nonparallelism that might be found.

EXAMPLE

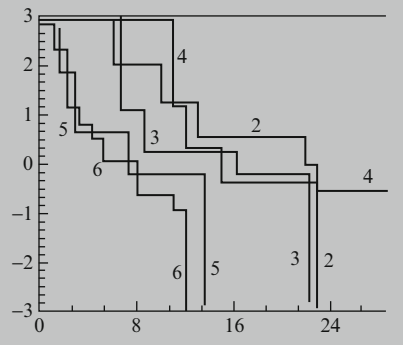
Remission Data:

Rx \ log WBC	Low	Medium	High
Treatment	✓	✓	✓
Placebo	✓	✓	✓

As an example of this strategy, suppose we use the remission data again and consider both *Rx* and log WBC together. Because we previously had two categories of *Rx* and three categories of log WBC, we get a total of six combined categories, consisting of treated subjects with low log WBC, placebo subjects with low log WBC, treated subjects with medium log WBC, and so on.

EXAMPLE: (continued)

Log–log $\widehat{\text{KM}}$ curves by six combinations of *Rx* by log WBC



Plots suggest PH not satisfied. However, the study is small, i.e., plots are unreliable.

The computer results are shown here for the log–log curves corresponding to each of the six combinations of *Rx* with log WBC. Notice that there are several points of intersection among the six curves. Therefore, these results suggest that the PH assumption is not satisfied when considering *Rx* and log WBC together.

However, the sample sizes used to estimate these curves are quite small, ranging between four subjects for group 4 (*Rx* = 1, log WBC = low) to 12 subjects for group 6 (*Rx* = 1, log WBC = high), with the total study size being 42. Thus, for this small study, the use of six log–log curves provides unreliable information for assessing the PH assumption.

Alternative strategy:

Adjust for predictors already satisfying PH assumption, i.e., use adjusted log–log $\hat{S}$ curves

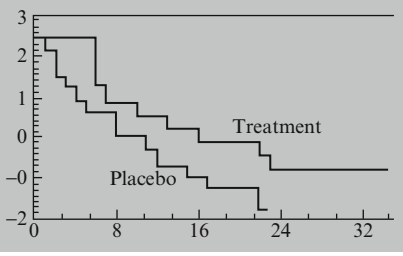
An alternative graphical strategy for considering several predictors together is to assess the PH assumption for one predictor adjusted for other predictors that are assumed to satisfy the PH assumption. Rather than using Kaplan–Meier curves, this involves a comparison of adjusted log–log survival curves.

EXAMPLE

Remission data:

- compare *Rx* categories adjusted for log WBC
- fit PH model for each *Rx* stratum
- obtain adjusted survival curves using overall mean of log WBC

Log–log $\hat{S}$ curves for *Rx* groups using PH model adjusted for log WBC

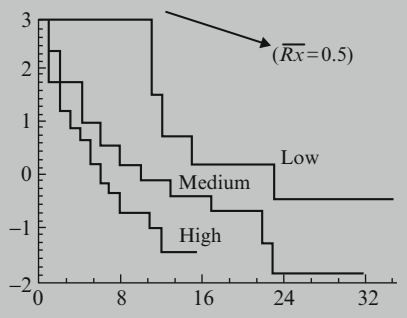


As an example, again we consider the remission data and the predictors *Rx* and log WBC. To assess the PH assumption for *Rx* adjusted for log WBC, we would compare adjusted log–log survival curves for the two treatment categories, where each adjusted curve is derived from a PH model containing log WBC as a predictor. In computing the adjusted survival curve, we need to stratify the data by treatment, fit a PH model in each stratum, and then obtain adjusted survival probabilities using the overall mean log WBC in the estimated survival curve formula for each stratum.

For the remission data example, the estimated log–log survival curves for the two treatment groups adjusted for log WBC are shown here. Notice that these two curves are roughly parallel, indicating that the PH assumption is satisfied for treatment.

EXAMPLE: (continued)

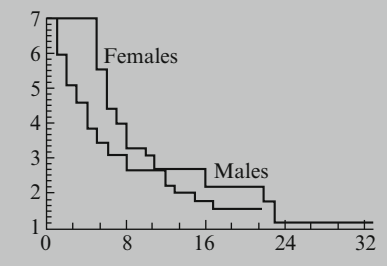
Log–log $\hat{S}$ curves for log WBC groups using PH model adjusted for *Rx*



Remission data:
Assess PH assumption for Sex:

- use PH model containing *Rx* and log WBC
- use *Rx* and log WBC in survival probability formula

Log–log $\hat{S}$ curves for Sex adjusted for *Rx* and log WBC



As another example, we consider adjusted log–log survival curves for three categories of log WBC, adjusted for the treatment status (*Rx*) variable. The adjusted survival probabilities in this case use the overall mean *Rx* score, i.e., 0.5, the proportion of the 42 total subjects that are in the placebo group (i.e., half the subjects have a score of *Rx* = 1).

The three log–log curves adjusted for treatment status are shown here. Although two of these curves intersect early in follow-up, they do not suggest a strong departure from parallelism overall, suggesting that the PH assumption is reasonable for log WBC, after adjusting for treatment status.

As a third example, again using the remission data, we assess the PH assumption for Sex, adjusting for both treatment status and log WBC in the model. This involves obtaining log–log survival curves for males and females separately, using a PH model that contains both treatment status and log WBC. The adjustment uses the overall mean treatment score and the overall mean log WBC score in the formula for the estimated survival probability.

The estimated log–log survival curves for Sex, adjusted for treatment and log WBC are shown here. These curves clearly cross, indicating that the PH assumption is not satisfied for Sex, after adjusting for treatment and log WBC.

✓ 1. log–log survival curves
2. observed versus expected survival curves

We have thus described and illustrated one of the two graphical approaches for checking the PH assumption, that is, using log–log survival plots. In the next section, we describe an alternative approach that compares "observed" with "expected" survival curves.

IV. Graphical Approach 2: Observed Versus Expected Plots

Graphical analog of GOF test

The use of observed versus expected plots to assess the PH assumption is the graphical analog of the goodness-of-fit (GOF) testing approach to be described later, and is therefore a reasonable alternative to the log–log survival curve approach.

Two strategies:

1. One-at-a-time: uses KM curves to obtain observed plots
2. Adjusting for other variables: uses stratified Cox PH model to obtain observed plots (see Chapter 5)

As with the log–log approach, the observed versus expected approach may be carried out using either or both of two strategies—(1) assessing the PH assumption for variables one-at-a-time, or (2) assessing the PH assumption after adjusting for other variables. The strategy which adjusts for other variables uses a stratified Cox PH model to form observed plots, where the PH model contains the variables to be adjusted and the stratified variable is the predictor being assessed. The stratified Cox procedure is described in Chapter 5.

Here, we describe only the one-at-a-time strategy, which involves using KM curves to obtain observed plots.

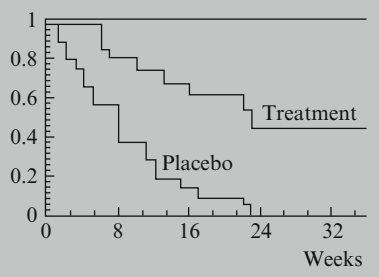
One-at-a-time:

- stratify data by categories of predictor
- obtain KM curves for each category

Using the one-at-a-time strategy, we first must stratify our data by categories of the predictor to be assessed. We then obtain observed plots by deriving the KM curves separately for each category.

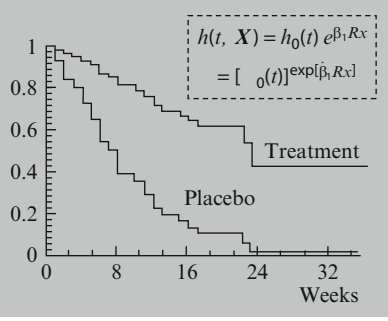
EXAMPLE: Remission Data

KM (Observed) Plots by *Rx* Group



As an example, for the remission data on 42 leukemia patients we have illustrated earlier, the KM plots for the treatment and placebo groups, with 21 subjects in each group, are shown here. These are the "observed" plots.

Expected Survival Plots by *Rx* Group Using PH Model

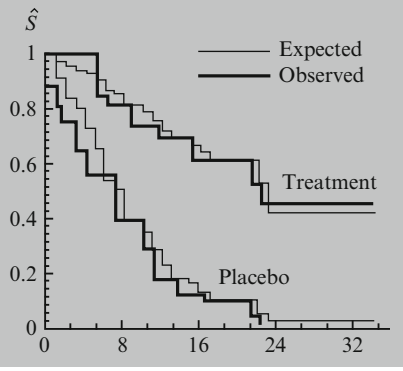


To obtain "expected" plots, we fit a Cox PH model containing the predictor being assessed. We obtain expected plots by separately substituting the value for each category of the predictor into the formula for the estimated survival curve, thereby obtaining a separate estimated survival curve for each category.

As an example, again using the remission data, we fit the Cox PH model with *Rx* as its only variable. Using the corresponding survival curve formula for this Cox model, as given in the box at the left, we then obtain separate expected plots by substituting the values of 0 (for treatment group) and 1 (for placebo group). The expected plots are shown here.

EXAMPLE: (continued)

Observed Versus Expected Plots by *Rx*



To compare observed with expected plots we then put both sets of plots on the same graph as shown here.

If observed and expected plots are:

- **close**, complies with PH assumption
- **discrepant**, PH assumption violated

If for each category of the predictor being assessed, the observed and expected plots are "close" to one another, we then can conclude that the PH assumption is satisfied. If, however, one or more categories show quite discrepant observed and expected plots, we conclude that the PH assumption is violated.

Example: Remission Data (continued)

Observed and expected plots are close for each treatment group.
Conclude PH assumption not violated.

For the example shown above, observed and expected curves appear to be quite close for each treatment group. Thus, we would conclude using this graphical approach that the treatment variable satisfies the PH assumption.

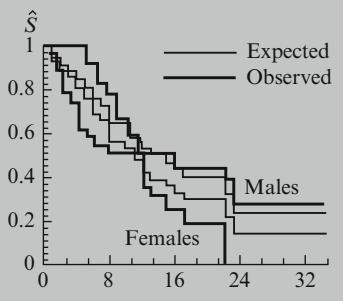
Drawback: How close is close?

Recommend: PH not satisfied *only* when plots are strongly discrepant.

An obvious drawback to this graphical approach is deciding "how close is close" when comparing observed versus expected curves for a given category. This is analogous to deciding "how parallel is parallel" when comparing log–log survival curves. Here, we recommend that the PH assumption be considered as not satisfied only when observed and expected plots are strongly discrepant.

Example: Remission Data

Observed Versus Expected Plots by Sex



PH assumption not satisfied for Sex. Same conclusion as with log–log curves.

As another example, again using the remission data, we consider observed versus expected plots by Sex, as shown here. Note that the observed plots for males and females, which are described by the thicker lines, cross at about 12 weeks, whereas the expected plots don't actually intersect, with the female plot lying below the male plot throughout follow-up. Moreover, for males and females separately, the observed and expected plots are quite different from one another.

Thus, the above plots suggest that the PH assumption is not satisfied for the variable Sex. We came to the same conclusion when using log–log survival curves, which crossed one another and were therefore clearly non-parallel.

Continuous variable:

- form strata from categories
- observed plots are KM curves for each category

When using observed versus expected plots to assess the PH assumption for a continuous variable, observed plots are derived, as for categorical variables, by forming strata from categories of the continuous variable and then obtaining KM curves for each category.

- two options for expected plots
  1. Use PH model with $k-1$ dummy variables X_c for k categories, i.e.,

$$h(t,\mathbf{X}) = h_0(t)\exp\left(\sum_{i=1}^{k-1}\beta_i X_{ci}\right)$$

Obtain adjusted survival curve:

$$\hat{S}(t,\mathbf{X}_c) = \left[\hat{S}_0(t)\right]^{\exp\left(\sum\hat{\beta}_i X_{ci}\right)}$$

where $\mathbf{X}_c = (X_{c1}, X_{c2}, \ldots, X_{c,k-1})$ gives values of dummy variables for category c.

However, for continuous predictors, there are two options available for computing expected plots. One option is to use a Cox PH model which contains $k-1$ dummy variables to indicate k categories. The expected plot for a given category is then obtained as an adjusted survival curve by substituting the values for the dummy variables that define the given category into the formula for the estimated survival curve, as shown here for category c.

Options for a continuous variable:

2. Use PH model:
   $h(t, \mathbf{X}) = h_0(t) \exp(\beta X)$
   so $\mathbf{X} = X$ (Continuous)
   Obtain adjusted survival curve:

$$\hat{S}(t, \overline{X}_c) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}\overline{X}_c\right)}$$

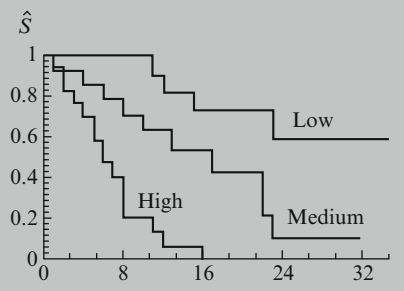
where $\overline{X}_c$ denotes the mean value for the variable X within category c.

The second option is to use a Cox PH model containing the continuous predictor (say, X) being assessed. Expected plots are then obtained as adjusted survival curves by specifying predictor values that distinguish categories, as, for example, when using mean predictor values for each category.

As an example to illustrate both options, we consider the continuous variable log WBC from the remission data example. To assess the PH assumption for this variable, we would first stratify log WBC into, say, three categories – low, medium, and high. The observed plots would then be obtained as KM curves for each of the three strata, as shown here.

Example: Remission Data

Observed (KM) Plots by log WBC Categories



Option 1:
$h(t, \mathbf{X}) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2)$

where $X_1 = \begin{cases} 1 & \text{if high} \\ 0 & \text{if other} \end{cases}$ $X_2 = \begin{cases} 1 & \text{if medium} \\ 0 & \text{if other} \end{cases}$

so that
high = (1, 0); medium = (0, 1); low = (0, 0)
Expected survival plots:

$$X_1 = 1, X_2 = 0: \hat{S}(t, X_{\text{high}}) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_1\right)}$$

$$X_1 = 0, X_2 = 1: \hat{S}(t, X_{\text{medium}}) = \left[\hat{S}_0(t)\right]^{\exp\left(\hat{\beta}_2\right)}$$

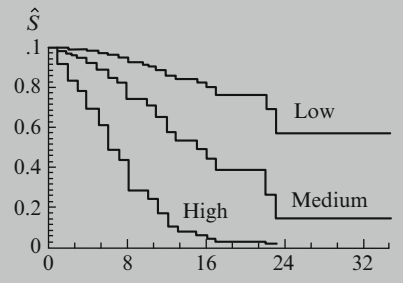
$$X_1 = 0, X_2 = 0: \hat{S}(t, X_{\text{low}}) = \left[\hat{S}_0(t)\right]$$

Using option 1, expected plots would be obtained by fitting a Cox PH model containing two dummy variables X_1 and X_2, as shown here, where X_1 takes the values 1 if high or 0 if other and X_2 takes the values 1 if medium or 0 if other. Thus, when log WBC is high, the values of X_1 and X_2 are 1 and 0, respectively; whereas when log WBC is medium, the values are 0 and 1, respectively; and when log WBC is low, the values are both 0.

The expected survival plots for high, medium, and low categories are then obtained by substituting each of the three specifications of X_1 and X_2 into the formula for the estimated survival curve, and then plotting the three curves.

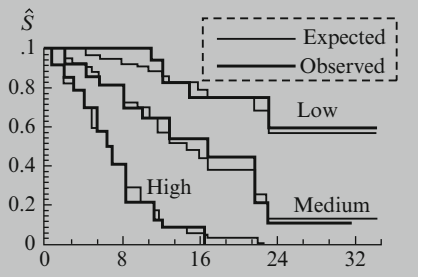
EXAMPLE: (continued)

Expected Plots for log WBC Using Option 1 (Dummy Variables)



The expected plots using option 1 (the dummy variable approach) are shown here for the three categories of log WBC.

Observed Versus Expected Plots Using Option 1



Here we put the observed and expected plots on the same graph. Although there are some discrepancies, particularly early in follow-up for the low log WBC category, these plots suggest overall that the PH assumption is satisfied for log WBC.

Option 2: Treat log WBC as continuous

$h(t, \mathbf{X}) = h_0(t)\exp[\beta(\log \text{WBC})]$

$\overline{\log \text{WBC}}_{\text{high}} = 3.83$:

$$\hat{S}(t, X_{\text{high}}) = [\hat{S}_0(t)]^{\exp[3.83\hat{\beta}]}$$

$\overline{\log \text{WBC}}_{\text{med}} = 2.64$:

$$\hat{S}(t, X_{\text{med}}) = [\hat{S}_0(t)]^{\exp[2.64\hat{\beta}]}$$

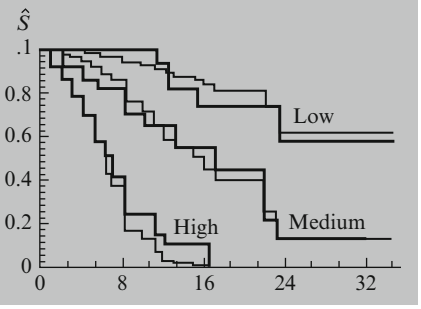
$\overline{\log \text{WBC}}_{\text{low}} = 1.71$:

$$\hat{S}(t, X_{\text{low}}) = [\hat{S}_0(t)]^{\exp[1.71\hat{\beta}]}$$

Using option 2, expected plots would be obtained by first fitting a Cox PH model containing the continuous variable log WBC, as shown here.

Adjusted survival curves are then obtained for specified values of log WBC that summarize the three categories used to form observed curves. Here, we find that the mean log WBC scores for low, medium, and high categories are, respectively, 1.71, 2.64, and 3.83. These values are substituted into the estimated survival curve formula as shown here.

Observed Versus Expected Plots for log WBC Using Option 2.



Here are the observed and expected plots using option 2. As with option 1, although there are some discrepancies within categories, overall, these plots suggest that the PH assumption is satisfied for the log WBC variable.

V. The Goodness of Fit (GOF) Testing Approach

Statistical test appealing

- Provides p-value
- More objective decision than when using graphical approach

The GOF testing approach is appealing because it provides a test statistic and p-value for assessing the PH assumption for a given predictor of interest. Thus, the researcher can make a more objective decision using a statistical test than is typically possible when using either of the two graphical approaches described above.

Test of Harrel and Lee (1986)

- Variation of test of Schoenfeld
- Uses Schoenfeld residuals

A number of different tests for assessing the PH assumption have been proposed in the literature. We present the test of Harrel and Lee (1986), a variation of a test originally proposed by Schoenfeld (1982) and based on the residuals defined by Schoenfeld, now called the Schoenfeld residuals.

Schoenfeld residuals defined for

- Each predictor in model
- Every subject who has event

Consider Cox PH model
$h(t,\mathbf{X}) = h_0(t)\exp(\beta_1 RX + \beta_2 \log WBC + \beta_3 SEX)$
3 predictors → 3 Schoenfeld residuals for each subject who has event

For each predictor in the model, Schoenfeld residuals are defined for every subject who has an event. For example, consider a Cox PH model with three predictors: RX, LOGWBC, and SEX. Then there are three Schoenfeld residuals defined for each subject who has an event, one for each of the three predictors.

Schoenfeld residual for ith subject for LOGWBC:
Observed LOGWBC
– LOGWBC weighted average

Weights are other subjects' hazard (from subjects still at risk)

Suppose subject i has an event at time t. Then her Schoenfeld residual for LOGWBC is her observed value of log white blood cell count minus a weighted average of the log white blood cell counts for the other subjects still at risk at time t. The weights are each subject's hazard.

Underlying idea of test
If PH holds then Schoenfeld residuals uncorrelated with time

The idea behind the statistical test is that **if the PH assumption holds for a particular covariate then the Schoenfeld residuals for that covariate will not be related to survival time**.

Steps for test implementation

1. Obtain Schoenfeld residuals
2. Rank failure times
3. Test correlation of residuals to ranked failure time H_0: $\rho = 0$

The implementation of the test can be thought of as a three-step process.

Step 1. Run a Cox PH model and obtain Schoenfeld residuals for each predictor.

Step 2. Create a variable that ranks the order of failures. The subject who has the first (earliest) event gets a value of 1, the next gets a value of 2, and so on.

Step 3. Test the correlation between the variables created in the first and second steps. The null hypothesis is that the correlation between the Schoenfeld residuals and ranked failure time is zero.

H_0 rejected
Conclude PH assumption violated

Rejection of the null hypothesis leads to a conclusion that the PH assumption is violated.

PH test in Stata, SAS, SPSS, R
shown in Computer Appendix

The implementation of the test for the PH assumption in Stata, SAS, SPSS, and R is shown in the Computer Appendix. Stata uses a slight variation of the test we just described in that it uses the scaled Schoenfeld residual rather than the Schoenfeld residual (Grambsch and Therneau, 1994). The tests typically (but not always) yield similar results.

Stata uses scaled Schoenfeld residuals rather than Schoenfeld residuals (typically similar results)

EXAMPLE: Remission Data

Column name.	Coeff.	StErr.	*P*(*PH*)
Rx	1.294	0.422	0.917
log WBC	1.604	0.329	0.944

Both variables satisfy PH assumption.

Note: *P*(*PH*) = 0.917 assesses PH for *Rx*, assuming PH OK for log WBC.

To illustrate the statistical test approach, we return to the remission data example. The printout on the left gives p-values P(PH) for treatment group and log WBC variables based on fitting a Cox PH model containing these two variables.

The P(PH) values are quite high for both variables, suggesting that both variables satisfy the PH assumption. Note that each of these p-values tests the assumption for one variable given that the other predictors are included in the model. For example, the P(PH) of 0.917 assesses the PH assumption for Rx, assuming the PH assumption is satisfied for log WBC.

EXAMPLE

Column name	Coeff.	StErr.	*P*(*PH*)
Rx	1.391	0.457	0.935
log WBC	1.594	0.330	0.828
Sex	0.263	0.449	0.038

log WBC and *Rx* satisfy PH.

Sex does not satisfy PH.

(Same conclusions using graphical approaches).

As another example, consider the computer results shown here for a Cox PH model containing the variable SEX in addition to log WBC and treatment group. The P(PH) values for log WBC and treatment group are still nonsignificant. However, the P(PH) value for SEX is significant below the 0.05 level. This result suggests that log WBC and treatment group satisfy the PH assumption, whereas SEX does not. We came to the same conclusion about these variables using the graphical procedures described earlier.

Statistical Tests

Null is never proven

- May say not enough evidence to reject

An important point concerning a testing approach is that the null hypothesis is never proven with a statistical test. The most that may be said is that there is not enough evidence to reject the null. A p-value can be driven by sample size. A gross violation of the null assumption may not be statistically significant if the sample is very small. Conversely, a slight violation of the null assumption may be highly significant if the sample is very large.

p-value can be driven by sample size

- Small sample – gross violation of null may not be significant
- Large sample – slight violation of null may be highly significant

Test – more objective
Graph – more subjective, but can detect specific violations
Recommend – Use both graphs and tests

A statistical test offers a more objective approach for assessing the PH assumption compared to the subjectivity of the graphical approach. However, the graphical approach enables the researcher to detect specific kinds of departures from the PH assumption; the researcher can see what is going on from the graph. Consequently, we recommend that when assessing the PH assumption, the investigator use both graphical procedures and statistical testing before making a final decision.

VI. Assessing the PH Assumption Using Time-Dependent Covariates

When time-dependent variables are used to assess the PH assumption for a time-independent variable, the Cox model is extended to contain product (i.e., interaction) terms involving the time-independent variable being assessed and some function of time.

Extended Cox model:
contains product terms of the form $X \times g(t)$, where $g(t)$ is a function of time.

One-at-a-time model:

$h(t,\mathbf{X}) = h_0(t)\exp[\beta X + \delta(X \times g(t))]$

When assessing predictors one-at-a-time, the extended Cox model takes the general form shown here for the predictor X.

Some choices for $g(t)$:

$g(t) = t$

$g(t) = \log t$

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases} \quad \text{(heaviside function)}$$

One choice for the function $g(t)$ is simply $g(t)$ equal to t, so that the product term takes the form $X \times t$. Other choices for $g(t)$ are also possible, for example, $\log t$, or the "heaviside function" shown at the left.

H_0: $\delta = 0$

Under H_0, the model reduces to:

$h(t,\mathbf{X}) = h_0(t)\exp[\beta X]$

Using the above one-at-a-time model, we assess the PH assumption by testing for the significance of the product term. The null hypothesis is therefore "δ equal to zero." Note that if the null hypothesis is true, the model reduces to a Cox PH model containing the single variable X.

Use either Wald statistic or likelihood ratio statistic:

χ^2 with 1 df under H_0

The test can be carried out using either a Wald statistic or a likelihood ratio statistic. In either case, the test statistic has a chi-square distribution with one degree of freedom under the null hypothesis.

Example

$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 \text{Sex} + \beta_2 (\text{Sex} \times t)]$

$\beta_2 \neq 0$ = PH assumption violated

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable Sex $\times$ t in addition to Sex. If the coefficient of the product term turns out to be significant, we can conclude that the PH assumption is violated for Sex.[2]

Strategies for assessing PH:

- one-at-a-time
- several predictors simultaneously
- for a given predictor adjusted for other predictors

In addition to a one-at-a-time strategy, the extended Cox model can also be used to assess the PH assumption for several predictors simultaneously as well as for a given predictor adjusted for other predictors in the model.

[2] In contrast, if the test for H_0: $\beta_2 = 0$ is nonsignificant, we can conclude only that the particular version of the extended Cox model being considered is not supported by the data.

Several predictors simultaneously:

$$h(t,\mathbf{X}) = h_0(t)\ \exp\left(\sum_{i=1}^{p}[\,\beta_i X_i + \delta_i(X_i \times g_i(t))]\right)$$

$g_i(t)$ = function of time for ith predictor

H_0: $\delta_1 = \delta_2 = \ldots = \delta_p = 0$

$$LR = -2\ \ln L_{\text{PH model}} - (-2\ \ln L_{\text{ext. Cox model}}) \dot{\sim} \chi^2_{\text{p}} \text{ under } H_0$$

Cox PH (reduced) model:

$$h(t,\mathbf{X}) = h_0(t)\ \exp\left(\sum_{i=1}^{p}\ \beta_i X_i\right)$$

To assess the PH assumption for several predictors simultaneously, the form of the extended model is shown here. This model contains the predictors being assessed as main effect terms and also as product terms with some function of time. Note that different predictors may require different functions of time; hence, the notation $g_i(t)$ is used to define the time function for the ith predictor.

With the above model, we test for the PH assumption simultaneously by assessing the null hypothesis that all the δ_i coefficients are equal to zero. This requires a likelihood ratio chi-square statistic with p degrees of freedom, where p denotes the number of predictors being assessed. The LR statistic computes the difference between the log likelihood statistic (i.e., $-2\ \ln L$) for the PH model and the log likelihood statistic for the extended Cox model. Note that under the null hypothesis, the model reduces to the Cox PH model shown here.

EXAMPLE: Remission Data

$$h(t,X) = h_0(t)\exp[\,\beta_1(Rx) + \beta_2(\log\ \text{WBC}) + \beta_3(\text{Sex}) + \delta_1(Rx \times g(t)) + \delta_2(\log\ \text{WBC} \times \text{g}(t)) + \delta_3(\text{Sex} \times g(t))]$$

where $g(t) = \begin{cases} 1 & \text{if t} \geq 7 \\ 0 & \text{if t} < 7 \end{cases}$

H_0: $\delta_1 = \delta_2 = \delta_3 = 0$

$LR \dot{\sim} \chi^2$ with 3 df under H_0

If test is significant, use backward elimination to find predictors not satisfying PH assumption.

As an example, we assess the PH assumption for the predictors *Rx*, log WBC, and Sex from the remission data considered previously. The extended Cox model is given as shown here, where the functions $g_i(t)$ have been chosen to be the same "heaviside" function defined by $g(t)$ equals 1 if t is 7 weeks or more and $g(t)$ equals 0 if t is less than 7 weeks. The null hypothesis is that all three δ coefficients are equals to zero. The test statistic is a likelihood-ratio chi-square with 3 degrees of freedom.

If the above test is found to be significant, then we can conclude that the PH assumption is not satisfied for at least one of the predictors in the model. To determine which predictor(s) do not satisfy the PH assumption, we could proceed by backward elimination of nonsignificant product terms until a final model is attained.

Heaviside function:

$$g(t) = \begin{cases} 1 & \text{if t} \geq 7 \\ 0 & \text{if t} < 7 \end{cases}$$

$h(t,\mathbf{X})$ differs for t $\geq$ 7 and t $<$ 7.

Properties of heaviside functions and numerical results are described in Chapter 6.

Note that the use of a heaviside function for $g(t)$ in the above example yields different expressions for the hazard function depending on whether t is greater than or equal to 7 weeks or t is less than 7 weeks. Chapter 6 provides further details on the properties of heaviside functions, and also provides numerical results from fitting extended Cox models.

Assessing PH for a given predictor adjusted for other predictors:

$$h(t,\mathbf{X}) = h_0(t)\exp\left[\sum_{i=1}^{p-1}\beta_i X_i + \beta^* X^* + \delta^*\left(X^* \times g(t)\right)\right]$$

X^* = Predictor of interest
H_0: $\delta^* = 0$
Wald or LR statistic $\dot{\sim}\chi^2$ with 1 df

We show here an extended Cox model that can be used to evaluate the PH assumption for a given predictor **adjusted for predictors already satisfying the PH assumption**. The predictor of interest is denoted as X^*, and the predictors considered to satisfy the PH assumption are denoted as X_i. The null hypothesis is that the coefficient δ^* of the product term $X^*g(t)$ is equal to zero. The test statistic can either be a Wald statistic or a likelihood ratio statistic, with either statistic having a chi-square distribution with 1 degree of freedom under the null hypothesis.

Example: Remission Data

for Sex, adjusted for *Rx* and log WBC:

$$h(t,\mathbf{X}) = \exp[\beta_1 Rx + \beta_2 \log\ \text{WBC} + \beta^* \text{Sex} + \delta^*(\text{Sex} \times \text{g(t)})]$$

As an example, suppose, again considering the remission data, we assess the PH assumption for the variable, Sex, adjusted for the variables *Rx* and log WBC, which we assume already satisfy the PH assumption. Then, the extended Cox model for this situation is shown here.

Two models for LR test of PH:

1. Cox PH model
2. extended Cox model

See Computer Appendix for SAS, Stata, SPSS, and R

To carry out the computations for any of the likelihood ratio tests described above, two different types of models, a PH model and an extended Cox model, need to be fit. See the Computer Appendix for details on how the extended Cox model is fit using SAS, Stata, SPSS, and R.

Drawback: choice of $g_i(t)$

Different choices may lead to different conclusions about PH assumption.

The primary drawback of the use of an extended Cox model for assessing the PH assumption concerns the choice of the functions $g_i(t)$ for the time-dependent product terms in the model. This choice is typically not clear-cut, and it is possible that different choices, such as $g(t)$ equal to t versus log t versus a heaviside function, may result in different conclusions about whether the PH assumption is satisfied.

Chapter 6: Time-dependent covariates

Further discussion of the use of time-dependent covariates in an extended Cox model is provided in Chapter 6.

This presentation:
Three methods for assessing PH.

i. graphical
ii. GOF
iii. time-dependent covariates

Recommend using at least two methods.

This presentation is now complete. We have described and illustrated three methods for assessing the PH assumption: graphical, goodness-of-fit (GOF), and time-dependent covariate methods. Each of these methods has both advantages and drawbacks. We recommend that the researcher use at least two of these approaches when assessing the PH assumption.

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Survival Curves and the Log–Rank Test
3. The Cox Proportional Hazards Model and Its Characteristics
✓4. Evaluating the Proportional Hazards Assumption

Next:

5. The Stratified Cox Procedure
6. Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

We suggest that the reader review this presentation using the detailed outline that follows. Then answer the practice exercises and the test that follow.

The next Chapter (5) is entitled "The Stratified Cox Procedure." There, we describe how to use a stratification procedure to fit a PH model when one or more of the predictors do not satisfy the PH assumption.

Detailed Outline

I. Background (pages 164–165)

A. The formula for the Cox PH model:

$$h(t, \mathbf{X}) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

B. Formula for hazard ratio comparing two individuals, $\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_p^*\right)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp\left[\sum_{i=1}^{p} \beta_i \left(X_i^* - X_i\right)\right]$$

C. Adjusted survival curves using the Cox PH model:

$$\mathcal{S}(t, \mathbf{X}) = [\mathcal{S}_0(t)]^{\exp[\sum \beta_1 X_1]}$$

i. To graph $\mathcal{S}(t, \mathbf{X})$, must specify values for $\mathbf{X} = (X_1, X_2, \ldots, X_p)$.

ii. To obtain "adjusted" survival curves, usually use overall mean values for the X's being adjusted.

D. The meaning of the PH assumption

i. Hazard ratio formula shows that hazard ratio is independent of time:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}$$

ii. Hazard ratio for two X's are proportional:

$$\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X})$$

II. Checking the proportional hazards assumption: overview (pages 165–167)

A. Three methods for checking the PH assumption:

i. Graphical: compare −ln −ln survival curves or observed versus predicted curves.

ii. Goodness-of-fit test: use a large sample Z statistic.

iii. Time-dependent covariates: use product (i.e., interaction) terms of the form $X \times g(t)$.

B. Abbreviated illustrations of each method are provided.

III. Graphical approach 1: log–log plots (pages 167–175)

A. A log–log curve is a transformation of an estimated survival curve, where the scale for a log–log curve is $-\infty$ to $+\infty$.

B. The log–log expression for the Cox model survival curve is given by

$$\ln[-\ln \mathcal{S}(t, \mathbf{X})] = \sum_{i=1}^{p} \beta_i X_i + \ln[-\ln S_0(t)]$$

C. For the Cox model, the log–log survival curve for individual $\mathbf{X}_1$ can be written as the log–log curve for individual $\mathbf{X}_2$ plus a linear sum term that is independent of time t. This formula is given by

$$\ln[-\ln \mathcal{S}(t, \mathbf{X}_1)]$$

$$= \ln[-\ln \mathcal{S}(t, \mathbf{X}_2)] + \sum_{i=1}^{p} \beta_i (X_{1i} - X_{2i})$$

D. The above log–log formula can be used to check the PH assumption as follows: the PH model is appropriate if "empirical" plots of log–log survival curves are parallel.

E. Two kinds of empirical plots for $-\ln -\ln \hat{\mathcal{S}}$:
   i. $\hat{\mathcal{S}}$ is a KM curve
   ii. $\hat{\mathcal{S}}$ is an adjusted survival curve where predictor being assessed is not included in the Cox regression model.

F. Several examples of log–log plots are provided using remission data from a clinical trial of leukemia patients.

G. Problems with log–log curves:
   i. How parallel is parallel?
   ii. How to categorize a continuous variable?
   iii. How to evaluate several variables simultaneously?

H. Recommendation about problems:
   i. Use small number of categories, meaningful choice, reasonable balance.
   ii. With several variables, two options:
      a. Compare log–log curves from combinations of categories.
      b. Adjust for predictors already satisfying PH assumption.

IV. Graphical approach 2: observed versus expected plots (pages 175–180)

A. Graphical analog of the GOF test.

B. Two strategies
   i. One-at-a-time: uses KM curves to obtain observed plots.
   ii. Adjusting for other variables: uses stratified Cox PH model to obtain observed plots.

C. Expected plots obtained by fitting a Cox model containing the predictor being assessed; substitute into the fitted model the value for each category of the predictor to obtain the expected value for each category.
D. If observed and expected plots are close, conclude PH assumption is reasonable.
E. Drawback: how close is close?
F. Recommend: conclude PH not satisfied *only* if plots are strongly discrepant.
G. Another drawback: what to do if assessing continuous variable.
H. Recommend for continuous variable:
   i. Form strata from categories.
   ii. Observed plots are KM curves for each category.
   iii. Two options for expected plots:
      a. Use PH model with $k - 1$ dummy variables for k categories.
      b. Use PH model with continuous predictor and specify predictor values that distinguish categories.

V. The goodness-of-fit (GOF) testing approach (pages 181–183)
A. Appealing approach because
   i. provides a test statistic (p-value).
   ii. researcher can make clear-cut decision.
B. References
   i. methodological: Schoenfeld (1982), Harrel and Lee (1986).
   ii. SAS and Stata use different GOF formulae.
C. The method:
   i. Schoenfeld residuals for each predictor uses a chi-square statistic with 1 df.
   ii. Correlations between Schoenfeld's residuals and ranked failure times.
   iii. If p-value small, then departure from PH.
D. Examples using remission data.
E. Drawbacks:
   i. global test: may fail to detect a specific kind of departure from PH; recommend using both graphical and GOF methods.
   ii. several strategies to choose from, with no one strategy clearly preferable (one-at-a-time, all variables, each variable adjusted for others).

VI. Assessing the PH assumption using time-dependent covariates (pages 183–187)

A. Use extended Cox model: contains product terms of form $X \times g(t)$, where $g(t)$ is function of time, e.g., $g(t) = t$, or $\log t$, or heaviside function.

B. One-at-a-time model:
$h(t, \mathbf{X}) = h_0(t) \exp[\beta X + \delta X g(t)]$
Test H_0: $\delta = 0$ using Wald or LR test (chi- square with 1 df).

C. Evaluating several predictors simultaneously:

$$h(t, \mathbf{X}) = h_0(t) \exp\left(\sum_{i=1}^{p} [\beta_i X_i + \delta_i X_i g_i(t)]\right)$$

where $g_i(t)$ is function of time for ith predictor.
Test H_0: $\delta_1 = \delta_2 = \cdots = \delta_p = 0$ using *LR* (chi-square) test with p df.

D. Examples using remission data.

E. Two computer programs, required for test:
   i. Cox PH model program.
   ii. Extended Cox model program.

F. Drawback: choice of $g(t)$ not always clear; different choices may lead to different conclusions about PH assumption.

Practice Exercises

The dataset "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (*The Statistical Analysis of Survival Time Data*, Wiley, 2002). The exposure variable of interest is treatment status (standard = 1, test = 2). Other variables of interest as control variables are cell type (four types, defined by dummy variables), performance status, disease duration, age, and prior therapy status. Failure status is defined by the status variable (0 if censored, 1 if died). A complete list of the variables is given below.

Column 1: Treatment (standard = 1, test = 2)
Column 2: Cell type 1 (large = 1, other = 0)
Column 3: Cell type 2 (adeno = 1, other = 0)
Column 4: Cell type 3 (small = 1, other = 0)
Column 5: Cell type 4 (squamous = 1, other = 0)
Column 6: Survival time (days)
Column 7: Performance status (0 = worst,. . ., 100 = best)

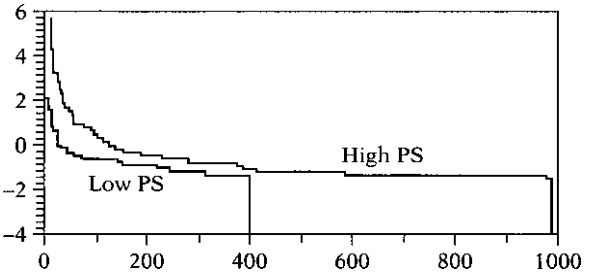
Column 8: Disease duration (months)
Column 9: Age
Column 10: Prior therapy (none = 0, some = 10)
Column 11: Status (0 = censored, 1 = died)

1. State the hazard function form of the Cox PH model that describes the effect of the treatment variable and controls for the variables, cell type, performance status, disease duration, age, and prior therapy. In stating this model, make sure to incorporate the cell type variable using dummy variables, but do not consider possible interaction variables in your model.
2. State three general approaches that can be used to evaluate whether the PH assumption is satisfied for the variables included in the model you have given in question 1.
3. The following printout is obtained from fitting a Cox PH model to these data. Using the information provided, what can you conclude about whether the PH assumption is satisfied for the variables used in the model? Explain briefly.

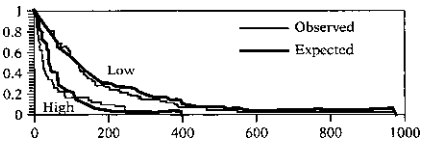
Cox regression	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]		*P*(*PH*)
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Performance status	−0.033	0.006	0.000	0.968	0.958	0.978	0.000
Disease duration	0.000	0.009	0.992	1.000	0.982	1.018	0.919
Age	−0.009	0.009	0.358	0.991	0.974	1.010	0.198
Prior therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.145

4. For the variables used in the PH model in question 1, describe a strategy for evaluating the PH assumption using log–log survival curves for variables considered one-at-a-time.
5. Again considering the variables used in question 1, describe a strategy for evaluating the PH assumption using log–log survival curves that are adjusted for other variables in the model.

6. For the variable "performance status," describe how you would evaluate the PH assumption using observed versus expected survival plots?
7. For the variable "performance status," log–log plots which compare high ($\geq$50) with low ($<$50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



8. What are some of the drawbacks of using the log–log approach for assessing the PH assumption and what do you recommend to deal with these drawbacks?
9. For the variable "performance status," observed versus expected plots that compare high ($\geq$50) with low ($<$50) are given by the following graph. Based on this graph, what do you conclude about the PH assumption with regard to this variable?



10. State the form of an extended Cox model that allows for the one-at-a-time assessment of the PH assumption for the variable "performance status," and describe how you would carry out a statistical test of the assumption for this variable.
11. State the form of an extended Cox model that allows for the simultaneous assessment of the PH assumption for the variables treatment, cell type, performance status, disease duration, age, and prior therapy. For this model, describe how you would carry out a statistical test of the PH assumption for these variables. Also, provide a strategy for assessing which of these variables satisfy the PH assumption and which do not using the extended Cox model approach.

12. Using any of the information provided above and any additional analyses that you perform with this dataset, what do you conclude about which variables satisfy the PH assumption and which variables do not? In answering this question, summarize any additional analyses performed.

Test

The following questions consider a dataset from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," *Med. J. Aust.*, 1991). These data comprise the times in days spent by heroin addicts from entry to departure from one of two methadone clinics. There are two additional covariates, namely, prison record and maximum methadone dose, believed to affect the survival times. The dataset name is **addicts.dat**. A listing of the variables is given below:

Column 1: Subject ID
Column 2: Clinic (1 or 2)
Column 3: Survival status (0 = censored, 1 = departed from clinic)
Column 4: Survival time in days
Column 5: Prison record (0 = none, 1 = any)
Column 6: Maximum methadone dose (mg/day)

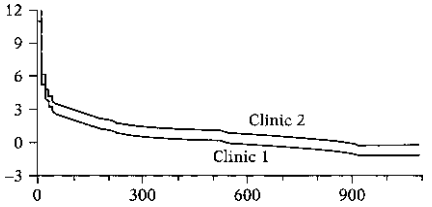
1. The following edited printout was obtained from fitting a Cox PH model to these data:

Cox regression Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]		*P*(*PH*)
Clinic	−1.009	0.215	0.000	0.365	0.239	0.556	0.001
Prison	0.327	0.167	0.051	1.386	0.999	1.924	0.332
Dose	−0.035	0.006	0.000	0.965	0.953	0.977	0.347

No. of subjects: 238 Log likelihood = −673.403

Based on the information provided in this printout, what do you conclude about which variables satisfy the PH assumption and which do not? Explain briefly.

2. Suppose that for the model fit in question 1, log–log survival curves for each clinic adjusted for prison and dose are plotted on the same graph. Assume that these curves are obtained by substituting into the formula for the estimated survival curve the values for each clinic and the overall mean values for the prison and dose variables. Below, we show these two curves. Are they parallel? Explain your answer.

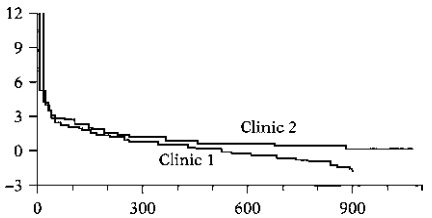


3. The following printout was obtained from fitting a stratified Cox PH model to these data, where the variable being stratified is clinic:

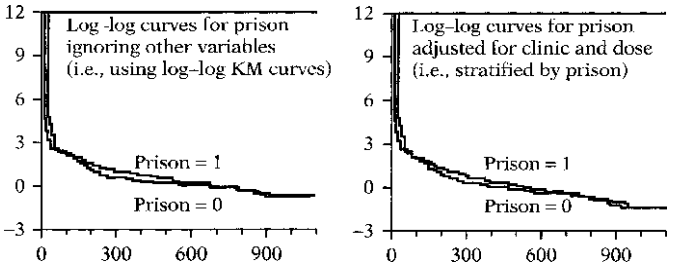
Stratified Cox regression Analysis time.t: survt (in days)	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Prison	0.389	0.169	0.021	1.475	1.059	2.054
Dose	−0.035	0.006	0.000	0.965	0.953	0.978

No. of subjects = 238 Log likelihood = −597.714 Stratified by clinic

Using the above fitted model, we can obtain the log–log curves below that compare the log–log survival for each clinic (i.e., stratified by clinic) adjusted for the variables prison and dose. Using these curves, what do you conclude about whether or not the clinic variable satisfies the PH assumption? Explain briefly.



4. Consider the two plots of log–log curves below that compare the log–log survival for the prison variable ignoring other variables and adjusted for the clinic and dose variables. Using these curves, what do you conclude about whether or not the prison variable satisfies the PH assumption? Explain briefly.



5. How do your conclusions from question 1 compare with your conclusions from question 4? If the conclusions differ, which conclusion do you prefer? Explain.
6. Describe briefly how you would evaluate the PH assumption for the variable maximum methadone dose using observed versus expected plots.
7. State an extended Cox model that would allow you to assess the PH assumption for the variables clinic, prison, and dose simultaneously. For this model, state the null hypothesis for the test of the PH assumption and describe how the likelihood ratio statistic would be obtained and what its degrees of freedom would be under the null hypothesis.
8. State at least one drawback to the use of the extended Cox model approach described in question 7.
9. State an extended Cox model that would allow you to assess the PH assumption for the variable clinic alone, assuming that the prison and dose variables already satisfy the PH assumption. For this model, state the null hypothesis for the test of the PH assumption, and describe how the likelihood ratio (LR) statistic would be obtained. What is the degrees of freedom of the LR test under the null hypothesis?

10. Consider the situation described in question 9, where you wish to use an extended Cox model that would allow you to assess the PH assumption for the variable clinic alone, assuming that the assumption is satisfied for the prison and dose variables. Suppose you use the following extended Cox model:

$$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(\text{prison}) + \beta_2(\text{dose}) + \beta_3(\text{clinic}) + \delta_1(\text{clinic})g(t)]$$

where $g(t)$ is defined as follows:

$$g(t) = \begin{cases} 1 & \text{if } t > 365 \text{ days} \\ 0 & \text{if } t \leq 365 \text{ days} \end{cases}$$

For the above model, what is the formula for the hazard ratio that compares clinic 1 to clinic 2 when t is greater than 365 days? when t is less than or equal to 365 days? In terms of the hazard ratio formulae just described, what specific departure from the PH assumption is being tested when the null hypothesis is H_0: $\delta_1 = 0$?

Answers to Practice Exercises

1. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(\text{treatment}) + \beta_2(\text{CT1}) + \beta_3(\text{CT2}) + \beta_4(\text{CT3}) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT})]$

   where *CTi* denotes the cell type *i* dummy variable, PS denotes the performance status variable DD denotes the disease duration variable, and PT denotes the prior therapy variable.

2. The three general approaches for assessing the PH model for the above model are:
   (a) graphical, using either log–log plots or observed versus expected plots;
   (b) statistical test;
   (c) an extended Cox model containing product terms involving the variables being assessed with some function(s) of time.

3. The *P*(*PH*) values given in the printout provide goodness-of-fit tests for each variable in the fitted model adjusted for the other variables in the model. The *P*(*PH*) values shown indicate that the large cell type variables and the performance status variable do not satisfy the PH assumption, whereas the treatment, age, disease duration, and prior therapy variables satisfy the PH assumption, and the adeno and small cell type variable are of borderline significance.

4. A strategy for evaluating the PH assumption using log–log survival curves for variables considered one-at-a-time is given as follows:

   For each variable separately, obtain a plot of log–log Kaplan–Meier curves for the different categories of that variable. For the cell type variable, this requires obtaining a plot of four log–log KM curves, one for each cell type. (Note that this is not the same as obtaining four separate plots of two log–log curves, where each plot corresponds to one of the dummy variables used in the model.) For the variables PS, DD, and Age, which are interval variables, each variable must be separately categorized into two or more groups – say, low versus high values – and KM curves are obtained for each group. For the variable PT, which is a dichotomous variable, two log–log curves are obtained which compare the "none" versus "some" groups.

   For each plot (i.e., one for each variable), those plots that are noticeably nonparallel indicate variables which do not satisfy the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

5. One strategy for evaluating the PH assumption for each variable adjusted for the others is to use adjusted log–log survival curves instead of KM curves separately for each of the variables in the model. That is, for each variable separately, a stratified Cox model is fit stratifying on the given variable while adjusting for the other variables. Those variables that yield adjusted log–log plots that are noticeably nonparallel are then to be considered as not satisfying the PH assumption. The remaining variables are assumed to satisfy the PH assumption.

   A variation of the above strategy uses adjusted log–log curves for only those variables *not satisfying* the PH assumption from a one-at-a-time approach, adjusting for those variables *satisfying* the PH assumption from the one-at-a-time approach. This second iteration would flag a subset of the one-at-a-time flagged variables for further iteration. At each new iteration, those variables found to satisfy the assumption get added to the list of variables previously determined to satisfy the assumption.

6. For the performance status (PS) variable, **observed plots** are obtained by categorizing the variable into strata (say, two strata: low versus high) and then obtaining KM survival plots for each stratum. **Expected plots** can be obtained by fitting a Cox

model containing the (continuous) PS variable and then obtaining estimated survival curves for values of the performance status (PS) variable that represent summary descriptive statistics for the strata previously identified. For example, if there are two strata, say, high (PS $>$ 50) and low (PS $\leq$ 50), then the values of PS to be used could be the mean or median PS score for persons in the high stratum and the mean or median PS score for persons in the low stratum.

An alternative method for obtaining expected plots involves first dichotomizing the PS variable – say, into high and low groups – and then fitting a Cox model containing the dichotomized PS variable instead of the original continuous variable. The expected survival plots for each group are estimated survival curves obtained for each value of the dichotomized PS variable.

Once observed and expected plots are obtained for each stratum of the PS variable, they are then compared on the same graph to determine whether or not corresponding observed and expected plots are "close." If it is determined that, overall, comparisons for each stratum are close, then it is concluded that the PH assumption is satisfied for the PH variable. In determining how close is close, the researcher should look for noticeably discrepant observed versus expected plots.

7. The log–log plots that compare high versus low PS groups (ignoring other variables) are *arguably* parallel early in follow-up, and are not comparable later because survival times for the two groups do not overlap after 400 days. These plots do not strongly indicate that the PH assumption is violated for the variable PS. This contradicts the conclusion previously obtained for the PS variable using the *P*(*PH*) results.

8. Drawbacks of the log–log approach are:

   - How parallel is parallel?
   - How to categorize a continuous variable?
   - How to evaluate several variables simultaneously?

   Recommendations about problems:

   - Look for noticeable nonparallelism; otherwise PH assumption is OK.
   - For continuous variables, use a small number of categories, a meaningful choice of categories, and a reasonable balance in sample size for categories.

- With several variables, there are two options:
  i. Compare log–log curves from combinations of categories.
  ii. Adjust for predictors already satisfying PH assumption.

9. The observed and expected plots are relatively close for low and high groups separately, although there is somewhat more discrepancy for the high group than for the low group. Deciding how close is close is quite subjective for these plots. Nevertheless, because there are no major discrepancies for either low or high groups, we consider the PH assumption satisfied for this variable.

10. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(\text{PS}) + \delta(\text{PS})g(t)]$

    where $g(t)$ is a function of t, such as $g(t) = t$, or $g(t) = \log t$, or a heaviside function. The PH assumption is tested using a 1 df Wald or LR statistic for H_0: $\delta = 0$.

11. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(\text{treatment}) + \beta_2(\text{CT1}) + \beta_3(\text{CT2}) + \beta_4(\text{CT3}) + \beta_5(\text{PS}) + \beta_6(\text{DD}) + \beta_7(\text{Age}) + \beta_8(\text{PT}) + \delta_1(\text{treatment} \times g(t)) + \delta_2(\text{CT1} \times g(t)) + \delta_3(\text{CT2} \times g(t)) + \delta_4(\text{CT3} \times g(t)) + \delta_5(\text{PS} \times g(t)) + \delta_6(\text{DD} \times g(t)) + \delta_7(\text{Age} \times g(t)) + \delta_8(\text{PT} \times g(t))]$

    where $g(t)$ is some function of time, such as $g(t) = t$, or $g(t) = \log t$, or a heaviside function. To test the PH assumption simultaneously for all variables, the null hypothesis is stated as H_0: $\delta_1 = \delta_2 = \ldots = \delta_8 = 0$. The test statistic is a likelihood-ratio statistic of the form

    $$LR = -2 \ln LR - (-2 \ln L_F)$$

    where R denotes the reduced (PH) model obtained when all δ's are 0, and F denotes the full model given above. Under H_0, the LR statistic is approximately chi-square with 8 df.

12. The question here is somewhat open-ended, leaving the reader the option to explore additional graphical, GOF, or extended Cox model approaches for evaluating the PH assumption for the variables in the model. The conclusions from the GOF statistics provided in question 3 are likely to hold up under further scrutiny, so that a reasonable conclusion is that cell type and performance status variables do not satisfy the PH assumption, with the remaining variables satisfying the assumption.

5 The Stratified Cox Procedure

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_5,
© Springer Science+Business Media, LLC 2012

Introduction

We begin with an example of the use of the stratified Cox procedure for a single predictor that does not satisfy the PH assumption. We then describe the general approach for fitting a stratified Cox model, including the form of the (partial) likelihood function used to estimate model parameters.

We also describe the assumption of no interaction that is typically incorporated into most computer programs that carry out the stratified Cox procedure. We show how the no-interaction assumption can be tested, and what can be done if interaction is found.

We conclude with a second example of the stratified Cox procedure in which more than one variable is stratified.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Preview** (page 204)
II. **An Example** (pages 204–208)
III. **The General Stratified Cox (SC) Model** (pages 208–209)
IV. **The No-Interaction Assumption and How to Test It** (pages 210–216)
V. **A Second Example Involving Several Stratification Variables** (pages 216–221)
VI. **A Graphical View of the Stratified Cox Approach** (pages 221–222)
VII. **The Stratified Cox Likelihood** (pages 223–225)
VIII. **Summary** (pages 225–227)

Objectives

Upon completing the chapter, the learner should be able to:

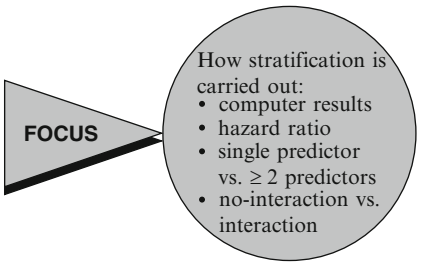
1. Recognize a computer printout for a stratified Cox procedure.
2. State the hazard form of a stratified Cox model for a given survival analysis scenario and/or a given set of computer results for such a model.
3. Evaluate the effect of a predictor of interest based on computer results from a stratified Cox procedure.
4. For a given survival analysis scenario and/or a given set of computer results involving a stratified Cox model.
   - state the no-interaction assumption for the given model;
   - describe and/or carry out a test of the no-interaction assumption;
   - describe and/or carry out an analysis when the no-interaction assumption is not satisfied.

Presentation

I. Preview

Stratified Cox model:

- modification of Cox PH model
- Stratification of predictor not satisfying PH
- includes predictors satisfying PH



The "stratified Cox model" is a modification of the Cox proportional hazards (PH) model that allows for control by "stratification" of a predictor that does not satisfy the PH assumption. Predictors that are assumed to satisfy the PH assumption are included in the model, whereas the predictor being stratified is not included.

In this presentation, we focus on how stratification is carried out by describing the analysis of computer results and the form of the hazard function for a stratified Cox model. We first consider stratifying on a single predictor and then later consider stratifying on two or more predictors. Further, we distinguish between the use of a "no-interaction" version of the stratified Cox model and an alternative approach that allows interaction.

II. An Example

EXAMPLE

Clinical trial: 42 leukemia patients
Response-days in remission

	Coef.	Std. Err.	*P*(*PH*)
log WBC	1.594	0.330	0.828
Rx	1.391	0.457	0.935
Sex	0.263	0.449	0.031

- log WBC and *Rx* satisfy PH
- Sex does not satisfy PH

(Same conclusions using graphical approaches)

Stratified Cox (SC):

- control for sex (stratified);
- simultaneously include log WBC and *Rx* in the model

Consider the computer results shown here for a Cox PH model containing the three variables, log WBC, treatment group (*Rx*), and SEX. These results derive from a clinical trial of 42 leukemia patients, where the response of interest is days in remission.

From the printout, the *P*(*PH*) values for log WBC and treatment group are nonsignificant. However, the *P*(*PH*) value for SEX is significant below the .05 level. These results indicate that log WBC and treatment group satisfy the PH assumption, whereas the SEX variable does not. The same conclusions regarding the PH assumption about these variables would also be made using the graphical procedures described earlier.

Because we have a situation where one of the predictors does not satisfy the PH assumption, we carry out a stratified Cox (SC) procedure for the analysis. Using SC, we can control for the SEX variable – which does not satisfy the PH assumption – by stratification while simultaneously including in the model the log WBC and treatment variables – which do satisfy the PH assumption.

EXAMPLE: (continued)

STATA OUTPUT USING SC:
Stratified Cox regression
Analysis time_t: survt

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.390	0.338	0.000	4.016	2.072	7.783
Rx	0.931	0.472	0.048	2.537	1.006	6.396

No. of subjects = 42 Log likelihood = −57.560 Stratified by sex

The computer results from a SC procedure are shown here. These results come from the Stata package. (See the Computer Appendix for running a SC procedure in Stata, SAS, SPSS, and R).

Appendix A illustrates SC procedures using Stata, SAS, SPSS, and R

- Log WBC and *Rx* are included in SC model.
- SC model is stratified by SEX.

The computer results show that the log WBC and *Rx* variables are included in the model listing, whereas the SEX variable is not included; rather, the model stratifies on the SEX variable, as indicated at the bottom of the output. Note that the SEX variable is being adjusted by stratification, whereas log WBC is being adjusted by its inclusion in the model along with *Rx*.

Effect of *Rx* adjusted for log WBC and SEX.

- Hazard ratio: $2.537 = e^{0.931}$
- Interpretation: Placebo group ($Rx = 1$) has 2.5 times the hazard as the treatment group (Rx = 0)

In the above output, we have also circled some key information that can be used to assess the effect of the *Rx* variable adjusted for both log WBC and SEX. In particular, we can see that the hazard ratio for the effect of *Rx* adjusted for log WBC and SEX is given by the value 2.537. This value can be obtained by exponentiating the coefficient 0.931 of the *Rx* variable. The hazard ratio value can be interpreted to mean that the placebo group (for which $Rx = 1$) has 2.5 times the hazard for going out of remission as the treatment group (for which $Rx = 0$).

Stratified Cox regression
Analysis time_t: survt

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.390	0.338	0.000	4.016	2.072	7.783
Rx	0.931	0.472	0.048	2.537	1.006	6.396

No. of subjects = 42 Log likelihood = −57.560 Stratified by sex

95% CI for *Rx* (1.006, 6.396) indicates considerable variability.
CI formula: $\exp(0.931 \pm 1.96 \times 0.472)$

Also, we can see from the output that a 95% confidence interval for the effect of the *Rx* variable is given by the limits 1.006 to 6.396. This is a fairly wide range, thus indicating considerable variability in the 2.537 hazard ratio point estimate. Note that these confidence limits can be obtained by exponentiating the quantity 0.931 plus or minus 1.96 times the standard error 0.472.

Wald test: P = 0.048 (two-tailed), significant at the 0.05 level.

From the above output, a test for the significance of the *Rx* variable adjusted for log WBC and SEX is given by the Wald statistic P-value of 0.048. This is a two-tailed P-value, and the test is (barely) significant at the 0.05 level.

EXAMPLE: (continued)

LR test: Output for reduced model
Stratified Cox regression
Analysis time _t: survt

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.456	0.320	0.000	4.289	2.291	8.03

No. of subjects = 42 Log likelihood = −59.648 Stratified by sex

$$\text{LR} = (-2 \times -59.648) - (-2 \times -57.560)$$
$$= 119.296 - 115.120 = 4.179\ (\text{P} < 0.05)$$

LR and Wald give same conclusion.

Hazard function for stratified Cox model:

$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$
$g = 1,2;$

g denotes stratum #.

SC model for males and females:
Females ($g = 1$):

$h_1(t,\mathbf{X}) = h_{01}(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$

Males ($g = 2$):

$h_2(t,\mathbf{X}) = h_{02}(t)\exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$

Rx and log WBC in the model
Sex *not* in the model (stratified)

$\widehat{HR}$ for effect of *Rx* adjusted for log WBC and sex:

$e^{\hat{\beta}_1}$

where β_1 is the coefficient of *Rx*.

An alternative test involves a likelihood ratio (*LR*) statistic that compares the above model (full model) with a reduced model that does not contain the *Rx* variable. The output for the reduced model is shown here. The log-likelihood statistic for the reduced model is −2 times −59.648, which is to be compared with the log-likelihood statistic of −2 times −57.560 for the full model.

The *LR* statistic is therefore 119.296 minus 115.120, which equals 4.179. Under H_0, this statistic has a chi-square distribution with one degree of freedom and is significant at the 0.05 level. Thus, the *LR* and Wald tests lead to the same conclusion.

So far, we have illustrated the results from a stratified Cox procedure without actually describing the model form being used. For the remission data example, we now present the hazard function form for the stratified Cox model, as shown here. This hazard function formula contains a subscript g that indicates the gth stratum.

Thus, in our remission data example, where we have stratified on SEX, g takes on one of two values, so that we have a different baseline hazard function for males and females.

Notice that the hazard function formula contains the variables *Rx* and log WBC, but does not contain the variable SEX. SEX is not included in the model because it doesn't satisfy the PH assumption. So, instead, the SEX variable is controlled by stratification.

Because the variables *Rx* and log WBC are included in the model, we can estimate the effect of each variable adjusted for the other variable and the SEX variable using standard exponential hazard ratio expressions. For example, the estimated hazard ratio for the effect of *Rx*, adjusted for log WBC and SEX, is given by e to the β_1 "hat," where β_1 is the coefficient of the *Rx* variable.

EXAMPLE: (continued)

Cannot estimate *HR* for SEX variable (SEX doesn't satisfy PH).

Nevertheless, because the SEX variable is not included in the model, it is not possible to obtain a hazard ratio value for the effect of SEX adjusted for the other two variables. This is the price to be paid for stratification on the SEX variable. Note that a single value for the hazard ratio for SEX is not appropriate if SEX doesn't satisfy the PH assumption, because the hazard ratio must then vary with time.

Different baseline hazard functions:

$h_{01}(t)$ for females and $h_{02}(t)$ for males.

Same coefficients β_1 and β_2 for both female and male models.

Notice also that the hazard functions for males and females differ only insofar as they have different baseline hazard functions, namely, $h_{01}(t)$ for females and $h_{02}(t)$ for males. However, the coefficients β_1 and β_2 are the same for both female and male models.

Different baselines $\begin{cases} h_{01}(t) \Rightarrow \text{Survival curve for females} \\ h_{02}(t) \Rightarrow \text{Survival curve for males} \end{cases}$

Because there are different baseline hazard functions, the fitted stratified Cox model will yield different estimated survival curves for females and males. These curves will be described shortly.

Females and males:
same β_1 and $\beta_2 \Rightarrow$ same $\widehat{HR}$'s, e.g., $e^{\hat{\beta}_1}$

No interaction assumption (see Section IV)

Note, however, that because the coefficients of *Rx* and log WBC are the same for females and males, estimates of hazard ratios, such as e to the β_1 "hat," are the same for both females and males. This feature of the stratified Cox model is called the "no-interaction" assumption. It is possible to evaluate whether this assumption is tenable and to modify the analysis if not tenable. We will discuss this assumption further in Section IV.

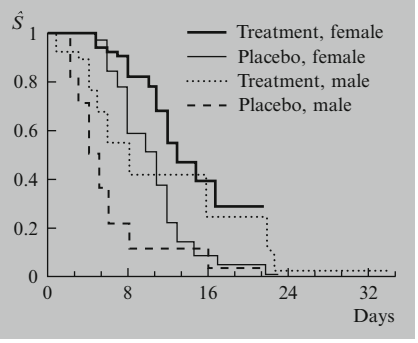
Estimates of β_1 and β_2

Maximize partial likelihood (L), where $L = L_1 \times L_2$
L_1 is the likelihood for females derived from $h_1(t)$,
and L_2 is the likelihood for males derived from $h_2(t)$.

To obtain estimates of β_1 and β_2, a (partial) likelihood function (L) is formed from the model and the data; this function is then maximized using computer iteration. The likelihood function (L) for the stratified Cox (SC) model is different from the nonstratified Cox model. For the SC model, L is obtained by multiplying together likelihood functions for each stratum. Thus, L is equal to the product of L_1 and L_2, where L_1 and L_2 denote the female and male likelihood functions, respectively, which are derived from their respective hazard functions $h_1(t)$ and $h_2(t)$.

EXAMPLE: (continued)

Adjusted Survival Curves for *Rx* from Stratified Cox Model (adjusted for log WBC)



As mentioned above, adjusted survival curves can be obtained for each stratum as shown here. Here we have shown *four* survival curves because we want to compare the survival for two treatment groups over each of two strata.

If we compare treatment and placebo group separately by sex, we can see that the treatment group has consistently better survival prognosis than the placebo group for females and males separately. This supports our findings about the hazard ratio for the treatment effect derived earlier from the computer results for the stratified Cox model.

III. The General Stratified Cox (SC) Model

In the previous example, we illustrated the SC model for one binary predictor not satisfying the PH assumption. We now describe the general form of the SC model that allows for stratification of several predictors over several strata.

Example: one binary predictor
↓
General: several predictors, several strata

We assume that we have k variables not satisfying the PH assumption and p variables satisfying the PH assumption. The variables not satisfying the PH assumption we denote as $Z_1, Z_2, \ldots, Z_k$; the variables satisfying the PH assumption we denote as $X_1, X_2, \ldots, X_p$.

$Z_1, Z_2, \ldots, Z_k$, do not satisfy PH
$X_1, X_2, \ldots, X_p$, satisfy PH

Define a single new variable Z^*:

1. categorize each Z_i
2. form combinations of categories (strata)
3. the strata are the categories of Z^*

To perform the stratified Cox procedure, we define a single new variable, which we call Z^*, from the Z's to be used for stratification. We do this by forming categories of each Z_i, including those Z_i that are interval variables. We then form combinations of categories, and these combinations are our strata. These strata are the categories of the new variable Z^*.

EXAMPLE

		Age		
		Young	Middle	Old
Treatment status	Placebo	1	2	3
	Treatment	4	5	6

Z^* = new variable with six categories stratify on Z^*

For example, suppose k is 2, and the two Z's are age (an interval variable) and treatment status (a binary variable). Then we categorize age into, say, three age groups – young, middle, and old. We then form six age group–by–treatment-status combinations, as shown here. These six combinations represent the different categories of a single new variable that we stratify on in our stratified Cox model. We call this new variable Z^*.

Z^* has k^* categories where k^* = total # of combinations (strata), e.g., $k^* = 6$ in above example.

In general, the stratification variable Z^* will have k^* categories, where k^* is the total number of combinations (or strata) formed after categorizing each of the Z's. In the above example, k^* is equal to 6.

The general SC model:
$h_g(t, \mathbf{X}) = h_{0g}(t)\exp[\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p]$
$g = 1, 2, \ldots, k^*$, strata defined from Z^*

We now present the general hazard function form for the stratified Cox model, as shown here. This formula contains a subscript g which indicates the gth stratum. The strata are defined as the different categories of the stratification variable Z^*, and the number of strata equals k^*.

Z^* not included in the model

$X_1, X_2, \ldots, X_p$ included in the model

Note that the variable Z^* is not explicitly included in the model but that the X's, which are assumed to satisfy the PH assumption, are included in the model.

Different baseline hazard functions: $h_{0g}(t)$, $g = 1, 2, \ldots, k^*$
Same coefficients: $\beta_1, \beta_2, \ldots, \beta_p$

Different baselines $\left\{ \begin{array}{lcl} \hat{h}_{01}(t) & \Rightarrow & \hat{S}_1(t) \\ \hat{h}_{02}(t) & \Rightarrow & \hat{S}_2(t) \\ \vdots & & \\ \hat{h}_{0k}(t) & \Rightarrow & \hat{S}_k(t) \end{array} \right\}$ Different survival curves

Note also that the baseline hazard function $h_{0g}(t)$ is allowed to be different for each stratum. However, the coefficients $\beta_1, \beta_2, \ldots, \beta_p$ are the same for each stratum.

As previously described by example, the fitted SC model will yield different estimated survival curves for each stratum because the baseline hazard functions are different for each stratum.

$\widehat{HR}$ same for each stratum

(no-interaction assumption, Section IV)

However, because the coefficients of the X's are the same for each stratum, estimates of hazard ratios are the same for each stratum. This latter feature of the SC model is what we previously have called the "no-interaction" assumption to be discussed further in Section IV.

(Partial) likelihood function:

$L = L_1 \times L_2 \times \ldots \times L_{k^*}$

Strata:	1	2	...	k^*
Likelihood:	L_1	L_2	...	L_{k^*}
Hazard:	$h_1(t, \mathbf{X})$	$h_2(t, \mathbf{X})$	...	$h_{k^*}(t, \mathbf{X}^*)$

To obtain estimates of the regression coefficients $\beta_1, \beta_2, \ldots, \beta_p$, we maximize a (partial) likelihood function L that is obtained by multiplying together likelihood functions for each stratum, as shown here. Thus, L is equal to the product of L_1 times L_2, and so on, up until L_{k^*}, where the subscripted L's denote the likelihood functions for different strata, with each of these L's being derived from its corresponding hazard function.

IV. The No-Interaction Assumption and How to Test It

We previously pointed out that the SC model contains regression coefficients, denoted as β's, that do not vary over the strata. We have called this property of the model the "no-interaction assumption." In this section, we explain what this assumption means. We also describe how to evaluate the assumption and what to do if the assumption is violated.

Stratified Cox model

$$h_g(t,\mathbf{X}) = h_{0g}(t)\exp\left[\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p\right]$$

β coefficients do not vary over strata (no-interaction assumption)

- how to evaluate
- what to do if violated

EXAMPLE

No-interaction SC model:
Stratified Cox regression
Analysis time_t: survt

	Coef.	Std. Err.	p>|z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.390	0.338	0.000	4.016	2.072	7.783
Rx	0.931	0.472	0.048	2.537	1.006	6.396

No. of subjects = 42 Log likelihood = −57.560 Stratified by sex

Interaction by fitting separate models:
Cox regression (Females)
Analysis time _t: survt

Column name	Coeff.	StErr.	p-value	*HR*	0.95	CI	*p* (*PH*)
log WBC	1.639	0.519	0.002	5.150	1.862	14.242	0.228
Rx	1.859	0.729	0.011	6.418	1.537	26.790	0.603

No. of subjects = 20 Log likelihood = −22.100

Cox regression (Males)
Analysis time _t: survt

Column name	Coeff.	StErr.	p-value	*HR*	0.95	CI	*P* (*PH*)
log WBC	1.170	0.499	0.019	3.222	1.213	8.562	0.674
Rx	0.267	0.566	0.637	1.306	0.431	3.959	0.539

No. of subjects = 22 Log likelihood = −33.736

Which model is more appropriate statistically?

We return to the SC output previously illustrated. Notice that only one set of coefficients, namely, 1.390 for log WBC and 0.931 for *Rx*, are provided, even though there are two strata, one for females and one for males. These results assume no interaction of the sex variable with either log WBC or *Rx*.

If we allow for interaction, then we would expect to obtain different coefficients for each of the (SEX) strata. This would happen if we fit separate hazard models to the female and male data, with each model containing the log WBC and Rx variables. The computer results from fitting separate models are shown here.

Notice that the coefficient of log WBC is 1.639 for females but is 1.170 for males. Also, the coefficient for *Rx* is 1.859 for females but 0.267 for males. These results show different coefficients for females than for males, particularly for the *Rx* variable.

But are corresponding coefficients statistically different? That is, which model is more appropriate statistically, the no-interaction model or the interaction model? To answer this question, we must first look at the hazard function model for the interaction situation.

EXAMPLE: (continued)

Interaction model:
(◆) $h_g(t,\mathbf{X})$
$= h_{0g}(t)\exp[\beta_{1g}\log \text{WBC} + \beta_{2g}Rx]$
where $g = 1$ (females), $g = 2$ (males)

No-interaction model:
$h_g(t,\mathbf{X})$
$= h_{0g}(t)\exp[\beta_1 \log \text{WBC} + \beta_2 Rx]$
where $g = 1$ (females), $g = 2$ (males)

Alternative interaction model:

$$(\star)h_g(t,\mathbf{X}) = h_{0g}(t)\exp\left[\beta_1^* \log \text{WBC} + \beta_2^* Rx + \beta_3^*(\text{SEX}\times\log \text{WBC}) + \beta_4^*\times(\text{SEX}\times Rx)\right]$$

$$\text{where SEX} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

$h_{0g}(t)$ are different for $g = 1, 2$
β^* coefficients do not involve g

Equivalence of models (◆) and (⋆):
$g = 1$ (females), so that sex = 1:

$$h_1(t,\boldsymbol{X}) = h_{01}(t)\exp\left[\beta_1^* \log \text{WBC} + \beta_2^* Rx + \beta_3^*(1\times\log \text{WBC}) + \beta_4^*(1\times Rx)\right]$$

$$= h_{01}(t)\exp\left[\boxed{(\beta_1^*+\beta_3^*)}\log \text{WBC} + \boxed{(\beta_2^*+\beta_4^*)}\,\text{Rx}\right]$$

$g = 2$ (males), so that sex = 0:

$$h_2(t,\boldsymbol{X}) = h_{02}(t)\exp\left[\beta_1^* \log \text{WBC} + \beta_2^* Rx + \beta_3^*(0\times\log \text{WBC}) + \beta_4^*(0\times Rx)\right]$$

$$= h_{02}(t)\exp\left[\boxed{\beta_1^*}\log \text{WBC} + \boxed{\beta_2^*}Rx\right]$$

Interaction models in same format:

Females ($g = 1$): $h_1(t,\mathbf{X})$
(◆) $= h_{01}(t)\exp[\beta_{11}\log \text{WBC} + \beta_{21}Rx]$
(⋆) $= h_{01}(t)\exp\left[(\beta_1^*+\beta_3^*)\log \text{WBC} + (\beta_2^*+\beta_4^*)Rx\right]$

Males ($g = 2$): $h_2(t,\mathbf{X})$
(◆) $= h_{02}(t)\exp[\beta_{12}\log \text{WBC} + \beta_{22}Rx]$
(⋆) $= h_{02}(t)\exp\left[\beta_1^*\log \text{WBC} + \beta_2^* Rx\right]$

One way to state the hazard model formula when there is **interaction** is shown here (◆). Notice that each variable in this model has a different coefficient for females than for males, as indicated by the subscript g in the coefficients β_{1g} and β_{2g}.

In contrast, in the **no-interaction** model, the coefficient (β_1) of log WBC is the same for females and for males; also, the coefficient (β_2) of Rx is the same for females and for males.

An alternative way to write the interaction model is shown here (⋆). This alternative form contains two product terms, SEX × log WBC and SEX × Rx, as well as the main effects of log WBC and Rx. We have coded the SEX so that 1 denotes female and 0 denotes male.

In this alternative model, note that although the baseline hazards $h_{0g}(t)$ are different for each sex, the β^* coefficients do not involve the subscript g and therefore are the same for each sex.

Nevertheless, this alternative formula (⋆) is equivalent to the interaction formula (◆) above. We show this by specifying the form that the model takes for $g = 1$ (females) and $g = 2$ (males).

Notice that the coefficients of log WBC are different in each formula, namely, $(\beta_1^*+\beta_3^*)$ for females versus β_1^* for males.

Similarly, the coefficients of Rx are different, namely, $(\beta_2^*+\beta_4^*)$ for females versus β_2^* for males.

The preceding formulae indicate that two seemingly different formulae for the interaction model – (◆) versus (⋆), shown earlier – can be written in the same format. We show these formulae here separately for females and males.

EXAMPLE: (continued)

$$\begin{array}{lll} & (\blacklozenge) & (\star) \\ \text{Females } (g=1): & \beta_{11} = \beta_1^* + \beta_3^* \\ & \beta_{21} = \beta_2^* + \beta_4^* \end{array}$$

$$\begin{array}{lll} & (\blacklozenge) & (\star) \\ \text{Males } (g=2): & \beta_{12} = \beta_1^* \\ & \beta_{22} = \beta_2^* \end{array}$$

Stratified Cox regression
Analysis time _t: survt

	Coef.	Std. Err:	p>|z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.170	0.499	0.019	3.222	1.213	8.562
Rx	0.267	0.566	0.637	1.306	0.431	3.959
Sex × log WBC	0.469	0.720	0.515	1.598	0.390	6.549
Sex × *Rx*	1.592	0.923	0.084	4.915	0.805	30.003

No. of subjects = 42 Log likelihood = −55.835 Stratified by sex

Females:

$$\log \text{ WBC} \begin{cases} \beta_{11} = \boxed{1.639} \\ \hat{\beta}_1^* + \hat{\beta}_3^* = 1.170 + 0.469 = \boxed{1.639} \end{cases}$$

$$Rx \begin{cases} \beta_{21} = \boxed{1.859} \\ \hat{\beta}_2^* + \hat{\beta}_4^* = 0.267 + 1.592 = \boxed{1.859} \end{cases}$$

Males:

$$\log \ WBC \ \hat{\beta}_{12} = \boxed{1.170} = \hat{\beta}_1^*$$

$$Rx \qquad \hat{\beta}_{22} = \boxed{0.267} = \hat{\beta}_2^*$$

Interaction model:

$$h_g(t, \boldsymbol{X}) = h_{0g}(t) \exp\left[\beta_1^* \log \text{WBC} + \beta_2^* Rx + \beta_3^* (\text{SEX} \times \log \text{WBC}) + \beta_4^* \times (\text{SEX} \times Rx)\right]$$

Notice that for females, the coefficient β_{11} in model $(\star)$ must be equivalent to $(\beta_1^* + \beta_3^*)$ in model $(\blacklozenge)$ because both models have the same format, and both β_{11} and $(\beta_1^* + \beta_3^*)$ are coefficients of the same variable, log WBC. Similarly, β_{21} in model $(\star)$ is equivalent to $(\beta_2^* + \beta_4^*)$ in model $(\blacklozenge)$ because both are coefficients of the same variable, *Rx*.

For males, it follows in an analogous way, that the coefficient β_{12} is equivalent to β_1^*, and, similarly, β_{22} equals β_2^*.

Here we provide computer results obtained from fitting the alternative interaction model $(\star)$. The estimated regression coefficients $\hat{\beta}_1^*$, $\hat{\beta}_2^*$, $\hat{\beta}_3^*$, , and $\hat{\beta}_4^*$, respectively, are circled.

We have indicated above that the sums $\hat{\beta}_1^* + \hat{\beta}_3^*$ and $\hat{\beta}_2^* + \hat{\beta}_4^*$ are equal to the coefficients $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$, respectively, in the original interaction model for females.

Also, we have indicated that $\hat{\beta}_1^*$ and $\hat{\beta}_2^*$ are equal to the coefficients $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$, respectively, in the original interaction model for the males. The numerical equivalences are shown here. Note again that the coefficients of log WBC and *Rx* for females are different from males, as is to be expected if sex interacts with each variable.

We have thus seen that the interaction model can be written in a format that contains product terms involving the variable being stratified, SEX, being multiplied by each of the predictors not being stratified. We show this model involving product terms again here. We will use this model to describe a test of the no-interaction assumption.

EXAMPLE: (continued)

Testing the no-interaction assumption:

$LR = -2 \ln L_R - (-2 \ln L_F)$
R = reduced (no-interaction) model
F = full (interaction) model

$LR \dot{\sim} \chi^2_{2\,df}$ under H_0: no interaction
(2 df because two product terms tested in interaction model)

No interaction (reduced model):

Output: –2 log *L*: 115.120

$-2 \ln L_R$

Interaction (full model):

Output: –2 log *L*: 111.670

$-2 \ln L_F$

$LR = 115.120 - 111.670 = 3.45$
($P > 0.05$ not significant).
Thus, the no-interaction model is acceptable.

The test is a likelihood ratio (LR) test which compares log-likelihood statistics for the interaction model and the no-interaction model. That is, the LR test statistic is of the form $-2 \ln L_R$ minus $-2 \ln L_F$, where R denotes the reduced model, which in this case is the no-interaction model, and F denotes the full model, which is the interaction model.

This LR test statistic has approximately a chi-square distribution with 2 degrees of freedom under the null hypothesis that the no-interaction model is correct. The degrees of freedom here is 2 because there are two product terms being tested in the interaction model.

The log-likelihood statistic for the reduced model comes from the computer output for the no-interaction model and is equal to -2 times -57.560, or 115.120.

The log-likelihood statistic for the full model comes from the computer results for the interaction model and is equal to -2 times -55.835, or 111.670.

The LR statistic is therefore 115.120 minus 111.670, which equals 3.45. This value is not significant at the 0.05 level for 2 degrees of freedom. Thus, it appears that despite the numerical difference between corresponding coefficients in the female and male models, there is no statistically significant difference. We can therefore conclude for these data that the no-interaction model is acceptable (at least at the 0.05 level).

Remission data example:

- described no-interaction assumption
- evaluated assumption using LR test
- provided interaction model if needed

Using the remission data example, we have described the no-interaction assumption, have shown how to evaluate this assumption using a likelihood ratio test, and have provided the form of an interaction model that should be used in case the no-interaction assumption does not hold. We now describe this process more generally for any stratified Cox analysis.

Now, we generalize this process.

No-interaction SC model:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p]$$
$g = 1, 2, \ldots, k^*$, strata defined from Z^*

Recall that the general form of the no-interaction model for the stratified Cox procedure is given as shown here. This model allows for several variables being stratified through the use of a newly defined variable called Z^*, whose strata consist of combinations of categories of the variables being stratified.

SC model allowing interaction:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_{1g} X_1 + \beta_{2g} X_2 + \ldots + \beta_{pg} X_p]$$
$g = 1, 2, \ldots, k^*$, strata defined from Z^*

If, in contrast, we allow for interaction of the Z^* variable with the X's in the model, we can write the model as shown here. Notice that in this interaction model, each regression coefficient has the subscript g, which denotes the gth stratum and indicates that the regression coefficients are different for different strata of Z^*.

Alternative SC interaction model:

- uses product terms involving Z^*
- define $k^* - 1$ dummy variables $Z_1^*, Z_2^*, \ldots, Z_{k^*-1}^*$, from Z^*
- products of the form $Z_i^* \times X_j$, where $i = 1, \ldots, k^* - 1$ and $j = 1, \ldots, p$.

An alternative way to write the interaction model uses product terms involving the variable Z^* with each of the predictors. However, to write this model correctly, we need to use $k^* - 1$ dummy variables to distinguish the k^* categories of Z^*; also, each of these dummy variables, which we denote as $Z_1^*, Z_2^*, \cdots Z_{k^*-1}^*$, needs to be involved in a product term with each of the X's.

$$\begin{aligned} h_g(t,\mathbf{X}) = h_{0g}(t)\exp\big[&\beta_1 X_1 + \cdots + \beta_p X_p \\ &+ \beta_{11}(Z_1^* \times X_1) + \cdots + \beta_{p1}(Z_1^* \times X_p) \\ &+ \beta_{12}(Z_2^* \times X_1) + \cdots + \beta_{p2}(Z_2^* \times X_p) \\ &+ \cdots + \beta_{1,k^*-1}(Z_{k^*-1}^* \times X_1) + \cdots \\ &+ \beta_{p,k^*-1}(Z_{k^*-1}^* \times X_p)\big] \end{aligned}$$
$g = 1, 2, \ldots, k^*$, strata defined from $Z*$

The hazard model formula alternative model is shown here. Notice that the first line of the formula contains the X's by themselves, the next line contains products of each X_j with Z_1^*, the third line contains the products with Z_2^*, and the last line contains products with $Z_{k^*-1}^*$. Note also that the subscript g occurs only with the baseline hazard function $h_{0g}(t)$, and is not explicitly used in the β coefficients.

EXAMPLE: (Remission Data)

$Z^* = \text{sex}, k^* = 2,$

$Z_1^* = \text{sex}(0, 1),$

$X_1 = \log \text{WBC}, X_2 = Rx \ (p = 2)$

$$\begin{aligned} h_g(t, \boldsymbol{X}) &= h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 \\ &\quad + \beta_{11}(Z_1^* \times X_1) \\ &\quad + \beta_{21}(Z_1^* \times X_2)] \\ &= h_{0g}(t) \exp[\beta_1^* \log \text{WBC} \\ &\quad + \beta_2^* \text{RX} + \beta_3^* (\text{sex} \times \log \text{WBC}) \\ &\quad + \beta_4^* (\text{sex} \times Rx)] \end{aligned}$$

$g = 1, 2.$

$\beta_1 = \beta_1^*, \ \beta_2 = \beta_2^*, \ \beta_{11} = \beta_3^*, \text{ and } \beta_{21} = \beta_4^*$

In our previous example involving the remission data, the stratification variable (Z^*) was the variable SEX, and k^* was equal to 2; thus, we have only one dummy variable Z_1^*, which uses a (0, 1) coding to indicate sex, and we have only (p equal to) two predictors: X_1 equal to log WBC and X_2 equal to Rx. The interaction model is then written in either of the forms shown here.

The latter version of the interaction model is what we previously presented for the remission data example. Because the two versions presented here are equivalent, it follows that $\beta_1^* = \beta_1, \ \beta_2 = \beta_2^*, \ \beta_{11} = \beta_3^*$, and $\beta_{21} = \beta_4^*$.

We have thus seen that the interaction model can be written in a format that contains product terms involving dummy variables (i.e., Z_i^*) for the variable being stratified being multiplied by each of the predictors (i.e., X_i) not being stratified. We will use this model to describe a test of the no-interaction assumption.

Testing the no-interaction assumption:

$LR = -2 \ln L_\text{R} - (-2 \ln L_\text{F})$

$R =$ reduced (no-interaction) model

$F =$ full (interaction) model contains product terms

$$H_0 : \begin{cases} \beta_{11} = \cdots = \beta_{p1} = 0 \\ \beta_{12} = \cdots = \beta_{p2} = 0 \\ \vdots \\ \beta_{1,k^*-1} = \cdots = \beta_{p,k^*-1} = 0 \end{cases}$$

The test is a likelihood ratio (LR) test which compares log likelihood statistics for the interaction model and the no-interaction model. That is, the LR test statistic is of the form $-2 \ln L_\text{R}$ minus $-2 \ln L_\text{F}$, where R denotes the reduced model, which in this case is the no-interaction model, and F denotes the full model, which is the interaction model.

The no-interaction model differs from the interaction model in that the latter contains additional product terms. Thus, one way to state the null hypothesis of no interaction is that the coefficients of each of these product terms are all zero.

$$LR \overset{\cdot}{\sim} \chi^2_{p(k^*-1) \text{ df}} \quad \text{under } H_0 : \text{no interaction}$$

The LR test statistic has approximately a chi-square distribution with $p(k^* - 1)$ degrees of freedom under the null hypothesis. The degrees of freedom here is $p(k^* - 1)$ because this value gives the number of product terms that are being tested in the interaction model.

$p(k^* - 1)$ gives number of product terms being tested in interaction model

EXAMPLE: (Remission Data)

$Z^* = \text{sex}, k^* = 2,$

$Z_1^* = \text{sex}(0, 1)$

$X_1 = \log \text{WBC}, X_2 = \text{Rx}\ (p = 2)$

$p(k^* - 1) = 2$, so

$LR \dot{\sim} \chi^2_{2\,\text{df}}$ under H_0: no interaction

Returning to the remission data example, for which $p = 2$ and $k^* = 2$, the value of $p(k^* - 1)$ is equal to two times $(2 - 1)$, which equals two. Thus, to test whether the SEX variable interacts with the log WBC and *Rx* predictors, the degrees of freedom for the *LR* statistic is two, as previously described.

V. A Second Example Involving Several Stratification Variables

EXAMPLE

vets.dat: survival time in days, $n = 137$

Veteran's Administration Lung Cancer Trial

Column 1: Treatment (standard = 1, test = 2)
Column 2: Cell type 1 (large = 1, other = 0)
Column 3: Cell type 2 (adeno = 1, other = 0)
Column 4: Cell type 3 (small = 1, other = 0)
Column 5: Cell type 4 (squamous = 1, other = 0)
Column 6: Survival time (days)
Column 7: Performance status (0 = worst,. . ., 100 = best)
Column 8: Disease duration (months)
Column 9: Age
Column 10: Prior therapy (none = 0, some = 10)
Column 11: Status (0 = censored, 1 = died)

The dataset "vets.dat" considers survival times in days for 137 patients from the Veteran's Administration Lung Cancer Trial cited by Kalbfleisch and Prentice in their text (*The Statistical Analysis of Survival Time Data*, Wiley, pp. 223–224, 1980). The exposure variable of interest is treatment status. Other variables of interest as control variables are cell type (four types, defined in terms of dummy variables), performance status, disease duration, age, and prior therapy status. Failure status is defined by the status variable. A complete list of the variables is shown here.

Cox regression
Analysis time_t: survt

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]	*P* (*PH*)
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Perf. Stat	−0.033	0.006	0.000	0.968	0.958	0.978	0.000
Dis. Durat.	0.000	0.009	0.992	1.000	0.982	1.018	0.919
Age	−0.009	0.009	0.358	0.991	0.974	1.010	0.198
Pr. Therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.145

No. of subjects = 137 Log likelihood = −475.180

Here we provide computer output obtained from fitting a Cox PH model to these data. Using the *P*(*PH*) information in the last column, we can see that at least four of the variables listed have *P*(*PH*) values below the 0.100 level. These four variables are labeled in the output as large cell (0.033), adeno cell (0.081), small cell (0.078), and Perf. Stat (0.000). Notice that the three variables, large cell, adeno cell, and small cell, are dummy variables that distinguish the four categories of cell type.

Variables not satisfying PH:

- cell type (3 dummy variables)
- performance status
- prior therapy (possibly)

Thus, it appears from the *P*(*PH*) results that the variables cell type (defined using dummy variables) and performance status do not satisfy the PH assumption.

SC model: stratifies on cell type and performance status

Based on the conclusions just made about the PH assumption, we now describe a stratified Cox analysis that stratifies on the variables, cell type and performance status.

EXAMPLE: (continued)

Z^* given by combinations of categories:

- cell type (four categories)
- performance status (interval) change to
- PSbin (two categories)

Z^* has $k^* = 4 \times 2 = 8$ categories

Four other variables considered as X's:

- treatment status
- disease duration
- age
- prior therapy

Here, we use treatment status and age as X's

Stratified Cox regression
Analysis time_t: survt

	Coef.	Std. Err.	p>|z|	Haz. Ratio	[95% Conf.	Interval]
Treatment	0.125	0.208	0.548	1.134	0.753	1.706
Age	−0.001	0.010	0.897	0.999	0.979	1.019

No. of subjects = 137 Log likelihood = −262.020 Stratified by Z^*

No-interaction model

$\widehat{HR} = 1.134 (P = 0.548)$

Treatment effect (adjusted for age and Z^*) is nonsignificant

No-interaction model:
$h_g(t,\mathbf{X})$
$= h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 \text{ Age}]$
$g = 1, 2, \ldots, 8$ (=# of strata defined from Z^*)

Interaction model:
$h_g(t,\mathbf{X})$
$= h_{0g}(t)\exp[\beta_{1g} \text{ Treatment} + \beta_{2g} \text{ Age}]$
$g = 1, 2, \ldots, 8$

Because we are stratifying on two variables, we need to form a single new categorical variable Z^* whose categories represent combinations of categories of the two variables. The cell type variable has four categories by definition. The performance status variable, however, is an interval variable ranging between 0 for worst to 100 for best, so it needs to be categorized. We categorize this variable into two groups using a cutpoint of 60, and we denote this binary variable as PSbin. Thus, the number of categories for our Z^* variable is 4×2, or 8; that is, $k^* = 8$.

In addition to the two stratification variables, cell type and performance status, there are four other variables to be considered as predictors in the stratified Cox model. These are treatment status, disease duration, age, and prior therapy.

For illustrative purposes here, we use only treatment status and age as predictors. The other two variables, disease duration and prior therapy, are considered in exercises following this presentation.

Here we show computer output from fitting a stratified Cox model that stratifies on cell type and performance status using the eight-category stratification variable Z^*. This model also includes treatment and age as predictors. These results consider a no-interaction model, because only one regression coefficient is provided for the treatment and age predictors. Notice that the estimated hazard ratio is 1.134 for the effect of the treatment variable adjusted for age and Z^*, the latter being adjusted by stratification. The p-value for this adjusted treatment effect is 0.548, which is highly nonsignificant.

The no-interaction model we have just described has the hazard function formula shown here.

To evaluate whether the no-interaction model is appropriate, we need to define an interaction model that allows different regression coefficients for different strata. One way to write this interaction model is shown here.

EXAMPLE: (continued)

Alternative interaction model:

$$h_g(t,X) = h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 \text{ Age} + \beta_{11}(Z_1^* \times \text{Treatment}) + \cdots + \beta_{17}(Z_7^* \times \text{Treatment}) + \beta_{21}(Z_1^* \times \text{Age}) + \cdots + \beta_{27}(Z_7^* \times \text{Age})]$$

$$g = 1, 2, \ldots, 8$$

An alternative version of this interaction model that involves product terms is shown here. This version uses seven dummy variables denoted as Z_1^*, Z_2^* up through Z_7^* to distinguish the eight categories of the stratification variable Z^*. The model contains the main effects of treatment and age plus interaction terms involving products of each of the seven dummy variables with each of the two predictors.

Another version of interaction model:
Replace $Z_1^*, \ldots, Z_7^*$ by

$Z_1^* =$ large cell (binary)
$Z_2^* =$ adeno cell (binary)
$Z_3^* =$ small cell (binary)
$Z_4^* =$ PSbin (binary)
$Z_5^* = Z_1^* \times Z_4^*$
$Z_6^* = Z_2^* \times Z_4^*$
$Z_7^* = Z_3^* \times Z_4^*$

Yet another version of the interaction model is to replace the seven dummy variables Z_1^* to Z_7^* by the seven variables listed here. These variables are three of the binary variables making up the cell type variable, the binary variable for performance status, plus three product terms involving each of the cell type dummy variables multiplied by the PSbin dummy variable (Z_4^*).

$$h_g(t,\boldsymbol{X}) = h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 \text{ Age} + \beta_{11}(\text{tr } Z_1^*) + \beta_{12}(\text{tr } Z_2^*) + \beta_{13}(\text{tr } Z_3^*) + \beta_{14}(\text{tr } Z_4^*) + \beta_{15}(\text{tr } Z_1^*Z_4^*) + \beta_{16}(\text{tr } Z_2^*Z_4^*) + \beta_{17}(\text{tr } Z_3^*Z_4^*) + \beta_{21}(\text{AGE } Z_1^*) + \beta_{22}(\text{AGE } Z_2^*) + \beta_{23}(\text{AGE } Z_3^*) + \beta_{24}(\text{AGE } Z_4^*) + \beta_{25}(\text{AGE } Z_1^*Z_4^*) + \beta_{26}(\text{AGE } Z_2^*Z_4^*) + \beta_{27}(\text{AGE } Z_3^*Z_4^*)]$$

The latter interaction model is shown here. In this model, the variable tr Z_1^* denotes the product of treatment status with the large cell dummy Z_1^*, the variable tr Z_2^* denotes the product of treatment status with the adeno cell variable Z_2^*, and so on. Also, the variable tr $Z_1^*Z_4^*$ denotes the triple product of treatment status times the large cell variable Z_1^* times the PSbin variable Z_4^*, and so on, for the other triple product terms involving treatment. Similarly, for the terms involving age, the variable Age Z_1^* denotes the product of age with Z_1^*, and the variable Age $Z_1^*Z_4^*$ denotes the triple product of age times Z_1^* times Z_4^*.

Note that we are only considering the interaction between the stratified variables and the predictors. We could also (but do not) consider the interaction between the two predictors, treatment, and age.

EXAMPLE: (continued)

Stratified Cox Regression Analysis on Variable: Z^*
Response: Surv. Time

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Treatment	0.286	0.664	0.667	1.331	0.362	4.893
Age	0.000	0.030	0.978	0.999	0.942	1.060
tr Z_1^*	2.351	1.772	0.184	10.495	0.326	337.989
tr Z_2^*	−1.158	0.957	0.226	0.314	0.048	2.047
tr Z_3^*	0.582	0.855	0.496	1.790	0.335	9.562
tr Z_4^*	−1.033	0.868	0.234	0.356	0.065	1.950
tr $Z_1^*Z_4^*$	−0.794	1.980	0.688	0.452	0.009	21.882
tr $Z_2^*Z_4^*$	2.785	1.316	0.034	16.204	1.229	213.589
tr $Z_3^*Z_4^*$	0.462	1.130	0.683	1.587	0.173	14.534
Age Z_1^*	0.078	0.064	0.223	1.081	0.954	1.225
Age Z_2^*	−0.047	0.045	0.295	0.954	0.873	1.042
Age Z_3^*	−0.059	0.042	0.162	0.943	0.868	1.024
Age Z_4^*	0.051	0.048	0.287	1.053	0.958	1.157
Age $Z_1^*Z_4^*$	−0.167	0.082	0.042	0.847	0.721	0.994
Age $Z_2^*Z_4^*$	−0.045	0.068	0.511	0.956	0.838	1.092
Age $Z_3^*Z_4^*$	0.041	0.061	0.499	1.042	0.924	1.175

No. of subjects = 137 Log likelihood = −249.972 Stratified by Z^*

Eight possible combinations of Z_1^* to Z_4^*:

$g = 1: \ Z_1^* = Z_2^* = Z_3^* = Z_4^* = 0$
$g = 2: \ Z_1^* = 1, \ Z_2^* = Z_3^* = Z_4^* = 0$
$g = 3: \ Z_2^* = 1, \ Z_1^* = Z_3^* = Z_4^* = 0$
$g = 4: \ Z_3^* = 1, \ Z_1^* = Z_2^* = Z_4^* = 0$
$g = 5: \ Z_1^* = Z_2^* = Z_3^* = 0, \ Z_4^* = 1$
$g = 6: \ Z_1^* = 1, \ Z_2^* = Z_3^* = 0, \ Z_4^* = 1$
$g = 7: \ Z_2^* = 1, \ Z_1^* = Z_3^* = 0, \ Z_4^* = 1$
$g = 8: \ Z_3^* = 1, \ Z_1^* = Z_2^* = 0, \ Z_4^* = 1$

$g = 1: \ Z_1^* = Z_2^* = Z_3^* = Z_4^* = 0$
(Squamous cell type and PSbin = 0)

All product terms are zero:
$h_1(t,\mathbf{X})$
$= h_{01}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 \text{ Age}]$,
where $\hat{\beta}_1 = 0.286$,
$\hat{\beta}_2 = 0.000$, so that
$\hat{h}_1(t,\mathbf{X}) = \hat{h}_{01}(t)\exp[(0.283)\text{ Treatment}]$

$g = 2: \ Z_1^* = 1, \ Z_2^* = Z_3^* = Z_4^* = 0$
(Large cell type and PSbin = 0)

Nonzero product terms	Coefficients
Age Z_1^* = Age	β_{21}
tr Z_1^* = Treatment	β_{11}

Here we provide the computer results from fitting the interaction model just described. Notice that the first two variables listed are the main effects of treatment status and age. The next seven variables are product terms involving the interaction of treatment status with the seven categories of Z^*. The final seven variables are product terms involving the interaction of age with the seven categories of Z^*. As defined on the previous page, the seven variables used to define Z^* consist of three dummy variables Z_1^*, Z_2^* and Z_3^* for cell type, a binary variable Z_4^* for performance status and products of Z_4^* with each of Z_1^*, Z_2^*, and Z_3^*. Note that once the variables Z_1^*, Z_2^* Z_3^*, and Z_4^* are specified, the values of the three product terms are automatically determined.

We can use these results to show that the interaction model being fit yields different regression coefficients for each of the eight categories defined by the subscript g for the stratification variable Z^*. These eight categories represent the possible combinations of the four variables Z_1^* to Z_4^*, as shown here.

Consider the hazard function when the variables Z_1^* through Z_4^* are all equal to zero. This stratum is defined by the combination of squamous cell type and a binary performance status value of 0. In this case, all product terms are equal to zero and the hazard model contains only the main effect terms treatment and age. The estimated hazard function for this stratum uses the coefficients 0.286 for treatment and 0.000 for age, yielding the expression shown here. Note that age drops out of the expression because its coefficient is zero to three decimal places.

Now consider the hazard function when the variable Z_1^* equals 1 and Z_2^* through Z_4^* are equal to zero. This stratum is defined by the combination of large cell type and a PSbin value of 0. In this case, the only nonzero product terms are Age Z_1^* and tr Z_1^*, whose coefficients are β_{21} and β_{11}, respectively.

EXAMPLE: (continued)

$h_2(t,\mathbf{X}) = h_{02}(t)\exp[(\beta_1 + \beta_{11})\,\text{Treatment} + (\beta_2 + \beta_{21})\,\text{Age}]$

$\hat{\beta}_1 = 0.286,\ \hat{\beta}_2 = 0.000$

$\hat{\beta}_{11} = 2.351,\ \hat{\beta}_{21} = 0.078$

The hazard function for this second stratum is shown here. Notice that the coefficients of the treatment and age variables are $(\beta_1 + \beta_{11})$ and $(\beta_2 + \beta_{21})$, respectively. The estimated values of each of these coefficients are given here.

Hazard functions for interaction model:

$g = 1: (Z_1^* = Z_2^* = Z_3^* = Z_4^* = 0):$
$\hat{h}_1(t,\mathbf{X}) = \hat{h}_{01}(t)\exp[(0.286)\,\text{Treatment}]$

$g = 2: (Z_1^* = 1, Z_2^* = Z_3^* = Z_4^* = 0):$
$\hat{h}_2(t,\mathbf{X}) = \hat{h}_{02}(t)\exp[(2.637)\,\text{Treatment} + (0.078)\text{Age}]$

The corresponding *estimated* hazard function for the second stratum (i.e., $g = 2$) is shown here. For comparison, we repeat the estimated hazard function for the first stratum.

$g = 3: (Z_2^* = 1, Z_1^* = Z_3^* = Z_4^* = 0):$
$\hat{h}_3(t,\mathbf{X}) = \hat{h}_{03}(t)\exp[(-0.872)\,\text{Treatment} + (-0.047)\text{Age}]$

$g = 4: (Z_3^* = 1, Z_1^* = Z_2^* = Z_4^* = 0):$
$\hat{h}_4(t,\mathbf{X}) = \hat{h}_{04}(t)\exp[(0.868)\,\text{Treatment} + (-0.059)\text{Age}]$

$g = 5: (Z_1^* = Z_2^* = Z_3^* = 0, Z_4^* = 1):$
$\hat{h}_5(t,\mathbf{X}) = \hat{h}_{05}(t)\exp[(0.747)\,\text{Treatment} + (-0.051)\text{Age}]$

$g = 6: (Z_1^* = 1, Z_2^* = Z_3^* = 0, Z_4^* = 1):$
$\hat{h}_6(t,\mathbf{X}) = \hat{h}_{06}(t)\exp[(0.810)\,\text{Treatment} + (-0.038)\text{Age}]$

$g = 7: (Z_2^* = 1, Z_1^* = Z_3^* = 0, Z_4^* = 1):$
$\hat{h}_7(t,\mathbf{X}) = \hat{h}_{07}(t)\exp[(0.880)\,\text{Treatment} + (-0.041)\text{Age}]$

$g = 8: (Z_3^* = 1, Z_1^* = Z_2^* = 0, Z_4^* = 1):$
$\hat{h}_8(t,\mathbf{X}) = \hat{h}_{08}(t)\exp[(0.297)\,\text{Treatment} + (0.033)\text{Age}]$

The estimated hazard functions for the remaining strata are provided here. We leave it up to the reader to verify these formulae. Notice that the coefficients of treatment are all different in the eight strata, and the coefficients of age also are all different in the eight strata.

LR test to compare no-interaction model with interaction model:

We have presented computer results for both the no-interaction and the interaction models. To evaluate whether the no-interaction assumption is satisfied, we need to carry out a likelihood ratio test to compare these two models.

H_0: no-interaction model acceptable, i.e., Treatment: $\beta_{11} = \beta_{12} = \cdots = \beta_{17} = 0$ and Age: $\beta_{21} = \beta_{22} = \cdots = \beta_{27} = 0$

The null hypothesis being tested is that the no-interaction model is acceptable. Equivalently, this null hypothesis can be stated by setting the coefficients of all product terms in the interaction model to zero. That is, the seven coefficients of product terms involving treatment and the seven coefficients of the product terms involving age are set equal to zero as shown here.

14 coefficients $\Rightarrow$ df = 14

$LR = -2\ln L_R - (2\ln L_F)$

R = reduced (no-interaction) model

F = full (interaction) model

Because the null hypothesis involves 14 coefficients, the degrees of freedom of the LR chi-square statistic is 14. The test statistic takes the usual form involving the difference, between log-likelihood statistics for the reduced and full models, where the reduced model is the no-interaction model and the full model is the interaction model.

EXAMPLE (continued)

$LR \dot{\sim} \chi^2_{14\,df}$ under H_0: no interaction

$LR = (-2 \times -262.020) - (-2 \times -249.972)$
$= 524.040 - 499.944 = 24.096$

$P = 0.045$ (significant at 0.05)

Conclusion:
Reject H_0: interaction model is preferred.

Might use further testing to simplify interaction model, e.g., test for seven products involving treatment or test for seven products involving age.

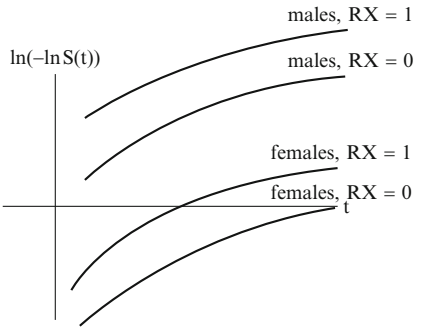
Thus, under the null hypothesis, the *LR* statistic is approximately chi-square with 14 degrees of freedom.

The computer results for the no-interaction and interaction models give log-likelihood values of 524.040 and 499.944, respectively. The difference is 24.096. A chi-square value of 24.096 with 14 degrees of freedom yields a p-value of 0.045, so that the test gives a significant result at the 0.05 level. This indicates that the no-interaction model is not acceptable and the interaction model is preferred.

Note, however, that it may be possible from further statistical testing to simplify the interaction model to have fewer than 14 product terms. For example, one might test for only the seven product terms involving treatment or only the seven product terms involving age.

VI. A Graphical View of the Stratified Cox Approach

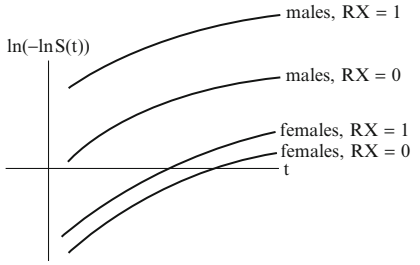
a. $h(t) = h_0(t)\exp(\beta_1 RX + \beta_2 SEX)$
$\ln(-\ln S(t)) = \ln(-\ln S_0(t)) + \beta_1 RX + \beta_2 SEX$



In this section we examine four log–log survival plots illustrating the assumptions underlying a stratified Cox model with or without interaction. Each of the four models considers two dichotomous predictors: treatment (coded RX = 1 for placebo and RX = 0 for new treatment) and SEX (coded 0 for females and 1 for males). The four models are as follows (see left).

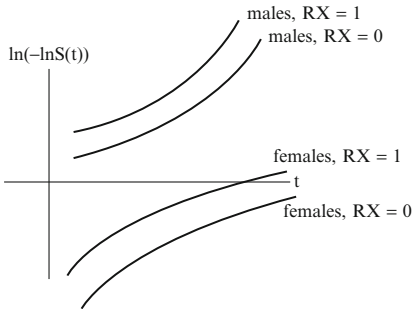
a. This model assumes the PH assumption for both RX and SEX and also assumes no interaction between RX and SEX. Notice all four log–log curves are parallel (PH assumption) and the effect of treatment is the same for females and males (no interaction). The effect of treatment (controlling for SEX) can be interpreted as the distance between the log–log curves from RX = 1 to RX = 0, for males and for females, separately.

b. $h(t) = h_0(t) \exp(\beta_1 RX + \beta_2 SEX + \beta_3 RX \times SEX)$
$\ln(-\ln S(t)) = \ln(-\ln S_0(t)) + \beta_1 RX + \beta_2 SEX + \beta_3 RX \times SEX$



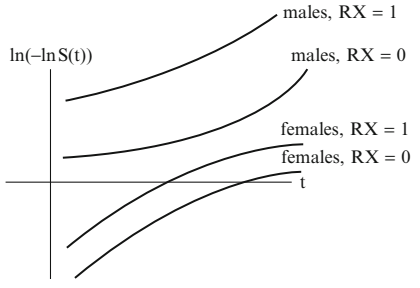
b. This model assumes the PH assumption for both RX and SEX and allows for interaction between these two variables. All four log–log curves are parallel (PH assumption) but the effect of treatment is larger for males than females as the distance from RX = 1 to RX = 0 is greater for males.

c. $h(t) = h_{0g}(t) \exp(\beta_1 RX)$
(g = 1 for males, g = 0 for females)
$\ln(-\ln S(t)) = \ln(-\ln S_{0g}(t)) + \beta_1 RX$



c. This is a stratified Cox model in which the PH assumption is not assumed for SEX. Notice the curves for males and females are not parallel. However, the curves for RX are parallel within each stratum of SEX indicating that the PH assumption is satisfied for RX. The distance between the log–log curves from RX = 1 to RX = 0 is the same for males and females indicating no interaction between RX and SEX.

d. $h(t) = h_{0g}(t) \exp(\beta_1 RX + \beta_2 RX \times SEX)$
(g = 1 for males, g = 0 for females)
$\ln(-\ln S(t)) = \ln(-\ln S_{0g}(t)) + \beta_1 RX + \beta_2 RX \times SEX$



d. This is a stratified Cox model allowing for interaction of RX and SEX. The curves for males and females are not parallel although the PH assumption is satisfied for RX within each stratum of SEX.The distance between the log–log curves from RX = 1 to RX = 0 is greater for males than females indicating interaction between RX and SEX.

VII. The Stratified Cox Likelihood

- Cox PH Likelihood (L) described in Chapter 3, Section VIII
- L now extended for SC model

At the end of the presentation from Chapter 3 (Section VIII), we illustrated the Cox likelihood using the dataset shown on the left. In this section, we extend that discussion to illustrate the likelihood for a stratified Cox model.

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

TIME = Survival time (in years)
STATUS = 1 for event, 0 for censorship
SMOKE = 1 for a smoker, 0 for a non-smoker

To review: The data indicate that Barry got the event at TIME = 2 years. Gary got the event at 3 years, Harry was censored at 5 years and Larry got the event at 8 years. Furthermore, Barry and Larry were smokers while Gary and Harry were non-smokers.

Cox PH model: $h(t) = h_0(t)e^{\beta_1 SMOKE}$

Cox PH Likelihood:

$$L = \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}}\right] \times \left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

In Chapter 3, we constructed the Cox likelihood with one predictor, SMOKE, in the model. The model and the likelihood are shown on the left.

The likelihood is a product of three terms, one term for each event time t_f (TIME = 2, 3, and 8). The denominator of each term is the sum of the hazards from the subjects still in the risk set at time t_f, including the censored subject, Harry. The numerator of each term is the hazard of the subject who got the event at t_f. The reader may wish to reread Section VIII of Chapter 3.

Stratified Cox model:

$h_g(t) = h_{0g}(t)e^{\beta_1 SMOKE}$

g=1 history of hypertension
g=2 no history of hypertension

Now consider a stratified Cox model, in which the stratified variable is a dichotomous indicator of whether the subject has or does not have a history of hypertension. The predictor SMOKE remains in the model (shown at left).

PH assumption:

- Violated overall
- Assumed to hold within categories of stratified variable

The model allows for a violation of the PH assumption. However, within each category of hypertension, the PH assumption is assumed to hold.

ID	TIME	STATUS	SMOKE	HT
Barry	2	1	1	1
Gary	3	1	0	2
Harry	5	0	0	1
Larry	8	1	1	2

The data are shown on the left. The additional variable is HT which is the variable we wish to stratify on. Barry and Harry have a history of hypertension (coded HT=1) while Gary and Larry do not (coded HT=2).

HT = History of hypertension
(1=yes, 2=no)

Formulate likelihood in pieces:

The stratified Cox likelihood is formulated in pieces. Each piece represents a stratified category. Within each piece, the likelihood is formulated similarly as the likelihood formulated for a Cox PH model.

Data for HT=1

ID	TIME	STATUS	SMOKE	HT
Barry	2	1	1	1
Harry	5	0	0	1

Data for HT=2

The first piece (stratified category) is defined among observations in which HT=1. The second piece is defined among observations in which HT=2.

ID	TIME	STATUS	SMOKE	HT
Gary	3	1	0	2
Larry	8	1	1	2

Stratified Cox model:

$h_g(t) = h_{0g}(t)e^{\beta_1 SMOKE}$

g=1 history of hypertension
g=2 no history of hypertension

Here, for convenience, we again show you the SC model for these data on the left.

Among HT=1
Barry has only event
Barry and Harry at risk at TIME=2

$$L_1 = \frac{h_{01}e^{\beta_1}}{h_{01}e^{\beta_1} + h_{01}e^{0}}$$

For the first stratum (HT=1), Barry gets an event at TIME=2. Barry and Harry are in the risk set when Barry gets his event. Barry is the only event from this stratum as Harry is censored at TIME=5. Therefore the likelihood for this piece, L_1 (shown at left), contains one term. That one term has Barry's hazard in the numerator. Barry's and Harry's hazard are summed in the denominator.

Among HT=2
Gary and Larry both get events
Gary and Larry at risk at TIME=3
Larry at risk at TIME=8

$$L_2 = \frac{h_{02}e^{0}}{h_{02}e^{0} + h_{02}e^{\beta_1}} \times \frac{h_{02}e^{\beta_1}}{h_{02}e^{\beta_1}}$$

For the second stratum (HT=2), Gary gets an event at TIME=3. Gary and Larry are in the risk set when Gary gets his event. Larry gets an event at TIME=8 and is the only one in the risk set at the time of his event. Therefore, the likelihood for this piece, L_2 (shown at left), is a product of two factors. Each of the two factors corresponds to an event.

$$L = L_1 \times L_2$$
$$= \left[\frac{h_{01}e^{\beta_1}}{h_{01}e^{\beta_1} + h_{01}e^{0}}\right]\left[\frac{h_{02}e^{0}}{h_{02}e^{0} + h_{02}e^{\beta_1}} \times \frac{h_{02}e^{\beta_1}}{h_{02}e^{\beta_1}}\right]$$

The stratified Cox likelihood can be formulated by taking the product of each piece (L_1 and L_2). Each piece was formulated within a stratum.

The baseline hazard cancels in L

$$L = \left[\frac{e^{\beta_1}}{e^{\beta_2} + e^{0}}\right] \times \left[\frac{e^{0}}{e^{0} + e^{\beta_1}} \times \frac{e^{\beta_1}}{e^{\beta_1}}\right]$$

Notice that the baseline hazard cancels from the likelihood. As with the Cox PH model, the stratified likelihood is just determined by the order of events (not the baseline hazard).

$$L = \left[\frac{e^{\beta_1}}{e^{\beta_2} + e^{0}}\right] \times \left[\frac{e^{0}}{e^{0} + e^{\beta_1}} \times \frac{e^{\beta_1}}{e^{\beta_1}}\right]$$

Same β_1 in each piece

One additional point about this model is that the β_1 in the first piece of the likelihood (L_1) is the same β_1 that is in the second piece of the likelihood (L_2). In other words, this is a no interaction model. The effect of smoking (expressed as β_1) does not depend on hypertension status.

No interaction model:

Effect of smoking is same for those with or without hypertension

VIII. Summary

We now summarize the most important features of the stratified Cox (SC) model described in this presentation.

Stratified Cox (SC) model:

- stratification of predictors not satisfying PH assumption
- includes predictors satisfying PH
- does not include stratified variables

The SC model is a modification of the Cox PH model to allow for control by "stratification" of predictors not satisfying the PH assumption. Variables that are assumed to satisfy the assumption are included in the model as predictors; the stratified variables are not included in the model.

Computer Results

The computer results for a SC model provides essentially the same type of output as provided for a Cox PH model without stratification. An example of SC output using the remission data is shown here. The variables included as predictors in the model are listed in the first column followed by their estimated coefficients, standard errors, p-values, hazard ratio values, and 95% confidence limits. Such information cannot be provided for the variables being stratified, because these latter variables are not explicitly included in the model.

Stratified Cox regression
Analysis time _t: survt

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.390	0.338	0.000	4.016	2.072	7.783
RX	0.931	0.472	0.048	2.537	1.006	6.396

No. of subjects = 42 | Log likelihood = −57.560 | Stratified by sex

Hazard function for no-interaction stratified Cox model:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p]$$

$g = 1, 2, \ldots, k^*$, strata defined from Z^*

Z^* has k^* categories

$X_1, X_2, \ldots, X_p$ satisfy PH

The general hazard function form for the no-interaction stratified Cox model is shown here. This formula contains a subscript g that indicates the gth stratum, where the strata are different categories of the stratification variable Z^* and the number of strata equals k^*. Notice that the baseline hazard functions are different in each stratum.

Stratification variable Z^*:

- identify $Z_1, Z_2, \ldots, Z_k$ not satisfying PH
- categorize each Z
- form combinations of categories (strata)
- each combination is a stratum of Z^*

The variable Z^* is defined by first identifying the Z_i variables not satisfying the PH assumption. We then categorize each Z and form combinations of categories of each of the Z's. Each combination represents a different stratum making up the variable Z^*.

No-interaction model:
Same coefficients $\beta_1, \beta_2, \ldots, \beta_p$ for each g, i.e., Z^* does not interact with the X's.

The above model is designated as a "no-interaction" model because the β's in the model are the same for each subscript g. The no-interaction assumption means that the variables being stratified are assumed *not* to interact with the X's in the model.

$$\text{Different baselines} \left\{ \begin{array}{lcl} h_{01}(t) & \Rightarrow & \hat{S}_1(t) \\ h_{02}(t) & \Rightarrow & \hat{S}_2(t) \\ & \vdots & \\ h_{0k}(t) & \Rightarrow & \hat{S}_{k^*}(t) \end{array} \right\} \text{Different survival curves}$$

For the no-interaction model, the fitted SC model will yield different estimated survival curves for each stratum because the baseline hazard functions are different for each stratum.

$\widehat{HR}$ same for each stratum

However, because the coefficients of the X's are the same for each stratum, estimates of hazard ratios are the same for each stratum.

(Partial) likelihood function:

$$L = L_1 \times L_2 \times \ldots \times L_{k^*}$$

Regression coefficients in the SC model are estimated by maximizing a partial likelihood function that is obtained by multiplying likelihood functions for each stratum.

Stratified Cox model allowing interaction:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_{1g} X_1 + \beta_{2g} X_2 + \ldots + \beta_{pg} X_p]$$

$g = 1, 2, \ldots, k^*$, strata defined from Z^*.

In order to evaluate the no-interaction assumption, we must define an interaction model for comparison. One version of the interaction model is shown here. This version shows regression coefficients with different subscripts in different strata; that is, each β coefficient has a subscript g.

Alternative stratified Cox interaction model:

- uses product terms involving Z^*
- define $k^* - 1$ dummy variables from Z^*
- products of the form $Z_i^* \times X_j$

An alternative way to write the interaction model uses product terms involving the Z^* variable with each predictor. This model uses $k^* - 1$ dummy variables to distinguish the k^* categories of Z^*. Each of these dummy variables is included as a product term with each of the X's.

Testing the no-interaction assumption:

$LR = -2 \ln L_R - (-2 \ln L_F)$
R = reduced (no-interaction) model
F = full (interaction) model contains product terms
$LR \dot{\sim} \chi^2_{p(k^*-1)\text{df}}$ under H_0: no interaction

To evaluate the no-interaction assumption, we can perform a likelihood ratio test that compares the (reduced) no-interaction model to the (full) interaction model. The null hypothesis is that the no-interaction assumption is satisfied. The test statistic is given by the difference between the log-likelihood statistics for the no-interaction and interaction models. This statistic is approximately chi-square under the null hypothesis. The degrees of freedom is $p(k^* - 1)$ where p denotes the number of X's and k^* is the number of categories making up Z^*.

Presentation Complete!

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Survival Curves and the Log–Rank Test
3. The Cox Proportional Hazards Model and Its Characteristics
4. Evaluating the Proportional Hazards Assumption
✓5. The Stratified Cox Procedure

Next:

6. Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

This presentation is now complete. We suggest that the reader review this presentation using the detailed outline that follows. Then answer the practice exercises and the test that follow.

The next Chapter (6) is entitled "Extension of the Cox PH Model for Time-Dependent Variables." There we show how an "extended" Cox model can be used as an alternative to the stratified Cox model when one or more predictors do not satisfy the PH assumption. We also discuss more generally what is a time-dependent variable, and show how such a variable can be evaluated using an extended Cox model.

Detailed Outline

I. Preview (page 204)

A. Focus on how stratified Cox (SC) procedure is carried out:
- analysis of computer results from SC procedure;
- hazard function for SC model;
- stratifying on a single predictor versus two or more predictors;
- no-interaction versus interaction models.

II. An Example (pages 204–208)

A. Cox PH results for remission data yield $P(PH) = 0.031$ for SEX.

B. SC model used: control for SEX (stratified); include log WBC and *Rx* in model.

C. Analysis of *Rx* effect from stratified Cox results: $\widehat{HR} = 2.537$; 95% CI: (1.006,6.396); LR and Wald tests: $P < 0.05$.

D. Hazard model: $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 \log \text{WBC} + \beta_2 Rx]$, $g = 1,2$
- different baseline hazard functions and survival curves for females and males;
- same coefficients β_1 and β_2 for both females and males (no-interaction assumption);
- obtain estimates by maximizing partial likelihood $L = L_1 \times L_2$.

E. Graph of four adjusted survival curves for *Rx* (adjusted for log WBC).

III. The General Stratified Cox (SC) Model (pages 208–209)

A. $\boxed{h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p],\ g = 1, 2,\ldots,k^*}$

where the strata are defined from the stratification variable Z^*.

B. Z^* defined from $Z_1,Z_2,\ldots,Z_k$ variables that do not satisfy PH:
- categorize each Z_i
- form combinations of categories
- each combination is a stratum of Z^*

C. Different baseline hazard functions and survival curves for each stratum.

D. Assumes no interaction: same coefficients $\beta_1,\beta_2,\ldots,\beta_p$ for each g; i.e., Z^* does not interact with the X's; i.e., estimated HR is same for each stratum.

E. Obtain estimates by maximizing partial likelihood $L = L_1 \times L_2 \times \ldots \times L_{k^*}$, where L_i is likelihood for ith stratum.

IV. The No-Interaction Assumption and How to Test It (pages 210–216)

A. Assumes same coefficients $\beta_1, \beta_2, \ldots, \beta_p$ for each g.

B. Interaction model:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_{1g}X_1 + \beta_{2g}X_2 + \ldots + \beta_{pg}X_p],$$

$g = 1, 2, \ldots, k^*$ strata defined from Z^*.

C. Alternative stratified Cox interaction model:

- uses product terms involving Z^*
- define $k^* - 1$ dummy variables $Z_1^*, Z_2^*, \ldots, Z_{k^*-1}^*$ from Z^*
- products of the form $Z_i^* \times X_j$, where $i = 1, \ldots, k^* - 1; j = 1, \ldots, p$
- hazard function: $g = 1, 2, \ldots, k^*$ strata defined from Z^*

$$\begin{aligned} h_g(t,\boldsymbol{X}) = h_{0g}(t)\exp\big[&\beta_1 X_1 + \cdots + \beta_p X_p + \beta_{11}(Z_1^* \times X_1) \\ &+ \cdots + \beta_{p1}(Z_1^* \times X_p) + \beta_{12}(Z_2^* \times X_1) + \cdots + \beta_{p2}(Z_2^* \times X_p) \\ &+ \cdots + \beta_{1,k^*-1}(Z_{k^*-1}^* \times X_1) + \cdots + \beta_{p,k^*-1}(Z_{k^*-1}^* \times X_p)\big] \end{aligned}$$

D. Testing the no-interaction assumption: use LR statistic given by $LR = -2 \ln L_R - (-2 \ln L_F)$ where R = reduced (no interaction) model and F = full (interaction) model
$LR \dot{\sim} \chi^2_{p(k^*-1)\text{df}}$ under H_0: no interaction, i.e., $\beta_{11} = \beta_{12} = \ldots = \beta_{p,k^*-1} = 0$

V. A Second Example Involving Several Stratification Variables (pages 216–221)

A. Dataset "vets.dat" from Veteran's Administration Lung Cancer Trial; $n = 137$; survival time in days.

B. Variables are: treatment status, cell type (four types), performance status, disease duration, age, and prior therapy status.

C. Cox PH results indicate [using $P(PH)$] that cell type and performance status do not satisfy PH assumption.

D. Example stratifies on cell type and performance status using four categories of cell type and two categories of performance status, so that Z^* has $k^* = 8$ strata.

E. X's considered in model are treatment status and age.

F. Computer results for no-interaction model: estimated HR for effect of treatment adjusted for age and Z^* is 1.134 ($P = 0.548$); not significant.

G. Hazard function for no-interaction model:

$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 \text{ Age}]$,
$g = 1, 2, \ldots, 8$

H. Hazard function for interaction model:

$h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_{1g} \text{ Treatment} + \beta_{2g} \text{ Age}]$,
$g = 1, 2, \ldots, 8$

I. Alternative version of interaction model:

$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_1 \text{ Treatment} + \beta_2 \text{ Age} + \beta_{11}(Z_1^* \times \text{Treatment}) + \cdots + \beta_{17}(Z_7^* \times \text{Treatment}) + \beta_{21}(Z_1^* \times \text{Age}) + \cdots + \beta_{27}(Z_7^* \times \text{Age})]$,
$g = 1, 2, \ldots 8$
where Z_1^* = large cell (binary), Z_2^* = adeno cell (binary), Z_3^* = small cell (binary), Z_4^* = PSbin (binary), $Z_5^* = Z_1^* \times Z_4^*$, $Z_6^* = Z_2^* \times Z_4^*$, $Z_7^* = Z_3^* \times Z_4^*$

J. Demonstration that alternative interaction version (in item I) is equivalent to original interaction formulation (in item H) using computer results for the alternative version.

K. Test of no-interaction assumption:
- null hypothesis: $\beta_{11} = \beta_{12} = \ldots = \beta_{17} = 0$ and $\beta_{21} = \beta_{22} = \ldots = \beta_{27} = 0$
- $LR \dot{\sim} \chi^2_{14\text{df}}$ under H_0: no interaction
- $LR = 524.040 - 499.944 = 24.096$ $(P = 0.045)$

  Conclusion: Reject null hypothesis; interaction model is preferred.

VI. A Graphical View of the Stratified Cox Approach (pages 221–222)
Comparison of log–log survival curves
1. Describe interaction of Rx and Sex.
2. Describe violation of PH assumption for Sex.

VII. The Stratified Cox Likelihood (pages 223–225)

VIII. Summary (pages 225–227)

Practice Exercises

The following questions derive from the dataset **vets.dat** concerning the Veteran's Administration Lung Cancer Trial that we previously considered in this chapter. Recall that survival times are in days and that the study size contains 137 patients. The exposure variable of interest is treatment status (standard = 1, test = 2). Other variables of interest as control variables are cell type (four types, defined in terms of dummy variables), performance status, disease duration, age, and prior therapy status. Failure status is defined by the status variable (0 = censored, 1 = died).

1. Consider the following two edited printouts obtained from fitting a Cox PH model to these data.

Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]	*P*(*PH*)
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Perf.Stat	−0.033	0.006	0.000	0.968	0.958	0.978	0.000
Dis.Durat.	0.000	0.009	0.992	1.000	0.982	1.018	0.919
Age	−0.009	0.009	0.358	0.991	0.974	1.010	0.198
Pr.Therapy	0.007	0.023	0.755	1.007	0.962	1.054	0.145

No. of subjects = 137 Log likelihood = −475.180

Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]	*P*(*PH*)
Treatment	0.298	0.197	0.130	1.347	0.916	1.981	0.739
Small cell	0.392	0.210	0.062	1.481	0.981	2.235	0.382
Perf.Stat	−0.033	0.005	0.000	0.968	0.958	0.978	0.000
Dis.Durat.	−0.001	0.009	0.887	0.999	0.981	1.017	0.926
Age	−0.006	0.009	0.511	0.994	0.976	1.012	0.211
Pr.Therapy	−0.003	0.023	0.884	0.997	0.954	1.042	0.146

No. of subjects = 137 Log likelihood = −487.770

How do the printouts differ in terms of what the *P*(*PH*) information says about which variables do not satisfy the PH assumption?

2. Based on the above information, if you were going to stratify on the cell type variable, how would you define the strata? Explain.

3. Consider a stratified analysis that stratifies on the variables Z_1 = "small cell" and Z_2 = "performance status." The small cell variable is one of the dummy variables for cell type defined above. The performance status variable is dichotomized into high (60 or above) and low (below 60) and is denoted as PSbin. The stratification variable which combines categories from Z_1 and Z_2 is denoted as SZ^* and consists of four categories. The predictors included (but not stratified) in the analysis are treatment status, disease duration, age, and prior therapy. The computer results are as follows:

Stratified Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Treatment	0.090	0.197	0.647	1.095	0.744	1.611
Dis.Durat.	0.000	0.010	0.964	1.000	0.982	1.019
Age	0.002	0.010	0.873	1.002	0.983	1.021
Pr.Therapy	−0.010	0.023	0.656	0.990	0.947	1.035

No. of subjects = 137 Log likelihood = −344.848 Stratified by SZ^*

Based on these results, describe the point and interval estimates for the hazard ratio for the treatment effect adjusted for the other variables, including SZ^*. Is this hazard ratio meaningfully and/or statistically significant? Explain.

4. State the form of the hazard function for the model being fit in question 3. Why does this model assume no interaction between the stratified variables and the predictors in the model?
5. State two alternative ways to write the hazard function for an "interaction model" that allows for the interaction of the stratified variables with the treatment status variable, but assumes no other type of interaction.
6. State two alternative versions of the hazard function for an interaction model that allows for the interaction of the stratified variables (small cell and performance status) with each of the predictors treatment status, disease duration, age, and prior therapy.
7. For the interaction model described in question 6, what is the formula for the hazard ratio for the effect of treatment adjusted for the other variables? Does this formula give a different hazard ratio for different strata? Explain.

8. State two alternative versions of the null hypothesis for testing whether the no-interaction assumption is satisfied for the stratified Cox model. Note that one of these versions should involve a set of regression coefficients being set equal to zero.
9. State the form of the likelihood ratio statistic for evaluating the no-interaction assumption. How is this statistic distributed under the null hypothesis, and with what degrees of freedom?
10. Provided below are computer results for fitting the interaction model described in question 6. In this printout the variable Z_1^* denotes the small cell variable and the variable Z_2^* denotes the PSbin variable. The variable DDZ_1^* denotes the product of Z_1^* with disease duration, and other product terms are defined similarly.

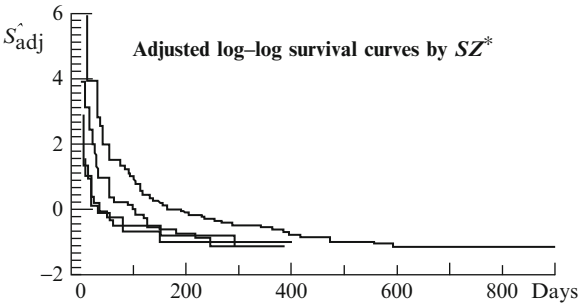
Stratified Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Treatment	0.381	0.428	0.374	1.464	0.632	3.389
Dis.Durat.	0.015	0.021	0.469	1.015	0.975	1.057
Age	0.000	0.017	0.994	1.000	0.968	1.033
Pr.Therapy	0.023	0.041	0.571	1.023	0.944	1.109
DDZ_1^*	−0.029	0.024	0.234	0.971	0.926	1.019
$AgeZ_1^*$	−0.055	0.037	0.135	0.946	0.880	1.018
PTZ_1^*	0.043	0.075	0.564	1.044	0.901	1.211
DDZ_2^*	0.025	0.032	0.425	1.026	0.964	1.092
$AgeZ_2^*$	0.001	0.024	0.956	1.001	0.956	1.049
PTZ_2^*	−0.078	0.054	0.152	0.925	0.831	1.029
$DDZ_1\ Z_2^*$	−0.071	0.059	0.225	0.931	0.830	1.045
$AgeZ_1Z_2^*$	0.084	0.049	0.084	1.088	0.989	1.196
$PTZ_1Z_2^*$	−0.005	0.117	0.963	0.995	0.791	1.250
trZ_1^*	0.560	0.732	0.444	1.751	0.417	7.351
trZ_2^*	−0.591	0.523	0.258	0.554	0.199	1.543
$trZ_1Z_2^*$	−0.324	0.942	0.731	0.723	0.114	4.583

No. of subjects = 137 Log likelihood = −335.591 Stratified by SZ^*

Use the above computer results to state the form of the **estimated** hazard model for each of the four strata of the stratification variable SZ^*. Also, for each strata, compute the hazard ratio for the treatment effect adjusted for disease duration, age, and prior therapy.

11. Carry out the likelihood ratio test to evaluate the no-interaction model described in question 4. In carrying out this test, make sure to state the null hypothesis in terms of regression coefficients being set equal to zero in the interaction model fitted in question 10. Also, determine the p-value for this test and state your conclusions about significance as well as which model you prefer, the no-interaction model or the interaction model.
12. The adjusted log–log survival curves for each of the four strata defined by the stratification variable SZ^* (adjusted for treatment status, disease duration, age, and prior therapy) are presented below.



Using this graph, what can you conclude about whether the PH assumption is satisfied for the variables, small cell type and PSbin?
13. Comment on what you think can be learned by graphing adjusted survival curves that compare the two treatment groups for each of the four strata of SZ^*.

Test

The following questions consider a dataset from a study by Caplehorn et al. ("Methadone Dosage and Retention of Patients in Maintenance Treatment," *Med. J. Aust.*, 1991). These data comprise the times in days spent by heroin addicts from entry to departure from one of two methadone clinics. Two other covariates, namely, prison record and maximum methadone dose, are believed to affect the survival times. The dataset name is **addicts.dat**. A listing of the variables is given below:

Column 1: Subject ID
Column 2: Clinic (1 or 2)

Column 3: Survival status (0 = censored, 1 = departed from clinic)
Column 4: Survival time in days
Column 5: Prison record (0 = none, 1 = any)
Column 6: Maximum methadone dose (mg/day)

1. The following edited printout was obtained from fitting a Cox PH model to these data:

Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]	*P*(*PH*)
Clinic	−1.009	0.215	0.000	0.365	0.239	0.556	0.001
Prison	0.327	0.167	0.051	1.386	0.999	1.924	0.332
Dose	−0.035	0.006	0.000	0.965	0.953	0.977	0.341

No. of subjects = 238 Log likelihood = −673.403

Based on the *P*(*PH*) information in the above printout, it appears that clinic does not satisfy the PH assumption; this conclusion is also supported by comparing log–log curves for the two clinics and noticing strong nonparallelism. What might we learn from fitting a stratified Cox (SC) model stratifying on the clinic variable? What is a drawback to using a SC procedure that stratifies on the clinic variable?

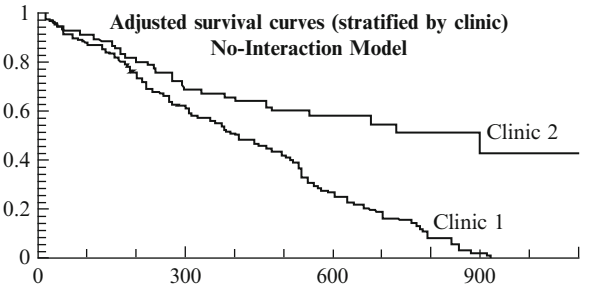
2. The following printout was obtained from fitting a SC PH model to these data, where the variable being stratified is clinic:

Stratified Cox regression
Analysis time _t:

survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Prison	0.389	0.169	0.021	1.475	1.059	2.054
Dose	−0.035	0.006	0.000	0.965	0.953	0.978

No. of subjects = 238 Log likelihood = −597.714 Stratified by clinic

Using the above fitted model, we can obtain the adjusted curves below that compare the adjusted survival probabilities for each clinic (i.e., stratified by clinic) adjusted for the variables, prison and maximum methadone dose.



Based on these adjusted survival curves, what conclusions can you draw about whether the survival experience is different between the two clinics? Explain.

3. State the hazard function model being estimated in the above computer results. Why is this model a no-interaction model?
4. Using the above computer results, provide point and interval estimates for the effect of prison adjusted for clinic and dose. Is this adjusted prison effect significant? Explain.
5. The following computer results consider a SC model that allows for interaction of the stratified variable clinic with each of the predictors, prison and dose. Product terms in the model are denoted as clinpr = clinic × prison and clindos = clinic × dose.

Stratified Cox regression
Analysis time _t: survt

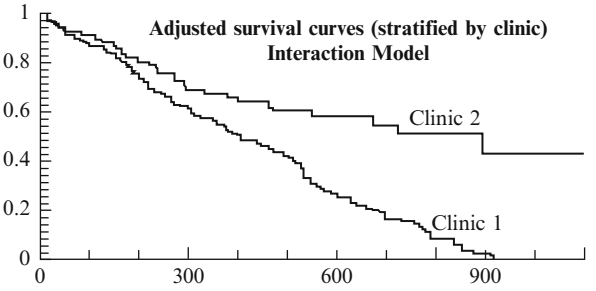
	Coef.	Std. Err.	P > |z|	Haz. Ratio	[95% Conf.	Interval]
prison	1.087	0.539	0.044	2.966	1.032	8.523
dose	−0.035	0.020	0.079	0.966	0.929	1.004
clinpr	−0.585	0.428	0.172	0.557	0.241	1.290
clindos	−0.001	0.015	0.942	0.999	0.971	1.028

No. of subjects = 238 Log likelihood = −596.779 Stratified by clinic

State two alternative versions of the interaction model being estimated by the above printout, where one of these versions should involve the product terms used in the above printout.

6. Using the computer results above, determine the estimated hazard models for each clinic. (Note that the clinics are coded as 1 or 2.)

7. below are the adjusted survival curves for each clinic based on the interaction model results above. These curves are adjusted for the prison and dose variables.



   Compare the survival curves by clinic obtained for the interaction model with the corresponding curves previously shown for the no-interaction model. Do both graphs indicate the similar conclusions about the clinic effect? Explain.
8. Carry out a likelihood ratio test to determine whether the no-interaction model is appropriate. In doing so, make use of the computer information described above, state the null hypothesis, state the form of the likelihood statistic and its distribution under the null hypothesis, and compute the value of the likelihood statistic and evaluate its significance. What are your conclusions?

Answers to Practice Exercises

1. The first printout indicates that the variables large cell, adeno cell, small cell, and performance status do not satisfy the PH assumption at the 0.10 level. The second printout considers a different model that does not contain the large cell and adeno cell variables. This latter printout indicates that small cell satisfies the PH assumption, in contrast to the first printout. The performance status variable, however, does not satisfy the PH assumption as in the first printout.
2. The cell type variable is defined to have four categories, as represented by the three dummy variables in the first printout. The "small cell" variable dichotomizes the cell type variable into the categories small cell type versus the rest. From the second printout, the small cell variable does not appear by itself to violate the PH assumption. This result conflicts with the results of the first printout, for which the cell type variable considered in four categories does not satisfy the PH assumption at the 0.10 level of significance.

We therefore think it is more appropriate to use a SC procedure only if four strata are to be used. A drawback to using four strata, however, is that the number of survival curves to be plotted is larger than for two strata; consequently, a large number of curves is more difficult to interpret graphically than when there are only two curves. Thus, for convenience of interpretation, we may choose to dichotomize the cell type variable instead of considering four strata. We may also consider dichotomies other than those defined by the small cell variable. For instance, we might consider dichotomizing on either the adeno or large cell variables instead of the small cell variable. Alternatively, we may combine categories so as to compare, say, large and adeno cell types with small and squamous types. However, a decision to combine categories should not be just a statistical decision, but should also be based on biologic considerations.

3. $\widehat{HR}_{\text{adj}} = 1.095$, 95% CI: (0.744,1.611), two-tailed p-value is 0.647, not significant. The estimated hazard ratio for treatment is neither meaningfully nor statistically significant. The point estimate is essentially 1, which says that there is no meaningful effect of treatment adjusted for the predictors in the model and for the stratified predictor SZ^*.

4. $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 DD + \beta_3 \text{ Age} + \beta_4 PT]$, $g = 1,\ldots, 4$, where the strata are defined from the stratification variable SZ^*, DD = disease duration, and PT = prior therapy. This model assumes no interaction because the coefficient of each predictor in the model is not subscripted by g, i.e., the regression coefficients are the same for each stratum.

5. Version 1: $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_{1g} \text{ Treatment} + \beta_2 DD + \beta_3 \text{ Age} + \beta_4 PT]$, $g = 1,\ldots, 4$.

   Version 2: $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1 \text{ Treatment} + \beta_2 DD + \beta_3 \text{Age} + \beta_4 PT + \beta_5(Z_1^* \times \text{Treatment}) + \beta_6(Z_2^* \times \text{Treatment}) + \beta_7(Z_1^* \times Z_2^* \times \text{Treatment})]$, where $Z_1^* =$ small cell type (0, 1), $Z_2^* =$ Psbin (0, 1), and $g = 1,\ldots, 4$.

6. Version 1: $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_{1g}\,\text{Treatment} + \beta_{2g}\,DD + \beta_{3g}\,\text{Age} + \beta_{4g}\,PT]$, $g = 1,\ldots, 4$.

   Version 2: $h_g(t,\mathbf{X}) = h_{0g}(t)\exp[\beta_1\,\text{Treatment} + \beta_2 DD + \beta_3\,\text{Age} + \beta_4 PT + \beta_5(Z_1^* \times \text{Treatment}) + \beta_6(Z_1^* \times DD) + \beta_7(Z_1^* \times \text{Age}) + \beta_8(Z_1^* \times PT) + \beta_9(Z_2^* \times \text{Treatment}) + \beta_{10}(Z_2^* \times DD) + \beta_{11}(Z_2^* \times \text{Age}) + \beta_{12}(Z_2^* \times PT) + \beta_{13}(Z_1^* \times Z_2^* \times \text{Treatment}) + \beta_{14}(Z_1^* \times Z_2^* \times DD) + \beta_{15}(Z_1^* \times Z_2^* \times \text{Age}) + \beta_{16}(Z_1^* \times Z_2^* \times PT)]$, $g = 1,\ldots,4$.

7. $HR_g = \exp(\beta_{1g})$, using version 1 model form. Yes, this formula gives different hazard ratios for different strata because the value of the hazard ratio changes with the subscript g.

8. H_0: No interaction assumption is satisfied.

   H_0: $\beta_{11} = \beta_{12} = \beta_{13} = \beta_{14}$, $\beta_{21} = \beta_{22} = \beta_{23} = \beta_{24}$, $\beta_{31} = \beta_{32} = \beta_{33} = \beta_{34}$, $\beta_{41} = \beta_{42} = \beta_{43} = \beta_{44}$ from version 1.
   H_0: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0$ from version 2.

9. $LR = -2\ln L_R - (-2\ln L_F)$, where R denotes the reduced (no-interaction) model and F denotes the full (interaction) model. Under the null hypothesis, LR is approximately a chi-square with 12 degrees of freedom.

10. Estimated hazard models for each stratum:

    $g = 1;\ Z_1^* = Z_2^* = 0:$
    $\hat{h}_1,\mathbf{X}) = \hat{h}_{01}(t)\exp[(0.381)\text{Treatment} + (0.015)DD + (0.000)\text{Age} + (0.023)PT]$

    $g = 2;\ Z_1^* = 1,\ Z_2^* = 0:$
    $\hat{h}_2(t,\mathbf{X}) = \hat{h}_{02}(t)\exp[(0.941)\text{Treatment} + (-0.014)DD + (-0.055)\text{Age} + (0.066)PT]$

    $g = 3;\ Z_1^* = 0,\ Z_2^* = 1:$
    $\hat{h}_3(t,\mathbf{X}) = \hat{h}_{03}(t)\exp[(-0.210)\text{Treatment} + (0.040)DD + (0.001)\text{Age} + (-0.055)PT]$

    $g = 4;\ Z_1^* = 1,\ Z_2^* = 1:$
    $\hat{h}_4(t,\mathbf{X}) = \hat{h}_{04}(t)\exp[(0.026)\text{Treatment} + (-0.060)DD + (0.030)\text{Age} + (-0.017)PT]$

Estimated hazard ratios for treatment effect adjusted for DD, Age, and PT:

$$g = 1 : \widehat{HR}_1 = \exp(0.381) = 1.464$$
$$g = 2 : \widehat{HR}_2 = \exp(0.941) = 2.563$$
$$g = 3 : \widehat{HR}_3 = \exp(-0.210) = 0.811$$
$$g = 4 : \widehat{HR}_4 = \exp(0.026) = 1.026$$

11. H_0: $\beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0$

    $LR = 689.696 - 671.182 = 18.514$, which is approximately chi-square with 12 df.

    $P = 0.101$, which is not significant below the .05 level. *Conclusion*: Accept the null hypothesis and conclude that the no-interaction model is preferable to the interaction model.

12. The three curves at the bottom of the graph appear to be quite non-parallel. Thus, the PH assumption is not satisfied for one or both of the variables, small cell type and PSbin. Note, however, that because both these variables have been stratified together, it is not clear from the graph whether only one of these variables fails to satisfy the PH assumption.

13. If we graph adjusted survival curves that compare the two treatment groups for each of the four strata, we will be able to see graphically how the treatment effect, if any, varies over time within each strata. The difficulty with this approach, however, is that eight adjusted survival curves will be produced, so that if all eight curves are put on the same graph, it may be difficult to see what is going on.

6

Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

D.G. Kleinbaum and M.Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_6,
© Springer Science+Business Media, LLC 2012

Introduction

We begin by defining a time-dependent variable and providing some examples of such a variable. We also state the general formula for a Cox model that is extended to allow time dependent variables, followed by a discussion of the characteristics of this model, including a description of the hazard ratio.

In the remainder of the presentation, we give examples of models with time-dependent variables, including models that allow for checking the PH assumption for time-independent variables. In particular, we describe a method that uses "heaviside functions" to evaluate the PH assumption for time-independent variables. We also describe two computer applications of the extended Cox model, one concerning a study on the treatment of heroin addiction and the other concerning the Stanford heart transplant study.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Preview** (page 244)
II. **Review of the Cox PH Model** (pages 244–246)
III. **Definition and Examples of Time-Dependent Variables** (pages 246–249)
IV. **The Extended Cox Model for Time-Dependent Variables** (pages 249–251)
V. **The Hazard Ratio Formula for the Extended Cox Model** (pages 251–253)
VI. **Assessing Time-Independent Variables That Do Not Satisfy the PH Assumption** (pages 254–259)
VII. **An Application of the Extended Cox Model to An Epidemiologic Study on the Treatment of Heroin Addiction** (pages 260–264)
VIII. **An Application of the Extended Cox Model to the Analysis of the Stanford Heart Transplant Data** (pages 265–269)
IX. **The Extended Cox Likelihood** (pages 269–274)
X. **Summary** (pages 274–277)

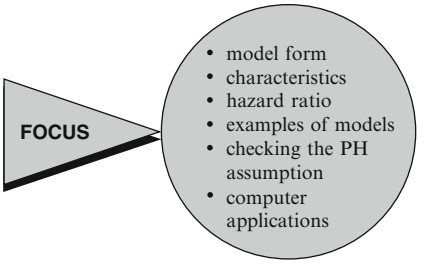
Objectives

Upon completing the chapter, the learner should be able to:

1. State or recognize the general form of the Cox model extended for time-dependent variables.
2. State the specific form of an extended Cox model appropriate for the analysis, given a survival analysis scenario involving one or more time-dependent variables.
3. State the formula for a designated hazard ratio of interest, given a scenario describing a survival analysis using an extended Cox model.
4. State the formula for an extended Cox model that provides a method for checking the PH assumption for one more of the time-independent variables in the model, given a scenario describing a survival analysis involving time-independent variables.
5. State the formula for an extended Cox model that uses one or more heaviside functions to check the PH assumption for one more of the time-independent variables in the model, given a scenario describing a survival analysis involving time-independent variables.
6. State the formula for the hazard ratio during different time interval categories specified by heaviside function(s) that are contained in an extended cox model.
7. Carry out an appropriate analysis of the data to evaluate the effect of one or more of the explanatory variables in the model(s) being used, given computer results for a survival analysis involving time-dependent variables. Such an analysis will involve:
   - computing and interpreting any hazard ratio(s) of interest;
   - carrying out and interpreting appropriate test(s) of hypotheses for effects of interest;
   - obtaining confidence intervals for hazard ratios of interest;
   - evaluating interaction and confounding involving one or more covariates.

Presentation

I. Preview



This presentation describes how the Cox proportional hazards (PH) model can be extended to allow time-dependent variables as predictors. Here, we focus on the model form, characteristics of this model, the formula for and interpretation of the hazard ratio, and examples of the extended Cox model. We also show how the extended Cox model can be used to check the PH assumption for time-independent variables, and we provide computer applications to illustrate different types of time-dependent variables. Finally, we describe the extended cox likelihood and how it contrasts with the Cox PH likelihood function.

II. Review of the Cox PH Model

$$h(t, \mathbf{X}) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

$\mathbf{X} = (X_1, X_2, \ldots, X_p)$
Explanatory/predictor variables

The general form of the Cox PH model is shown here. This model gives an expression for the hazard at time t for an individual with a given specification of a set of explanatory variables denoted by the bold $\mathbf{X}$. That is, the bold $\mathbf{X}$ represents a collection (sometimes called a "vector") of predictor variables that is being modeled to predict an individual's hazard.

$$h_0(t) \times \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

Baseline hazard	Exponential
Involves t but not X's	Involves X's but not t (X's are time-independent)

The Cox model formula says that the hazard at time t is the product of two quantities. The first of these, $h_0(t)$, is called the **baseline hazard** function. The second quantity is the exponential expression e to the linear sum of $\beta_i X_i$, where the sum is over the p explanatory X variables.

An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t but does not involve the X's, whereas the exponential expression involves the X's but does not involve t. The X's here are called **time-independent** X's.

X's involving t: time-dependent

Requires extended Cox model (no PH)

It is possible, nevertheless, to consider X's that do involve t. Such X's are called **time-dependent** variables. If time-dependent variables are considered, the Cox model form may still be used, but such a model no longer satisfies the PH assumption and is called the **extended Cox model**. We will discuss time-dependent variables and the corresponding extended Cox model beginning in the next section.

Hazard ratio formula:

$$\widehat{HR} = \exp\left[\sum_{i=1}^{p} \hat{\beta}_i (X_i^* - X_i)\right]$$

where $\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_p^*\right)$ and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ denote the two sets of X's.

From the Cox PH model, we can obtain a general formula, shown here, for estimating a hazard ratio that compares two specifications of the X's, defined as $\mathbf{X}^*$ and $\mathbf{X}$.

PH assumption:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta} \text{ (a constant over } t)$$

i.e., $\hat{h}(t, \mathbf{X}^*) = \hat{\theta}\hat{h}(t, \mathbf{X})$

The (PH) assumption underlying the Cox PH model is that the hazard ratio comparing any two specifications of $\mathbf{X}$ predictors is constant over time. Equivalently, this means that the hazard for one individual is proportional to the hazard for any other individual, where the proportionality constant is independent of time.

Hazards cross $\Rightarrow$ PH not met

Hazards don't cross $\not\Rightarrow$ PH met

An example of when the PH assumption is not met is given by any study situation in which the hazards for two or more groups cross when graphed against time. However, even if the hazard functions do not cross, it is possible that the PH assumption is not met.

Three approaches:

- graphical
- time-dependent variables
- goodness-of-fit test

As described in more detail in Chapter 4, there are three general approaches for assessing the PH assumption. These are

- a graphical approach;
- the use of time-dependent variables in an extended Cox model; and
- the use of a goodness-of-fit test.

Time-dependent covariates used to assess PH for time-indep. X's

$\Downarrow$

Extend Cox model: add product term(s) involving function of t

When time-dependent variables are used to assess the PH assumption for a time-independent variable, the Cox model is extended to contain **product** (i.e., interaction) **terms** involving the time-independent variable being assessed and some function of time.

EXAMPLE

$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 \text{sex} + \beta_2(\text{sex} \times t)]$
H_0: $\beta_2 = 0 \Rightarrow$ PH assumption satisfied

For example, if the PH assumption is being assessed for sex, a Cox model might be extended to include the variable sex $\times$ t in addition to sex. If the coefficient of the product term turns out to be non-significant, we can conclude that the PH assumption is satisfied for sex provided that the variable sex $\times$ t is an appropriate choice of time-dependent variable.

Options when PH assumption not satisfied:

- Use a stratified Cox (SC) model.
- Use time-dependent variables.

There are two options to consider if the PH assumption is not satisfied for one or more of the predictors in the model. In Chapter 5, we described the option of using a stratified Cox (SC) model, which stratifies on the predictor(s) not satisfying the PH assumption, while keeping in the model those predictors that satisfy the PH assumption. In this chapter, we describe the other option, which involves using time-dependent variables.

Time-dependent variables may be:

- inherently time-dependent
- defined to analyze a time-independent predictor not satisfying the PH assumption.

Note that a given study may consider predictors that are inherently defined as time-dependent, as we will illustrate in the next section. Thus, in addition to considering time-dependent variables as an option for analyzing a time-independent variable not satisfying the PH assumption, we also discuss predictors which are inherently defined as time-dependent.

III. Definition and Examples of Time-Dependent Variables

Definition:

A time-dependent variable is defined as any variable whose value for a given subject may differ over time (t). In contrast, a time-independent variable is a variable whose value for a given subject remains constant over time.

Time-dependent	Time-independent
Value of variable differs over time	Value of variable is constant over time
Example:	
Race $\times$ t	Race

As a simple example, the variable RACE is a time-independent variable, whereas the variable RACE $\times$ time is a time-dependent variable.

EXAMPLE OF DEFINED VARIABLES

Defined variable: RACE $\times$ t

Time-independent

Race $= 1 \Rightarrow$ Race $\times\, t = t$

Race $= 0 \Rightarrow$ Race $\times\, t = 0$ (at any t)

$E \times \underbrace{(\log t - 3)}$

Function of t

[E denotes a (0,1) exposure variable].

$E \times g(t)$ where $g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$

Heaviside function

$t \geq t_0 : E \times \overbrace{g(t)}^{\boxed{1}} = E$

$t < t_0 : E \times \overbrace{g(t)}^{\boxed{0}} = 0$

Heaviside functions used when PH assumptions not met.

The variable RACE $\times$ time is an example of what is called a "defined" time-dependent variable. Most defined variables are of the form of the product of a time-independent variable (e.g., RACE) multiplied by time or some function of time. Note that after RACE is determined for a given subject, all the values of the RACE $\times$ time variable are completely defined over a specified time interval of study.

A second example of a defined variable is given by $E \times (\log t - 3)$, where E denotes, say, a (0,1) exposure status variable determined at one's entry into the study. Notice that here we have used a function of time, $\log t - 3$, rather than time t alone.

Yet another example of a defined variable, which also involves a function of time, is given by $E \times g(t)$, where $g(t)$ is defined to take on the value 1 if t is greater than or equal to some specified value of t, called t_0, and takes on the value 0 if t is less than t_0.

The function $g(t)$ is called a "heaviside" function. Note that whenever t is greater than or equal to t_0, $g(t)$ equals 1, so $E \times g(t) = E$; however, whenever t is less than t_0, $g(t) = 0$, so the value of $E \times g(t)$ is always 0. We will later return to illustrate how heaviside functions may be used as one method for the analysis when a time-independent variable like E does not satisfy the proportional hazards assumption.

Internal variable:

EXAMPLES OF INTERNAL VARIABLES

$E(t)$, $EMP(t)$, $SMK(t)$, $OBS(t)$,

Values change because of "internal" characteristics or behavior of the individual.

Another type of time-dependent variable is called an "internal" variable. Examples of such a variable include exposure level E at time t, employment status (EMP) at time t, smoking status (SMK) at time t, and obesity level (OBS) at time t.

All these examples consider variables whose values may change over time for any subject under study; moreover, for internal variables, the reason for a change in value depends on "internal" characteristics or behavior specific to the individual.

"Ancillary" variable:
Value changes because of "external" characteristics.

In contrast, a variable is called an "ancillary" variable if its value changes primarily because of "external" characteristics of the environment that may affect several individuals simultaneously. An example of an ancillary variable is air pollution index at time t for a particular geographical area. Another example is employment status (EMP) at time t, if the primary reason for whether someone is employed or not depends more on general economic circumstances than on individual characteristics.

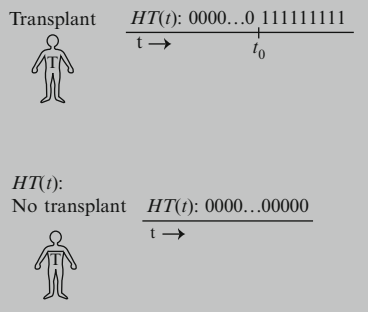
EXAMPLES OF ANCILLARY VARIABLES

Air pollution index at time t; $EMP(t)$

ANOTHER EXAMPLE

Heart transplant status at time t:

$$HT(t) = \begin{cases} 1 & \text{if received transplant at some time } t_0 \le t \\ 0 & \text{if did not receive transplant by time } t \end{cases}$$



As another example, which may be part internal and part ancillary, we consider heart transplant status (HT) at time t for a person identified to have a serious heart condition, making him or her eligible for a transplant. The value of this variable HT at time t is 1 if the person has already received a transplant at some time, say t_0, prior to time t. The value of HT is 0 at time t if the person has not yet received a transplant by time t.

Note that once a person receives a transplant, at time t_0, the value of HT remains at 1 for all subsequent times. Thus, for a person receiving a transplant, the value of HT is 0 up to the time of transplant, and then remains at 1 thereafter. In contrast, a person who never receives a transplant has HT equal to 0 for all times during the period he or she is in the study.

The variable "heart transplant status," $HT(t)$, can be considered essentially an internal variable, because individual traits of an eligible transplant recipient are important determinants of the decision to carry out transplant surgery. Nevertheless, the availability of a donor heart prior to tissue and other matching with an eligible recipient can be considered an "ancillary" characteristic external to the recipient.

Computer commands differ for defined vs. internal vs. ancillary.

But, the form of extended Cox model and procedures for analysis are the same regardless of variable type.

The primary reason for distinguishing among defined, internal, or ancillary variables is that the computer commands required to define the variables for use in an extended Cox model are somewhat different for the different variable types, depending on the computer program used. Nevertheless, the form of the extended Cox model is the same regardless of variable type, and the procedures for obtaining estimates of regression coefficients and other parameters, as well as for carrying out statistical inferences, are also the same.

IV. The Extended Cox Model for Time-Dependent Variables

$$h(t,\mathbf{X}(t)) = h_0(t)\exp\left[\sum_{i=1}^{p_1}\beta_i X_i + \sum_{j=1}^{p_2}\delta_j X_j(t)\right]$$

$$\mathbf{X}(t) = \underbrace{(X_1, X_2, \ldots X_{p_1}}_{\text{Time-independent}}, \underbrace{X_1(t), X_2(t), \ldots X_{p_2}(t))}_{\text{Time-dependent}}$$

Given a survival analysis situation involving both time-independent and time-dependent predictor variables, we can write the extended Cox model that incorporates both types as shown here at the left. As with the Cox PH model, the extended model contains a baseline hazard function $h_0(t)$ which is multiplied by an exponential function. However, in the extended model, the exponential part contains both time-independent predictors, as denoted by the X_i variables, and time-dependent predictors, as denoted by the $X_j(t)$ variables. The entire collection of predictors at time t is denoted by the bold $\mathbf{X}(t)$.

EXAMPLE

$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta(E \times t)]$,
$p_1 = 1, p_2 = 1$,
$\mathbf{X}(t) = (X_1 = E, X_1(t) = E \times t)$

As a simple example of an extended Cox model, we show here a model with one time-independent variable and one time-dependent variable. The time-independent variable is exposure status E, say a (0,1) variable, and the time-dependent variable is the product term $E \times t$.

Estimating regression coefficients:
ML procedure:
Maximize (partial) L.
Risk sets more complicated than for PH model.

As with the simpler Cox PH model, the regression coefficients in the extended Cox model are estimated using a maximum likelihood (ML) procedure. ML estimates are obtained by maximizing a (partial) likelihood function L. However, the computations for the extended Cox model are more complicated than for the Cox PH model, because the risk sets used to form the likelihood function are more complicated with time-dependent variables. The extended Cox likelihood is described later in this chapter.

Computer programs for the extended Cox model:

Stata (Stcox)
SAS (PHREG)
SPSS (COXREG)
R
} Computer Appendix

Computer packages that include programs for fitting the extended Cox model include Stata, SAS, SPSS, and R. See the Computer Appendix at the end of this text for a comparison of the Stata, SAS, SPSS, and R procedures applied to the same dataset.

Statistical inferences:
Wald and/or *LR* tests
Large sample confidence intervals

Methods for making statistical inferences are essentially the same as for the PH model. That is, one can use Wald and/or likelihood ratio (*LR*) tests and large sample confidence interval methods.

Assumption of the model:
The hazard at time t depends on the value of $X_j(t)$ at that same time.

An important assumption of the extended Cox model is that the effect of a time-dependent variable $X_j(t)$ on the survival probability at time t depends on the value of this variable at that *same* time t, and not on the value at an earlier or later time.

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right]$$

One coefficient for $X_j(t)$

Note that even though the values of the variable $X_j(t)$ may change over time, the hazard model provides only one coefficient for each time-dependent variable in the model. Thus, at time t, there is only one value of the variable $X_j(t)$ that has an effect on the hazard, that value being measured at time t.

Can modify for lag-time effect

It is possible, nevertheless, to modify the definition of the time-dependent variable to allow for a "lag-time" effect.

Lag-time effect:

EXAMPLE

$EMP(t)$ = employment status at week t

Model without lag-time:
$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta EMP(t)]$
↑ Same week

Model with 1-week lag-time:
$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta^* EMP(t-1)]$
↑ One-week earlier

To illustrate the idea of a lag-time effect, suppose, for example, that employment status, measured weekly and denoted as $EMP(t)$, is the time-dependent variable being considered. Then, an extended Cox model that does *not* consider lag-time assumes that the effect of employment status on the probability of survival at week t depends on the observed value of this variable at the same week t, and not, for example, at an earlier week.

However, to allow for, say, a time-lag of one week, the employment status variable may be modified so that the hazard model at time t is predicted by the employment status at week $t-1$. Thus, the variable $EMP(t)$ is replaced in the model by the variable $EMP(t-1)$.

General lag-time extended model:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp\left[\sum_{i=1}^{p_1}\beta_i X_i + \sum_{j=1}^{p_2}\delta_j X_j\big(t-L_j\big)\right]$$

$X_j(t-L_j)$ replaces $X_j(t)$

More generally, the extended Cox model may be alternatively written to allow for a lag-time modification of any time-dependent variable of interest. If we let L_j denote the lag-time specified for time-dependent variable j, then the general "lag-time extended model" can be written as shown here. Note that the variable $X_j(t)$ in the earlier version of the extended model is now replaced by the variable $X_j(t - L_j)$.

V. The Hazard Ratio Formula for the Extended Cox Model

PH assumption is not satisfied for the extended Cox model.

We now describe the formula for the hazard ratio that derives from the extended Cox model. The most important feature of this formula is that the proportional hazards assumption is no longer satisfied when using the extended Cox model.

$$\widehat{HR}(t) = \frac{\hat{h}(t,\mathbf{X}^*(t))}{\hat{h}(t,\mathbf{X}(t))} = \exp\left[\sum_{i=1}^{p_1}\hat{\beta}_i\left[X_i^* - X_i\right] + \sum_{j=1}^{p_2}\delta_j\left[X_j^*(t) - X_j(t)\right]\right]$$

The general hazard ratio formula for the extended Cox model is shown here. This formula describes the ratio of hazards at a particular time t, and requires the specification of two sets of predictors at time t. These two sets are denoted as bold $\mathbf{X}^*(t)$ and bold $\mathbf{X}(t)$.

Two sets of predictors:

$$\mathbf{X}^*(t) = \left(X_1^*, X_2^*, \ldots, X_{p_1}^*, X_1^*(t), X_2^*(t), \ldots, X_{p_2}^*(t)\right)$$

$$\mathbf{X}(t) = \left(X_1, X_2, \ldots, X_{p_1}, X_1(t), X_2(t), \ldots, X_{p_2}(t)\right)$$

The two sets of predictors, $\mathbf{X}^*(t)$ and $\mathbf{X}(t)$, identify two specifications at time t for the combined set of predictors containing both time-independent and time-dependent variables. The individual components for each set of predictors are shown here.

EXAMPLE

$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta(E \times t)]$

$$E = \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if unexposed} \end{cases}$$

$\mathbf{X}^*(t) = (E = 1, E \times t = t)$
$\mathbf{X}(t) = (E = 0, E \times t = 0)$

$$\widehat{HR}(t) = \frac{\hat{h}(t,E=1)}{\hat{h}(t,E=0)} = \exp\left[\hat{\beta}(1-0) + \hat{\delta}((1\times t) - (0\times t))\right] = \exp\left[\hat{\beta} + \hat{\delta}t\right]$$

$\hat{\delta}_0 \Rightarrow \widehat{HR}(t) \uparrow as\ t \uparrow$
PH assumption not satisfied

As a simple example, suppose the model contains only one time-independent predictor, namely, exposure status E, a (0,1) variable, and one time-dependent predictor, namely, $E \times t$. Then, to compare exposed persons, for whom $E = 1$, with unexposed persons, for whom $E = 0$, at time t, the bold $\mathbf{X}^*(t)$ set of predictors has as its two components $E = 1$ and $E \times t = t$; the bold $\mathbf{X}(t)$ set has as its two components $E = 0$ and $E \times t = 0$.

If we now calculate the estimated hazard ratio that compares exposed to unexposed persons at time t, we obtain the formula shown here; that is, *HR* "hat" equals the exponential of β "hat" plus δ "hat" times t. This formula says that the hazard ratio is a function of time; in particular, if δ "hat" is positive, then the hazard ratio increases with increasing time. Thus, the hazard ratio in this example is certainly not constant, so that the PH assumption is not satisfied for this model.

$$\widehat{HR}(t) = \exp\left[\sum_{i=1}^{p_1} \hat{\beta}_i\left[X_i^* - X_i\right] + \sum_{j=1}^{p_2} \boxed{\hat{\delta}_j}\boxed{\left[X_j^*(t) - X_j(t)\right]}\right]$$

A function of time

More generally, because the general hazard ratio formula involves differences in the values of the time-dependent variables at time t, this hazard ratio is a function of time. Thus, in general, the extended Cox model does not satisfy the PH assumption if any δ_j is not equal to zero.

In general, PH assumption not satisfied for extended Cox model.

$\hat{\delta}_j$ is not time-dependent.
$\hat{\delta}_j$ represents "overall" effect of $X_j(t)$.

Note that, in the hazard ratio formula, the coefficient δ_j "hat" of the difference in values of the jth time-dependent variable is itself not time-dependent. Thus, this coefficient represents the "overall" effect of the corresponding time-dependent variable, considering all times at which this variable has been measured in the study.

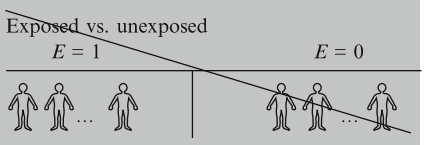
EXAMPLE

$E(t)$ = chemical exposure status at time t (weekly)

$$= \begin{cases} 0 & \text{if unexposed at time } t \\ 1 & \text{if exposed at time } t \end{cases}$$

$E(t)$	0 1 0 1 1
t	1 2 3 4 5…

$E(t)$	1 1 0 1 0
t	1 2 3 4 5…



$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta E(t)]$$

One coefficient

δ represents the overall effect of $E(t)$.

$$\widehat{HR}(t) = \frac{\hat{h}(t, E(t)=1)}{\hat{h}(t, E(t)=0)} = \exp\left[\hat{\delta}[1-0]\right] = e^{\hat{\delta}}, \text{ a fixed number}$$

But, PH is *not* satisfied:
$\widehat{HR}(t)$ is time-dependent because $E(t)$ is time-dependent.

As another example to illustrate the formula for the hazard ratio, consider an extended Cox model containing only one variable, say a weekly measure of chemical exposure status at time t. Suppose this variable, denoted as $E(t)$, can take one of two values, 0 or 1, depending on whether a person is unexposed or exposed, respectively, at a given weekly measurement.

As defined, the variable $E(t)$ can take on different patterns of values for different subjects. For example, for a 5-week period, subject A's values may be 01011, whereas subject B's values may be 11010.

Note that in this example, we do not consider two separate groups of subjects, with one group always exposed and the other group always unexposed throughout the study. This latter situation would require a (0,1) time-independent variable for exposure, whereas our example involves a time-dependent exposure variable.

The extended Cox model that includes only the variable $E(t)$ is shown here. In this model, the values of the exposure variable may change over time for different subjects, but there is only one coefficient, δ, corresponding to the one variable in the model. Thus, δ represents the overall effect on survival time of the time-dependent variable $E(t)$.

Notice, also, that the hazard ratio formula, which compares an exposed person to an unexposed person at time t, yields the expression e to the δ "hat."

Although this result is a fixed number, the PH assumption is not satisfied. The fixed number gives the hazard ratio at a given time, assuming that the exposure status at that time is 1 in the numerator and is 0 denominator. Thus, the hazard ratio is time-dependent, because exposure status is time-dependent, even though the formula yields a single fixed number.

VI. Assessing Time-Independent Variables That Do Not Satisfy the PH Assumption

Use an extended Cox model to

- check PH assumption;
- assess effect of variable not satisfying PH assumption.

We now discuss how to use an extended Cox model to check the PH assumption for time-independent variables and to assess the effect of a variable that does not satisfy the PH assumption.

Three methods for checking PH assumption:

1. graphical
2. extended Cox model
3. GOF test

As described previously (see Chapter 4), there are three methods commonly used to assess the PH assumption: (1) graphical, using, say, log–log survival curves; (2) using an extended Cox model; and (3) using a goodness-of-fit (GOF) test. We have previously (in Chapter 4) discussed items 1 and 3, but only briefly described item 2, which we focus on here.

Cox PH model for p time-independent X's:

$$h(t, \mathbf{X}) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

If the dataset for our study contains several, say p, time-independent variables, we might wish to fit a Cox PH model containing each of these variables, as shown here.

Extended Cox model:
Add product terms of the form: $X_i \times g_i(t)$

However, to assess whether such a PH model is appropriate, we can extend this model by defining several product terms involving each time-independent variable with some function of time. That is, if the ith time-independent variable is denoted as X_i, then we can define the ith product term as $X_i \times g_i(t)$ where $g_i(t)$ is some function of time for the ith variable.

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i X_i g_i(t)\right]$$

The extended Cox model that simultaneously considers all time-independent variables of interest is shown here.

EXAMPLE

$g_i(t) = 0$ for all i implies no time-dependent variable involving X_i, i.e.,

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

In using this extended model, the crucial decision is the form that the functions $g_i(t)$ should take. The simplest form for $g_i(t)$ is that all $g_i(t)$ are identically 0 at any time; this is another way of stating the original PH model, containing no time-dependent terms.

EXAMPLE 2

$g_i(t) = t \Rightarrow X_i g(t) = X_\mathrm{i} \times \mathrm{t}$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i (X_i \times t)\right]$$

Another choice for the $g_i(t)$ is to let $g_i(t) = t$. This implies that for each X_i in the model as a main effect, there is a corresponding time-dependent variable in the model of the form $X_\mathrm{i} \times \mathrm{t}$. The extended Cox model in this case takes the form shown here.

EXAMPLE 3: one variable at a time

$$X_L \text{ only} \Rightarrow \begin{cases} g_L(t) = t, \\ g_i(t) = 0 \text{ for other } i \end{cases}$$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{\substack{\mathrm{i}=1 \\ \mathrm{i} \neq \mathrm{L}}}^{p} \beta_i \mathbf{X}_i + \delta_L (\mathbf{X}_L \times t)\right]$$

Suppose, however, we wish to focus on a particular time-independent variable, say, variable X_L. Then $g_i(t) = t$ for $i = L$, but equals 0 for all other i. The corresponding extended Cox model would then contain only one product term $X_L \times \mathrm{t}$, as shown here.

EXAMPLE 4

$g_i(t) = \ln t \Rightarrow X_i g(t) = X_i \times \ln t$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i (X_i \times \ln t)\right]$$

Another choice for the $g_i(t)$ is the log of t, rather than simply t, so that the corresponding time-dependent variables will be of the form $X_i \times \ln t$.

EXAMPLE 5: Heaviside function

$$g_i(t) = \begin{cases} 0 & \text{if } t \geq t_0 \\ 1 & \text{if } t > \leq t_0 \end{cases}$$

And yet another choice would be to let $g_i(t)$ be a "heaviside function" of the form $g_i(t) = 1$ when t is at or above some specified time, say t_0, and $g_i(t) = 0$ when t is below t_0. We will discuss this choice in more detail shortly.

Extended Cox model:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i X_i g_i(t)\right]$$

Given a particular choice of the $g_i(t)$, the corresponding extended Cox model, shown here again in general form, may then be used to check the PH assumption for the time-independent variables in the model. Also, we can use this extended Cox model to obtain a hazard ratio formula that considers the effects of variables not satisfying the PH assumption.

- Check PH assumption.
- Obtain hazard ratio when PH assumption not satisfied.

H_0: $\delta_1 = \delta_2 = \ldots = \delta_p = 0$

To check the PH assumption using a statistical test, we consider the null hypothesis that all the δ terms, which are coefficients of the $X_i g_i(t)$ product terms in the model, are zero.

Under H_0, the model reduces to PH model:

$$h(t, \mathbf{X}) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i\right]$$

Under this null hypothesis, the model reduces to the PH model.

$$LR = -2 \ln L_{\text{PH model}} - (-2 \ln L_{\text{ext. Cox model}}) \dot{\sim} \chi^2_p \text{ under } H_0$$

This test can be carried out using a likelihood ratio (LR) test which computes the difference between the log likelihood statistic, $-2 \ln L$, for the PH model and the log likelihood statistic for the extended Cox model. The test statistic thus obtained has approximately a chi-square distribution with p degrees of freedom under the null hypothesis, where p denotes the number of parameters being set equal to zero under H_0.

EXAMPLE

$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta(E \times t)]$
H_0: $\delta = 0$ (i.e., PH assumption is satisfied)

Reduced model:
$h(t,\mathbf{X}) = h_0(t)\exp[\beta E]$

$LR = -2 \ln L_R - (-2 \ln L_F) \dot{\sim} \chi^2$ with 1 df under H_0

F = full (extended), R = reduced (PH)

As an example of this test, suppose we again consider an extended Cox model that contains the product term $E \times t$ in addition to the main effect of E, where E denotes a (0,1) time-independent exposure variable.

For this model, a test for whether or not the PH assumption is satisfied is equivalent to testing the null hypothesis that $\delta = 0$. Under this hypothesis, the reduced model is given by the PH model containing the main effect E only. The likelihood ratio statistic, shown here as the difference between log-likelihood statistics for the full (i.e., extended model) and the reduced (i.e., PH) model, will have an approximate chi-square distribution with one degree of freedom in large samples.

SAS: **PHREG** fits both PH and extended Cox models.
Stata: **Stcox** fits both PH and extended Cox models.

Note that to carry out the computations for this test, two different types of models, a PH model and an extended Cox model, need to be fit. Nevertheless, some computer packages such as SAS and Stata use the same computer program to fit both models.

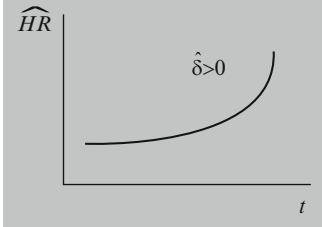
If PH test significant: Extended Cox model is preferred; HR is time-dependent.

If the result of the test for the PH assumption is significant, then the extended Cox model is preferred to the PH model. Thus, the hazard ratio expression obtained for the effect of an exposure variable of interest is time-dependent. That is, the effect of the exposure on the outcome cannot be summarized by a single HR value, but can only be expressed as a function of time.

EXAMPLE

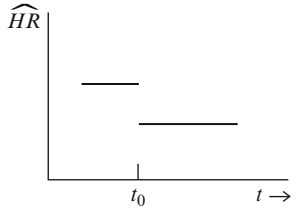
$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta(E \times t)]$

$\widehat{HR} = \exp\left[\hat{\beta} + \hat{\delta}t\right]$



We again consider the previous example, with the extended Cox model shown here. For this model, the estimated hazard ratio for the effect of exposure is given by the expression e to the quantity β "hat" plus δ "hat" times t. Thus, depending on whether δ "hat" is positive or negative, the estimated hazard ratio will increase or decrease exponentially as t increases. The graph shown here gives a sketch of how the hazard ratio varies with time if δ "hat" is positive.

Heaviside function:



We now provide a description of the use of a "heaviside" function. When such a function is used, the hazard ratio formula yields constant hazard ratios for different time intervals, as illustrated in the accompanying graph.

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t \leq t_0 \end{cases}$$

$h(t, \mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta E g(t)]$

Recall that a heaviside function is of the form $g(t)$, which takes on the value 1 if t is greater than or equal to some specified value of t, called t_0, and takes on the value 0 if t is less than t_0. An extended Cox model which contains a single heaviside function is shown here.

$t \geq t_0$: $g(t) = 1 \Rightarrow E \times g(t) = E$

$h(t, \mathbf{X}) = h_0(t)\exp[(\beta + \delta)E]$

$\widehat{HR} = \exp\left[\hat{\beta} + \hat{\delta}\right]$

Note that if $t \geq t_0$, $g(t) = 1$, so the value of $E \times g(t) = E$; the corresponding hazard function is of the form $h_0(t) \times e$ to the quantity $(\beta + \delta)$ times E, and the estimated hazard ratio for the effect of E has the form e to the sum of β "hat" plus δ "hat."

$t < t_0$: $g(t) = 0 \Rightarrow E \times g(t) = 0$

$h(t,\mathbf{X}) = h_0(t)\exp[\beta E]$

$\widehat{HR} = \exp\left[\hat{\beta}\right]$

If $t < t_0$, $g(t) = 0$, the corresponding hazard ratio is simplified to e to the β "hat."

A single heaviside function in the model

$h(t,\mathbf{X})$

$= h_0(t)\exp[\beta E + \delta(E \times g(t))]$

yields two hazard ratios:

$t \geq t_0$: $\widehat{HR} = \exp\left[\hat{\beta} + \hat{\delta}\right]$

$t \leq t_0$: $\widehat{HR} = \exp\left[\hat{\beta}\right]$

Thus, we have shown that the use of a single heaviside function results in an extended Cox model which gives two hazard ratio values, each value being constant over a fixed time interval.

Alternative model with two heaviside functions:

$$h(t,\mathbf{X}) = h_0(t)\exp[\delta_1(E \times g_1(t)) + \delta_2(E \times g_2(t))]$$

$$g_1(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t \leq t_0 \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } t \leq t_0 \\ 0 & \text{if } t \geq t_0 \end{cases}$$

There is actually an equivalent way to write this model that uses two heaviside functions in the same model. This alternative model is shown here. The two heaviside functions are called $g_1(t)$ and $g_2(t)$. Each of these functions are in the model as part of a product term with the exposure variable E. Note that this model does not contain a main effect term for exposure.

Note: Main effect for E not in model.

Two *HR*'s from the alternative model:

$$t \geq t_0 : \; g_1(t) = 1, \; g_2(t) = 0$$

$$h(t,\mathbf{X}) = h_0(t)\exp[\delta_1(E \times 1) + \delta_2(E \times 0)] = h_0(t)\exp[\delta_1 E]$$

so that $\widehat{HR} = \exp\left(\hat{\delta}_1\right)$

For this alternative model, as for the earlier model with only one heaviside function, two different hazard ratios are obtained for different time intervals. To obtain the first hazard ratio, we consider the form that the model takes when $t \geq t_0$. In this case, the value of $g_1(t)$ is 1 and the value of $g_2(t)$ is 0, so the exponential part of the model simplifies to $\delta_1 \times E$; the corresponding formula for the estimated hazard ratio is then e to the δ_1 "hat."

$$t > t_0 : g_1(t) = \; 0, \; g_2(t) = 1$$

$$h(t,\mathbf{X}) = h_0(t)\exp[\delta_1(E \times 0) + \delta_2(E \times 1)] = h_0(t)\exp[\delta_2 E]$$

so that $\widehat{HR} = \exp\left(\hat{\delta}_2\right)$

When $t < t_0$, the value of $g_1(t)$ is 0 and the value of $g_2(t)$ is 1. Then, the exponential part of the model becomes $\delta_2 \times E$, and the corresponding hazard ratio formula is e to the δ_2 "hat."

Alternative model:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta_1(E \times g_1(t)) + \delta_2(E \times g_2(t))]$$

Original model:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta(E \times g(t))]$$

$$t \geq t_0 : \widehat{HR} = \exp\left(\hat{\delta}_1\right) = \exp\left(\hat{\beta} + \hat{\delta}\right)$$

$$t \leq t_0 : \widehat{HR} = \exp\left(\hat{\delta}_2\right) = \exp\left(\hat{\beta}\right)$$

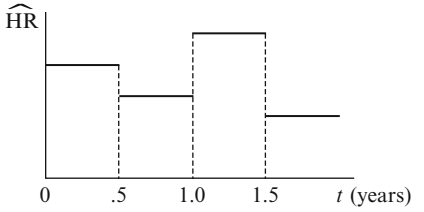
Thus, using the alternative model, again shown here, we obtain two distinct hazard ratio values. Mathematically, these are the same values as obtained from the original model containing only one heaviside function. In other words, δ_1 "hat" in the alternative model equals β "hat" plus δ "hat" in the original model (containing one heaviside function), and δ_2 "hat" in the alternative model equals β "hat" in the original model.

Heaviside functions:

- two $\widehat{HR}$'s constant within two time intervals
- *extension*: several $\widehat{HR}$'s constant within several time intervals

We have thus seen that heaviside functions can be used to provide estimated hazard ratios that remain constant within each of two separate time intervals of follow-up. We can also extend the use of heaviside functions to provide several distinct hazard ratios that remain constant within several time intervals.

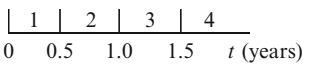
Four time intervals:



Suppose, for instance, that we wish to separate the data into *four* separate time intervals, and for each interval, we wish to obtain a different hazard ratio estimate as illustrated in the graph shown here.

Extended Cox model contains either

- E, $E \times g_1(t)$, $E \times g_2(t)$, $E \times g_3(t)$ or
- $E \times g_1(t)$, $E \times g_2(t)$, $E \times g_3(t)$, $E \times g_4(t)$



We can obtain four different hazard ratios using an extended Cox model containing *a main effect of exposure and three heaviside functions* in the model as products with exposure. Or, we can use a model containing *no main effect* exposure term, but with product terms involving exposure with *four heaviside functions*.

To illustrate the latter model, suppose, as shown on the graph, that the first time interval goes from time 0 to 0.5 of a year; the second time interval goes from 0.5 to 1 year; the third time interval goes from 1 year to a year and a half; and the fourth time interval goes from a year and a half onward.

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta_1 E g_1(t) + \delta_2 E g_2(t) + \delta_3 E g_3(t) + \delta_4 E g_4(t)]$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } 0 \le t < 0.5 \text{ year} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } 0.5 \text{ year } \le t < 1.0 \text{ year} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_3(t) = \begin{cases} 1 & \text{if } 1.0 \text{ year } \le t < 1.5 \text{ years} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_4(t) = \begin{cases} 1 & \text{if } t \ge 1.5 \text{ years} \\ 0 & \text{if otherwise} \end{cases}$$

Then, an appropriate extended Cox model containing the four heaviside functions $g_1(t)$, $g_2(t)$, $g_3(t)$, and $g_4(t)$ is shown here. This model assumes that there are four different hazard ratios identified by three cutpoints at half a year, one year, and one and a half years. The formulae for the four hazard ratios are given by separately exponentiating each of the four estimated coefficients, as shown below:

$$4\,\widehat{\text{HR}}\text{'s} \begin{cases} 0 \le t < 0.5: \ \widehat{HR} = \exp\left(\hat{\delta}_1\right) \\ 0.5 \le t < 1.0: \ \widehat{HR} = \exp\left(\hat{\delta}_2\right) \\ 1.0 \le t < 1.5: \ \widehat{HR} = \exp\left(\hat{\delta}_3\right) \\ t \ge 1.5: \ \widehat{HR} = \exp\left(\hat{\delta}_4\right) \end{cases}$$

VII. An Application of the Extended Cox Model to An Epidemiologic Study on the Treatment of Heroin Addiction

EXAMPLE

1991 Australian study (Caplehorn et al.) of heroin addicts

- two methadone treatment clinics
- T = days remaining in treatment (= days until drop out of clinic)
- clinics differ in treatment policies

Dataset name: ADDICTS
Column 1: Subject ID
Column 2: Clinic (1 or 2) ← E
Column 3: Survival status (0 = censored, 1 = departed clinic)
Column 4: Survival time in days
Column 5: Prison Record (0 = none, 1 = any) ← covariates
Column 6: Maximum Methadone Dose (mg/day)

A 1991 Australian study by Caplehorn et al., compared retention in two methadone treatment clinics for heroin addicts. A patient's survival time (T) was determined as the time in days until the patient dropped out of the clinic or was censored at the end of the study clinic. The two clinics differed according to their overall treatment policies.

A listing of some of the variables in the dataset for this study is shown here. The dataset name is called "ADDICTS," and survival analysis programs in the Stata package are used in the analysis. Note that the survival time variable is listed in column 4 and the survival status variable, which indicates whether a patient departed from the clinic or was censored, is listed in column 3. The primary exposure variable of interest is the clinic variable, which is coded as 1 or 2. Two other variables of interest are prison record status, listed in column 5 and coded as 0 if none and 1 if any, and maximum methadone dose, in milligrams per day, which is listed in column 6. These latter two variables are considered as covariates.

$$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose})]$$

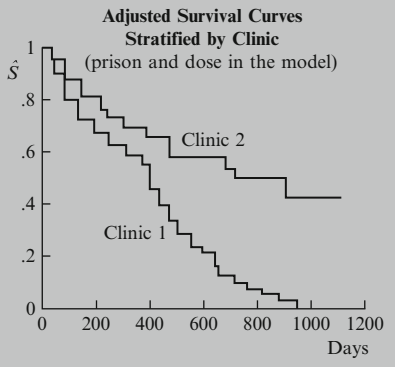
	Coef.	Std. Err.	$p > \|z\|$	Haz. Ratio	P (PH)
Clinic	−1.009	0.215	0.000	0.365	0.001
Prison	0.327	0.167	0.051	1.386	0.332
Dose	−0.035	0.006	0.000	0.965	0.347

One of the first models considered in the analysis of the addicts dataset was a Cox PH model containing the three variables, clinic, prison record, and dose. An edited printout of the results for this model is shown here. What stands out from this printout is that the $P(PH)$ value for the clinic variable is zero to three significant places, which indicates that the clinic variable does not satisfy the proportional hazard assumption.

$P(PH)$ for the variables prison and dose are nonsignificant ⇒ remain in model.

Since the $P(PH)$ values for the other two variables in the model are highly nonsignificant, this suggests that these two variables, namely, prison and dose, can remain in the model.

EXAMPLE: (continued)



Results:

- Curve for clinic 2 consistently lies above curve for clinic 1.
- Curves diverge, with clinic 2 being vastly superior after one year.

Stratifying by **clinic**: cannot obtain hazard ratio for **clinic**

Hazard ratio for **clinic** requires **clinic** in the model.

Extended Cox model:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta(\text{clinic})\,g(t)]$$

where

$$g(t) = \begin{cases} 1 & \text{if } t \geq 365 \text{ days} \\ 0 & \text{if } t < 365 \text{ days} \end{cases}$$

and

$$\text{clinic} = \begin{cases} 1 & \text{if clinic 1} \\ 0 & \text{if clinic 2} \end{cases}$$

None Previously clinic = 2 for clinic 2

$t \geq 365$ days: $HR = \exp\left(\hat{\beta}_1 + \hat{\delta}\right)$

$t < 365$ days: $HR = \exp\left(\hat{\beta}_1\right)$

Further evidence of the PH assumption not being satisfied for the clinic variable can be seen from a graph of adjusted survival curves stratified by clinic, where the prison and dose variables have been kept in the model. Notice that the two curves are much closer together at earlier times, roughly less than 1 year (i.e., 365 days), but the two curves diverge greatly after 1 year. This indicates that the hazard ratio for the clinic variable will be much closer to one at early times but quite different from one later on.

The above graph, nevertheless, provides important results regarding the comparison of the two clinics. The curve for clinic 2 consistently lies above the curve for clinic 1, indicating that clinic 2 does better than clinic 1 in retaining its patients in methadone treatment. Further, because the two curves diverge after about a year, it appears that clinic 2 is vastly superior to clinic 1 after one year but only slightly better than clinic 1 prior to one year.

Unfortunately, because the clinic variable has been stratified in the analysis, we cannot use this analysis to obtain a hazard ratio expression for the effect of clinic, adjusted for the effects of prison and dose. We can only obtain such an expression for the hazard ratio if the clinic variable is in the model.

Nevertheless, we can obtain a hazard ratio using an alternative analysis with an extended Cox model that contains a heaviside function, $g(t)$, together with the clinic variable, as shown here. Based on the graphical results shown earlier, a logical choice for the cutpoint of the heaviside function is one year (i.e., 365 days). The corresponding model then provides two hazard ratios: one that is constant above 365 days and the other that is constant below 365 days.

Note that in the extended Cox model here, we have coded the clinic variable as 1 if clinic 1 and 0 if clinic 2, whereas previously we had coded clinic 2 as 2. The reason for this change in coding, as illustrated by computer output below, is to obtain hazard ratio estimates that are greater than unity.

EXAMPLE: (continued)

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \beta_1(\text{clinic})g_1(t) + \delta_2(\text{clinic})g_2(t)]$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } t > 365 \text{ days} \\ 0 & \text{if } t \geq 365 \text{ days} \end{cases}$$

and

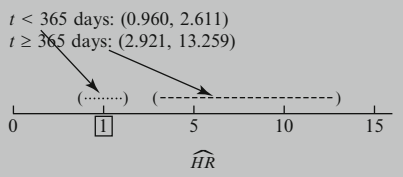
$$g_2(t) = \begin{cases} 1 & \text{if } t \geq 365 \text{ days} \\ 0 & \text{if } t > 365 \text{ days} \end{cases}$$

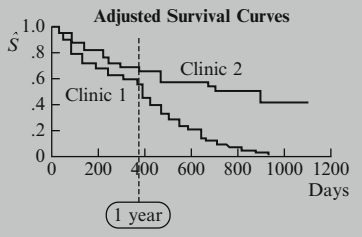
$t < 365$ days: $\widehat{HR} = \exp(\hat{\delta}_1)$
$t \geq 365$ days: $\widehat{HR} = \exp(\hat{\delta}_2)$

	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Prison	0.378	0.168	0.025	1.459	1.049	2.029
Dose	−0.036	0.006	0.000	0.965	0.953	0.977
Clinic × g_1	0.460	0.255	0.072	1.583	0.960	2.611
Clinic × g_2	1.828	0.386	0.000	6.223	2.921	13.259

$t < 365$ days: $\widehat{HR} = e^{0.460} = 1.583$
$t \geq 365$ days: $[\widehat{HR} = e^{1.828} = 6.223$

95% confidence intervals for clinic effect:





An equivalent way to write the model is to use two heaviside functions, $g_1(t)$ and $g_2(t)$, as shown here. This latter model contains product terms involving clinic with each heaviside function, and there is no main effect of clinic.

Corresponding to the above model, the effect of clinic is described by two hazard ratios, one for time less than 365 days and the other for greater than 365 days. These hazard ratios are obtained by separately exponentiating the coefficients of each product term, yielding e to the δ_1 "hat" and e to the δ_2 "hat," respectively.

A printout of results using the above model with two heaviside functions is provided here. The results show a borderline nonsignificant hazard ratio ($P = 0.072$) of 1.6 for the effect of clinic when time is less than 365 days in contrast to a highly significant (P = 0.000 to three decimal places) hazard ratio of 6.2 when time exceeds 365 days.

Note that the estimated hazard ratio of 1.583 from the printout is computed by exponentiating the estimated coefficient 0.460 of the product term "clinic × g_1" and that the estimated hazard ratio of 6.223 is computed by exponentiating the estimated coefficient 1.828 of the product term "clinic × g_2".

Note also that the 95% confidence interval for the clinic effect prior to 365 days (that is, for the product term "clinic × $g_1(t)$") is given by the limits 0.960 and 2.611, whereas the corresponding confidence interval after 365 days (that is, for the product term "clinic × $g_2(t)$") is given by the limits 2.921 and 13.259. The latter interval is quite wide, showing a lack of precision when t exceeds 365 days; however, when t precedes 365 days, the interval includes the null hazard ratio of 1, suggesting a chance effect for this time period.

The results we have just shown support the observations obtained from the graph of adjusted survival curves. That is, these results suggest a large difference in clinic survival times after 1 year in contrast to a small difference in clinic survival times prior to 1 year, with clinic 2 always doing better than clinic 1 at any time.

EXAMPLE: (continued)

One other analysis:
Use an extended Cox model that provides for diverging survival curves

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta(\text{clinic} \times \text{t})]$$

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}\text{t}\right)$$

$\widehat{\text{HR}}$ changes over time.

$t = 91$ days

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3\,(\text{dose}) + \delta(\text{clinic})(91)]$$

So

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + 91\hat{\delta}\right)$$

$t = 274$:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta(\text{clinic})(274)]$$

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + 274\hat{\delta}\right)$$

$t = 458.5$:

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + 458.5\hat{\delta}\right)$$

$t = 639$:

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + 639\hat{\delta}\right)$$

$t = 821.5$:

$$\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + 821.5\hat{\delta}\right)$$

$\hat{\delta} > 0 \Rightarrow \widehat{HR}\uparrow$ as time $\uparrow$

There is, nevertheless, at least one other approach to the analysis using time-dependent variables that we now describe. This approach considers our earlier graphical observation that the survival curves for each clinic continue to diverge from one another even after 1 year. In other words, it is reasonable to consider an extended Cox model that allows for such a divergence, rather than a model that assumes the hazard ratios are constant before and after 1 year.

One way to define an extended Cox model that provides for diverging survival curves is shown here. This model includes, in addition to the clinic variable by itself, a time-dependent variable defined as the product of the clinic variable with time (i.e., clinic $\times$ t). By including this product term, we are able to estimate the effect of clinic on survival time, and thus the hazard ratio, for any specified time t.

To demonstrate how the hazard ratio changes over time for this model, we consider what the model and corresponding estimated hazard ratio expression are for different specified values of t.

For example, if we are interested in the effect of clinic on survival on day 91, so that t = 91, the exponential part of the model simplifies to terms for the prison and dose variables plus β_1 times the clinic variable plus δ times the clinic variable times 91: the corresponding estimated hazard ratio for the clinic effect is then e to the power β_1 "hat" plus δ "hat" times t = 91.

At 274 days, the exponential part of the model contains the prison, dose, and clinic main effect terms as before, plus δ times the clinic variable times 274: the corresponding hazard ratio for the clinic effect is then e to β_1 "hat" plus 274 δ "hat".

The formulae for the estimated hazard ratio for other specified days are shown here. Notice that the estimated hazard ratio appears to be increase over the length of the follow-up period. Thus, if δ "hat" is a positive number, then the estimated hazard ratios will increase over time.

EXAMPLE: (continued)

Computer results for extended Cox model involving $T(t)$:

	Coef.	Std. Err.	P>|z|	Haz. Ratio	[95% Conf.	Interval]
prison	0.390	0.169	0.021	1.476	1.060	2.056
dose	−0.035	0.006	0.000	0.965	0.953	0.978
clinic	(0.0183)	0.347	0.958	0.982	0.497	1.939
clinic × t	(0.003)	0.001	0.001	(1.003)	1.001	1.005

$\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\delta}\right) = -.000259$ Log likelihood = −667.642

$\hat{\beta}_1 = -0.0183 \qquad \hat{\delta} = 0.003$

$\widehat{\text{HR}}$ depends on $\hat{\beta}_1$ and $\hat{\delta}$.

$t = 91.5$: $\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}t\right) = 1.292$

$t = 274$: $\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}t\right) = 2.233$

$t = 458.5$: $\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}t\right) = 3.862$

$t = 639$: $\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}t\right) = 6.677$

$t = 821.5$: $\widehat{\text{HR}} = \exp\left(\hat{\beta}_1 + \hat{\delta}t\right) = 11.544$

$$\exp\left[\hat{\beta}_1 + \hat{\delta}t \pm 1.96\sqrt{\widehat{\text{var}}\left(\hat{\beta}_1 + \hat{\delta}t\right)}\right]$$

$$\text{Var}\left(\hat{\beta}_1 + \hat{\delta}t\right) = s^2_{\hat{\beta}_1} + t^2 s^2_{\hat{\delta}} + 2t\,\widehat{\text{cov}}\left(\hat{\beta}_1, \hat{\delta}\right)$$

$(0.347)^2 \quad (0.001)^2 \quad (-.000259)$

Time (days)	$\widehat{\text{HR}}$	95% CI
91.5	1.292	(0.741, 2.250)
274	2.233	(1.470, 3.391)
458.5	3.862	(2.298, 6.491)
639	6.677	(3.102, 14.372)
821.5	11.544	(3.976, 33.513)

We now show edited results obtained from fitting the extended Cox model we have just been describing, which contains the product of clinic with time. The covariance estimate shown at the bottom of the table will be used below to compute confidence intervals.

From these results, the estimated coefficient of the clinic variable is β_1 "hat" equals −0.0183, and the estimated coefficient δ "hat" obtained for the product term equals 0.003. For the model being fit, the hazard ratio depends on the values of both β_1 "hat" and δ "hat."

On the left, the effect of the variable clinic is described by five increasing hazard ratio estimates corresponding to each of five different values of t. These values, which range between 1.292 at 91.5 days to 11.544 at 821.5 days, indicate how the effect of clinic diverges over time for the fitted model.

We can also obtain 95% confidence intervals for each of these hazard ratios using the large sample formula shown here. The variance expression in the formula is computed using the variances and covariances which can be obtained from the computer results given above. In particular, the variances are $(0.347)^2$ and $(0.001)^2$ for β_1 "hat" and δ "hat," respectively; the covariance value is −0.000259.

A table showing the estimated hazard ratios and their corresponding 95% confidence intervals for the clinic effect is given here. Note that all confidence intervals are quite wide.

VIII. An Application of the Extended Cox Model to the Analysis of the Stanford Heart Transplant Data

EXAMPLE

Patients identified as eligible for heart transplant:
T = time until death or censorship
65 patients receive transplants
38 patients do not receive transplants
n = 103 patients

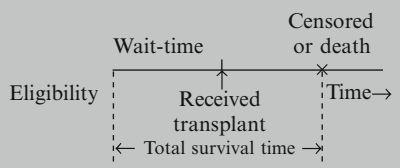
Goal: Do patients receiving transplants survive longer than patients not receiving transplants?

We now consider another application of the extended Cox model which involves the use of an internally defined time-dependent variable. In a 1977 report (Crowley and Hu, *J. Amer. Statist. Assoc.*) on the Stanford Heart Transplant Study, patients identified as being eligible for a heart transplant were followed until death or censorship. Sixty-five of these patients received transplants at some point during follow-up, whereas thirty-eight patients did not receive a transplant. There were, thus, a total of n = 103 patients. The goal of the study was to assess whether patients receiving transplants survived longer than patients not receiving transplants.

One approach:
Compare two separate groups: 65 transplants vs. 38 nontransplants

One approach to the analysis of these data was to separate the dataset into two separate groups, namely, the 65 heart transplant patients and the 38 patients not receiving transplants, and then to compare survival times for these groups.

Problem:



Note: Wait-time contributes to survival time for nontransplants.

A problem with this approach, however, is that those patients who received transplants had to wait from the time they were identified as eligible for a transplant until a suitable transplant donor was found. During this "wait-time" period, they were at risk for dying, yet they did not have the transplant. Thus, the wait-time accrued by transplant patients contributes information about the survival of nontransplant patients. Yet, this wait-time information would be ignored if the *total* survival time for each patient were used in the analysis.

Covariates:

Tissue mismatch score } prognostic only
Age at transplant } for transplants

Age at eligibility: not considered prognostic for nontransplants

Another problem with this approach is that two covariates of interest, namely, *tissue mismatch score* and *age at transplant*, were considered as prognostic indicators of survival only for patients who received transplants. Note that *age at eligibility* was not considered an important prognostic factor for the nontransplant group.

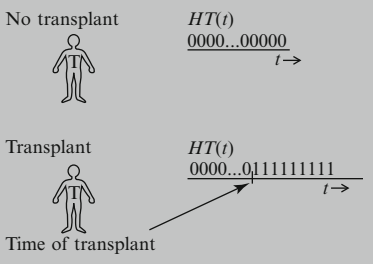
EXAMPLE: (continued)

Problems:
- wait-time of transplant recipients
- prognostic factors for transplants only

Alternative approach:
Uses an extended Cox model

Exposure variable:
Heart transplant status at time t, defined as

$$HT(t) = \begin{cases} 0 & \text{if did not receive transplant by time } t, \text{ i.e., if } t < \text{wait-time} \\ 1 & \text{if received transplant prior to time } t, \text{ i.e., if } t \geq \text{wait-time} \end{cases}$$



Wait-time for transplants contributes to survival for nontransplants.

In addition to $HT(t)$, two time-dependent covariates included in model.

Because of the problems just described, which concern the wait-time of transplants and the effects of prognostic factors attributable to transplants only, an alternative approach to the analysis is recommended. This alternative involves the use of time-dependent variables in an extended Cox model.

The exposure variable of interest in this extended Cox model is heart transplant status at time t, denoted by $HT(t)$. This variable is defined to take on the value 0 at time t if the patient has not received a transplant at this time, that is, if t is less than the wait-time for receiving a transplant. The value of this variable is 1 at time t if the patient has received a transplant prior to or at time t, that is, if t is equal to or greater than the wait-time.

Thus, for a patient who did not receive a transplant during the study, the value of $HT(t)$ is 0 at all times. For a patient receiving a transplant, the value of $HT(t)$ is 0 at the start of eligibility and continues to be 0 until the time at which the patient receives the transplant; then, the value of $HT(t)$ changes to 1 and remains 1 throughout the remainder of follow-up.

Note that the variable $HT(t)$ has the property that the wait-time for transplant patients contributes to the survival experience of nontransplant patients. In other words, this variable treats a transplant patient as a nontransplant patient prior to receiving the transplant.

In addition to the exposure variable $HT(t)$, two other time-dependent variables are included in our extended Cox model for the transplant data. These variables are covariates to be adjusted for in the assessment of the effect of the $HT(t)$ variable.

EXAMPLE: (continued)

Covariates:

$$TMS(t) = \begin{cases} 0 \text{ if } t < \text{wait-time} \\ TMS \text{ if } t \geq \text{wait-time} \end{cases}$$

$$AGE(t) = \begin{cases} 0 \text{ if } t < \text{wait-time} \\ AGE \text{ if } t \geq \text{wait-time} \end{cases}$$

These covariates are denoted as $TMS(t)$ and $AGE(t)$ and they are defined as follows: $TMS(t)$ equals 0 if t is less than the wait-time for a transplant but changes to the "tissue mismatch score" (TMS) at the time of the transplant if t is equal to or greater than the wait-time. Similarly, $AGE(t)$ equals 0 if t is less than the wait-time but changes to AGE at time of transplant if t is equal to or greater than the wait-time.

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta_1 HT(t) + \delta_2 TMS(t) + \delta_3 AGE(t)]$$

The extended Cox model for the transplant data is shown here. The model contains the three time-dependent variables $HT(t)$, $TMS(t)$ and $AGE(t)$ as described above.

Focus:
Asscesing the effect of $HT(t)$ adjusted for $TMS(t)$ and $AGE(t)$

Note: $HT(t)$ does not satisfy PH assumption.

For this model, since $HT(t)$ is the exposure variable of interest, the focus of the analysis concerns assessing the effect of this variable adjusted for the two covariates. Note, however, that because the $HT(t)$ variable is time-dependent by definition, this variable does not satisfy the PH assumption, so that any hazard ratio estimate obtained for this variable is technically time-dependent.

Variable	Coef.	Std. Err.	P>|z|	Haz. Ratio
$HT(t)$	−3.1718	1.1861	0.008	0.0417
$TMS(t)$	0.4442	0.2802	0.112	1.5593
$AGE(t)$	0.0552	0.0226	0.014	1.0567

A summary of computer results for the fit of the above extended Cox model is shown here. These results indicate that the exposure variable $HT(t)$ is significant below the one percent significance level (i.e., the two-sided p-value is 0.008). Thus, transplant status appears to be significantly associated with survival.

$$\widehat{\text{HR}} = e^{-3.1718} = 0.0417 = \frac{1}{23.98}$$

$$\widehat{\text{HR}} = \frac{\hat{h}(\text{transplants})}{\hat{h}(\text{nontransplants})} \approx \frac{1}{24}?$$

Not appropriate!

To evaluate the strength of the association, note that e to the coefficient of $HT(t)$ equals 0.0417. Since 1 over 0.0417 is 23.98, it appears that there is a 24-fold increase in the hazard of nontransplant patients to transplant patients. The preceding interpretation of the value 0.0417 as a hazard ratio estimate is not appropriate, however, as we shall now discuss further.

EXAMPLE (continued)

23.98 is inappropriate as a $\widehat{HR}$:

- does not compare two *separate* groups
- exposure variable is *not* time-independent
- wait-time on transplants contributes to survival on nontransplants

First, note that the value of 23.98 inappropriately suggests that the hazard ratio is comparing two separate groups of patients. However, the exposure variable in this analysis is *not* a time-independent variable that distinguishes between two separate groups. In contrast, the exposure variable is time-dependent, and uses the wait-time information on transplants as contributing to the survival experience of nontransplants.

Alternative interpretation:
At time t,
h("not yet received transplant")
$\approx 24\, \hat{h}$ ("already received transplant")

Since the exposure variable is time-dependent, an alternative interpretation of the hazard ratio estimate is that, at any given time t, the hazard for a person *who has not yet received a transplant* (but may receive one later) is approximately 24 times the hazard for a person *who already has received a transplant by that time*.

More appropriate:

Hazard ratio formula should account for *TMS* and *AGE*.

Transplant?	$HT(t)$	$TMS(t)$	$AGE(t)$
Yes	1	*TMS*	*AGE*
No	0	0	0

Actually, we suggest that a more appropriate hazard ratio expression is required to account for a transplant's *TMS* and *AGE* score. Such an expression would compare, at time t, the values of each of the three time-dependent variables in the model. For a person who received a transplant, these values are 1 for $HT(t)$ and *TMS* and *AGE* for the two covariates. For a person who has not received a transplant, the values of all three variables are 0.

i denotes ith transplant patient

$\mathbf{X}^*(t) = (HT(t) = 1, TMS(t) = TMS_i, \text{AGE}(t) = AGE_i)$
$\mathbf{X}(t) = (HT(t) = 0, TMS(t) = 0, AGE(t) = 0)$

Using this approach to compute the hazard ratio, the $\mathbf{X}^*(t)$ vector, which specifies the predictors for a patient i who received a transplant at time t, has the values 1, TMS_i and AGE_i for patient i; the $\mathbf{X}(t)$ vector, which specifies the predictors at time t for a patient who has not received a transplant at time t, has values of 0 for all three predictors.

$$\begin{aligned}\widehat{HR}(t) &= \exp\left[\hat{\delta}_1(1-0) + \hat{\delta}_2(TMS_i - 0) + \hat{\delta}_3(AGE_i - 0)\right] \\ &= \exp\left[\hat{\delta}_1 + \hat{\delta}_2\, TMS_i + \hat{\delta}_3\, AGE_i\right] \\ &= \exp[-3.1718 + 0.4442\, TMS_i + 0.0552\, AGE_i]\end{aligned}$$

The hazard ratio formula then reduces to e to the sum of δ_1 "hat" plus δ_2 "hat" times TMS_i plus δ_3 "hat" times AGE_i, where the δ "hat's" are the estimated coefficients of the three time-dependent variables. Substituting the numerical values for these coefficients in the formula gives the exponential expression circled here.

EXAMPLE: (continued)

$\widehat{HR}(t)$ is time-dependent, i.e., its value at time t depends on TMS_i and AGE_i at time t

TMS range: (0–3.05)
AGE range: (12–64)

The resulting formula for the hazard ratio is time-dependent in that its value depends on the *TMS* and *AGE* values of the *i*th patient at the time of transplant. That is, different patients can have different values for *TMS* and *AGE* at time of transplant. Note that in the dataset, *TMS* ranged between 0 and 3.05 and *AGE* ranged between 12 and 64.

We end our discussion of the Stanford Heart Transplant Study at this point. For further insight into the analysis of this dataset, we refer the reader to the 1977 paper by Crowley and Hu (*J. Amer. Statist. Assoc.*).

IX. The Extended Cox Likelihood

- Cox PH likelilhood (L) described in Chapter 3, Section VIII
- L now extended for extended Cox model

At the end of the presentation from Chapter 3 (Section VIII), we illustrated the Cox likelihood using the dataset shown on the left. In this section, we extend that discussion to illustrate the Cox likelihood with a time-dependent variable.

ID	TIME	STATUS	SMOKE
Barry	2	1	1
Gary	3	1	0
Harry	5	0	0
Larry	8	1	1

TIME = Survival time (in years)
STATUS = 1 for event, 0 for censorship
SMOKE = 1 for a smoker, 0 for a nonsmoker

To review: The data indicate that Barry got the event at TIME = 2 years. Gary got the event at 3 years, Harry was censored at 5 years and Larry got the event at 8 years. Furthermore, Barry and Larry were smokers while Gary and Harry were non-smokers.

Cox PH model: $h(t)=h_0(t)e^{\beta_1 SMOKE}$

In Chapter 3, we constructed the Cox likelihood with one predictor, SMOKE, in the model. The model and the likelihood are shown on the left.

Cox PH Likelihood:

$$L=\left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}+h_0(t)e^{0}+h_0(t)e^{0}+h_0(t)e^{\beta_1}}\right]\times\left[\frac{h_0(t)e^{0}}{h_0(t)e^{0}+h_0(t)e^{0}+h_0(t)e^{\beta_1}}\right]\times\left[\frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}}\right]$$

The likelihood is a product of three terms, one term for each event time t_f (TIME = 2, 3, and 8). The denominator of each term is the sum of the hazards from the subjects still in the risk set at time t_f, including the censored subject, Harry. The numerator of each term is the hazard of the subject who got the event at t_f. The reader may wish to reread Section VIII of Chapter 3.

Cox extended model:

$$h(t) = h_0(t)e^{\beta_1 SMOKE+\beta_2 SMOKE\times TIME}$$

Time-dependent covariate
(its value changes over time)

Now consider an extended Cox model, which contains the predictor SMOKE and a time-dependent variable SMOKE × TIME. For this model, it is not only the baseline hazard that may change over time but also the value of the predictor variables. This can be illustrated by examining Larry's hazard at each event time.

Larry got the event at TIME = 8

Larry's hazard at each event time

TIME	Larry's Hazard
2	$h_0(t)e^{\beta_1+2\beta_2}$
3	$h_0(t)e^{\beta_1+3\beta_2}$
8	$h_0(t)e^{\beta_1+8\beta_2}$

Larry, a smoker, got the event at TIME = 8. However at TIME = 2, 3, and 8, the covariate SMOKE × TIME changes values, thus impacting Larry's hazard at each event time (see left). Understanding how the expression for an individual's hazard changes over time is the key addition toward understanding how the Cox extended likelihood differs from the Cox PH likelihood.

Cox extended model:

$$h(t) = h_0(t)e^{\beta_1 SMOKE + \beta_2 SMOKE \times TIME}$$

$$L = \left[\frac{h_0(t)e^{\beta_1+2\beta_2}}{h_0(t)e^{\beta_1+2\beta_2} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1+2\beta_2}}\right] \times \left[\frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1+3\beta_2}}\right] \times \left[\frac{h_0(t)e^{\beta_1+8\beta_2}}{h_0(t)e^{\beta_1+8\beta_2}}\right]$$

The extended Cox model for these data is shown again on the left.

The extend cox likehood (L) for these data is shown on the left. This likehood is constructed in a similar mannar as the likehood for the Cox PH model. Yhe difference is that the expression for the subject's hazard is allowed to vary over time.

Likelihood is product of three terms:

$$L = L_1 \times L_2 \times L_3$$

Barry ($t = 2$) Gary ($t = 3$) Larry ($t = 8$)

Just as with the Cox PH likelihood shown on the previous page, the extended Cox likelihood is also a product of three terms, corresponding to the three event times ($L = L_1 \times L_2 \times L_3$). Barry got the event first at t = 2, then Gary at t = 3, and finally Larry at t = 8. Harry, who was censored at t = 5, was still at risk when Barry and Garry got the event. Therefore, Harry's hazard is still in the denominator of L_1 and L_2.

SMOKE × TIME = 0 for nonsmokers

SMOKE × TIME changes over time for smokers

Larry's hazard changes over L_1, L_2, L_3.

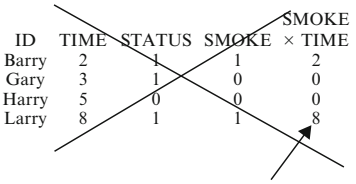
The inclusion of the time-varying covariate SMOKE × TIME does not change the expression for the hazard for the nonsmokers (Gary and Harry) since SMOKE is coded 0 for nonsmokers. However, for smokers (Barry and Larry), the expression for the hazard changes with time. Notice how Larry's hazard changes in the denominator of L_1, L_2, and L_3.

$h_0(t)$ cancels in L

$$L = \left[\frac{e^{\beta_1+2\beta_2}}{e^{\beta_1+2\beta_2} + e^0 + e^0 + e^{\beta_1+2\beta_2}}\right] \times \left[\frac{e^0}{e^0 + e^0 + e^{\beta_1+3\beta_2}}\right] \times \left[\frac{e^{\beta_1+8\beta_2}}{e^{\beta_1+8\beta_2}}\right]$$

The baseline hazard cancels in the extended Cox likelihood as it does with the Cox PH likelihood. Thus, the form of the baseline hazard need not be specified, as it plays no role in the estimation of the regression parameters.

Caution: Incorrect Coding of SMOKE × TIME



A word of caution for those planning to run a model with a time-varying covariate: It is incorrect to create a product term with TIME in the data step by multiplying each individual's value for SMOKE with their survival time. In other words, SMOKE×TIME should not be coded like the typical interaction term. In fact, if SMOKE×TIME was coded as it is on the left, then SMOKE×TIME would be a time-independent variable. Larry's value for SMOKE×TIME is incorrectly coded at a constant value of 8 even though Larry's value for SMOKE×TIME changes in the likelihood over L_1, L_2, and L_3.

Incorrectly coded SMOKE×TIME
- Time independent
- Probably highly significant
- Survival time should predict survival time
- But not meaningful

If the incorrectly coded time independent SMOKE×TIME was included in a Cox model, it would not be surprising if the coefficient estimate was highly significant even if the PH assumption was not violated. It would be expected that a product term with each individual's survival time would predict the outcome (their survival time), but it would not be meaningful. Nevertheless, this is a common mistake.

Correctly coding SMOKE×TIME
- Time dependent
- Computer packages typically allow definition in the analytic procedure

To obtain a correctly defined SMOKE×TIME time-dependent variable, computer packages typically allow the variable to be defined within the analytic procedure.

Alternatively can code using **CP format**

Alternatively, **the (start, stop) or counting process (CP) data layout** can be used, to explicitly define a time-dependent variable in the data.

- CP format introduced in Chapter 1
- Multiple observations per subject
- Time intervals at risk subdivided
- Covariate value can change from interval to interval for the same subject

The CP data format was introduced in Chapter 1. This data layout provides a straightforward approach for expressing a time-dependent covariate by allowing multiple observations to correspond to the same individual. With this format, an individual's total at-risk follow-up time is sub-divided into smaller time intervals providing a way for values of variables to change from time interval-to-interval for the same individual.

CP format with time-dependent variable SMOKE × TIME:

ID	START	STOP	STATUS	SMOKE	SMOKE × TIME
Barry	0	2	1	1	2
Gary	0	2	0	0	0
Gary	2	3	1	0	0
Harry	0	2	0	0	0
Harry	2	3	0	0	0
Harry	3	5	0	0	0
Larry	0	2	0	1	2
Larry	2	3	0	1	3
Larry	3	8	1	1	8

Coded as time dependent

START = Beginning of interval (in months)
STOP = End of interval (in months)
STATUS = 1 for event, 0 for censorship

The data layout on the left illustrates the CP approach with the (start, stop) time intervals defined by each event time (t=2, t=3, and t=8) from Barry, Gary, and Larry.

If we look at the final three observations, we can see that Larry's total time at risk is subdivided into three time intervals. Larry got the event at t=8 (STOP=8 and STATUS=1). The two previous observations indicate that Larry did not get an event (STATUS=0) over the time intervals (0, 2) or (2, 3). Over the last three observations, the time-dependent variable SMOKE×TIME changes values for Larry from 2 to 3 to 8.

Alternative CP format:
Gary and Harry do not need multiple observations since SMOKE × TIME does not vary for them (same info as above)

ID	START	STOP	STATUS	SMOKE	SMOKE × TIME
Barry	0	2	1	1	2
Gary	0	3	1	0	0
Harry	0	5	0	0	0
Larry	0	2	0	1	2
Larry	2	3	0	1	3
Larry	3	8	1	1	8

An alternative CP data layout is shown below on the left. Since Gary and Harry are nonsmokers and the coding for nonsmokers is SMOKE=0, their values for the SMOKE×TIME variable stays at zero throughout their time at risk. Therefore, it is not necessary to have multiple observations for Gary and Harry (although it is not incorrect to do so).

2 reasons to include time-varying covariate:
1) To account for PH violation
2) The values actually change over time regardless of the PH assumption

There are two main reasons why a time-varying covariate might be included in a Cox model: (1) To account for a violation of the proportional hazards assumption (usually formulated as a product term with some function of time) and (2) The covariate may actually change its values over time regardless of the PH assumption.

SMOKE×TIME
defined to account for PH violation

The use of the SMOKE×TIME variable in the last example was of the first type (to account for a PH violation). An example of the second type could be the changing of an individual's dosage level of medication over time as illustrated in the next example.

DOSE changes at 3 time points for Jane

ID	MONTHS	STATUS	DOSE1	TIME1	DOSE2	TIME2	DOSE3	TIME3
Jane	49	1	60	0	120	12	150	30

The data on the left contains one observation for Jane who had an event at 49 months (MONTHS=49 and STATUS=1). Her dose of medication at the beginning of follow-up was 60 milligrams (DOSE1=60 and TIME1=0). At the 12th month of follow-up, her dose was changed to 120 milligrams (DOSE2=120 and TIME2=12). At the 30th month of follow-up, her dose was changed to 150 mg (DOSE3=120 and TIME3=30).

MONTHS = Survival time (in months)
STATUS = 1 for event, 0 for censorship

Same info as above using CP format

(3 observations instead of 1)

ID	START	STOP	STATUS	DOSE
Jane	0	12	0	60
Jane	12	30	0	120
Jane	30	49	1	150

The same information can be expressed using the counting process data layout. On the left, the data is transposed to contain three observations for Jane allowing DOSE to be represented as a time-dependent variable. For the first time interval (START=0, STOP=12), Jane's dose was 60 mg. For the second time interval (12–30 months), Jane's dose was 120 milligrams. For the third time interval (30–49 months), Jane's dose was 150 milligrams. The data indicates that Jane had an event at 49 months (STOP=49 and STATUS=1).

START = Beginning of interval (in months)
STOP = End of interval (in months)
STATUS = 1 for event, 0 for censorship
DOSE = Dose in milligrams

Multiple observations per subject: revisited in Chapter 8 (recurrent events)

The counting process data layout is further discussed in Chapter 8 on recurrent events. With recurrent event data, subjects may remain at risk for subsequent events after getting an event.

Coding SMOKE × TIME as time-dependent

When a time-dependent variable is defined within the Cox analytic procedure, the variable is defined internally such that the user may not see the time-dependent variable in the dataset. However, the dataset on the left will provide a clearer idea of the correct definition of SMOKE × TIME. The dataset contains multiple observations per subject. Barry was at risk at t = 2 and got the event at that time. Gary was at risk at t = 2 and t = 3. Gary didn't get the event at t = 2 but did get the event at t = 3. Harry was at risk at t = 2, t = 3, t = 5 and didn't get the event. Larry was at risk at t = 2, t = 3, t = 5, t = 8 and got the event at t = 8. Notice how the SMOKE × TIME variable changes values for Larry over time.

Multiple Observations per Subject

ID	TIME	STATUS	SMOKE	SMOKE × TIME
Barry	2	1	1	2
Gary	2	0	0	0
Gary	3	1	0	0
Harry	2	0	0	0
Harry	3	0	0	0
Harry	5	0	0	0
Larry	2	0	1	2
Larry	3	0	1	3
Larry	5	0	1	5
Larry	8	1	1	8

↑ Coded as time-dependent

Multiple observations per subject: revisited in Chapter 8 (recurrent events)

Survival analysis datasets containing multiple observations per subject are further discussed in Chapter 8 on recurrent events. With recurrent event data, subjects may remain at risk for subsequent events after getting an event.

X. Summary

A summary of this presentation on time-dependent variables is now provided. We began by reviewing the main features of the Cox PH model. We then defined a time-dependent variable and illustrated three types of these variables–defined, internal, and ancillary.

Review Cox PH model.

Define time-dependent variable: defined, internal, ancillary.

Extended Cox model:

Next, we gave the form of the "extended Cox model," shown here again, which allows for time-dependent as well as time-independent variables.

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i \mathbf{X}_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right]$$

$$\widehat{HR}(t) = \exp\left[\sum_{i=1}^{p_1} \hat{\beta}_i \left[X_i^* - X_i\right] + \sum_{j=1}^{p_2} \hat{\delta}_j \left[X_j^*(t) - X_j(t)\right]\right]$$

↑ Function of time

We then described various characteristics of this extended Cox model, including the formula for the hazard ratio. The latter formula is time-dependent so that the PH assumption is not satisfied.

Model for assessing PH assumption:

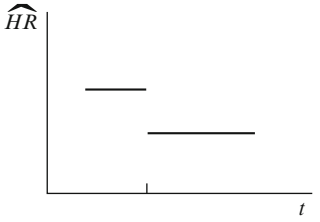
$$h(t,\mathbf{X}(t)) = h_0(t)\exp\left[\sum_{i=1}^{p}\beta_i\mathbf{X}_i + \sum_{i=1}^{p}\delta_i X_i g_i(t)\right]$$

We also showed how to use time-dependent variables to assess the PH assumption for time-independent variables. A general formula for an extended Cox model that simultaneously considers all time-independent variables of interest is shown here.

Examples of $g_i(t)$:
t, log t, heaviside function

The functions $g_i(t)$ denote functions of time for the ith variable that are to be determined by the investigator. Examples of such functions are $g_i(t) = t$, log t, or a heaviside function.

Heaviside functions:



The use of heaviside functions were described and illustrated. Such functions allow for the hazard ratio to be constant within different time intervals.

$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta E + \delta E g(t)]$

where

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

For two time intervals, the model can take either one of two equivalent forms as shown here. The first model contains a main effect of exposure and only one heaviside function. The second model contains two heaviside functions without a main effect of exposure. Both models yield two distinct and equivalent values for the hazard ratio.

$h(t,\mathbf{X}(t))$
$= h_0(t)\exp[\beta_1 E g_1(t) + \beta_2 E g_2(t)]$
where

$$g_1(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } < t_0 \end{cases}$$

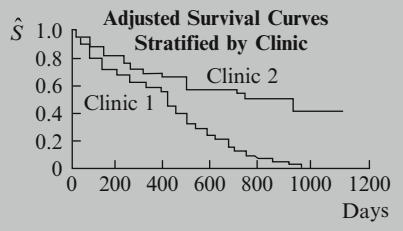
$$g_2(t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{if } t \geq t_0 \end{cases}$$

EXAMPLE 1

1991 Australian study of heroin addicts
- two methadone maintenance clinics
- *addicts* dataset file
- clinic variable did not satisfy PH assumption

We illustrated the use of time-dependent variables through two examples. The first example considered the comparison of two methadone maintenance clinics for heroin addicts. The dataset file was called *addicts*. In this example, the clinic variable, which was a dichotomous exposure variable, did not satisfy the PH assumption.

EXAMPLE: (continued)



Two distinct HR's:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta_1(\text{clinic})g_1(t) + \delta_2(\text{clinic})g_2(t)]$$

Heaviside functions

Diverging HR's:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \beta_1(\text{clinic}) + \delta(\text{clinic} \times t)]$$

Adjusted survival curves stratified by clinic showed clinic 2 to have consistently higher survival probabilities than clinic 1, with a more pronounced difference in clinics after one year of follow-up. However, this stratification did not allow us to obtain a hazard ratio estimate for clinic. Such an estimate was possible using an extended Cox model containing interaction terms involving clinic with time.

Two extended Cox models were considered. The first used heaviside functions to obtain two distinct hazard ratios, one for the first year of follow-up and the other for greater than one year of follow-up. The model is shown here.

The second extended Cox model used a time-dependent variable that allowed for the two survival curves to diverge over time. This model is shown here.

Both models yielded hazard ratio estimates that agreed reasonably well with the graph of adjusted survival curves stratified by clinic.

EXAMPLE 2: Stanford Heart Transplant Study

Goals: Do patients receiving transplants survive longer than patients not receiving transplants?

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta_1 HT(t) + \delta_2 TMS(t) + \delta_3 AGE(t)]$$

Exposure variable

The second example considered results obtained in the Stanford Heart Transplant Study. The goal of the study was to assess whether patients receiving transplants survived longer than patients not receiving transplants.

The analysis of these data involved an extended Cox model containing three time-dependent variables. One of these, the exposure variable, and called $HT(t)$, was an indicator of transplant status at time t. The other two variables, $TMS(t)$ and $AGE(t)$, gave tissue mismatch scores and age for transplant patients when time t occurred after receiving a transplant. The value of each of these variables was 0 at times prior to receiving a transplant.

EXAMPLE: (continued)

Results: $HT(t)$ highly significant, i.e., transplants have better prognosis than nontransplants.
Hazard ratio estimate problematic:

$$\widehat{HR} = e^{\hat{\delta}_1} = \frac{1}{23.98}$$

More appropriate formula:

$$\widehat{HR} = \exp[-3.1718 + 0.4442\ TMS_i + 0.0552\ AGE_i]$$

The results from fitting the above extended Cox model yielded a highly significant effect of the exposure variable, thus indicating that survival prognosis was better for transplants than for nontransplants.

From these data, we first presented an inappropriate formula for the estimated hazard ratio. This formula used the exponential of the coefficient of the exposure variable, which gave an estimate of 1 over 23.98. A more appropriate formula considered the values of the covariates $TMS(t)$ and $AGE(t)$ at time t. Using the latter, the hazard ratio estimate varied with the tissue mismatch scores and age of each transplant patient.

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Curves and the Log–Rank Test
3. The Cox Proportional Hazards Model
4. Evaluating the Proportional Hazards Assumption
5. The Stratified Cox Procedure
6. ✓ Extension of the Cox Proportional Hazards Model for Time-Dependent Variables

Next:

7. Parametric models

This presentation is now complete. We suggest that the reader review the detailed outline that follows and then answer the practice exercises and test that follow the outline.

A key property of Cox models is that the distribution of the outcome, survival time, is unspecified. In the next chapter, parametric models are presented in which the underlying distribution of the outcome is specified. The exponential, Weibull, and log-logistic models are examples of parametric models.

Detailed Outline

I. Preview (page 244)

II. Review of the Cox PH Model (pages 244–246)

A. The formula for the Cox PH model:

$$h(t,\mathbf{X}) = h_0(t)\exp\left[\sum_{i=1}^{p}\beta_i X_i\right]$$

B. Formula for hazard ratio comparing two individuals:

$$\mathbf{X}^* = \left(X_1^*, X_2^*, \ldots, X_p^*\right) \text{ and } \mathbf{X} = \left(X_1, X_2, \ldots, X_p\right):$$

$$\frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp\left[\sum_{i=1}^{p}\beta_i\left(X_i^* - X_i\right)\right]$$

C. The meaning of the PH assumption:

- Hazard ratio formula shows that the hazard ratio is independent of time:

$$\frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \theta$$

- Hazard ratio for two X's are proportional:

$$h(t,\mathbf{X}^*) = \theta h(t,\mathbf{X})$$

D. Three methods for checking the PH assumption:

i. *Graphical:* Compare ln–ln survival curves or observed versus predicted curves

ii. *Time-dependent covariates:* Use product (i.e., interaction) terms of the form $X \times g(t)$.

iii. *Goodness-of-fit test:* Use a large sample Z statistic.

E. Options when the PH assumption is not met:

i. Use a stratified Cox procedure.

ii. Use an extended Cox model containing a time-dependent variable of the form $X \times g(t)$.

III. Definition and Examples of Time-Dependent Variables (pages 246–249)

A. Definition: any variable whose values differ over time

B. Examples of defined, internal, and ancillary time-dependent variables

IV. The Extended Cox Model for Time-Dependent Variables (pages 249–251)

A. $$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right]$$

where $\mathbf{X}(t) = (X_1, X_2, \ldots, X_{p_1}, X_1(t), X_2(t), \ldots, X_{p_2}(t))$ denotes the entire collection of predictors at time t, X_i denotes the ith time-independent variable, and $X_j(t)$ denotes the jth time-dependent variable.

B. ML procedure used to estimate regression coefficients.

C. List of computer programs for the extended Cox model.

D. Model assumes that the hazard at time t depends on the value of $X_j(t)$ at the *same* time.

E. Can modify model for lag-time effect.

V. The Hazard Ratio Formula for the Extended Cox Model (pages 251–253)

A. $$HR(t) = \exp\left[\sum_{i=1}^{p_1} \beta_i \left[X_i^* - X_i\right] + \sum_{j=1}^{p_2} \delta_j \left[X_j^*(t) - X_j(t)\right]\right]$$

B. Because $HR(t)$ is a function of time, the PH assumption is not satisfied.

C. The estimated coefficient of $X_j(t)$ is time-independent, and represents an "overall" effect of $X_j(t)$.

VI. Assessing Time-Independent Variables That Do Not Satisfy the PH Assumption (pages 254–259)

A. General formula for assessing PH assumption:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp\left[\sum_{i=1}^{p} \beta_i X_i + \sum_{i=1}^{p} \delta_i X_i g_i(t)\right]$$

B. $g_i(t)$ is a function of time corresponding to X_i

C. Test H_0: $\delta_1 = \delta_2 = \ldots = \delta_p = 0$

D. Heaviside function:

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

E. The model with a single heaviside function:

$$\boxed{h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta E g(t)]}$$

F. The model with two heaviside functions:

$$\boxed{h(t, \mathbf{X}(t)) = h_0(t) \exp[\delta_1 E g_1(t) + \delta_2 E g_2(t)]}$$

where

$$g_1(t) = \begin{cases} 1 \text{ if } t \geq t_0 \\ 0 \text{ if } t < t_0 \end{cases} \quad \text{and} \quad g_2(t) = \begin{cases} 1 \text{ if } t < t_0 \\ 0 \text{ if } t \geq t_0 \end{cases}$$

G. The hazard ratios:

$$t \geq t_0 : \widehat{HR} = \exp\left[\hat{\beta} + \hat{\delta}\right] = \exp\left[\hat{\delta}_1\right]$$
$$t < t_0 : \widehat{HR} = \exp\left[\hat{\beta}\right] = \exp\left[\hat{\delta}_2\right]$$

H. Several heaviside functions: examples given with four time-intervals:

- Extended Cox model contains either $\{E, E \times g_1(t), E \times g_2(t), E \times g_3(t)\}$ or $\{E \times g_1(t), E \times g_2(t), E \times g_3(t), E \times g_4(t)\}$
- The model using four product terms and no main effect of E:

$$\boxed{\begin{aligned} h(t, \mathbf{X}(t)) = h_0(t) \exp[&\delta_1 E g_1(t) + \delta_2 E g_2(t) \\ &+ \delta_3 E g_3(t) + \delta_4 E g_4(t)] \end{aligned}}$$

where

$$g_i(t) = \begin{cases} 1 & \text{if } t \text{ is within interval } i \\ 0 & \text{if otherwise} \end{cases}$$

VII. An Application of the Extended Cox Model to An Epidemiologic Study on the Treatment of Heroin Addiction (pages 260–264)

A. 1991 Australian study of heroin addicts
   - two methadone maintenance clinics
   - *addicts* dataset file
   - clinic variable did not satisfy PH assumption

B. Clinic 2 has consistently higher retention probabilities than clinic 1, with a more pronounced difference in clinics after 1 year of treatment.

C. Two extended Cox models were considered:
   - Use heaviside functions to obtain two distinct hazard ratios, one for less than 1 year and the other for greater than 1 year.
   - Use a time-dependent variable that allows for the two survival curves to diverge over time.

VIII. An Application of the Extended Cox Model to the Analysis of the Stanford Heart Transplant Data (pages 265–269)

A. The goal of the study was to assess whether patients receiving transplants survived longer than patients not receiving transplants.

B. We described an extended Cox model containing three time-dependent variables:

$$h(t,\mathbf{X}(t)) = h_0(t)\exp[\delta_1 HT(t) + \delta_2 TMS(t) + \delta_3 AGE(t)]$$

C. The exposure variable, called $HT(t)$, was an indicator of transplant status at time t. The other two variables, $TMS(t)$ and AGE(t), gave tissue mismatch scores and age for transplant patients when time t occurred after receiving a transplant.

D. The results yielded a highly significant effect of the exposure variable.

E. The use of a hazard ratio estimate for this data was problematical.

- An inappropriate formula is the exponential of the coefficient of $HT(t)$, which yields 1/23.98.
- An alternative formula considers the values of the covariates $TMS(t)$ and $AGE(t)$ at time t.

IX. The Extended Cox Likelihood (pages 269–274)

A. Review of PH likelihood (Chapter 3).

B. Barry, Gary, Larry, example of Cox likelihood.

X. Summary (pages 274–277)

Practice Exercises

The following dataset called "anderson.dat" consists of remission survival times on 42 leukemia patients, half of whom receive a new therapy and the other half of whom get a standard therapy (Freireich et al., *Blood*, 1963). The exposure variable of interest is treatment status (Rx = 0 if new treatment, Rx = 1 if standard treatment). Two other variables for control are log white blood cell count (i.e., log WBC) and sex. Failure status is defined by the relapse variable (0 if censored, 1 if failure). The dataset is listed as follows:

Subj	Surv	Relapse	Sex	log WBC	*Rx*
1	35	0	1	1.45	0
2	34	0	1	1.47	0
3	32	0	1	2.2	0
4	32	0	1	2.53	0
5	25	0	1	1.78	0
6	23	1	1	2.57	0

(Continued on next page)

(Continued)

Subj	Surv	Relapse	Sex	log WBC	*Rx*
7	22	1	1	2.32	0
8	20	0	1	2.01	0
9	19	0	0	2.05	0
10	17	0	0	2.16	0
11	16	1	1	3.6	0
12	13	1	0	2.88	0
13	11	0	0	2.6	0
14	10	0	0	2.7	0
15	10	1	0	2.96	0
16	9	0	0	2.8	0
17	7	1	0	4.43	0
18	6	0	0	3.2	0
19	6	1	0	2.31	0
20	6	1	1	4.06	0
21	6	1	0	3.28	0
22	23	1	1	1.97	1
23	22	1	0	2.73	1
24	17	1	0	2.95	1
25	15	1	0	2.3	1
26	12	1	0	1.5	1
27	12	1	0	3.06	1
28	11	1	0	3.49	1
29	11	1	0	2.12	1
30	8	1	0	3.52	1
31	8	1	0	3.05	1
32	8	1	0	2.32	1
33	8	1	1	3.26	1
34	5	1	1	3.49	1
35	5	1	0	3.97	1
36	4	1	1	4.36	1
37	4	1	1	2.42	1
38	3	1	1	4.01	1
39	2	1	1	4.91	1
40	2	1	1	4.48	1
41	1	1	1	2.8	1
42	1	1	1	5	1

The following edited printout gives computer results for fitting a Cox PH model containing the three predictives *Rx*, log WBC, and Sex.

Cox regression Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]		*P*(*PH*)
Sex	0.263	0.449	0.558	1.301	0.539	3.139	0.042
log WBC	1.594	0.330	0.000	4.922	2.578	9.397	0.714
Rx	1.391	0.457	0.002	4.018	1.642	9.834	0.500

No. of subjects = 42 Log likelihood = −72.109

1. Which of the variables in the model fitted above are time-independent and which are time-dependent?
2. Based on this printout, is the PH assumption satisfied for the model being fit? Explain briefly.
3. Suppose you want to use an extended Cox model to assess the PH assumption for all three variables in the above model. State the general form of an extended Cox model that will allow for this assessment.
4. Suppose you wish to assess the PH assumption for the Sex variable using a heaviside function approach designed to yield a constant hazard ratio for less than 15 weeks of follow-up and a constant hazard ratio for 15 weeks or more of follow-up. State two equivalent alternative extended Cox models that will carry out this approach, one model containing one heaviside function and the other model containing two heaviside functions.
5. The following is an edited printout of the results obtained by fitting an extended Cox model containing two heaviside functions:

   Time-Dependent Cox Regression Analysis

Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio		[95% Conf. Interval]
log WBC	1.567	0.333	0.000	4.794	2.498	9.202
Rx	1.341	0.466	0.004	3.822	1.533	9.526
0–15 wks	0.358	0.483	0.459	1.430	0.555	3.682
15+ wks	−0.182	0.992	0.855	0.834	0.119	5.831

   No. of subjects = 42 Log likelihood = −71.980

   Using the above computer results, carry out a test of hypothesis, estimate the hazard ratio, and obtain 95% confidence interval for the treatment effect adjusted for log WBC and the time-dependent Sex variables. What conclusions do you draw about the treatment effect?
6. We now consider an alternative approach to controlling for Sex using an extended Cox model. We define an interaction term between sex and time that allows for diverging survival curves over time.

   For the situation just described, write down the extended Cox model, which contains *Rx*, log WBC, and Sex as main effects plus the product term Sex × time.

7. Using the model described in question 6, express the hazard ratio for the effect of Sex adjusted for *Rx* and log WBC at 8 and 16 weeks.

8. The following is an edited printout of computer results obtained by fitting the model described in question 6.

Time-Dependent Cox Regression Analysis

Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
Sex	1.820	1.012	0.072	6.174	0.849	44.896
log WBC	1.464	0.336	0.000	4.322	2.236	8.351
Rx	1.093	0.479	0.022	2.984	1.167	7.626
Sex × Time	−0.345	0.199	0.083	0.708	0.479	1.046

No. of subjects = 42 Log likelihood = −70.416

Based on the above results, describe the hazard ratio estimate for the treatment effect adjusted for the other variables in the model, and summarize the results of the significance test and interval estimate for this hazard ratio. How do these results compare with the results previously obtained when a heaviside function approach was used? What does this comparison suggest about the drawbacks of using an extended Cox model to adjust for variables not satisfying the PH assumption?

9. The following gives an edited printout of computer results using a stratified Cox procedure that stratifies on the Sex variable but keeps *Rx* and log WBC in the model.

Stratified Cox regression

Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf.	Interval]
log WBC	1.390	0.338	0.000	4.016	2.072	7.783
Rx	0.931	0.472	0.048	2.537	1.006	6.396

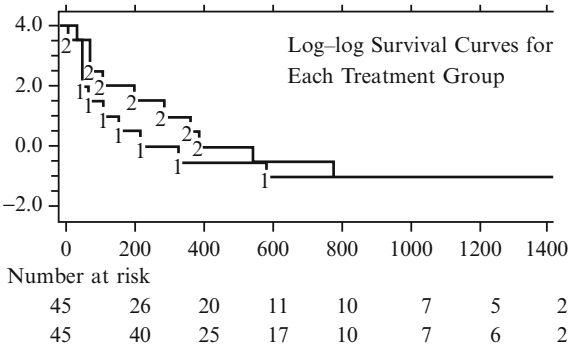
No. of subjects = 42 Log likelihood = −57.560 Stratified by sex

Compare the results of the above printout with previously provided results regarding the hazard ratio for the effect of *Rx*. Is there any way to determine which set of results is more appropriate? Explain.

Test

The following questions consider the analysis of data from a clinical trial concerning gastric carcinoma, in which 90 patients were randomized to either chemotherapy (coded as 2) alone or to a combination of chemotherapy and radiation (coded as 1). See Stablein et al., "Analysis of Survival Data with Nonproportional Hazard Functions," *Controlled Clinical Trials*, vol. 2, pp. 149–159 (1981).

1. A plot of the log–log Kaplan–Meier curves for each treatment group is shown below. Based on this plot, what would you conclude about the PH assumption regarding the treatment group variable? Explain.



2. The following is an edited printout of computer results obtained when fitting the PH model containing only the treatment group variable. Based on these results, what would you conclude about the PH assumption regarding the treatment group variable? Explain.

Cox regression Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]		*P*(*PH*)
Tx	−0.267	0.233	0.253	0.766	0.485	1.21	0
No. of subjects 90				Log likelihood = −282.744			

3. The following printout shows the results from using a heaviside function approach with an extended Cox model to fit these data. The model used product terms of the treatment variable (*Tx*) with each of three heaviside functions. The first product term (called Time1) involves a heaviside function for the period from 0 to 250 days, the second product term (i.e., Time2) involves the period from 250 to 500 days, and the third product term (i.e., Time3) involves the open-ended period from 500 days and beyond.

Time-Dependent Cox Regression Analysis

Analysis time_t: survt	Coef.	Std. Err.	p > |z|	Haz. Ratio	[95% Conf. Interval]	
Time1	−1.511	0.461	0.001	0.221	0.089	0.545
Time2	0.488	0.450	0.278	1.629	0.675	3.934
Time3	0.365	0.444	0.411	1.441	0.604	3.440

No. of subjects = 90 Log likelihood = −275.745

Write down the hazard function formula for the extended Cox model being used, making sure to explicitly define the heaviside functions involved.

4. Based on the printout, describe the hazard ratios in each of the three time intervals, evaluate each hazard ratio for significance, and draw conclusions about the extent of the treatment effect in each of the three time intervals considered.
5. Inspection of the printout provided in question 3 indicates that the treatment effect in the second and third intervals appears quite similar. Consequently, another analysis was considered that uses only two intervals, from 0 to 250 days versus 250 days and beyond. Write down the hazard function formula for the extended Cox model that considers this situation (i.e., containing two heaviside functions). Also, write down an equivalent alternative hazard function formula which contains the main effect of treatment group plus one heaviside function variable.
6. For the situation described in question 5, the computer results are provided below. Based on these results, describe the hazard ratios for the treatment effect below and above 250 days, summarize the inference results for each hazard ratio, and draw conclusions about the treatment effect within each time interval.

Time-Dependent Cox Regression Analysis

Analysis time_t: survt Column name	Coeff	StErr	p-value	*HR*	0.95	CI
Time1	−1.511	0.461	0.001	0.221	0.089	0.545
Time2	0.427	0.315	0.176	1.532	0.826	2.842

No. of subjects = 90 Log likelihood = −275.764

Answers to Practice Exercises

1. All three variables in the model are time-independent variables.
2. The computer results indicate that the Sex variables do not satisfy the PH assumption because the $P(PH)$ value is 0.042, which is significant at the 0.05 level.
3. $h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{sex}) + \beta_2(\log\ \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex})g_1(t) + \delta_2(\log\ \text{WBC})g_2(t) + \delta_3(Rx)g_3(t)]$
   where the $g_i(t)$ are functions of time.
4. Model 1 (one heaviside function)

   $$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{sex}) + \beta_2(\log\ \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex})g_1(t)]$$

   where

   $$g_1(t) = \begin{cases} 1 \text{ if } 0 \leq t < 15 \text{ weeks} \\ 0 \text{ if } t \geq 15 \text{ weeks} \end{cases}$$

   Model 2 (two heaviside functions):

   $h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_2(\log\ \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex})g_1(t) + \delta_2(\text{sex})g_2(t)]$

   where

   $$g_1(t) = \begin{cases} 1 \text{ if } 0 \leq t < 15 \text{ weeks} \\ 0 \text{ if } t \geq 15 \text{ weeks} \end{cases}$$

   and

   $$g_2(t) = \begin{cases} 0 \text{ if } t \geq 15 \text{ weeks} \\ 1 \text{ if } 0 \leq t < 15 \text{ weeks} \end{cases}$$

5. The estimated hazard ratio for the effect of Rx is 3.822; this estimate is adjusted for log WBC and for the Sex variable considered as two time-dependent variables involving heaviside functions. The Wald test for significance of Rx has a p-value of 0.004, which is highly significant. The 95% confidence interval for the treatment effect ranges between 1.533 and 9.526, which is quite wide, indicating considerable unreliability of the 3.822 point estimate. Nevertheless, the results estimate a statistically significant treatment effect of around 3.8.
6. $h(t,\mathbf{X}(t)) = h_0(t)\exp[\beta_1(\text{sex}) + \beta_2(\log\ \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex} \times \text{t})]$

7. The hazard ratio for the effect of Sex in each time interval, controlling for *Rx* and log WBC is given as follows:

$$t = 8 \text{ weeks} \quad \widehat{HR} = \exp\left[\hat{\beta}_1 + 8\hat{\delta}_1\right]$$
$$t = 16 \text{ weeks} \quad \widehat{HR} = \exp\left[\hat{\beta}_1 + 16\hat{\delta}_1\right]$$

8. Using the model containing Sex, log WBC, *Rx*, and Sex $\times$ Time, the estimated hazard ratio for the treatment effect is given by 2.984, with a p-value of 0.022 and a 95% confidence interval ranging between 1.167 and 7.626. The point estimate of 2.984 is quite different from the point estimate of 3.822 for the heaviside function model, although the confidence intervals for both models are wide enough to include both estimates. The discrepancy between point estimates demonstrates that when a time-dependent variable approach is to be used to account for a variable not satisfying the PH assumption, different results may be obtained from different choices of time-dependent variables.

9. The stratified Cox analysis yields a hazard ratio of 2.537 with a p-value of 0.048 and a 95% CI ranging between 1.006 and 6.396. The point estimate is much closer to the 2.984 for the model containing the Sex $\times$ Time product term than to the 3.822 for the model containing two heaviside functions. One way to choose between models would be to compare goodness-of-fit test statistics for each model; another way is to compare graphs of the adjusted survival curves for each model and determine by eye which set of survival curves fits the data better.

7 Parametric Survival Models

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_7,
© Springer Science+Business Media, LLC 2012

Introduction

The Cox model is the most widely used survival model in the health sciences, but it is not the only model available. In this chapter we present a class of survival models, called parametric models, in which the distribution of the outcome (i.e., the time to event) is specified in terms of unknown parameters. Many parametric models are acceleration failure time models in which survival time is modeled as a function of predictor variables. We examine the assumptions that underlie accelerated failure time models and compare the acceleration factor as an alternative measure of association to the hazard ratio. We present examples of the exponential, Weibull, and log-logistic models and give a brief description of other parametric approaches. The parametric likelihood is constructed and described in relation to left, right, and interval-censored data. Binary regression is presented as an alternative approach for modeling interval-censored outcomes. The chapter concludes with a discussion of frailty models.

Abbreviated Outline

The outline below gives the user a preview of the material covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Overview** (pages 292–294)
II. **Probability Density Function in Relation to the Hazard and Survival Function** (pages 294–295)
III. **Exponential Example** (pages 295–297)
IV. **Accelerated Failure Time Assumption** (pages 298–300)
V. **Exponential Example Revisited** (pages 300–304)
VI. **Weibull Example** (pages 304–309)
VII. **Log-Logistic Example** (pages 309–314)
VIII. **A More General Form of the AFT Model** (pages 314–316)
IX. **Other Parametric Models** (pages 316–318)
X. **The Parametric Likelihood** (pages 318–321)
XI. **Interval-Censored Data** (pages 321–326)
XII. **Frailty Models** (pages 326–340)
XIII. **Summary** (pages 341–344)

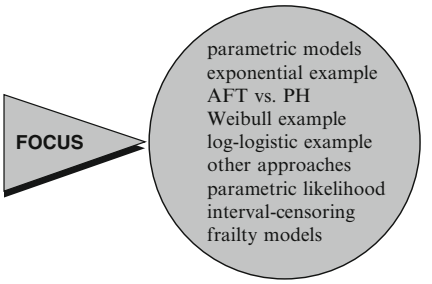
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize the form of a parametric survival model and contrast it with a Cox model.
2. State common distributions used for parametric survival models.
3. Contrast an AFT model with a PH model.
4. Interpret output from an exponential survival model.
5. Interpret output from a Weibull survival model.
6. Interpret output from a log-logistic survival model.
7. State or recognize the formulation of a parametric likelihood.
8. State or recognize right-censored, left-censored, and interval-censored data.
9. State or recognize the form of a frailty model and the purpose of including a frailty component.
10. Interpret the output obtained from a frailty model.

Presentation

I. Overview



In this chapter we present parametric survival models and the assumptions that underlie these models. Specifically we examine the accelerated failure time (AFT) assumption and contrast it to the proportional hazards (PH) assumption. We present examples of several parametric models, including the exponential model, the Weibull model, and the log-logistic model. The parametric likelihood is discussed and how it accommodates left-, right-, and interval-censored data. We also consider models that include a frailty component to account for unobserved heterogeneity.

Parametric Modeling

- Outcome assumed to follow some family of distributions
- Exact distribution is unknown if parameters are unknown
- Data used to estimate parameters
- Examples of parametric models:
  - Linear regression
  - Logistic regression
  - Poisson regression

Linear regression, logistic regression, and Poisson regression are examples of parametric models that are commonly used in the health sciences. With these models, the **outcome is assumed to follow some distribution** such as the normal, binomial, or Poisson distribution. Typically, what is actually meant is that the outcome follows some family of distributions of similar form with unknown parameters. It is only when the value of the parameter(s) is known that the exact distribution is fully specified. For example, if one distribution is normal with a mean of 3 and another distribution is normal with a mean of 7, the distributions are of the same family (i.e., normal) but they are not exactly the same distribution. For parametric regression models, the data are typically used to estimate the values of the parameters that fully specify that distribution.

Distributions commonly used for parametric survival models:

- Weibull
- Exponential
- Log-logistic
- Lognormal
- Generalized gamma

A **parametric survival model** is one in which survival time (the outcome) is assumed to follow a known distribution. Examples of distributions that are commonly used for survival time are: the **Weibull**, the **exponential** (a special case of the Weibull), the **log-logistic**, the **lognormal**, and the **generalized gamma**, which are supported by SAS, Stata, and R software.

Parametric survival models → Distribution specified for time

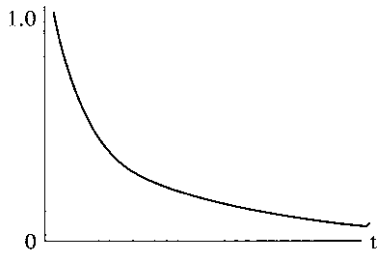
Cox model is semiparametric: → Baseline survival not specified

The Cox proportional hazards model, by contrast, is not a fully parametric model. Rather it is a **semi-parametric model** because even if the regression parameters (the betas) are known, the distribution of the outcome remains unknown. The baseline survival (or hazard) function is not specified in a Cox model.

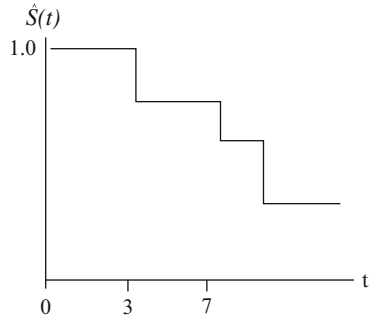
Cox model widely popular:

- No reliance on assumed distribution
- Computer packages can output Cox-adjusted survival estimates using algorithm that generalizes KM
- Baseline not necessary for estimation of hazard ratio

A key reason why the Cox model is widely popular is that it does not rely on distributional assumptions for the outcome. Although the baseline survival function is not estimated with a Cox model, computer packages such as SAS, Stata, SPSS, and R can output Cox-adjusted survival estimates (see Computer Appendix) by using a complicated algorithm that generalizes the Kaplan–Meier (KM) approach while making use of estimated regression coefficients obtained from a Cox model (Kalbfleisch and Prentice, 1980). Also, an estimation of the baseline hazard is not necessary for the estimation of a hazard ratio because the baseline hazard cancels in the calculation.



Theoretical $S(t)$



Step function (nondistributional estimates)

In theory, as time ranges from 0 to infinity, the survival function can be graphed as a smooth curve from $S(0) = 1$ to $S(\infty) = 0$ (see Chapter 1). Kaplan–Meier and Cox-adjusted survival estimates use empirical nondistributional methods that typically graph as step functions, particularly if the sample size is small. If in the data, for example, an event occurred at 3 weeks and the next event occurred at 7 weeks, then the estimated survival curve would be flat between 3 and 7 weeks using these nondistributional approaches. Moreover, if the study ends with subjects still remaining at risk, then the estimated survival function would not go all the way down to zero.

Appeal of Parametric Survival Models

- More consistent with theoretical S(t) than nondistributional approaches
- Simplicity
- Completeness — h(t) and S(t) specified

Survival estimates obtained from parametric survival models typically yield plots more consistent with a theoretical survival curve. If the investigator is comfortable with the underlying distributional assumption, then parameters can be estimated that completely specify the survival and hazard functions. This simplicity and completeness are the main appeals of using a parametric approach.

II. Probability Density Function in Relation to the Hazard and Survival Function

Probability density function known
⇓
Survival and hazard functions

For parametric survival models, time is assumed to follow some distribution whose probability density function f(t) can be expressed in terms of unknown parameters. Once a probability density function is specified for survival time, the corresponding survival and hazard functions can be determined. The survival function $S(t) = P(T > t)$ can be ascertained from the probability density function by integrating over the probability density function from time t to infinity. The hazard can then be found by dividing the negative derivative of the survival function by the survival function (see left).

$f(t) = dF(t)/dt$ where
$F(t) = Pr(T \leq t)$

$$S(t) = P(T > t) = \int_t^\infty f(u)du$$

$$h(t) = \frac{-d[S(t)]/dt}{S(t)}$$

Survival in terms of hazard

$$S(t) = \exp\left(-\int_0^t h(u)du\right)$$

The survival function can also be expressed in terms of the hazard function (see Chapter 1) by exponentiating the negative of the cumulative hazard function. The cumulative hazard function is the integral of the hazard function between integration limits of 0 and t.

Cumulative hazard: $\int_0^t h(u)du$

$f(t) = h(t)S(t)$

Finally, the probability density function can be expressed as the product of the hazard and the survival functions, $f(t) = h(t)S(t)$.

Key Point

Specifying one of f(t), S(t), or h(t) specifies all three functions

The key point is that specifying **any one of the probability density function, survival function, or hazard function allows the other two functions to be ascertained** by using the formulas shown on the left.

Survival and Hazard Functions for Selected Distributions

Distribution	$S(t)$	$h(t)$
Exponential	$\exp(-\lambda t)$	λ
Weibull	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{1}{1+\lambda t^p}$	$\frac{\lambda p t^{p-1}}{1+\lambda t^p}$

On the left is a table containing the survival and hazard functions for three of the more commonly used distributions for survival models: the exponential, Weibull, and log-logistic distributions.

The exponential is a one-parameter distribution with a constant hazard λ. The Weibull and log-logistic distributions have two parameters λ and p. Notice that the Weibull distribution reduces to the exponential if $p = 1$. The probability density function for these distributions can be found by multiplying h(t) and S(t). As an example, the Weibull probability density function is shown on the left.

$f(t) = h(t)S(t)$

For example, Weibull:
$f(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$
because $h(t) = \lambda p t^{p-1}$ and
$S(t) = \exp(-\lambda t^p)$

Typically in parametric models:

- λ reparameterized for regression
- p held fixed

Typically for parametric survival models, the parameter λ is reparameterized in terms of predictor variables and regression parameters and the parameter p (sometimes called the shape parameter) is held fixed. This is illustrated in the examples to come.

III. Exponential Example

Simplest parametric survival model:
Hazard function: $h(t) = \lambda$
(where λ is a constant)

EXAMPLE

Remission data (n = 42)

21 patients given treatment (TRT = 1)
21 patients given placebo (TRT = 0)

The first example we consider is the exponential model, which is the simplest parametric survival model in that the hazard is constant over time (i.e., $h(t) = \lambda$). The model is applied to the remission data (Freireich et al., 1963), in which 42 leukemia patients were followed until remission or censorship. Twenty-one patients received an experimental treatment (coded TRT = 1) and the other 21 received a placebo (coded TRT = 0). The data are listed in Chapter 1. The variable TRT is just a reverse coding of the variable RX presented in Chapter 3.

$h(t) = \lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$

$\text{TRT} = 1$: $h(t) = \exp(\beta_0 + \beta_1)$
$\text{TRT} = 0$: $h(t) = \exp(\beta_0)$

$\text{HR}(\text{TRT} = 1 \text{ vs. } \text{TRT} = 0)$

$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

For simplicity, we demonstrate an exponential model that has TRT as the only predictor. We state the model in terms of the hazard by reparameterizing λ as $\exp(\beta_0 + \beta_1 \text{TRT})$. With this model, the hazard for subjects in the treated group is $\exp(\beta_0 + \beta_1)$ and the hazard for the placebo group is $\exp(\beta_0)$. The hazard ratio comparing the treatment and placebo (see left side) is the ratio of the hazards $\exp(\beta_1)$. The exponential model is a proportional hazards model.

Constant Hazards
$\Rightarrow$ Proportional Hazards

Proportional Hazards
$\not\Rightarrow$ Constant Hazards

The assumption that the hazard is constant for each pattern of covariates is a **much stronger assumption** than the PH assumption. If the hazards are constant, then of course the ratio of the hazards is constant. However, the **hazard ratio being constant does not necessarily mean that each hazard is constant**. In a Cox PH model the baseline hazard is not assumed constant. In fact, the form of the baseline hazard is not even specified.

Exponential Model — Hazards are constant

Cox PH Model — Hazards are proportional not necessarily constant

Remission Data

Exponential regression
log hazard form

_t	Coef.	Std. Err.	z	P>|z|
trt	−1.527	.398	−3.83	0.00
_cons	−2.159	.218	−9.90	0.00

Output from running the exponential model is shown on the left. The model was run using Stata software. The parameter estimates are listed under the column called Coef. The parameter estimate for the coefficient of TRT (β_1) is −1.527. The estimate of the intercept (called cons in the output) is −2.159. The standard errors (Std. Err.), Wald test statistics (z), and p-values for the Wald test are also provided. The output indicates that the z test statistic for TRT is statistically significant with a p-value <0.005 (rounds to 0.00 in the output).

Coefficient estimates obtained by MLE
↘
asymptotically normal

The regression coefficients are estimated **using maximum likelihood estimation (MLE)**, and are **asymptotically normally** distributed.

$\text{TRT} = 1$: $\hat{h}(t) = \exp(-2.159 + (-1.527)) = 0.025$
$\text{TRT} = 0$: $\hat{h}(t) = \exp(-2.159) = 0.115$

$\widehat{HR}\,(\text{TRT} = 1 \text{ vs } .0) = \exp(-1.527) = 0.22$
$95\%\text{CI} = \exp[-1.527 \pm 1.96(0.398)] = (0.10, 0.47)$

The estimated hazards for TRT = 1 and TRT = 0 are shown on the left. The estimated hazard ratio of 0.22 is obtained by exponentiating the estimated coefficient (−1.527) of the TRT variable. A 95% confidence interval can be calculated exp[−1.527 ± 1.96(0.398)] yielding a CI of (0.10, 0.47). These results suggest that the experimental treatment delays remission.

Results: suggest treatment lowers hazard

Parametric models

- Need not be PH models
- Many are AFT models

Up to this point in the book, the key assumption for survival models has been the proportional hazard assumption. However, parametric survival models need not be PH models. **Many parametric models are acceleration failure time (AFT) models rather than PH models**. The exponential and Weibull distributions can accommodate both the PH and AFT assumptions.

Exponential and Weibull

- Accommodate PH and AFT assumptions

Remission Data

Exponential regression accelerated failure-time form

_t	Coef.	Std. Err.	z	p > |z|
trt	1.527	.398	3.83	0.00
_cons	2.159	.218	9.90	0.00

On the left is Stata output from the AFT form of the exponential model with TRT as the only predictor. Stata can output both the PH or AFT form of an exponential or Weibull model (see Computer Appendix). SAS only runs the AFT form of parametric models and SPSS does not yet provide commands to run parametric models.

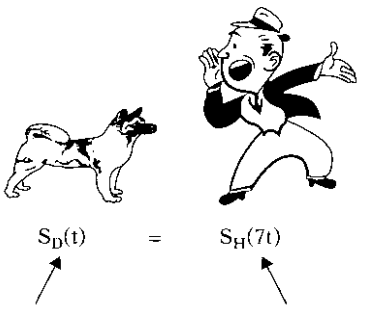
AFT vs. *PH*

- Different interpretation of parameters
- AFT applies to comparison of survival times
- PH applies to comparison of hazards

The interpretation of parameters differs for AFT and PH models. The AFT assumption is applicable for a comparison of survival times whereas the PH assumption is applicable for a comparison of hazards. In the following sections, we discuss the AFT assumption and then revisit this example and discuss the AFT form of this model.

IV. Accelerated Failure Time Assumption

AFT — Multiplicative effect with survival time
PH — Multiplicative effect with hazard



AFT models:
Describe "stretching out" or contraction of survival time

The underlying assumption for AFT models is that the effect of covariates is **multiplicative (proportional) with respect to survival time**, whereas for PH models the underlying assumption is that the effect of covariates is **multiplicative with respect to the hazard**.

To illustrate the idea underlying the AFT assumption, consider the lifespan of dogs. It is often said that dogs grow older seven times faster than humans. So a 10-year-old dog is in some way equivalent to a 70-year-old human. In AFT terminology we might say the probability of a dog surviving past 10 years equals the probability of a human surviving past 70 years. Similarly, we might say the probability of a dog surviving past 6 years equals the probability of a human surviving past 42 years because 42 equals 6 times 7. More generally we can say $SD(t) = SH(7t)$, where $SD(t)$ and $SH(t)$ are the survival functions for dogs and humans, respectively. In this framework dogs can be viewed, on average, as accelerating through life 7 times faster than humans. Or from the other perspective, the lifespan of humans, on average, is stretched out 7 times longer than the lifespan of dogs. **AFT models describe this "stretching out" or contraction of survival time as a function of predictor variables**.

Second Illustration

$S_1(t)$ — Survival function for smokers
$S_2(t)$ — Survival function for nonsmokers

AFT assumption:
$S_2(t) = S_1(\gamma t)$ for $t \geq 0$
γ is the acceleration factor

If $\gamma = \exp(\beta)$
$S_2(t) = S_1([\exp(\alpha)]t)$
or
$S_2([\exp(-\alpha)]t) = S_1(t)$

For a second illustration of the accelerated failure time assumption consider a comparison of survival functions among smokers $S_1(t)$ and non-smokers $S_2(t)$. The AFT assumption can be expressed as $S_2(t) = S_1(\gamma t)$ for $t \geq 0$, where γ is a constant called the **acceleration factor** comparing smokers to nonsmokers. In a regression framework the acceleration factor γ could be parameterized as $\exp(\alpha)$ where α is a parameter to be estimated from the data. With this parameterization, the AFT assumption can be expressed as $S_2(t) = S_1(\exp(\alpha)t)$ or equivalently: $S_2(\exp(-\alpha)t) = S_1(t)$ for $t \geq 0$.

Suppose $\exp(\alpha) = 0.75$
then
$S_2(80) = S_1(60)$
$S_2(40) = S_1(30)$

More generally
$S_2(t) = S_1(0.75t)$

Suppose $\exp(\alpha) = 0.75$; then the probability of a nonsmoker surviving 80 years equals the probability of a smoker surviving 80(0.75) or 60 years. Similarly, the probability of a nonsmoker surviving 40 years equals the probability of a smoker surviving 30 years. More generally, the probability of a nonsmoker surviving t years equals the probability of a smoker surviving 0.75 times t years (i.e., $S_2(t) = S_1(0.75t)$).

T_1 — Survival time for smokers
T_2 — Survival time for nonsmokers

AFT assumption in terms of random variables:
$T_1 = \gamma T_2$

The AFT assumption can also be expressed in terms of random variables for survival time rather than the survival function. If T_2 is a random variable (following some distribution) representing the survival time for nonsmokers and T_1 is a random variable representing the survival time for smokers, then the AFT assumption can be expressed as $T_1 = \gamma\, T_2$.

Acceleration factor
Measure of association on survival time

Hazard ratio
Measure of association on the hazard

The **acceleration factor** is the key measure of association obtained in an AFT model. It allows the investigator to evaluate the effect of predictor variables on survival time just as the hazard ratio allows the evaluation of predictor variables on the hazard.

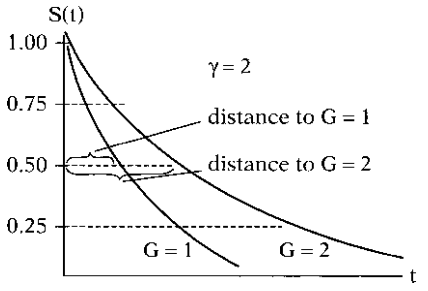
Acceleration factor (γ)

- Describes stretching or contraction of S(t)
- Ratio of times to any fixed value of S(t)

Suppose $\gamma = 2.0$
(Group 2 vs. Group 1)

- Time to $S(t) = 0.50$ (median) is double for Group 2
- Time to $S(t) = 0.20$ is double for Group 2
- Time to $S(t) = 0.83$ is double for Group 2
- Time to $S(t) = 0.98$ is double for Group 2
- Time to $S(t) =$ **q** is double for Group 2 (**generalization**)

The acceleration factor describes the "stretching out" or contraction of survival functions when comparing one group to another. More precisely, the **acceleration factor is a ratio of survival times corresponding to any fixed value of S(t)**. For example, if the acceleration factor comparing subjects in Group 2 vs. Group 1 is $\gamma = 2.0$, then the median survival time (value of t when $S(t) = 0.5$) for Group 2 is double the median survival time for Group 1. Moreover, the time it takes for S(t) to equal 0.2 or 0.83 or 0.98 is double for Group 2 compared to Group 1 for the same value of S(t). In general, the acceleration factor is a ratio of survival times corresponding to any quantile of survival time ($S(t) = q$).



Survival curves for Group 1 (G = 1) and Group 2 (G = 2)

Horizontal lines are twice as long to G = 2 compared to G = 1 because $\gamma = 2$

This idea is graphically illustrated by examining the survival curves for Group 1 (G = 1) and Group 2 (G = 2) shown on the left. For any fixed value of S(t), the distance of the horizontal line from the S(t) axis to the survival curve for G = 2 is double the distance to the survival curve for G = 1. Notice the median survival time (as well as the 25th and 75th percentiles) is double for G = 2. **For AFT models, this ratio of survival times is assumed constant for all fixed values of S(t)**.

V. Exponential Example Revisited

Remission data (n = 42)

21 patients given treatment (TRT = 1)
21 patients given placebo (TRT = 0)

We return to the exponential example applied to the remission data with treatment status (TRT) as the only predictor. In Section III, results from the PH form of the exponential model were discussed. In this section we discuss the AFT form of the model.

Previously discussed PH form of model
Now discuss AFT form of model

Exponential survival and hazard functions:

$S(t) = \exp(-\lambda t)$
$h(t) = \lambda$

The exponential survival and hazard functions are shown on the left. Recall that the exponential hazard is constant and can be reparameterized as a PH model, $h(t) = \lambda = \exp(\beta_0 + \beta_1 \text{TRT})$. In this section we show how S(t) can be reparameterized as an AFT model.

Recall for PH model:

$h(t) = \lambda = \exp(\beta_0 + \beta_1 \text{TRT})$

AFT assumption
(comparing 2 levels of TRT)

- Ratio of times is constant for all fixed S(t)

Strategy for developing the model:

- Solve for t in terms of S(t)
- Scale t in terms of the predictors

The underlying AFT assumption, for comparing the two levels of the TRT covariate, is that the ratio of times to any fixed value of $S(t) = q$ is constant for any probability q. We develop the AFT model with the survival function by solving for t in terms of S(t). We then scale t in terms of the predictors.

$$S(t) = \exp(-\lambda t)$$

$$t = [-\ln(S(t)] \times \frac{1}{\lambda}$$

$$\text{let } \frac{1}{\lambda} = \exp(\alpha_0 + \alpha_1 \text{TRT})$$

$$t = [-\ln(S(t)] \times \mathbf{exp}(\alpha_0 + \alpha_1 \mathbf{TRT})$$

Scaling of t

Median survival time, $S(t) = 0.5$:

$$t_m = [-\ln(0.5)] \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

The exponential survival function is $S(t) = \exp(-\lambda t)$. By solving for t, we can obtain a formula for t in terms of S(t). Taking the natural log, multiplying by negative 1, and then multiplying by the reciprocal of λ, yields the expression for t shown on the left. By reparameterizing $1/\lambda = \exp(\alpha_0 + \alpha_1 \text{TRT})$, or equivalently $\lambda = \exp[-(\alpha_0 + \alpha_1 \text{TRT})]$, it can be seen how the predictor variable TRT is used to scale the time to any fixed value of S(t) (see left). For example, to find an expression for the median survival time t_m, substitute $S(t) = 0.5$ (see left).

Let $S(t) = q$

$$t = [-\ln(q)] \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

Acceleration Factor:
γ(TRT = 1 vs. TRT = 0)

$$\gamma = \frac{[-\ln(q)]\exp(\alpha_0 + \alpha_1)}{[-\ln(q)]\exp(\alpha_0)}$$

$$= \exp(\alpha_1)$$

The expression for t is restated on the left in terms of any fixed probability $S(t) = q$. The acceleration factor γ is found by taking the ratio of the times to $S(t) = q$ for TRT = 1 and TRT = 0. After canceling, γ reduces to $\exp(\alpha_1)$.

Thus, the exponential hazard function for the remission data involving a single (0,1) TRT predictor satisfies the AFT model assumption shown at the left.

$$S_{\text{TRT}=0}(t) = S_{\text{TRT}=1}(\gamma t)$$
where $\gamma = \exp(\alpha_1)$

$$\text{AFT} : \frac{1}{\lambda} = \exp(\alpha_0 + \alpha_1 \text{TRT})$$

$$\text{PH} : \lambda = \exp(\beta_0 + \beta_1 \text{TRT})$$

The formula for λ as the AFT model form is compared at the left with the (different) formula for λ in the PH model form.

Remission Data

Exponential regression accelerated failure-time form

_t	Coef.	Std. Err.	z	P>|z|
trt	1.527	.398	3.83	0.00
_cons	2.159	.218	9.90	0.00

$\hat{\gamma} = \exp(1.527) = 4.60$

95% CI: $\exp[1.527 \pm 1.96(0.398)] = (2.11, 10.05)$

On the left is Stata output from the AFT form of the exponential model with TRT as the only predictor. The estimate of the coefficient for TRT is 1.527 with a standard error of 0.398. An estimate of the acceleration factor for treatment is $\hat{\gamma} = \exp(1.527) = 4.60$. A 95% confidence interval for γ is calculated as $\exp[1.527 \pm 1.96(0.398)]$ yielding a CI of (2.11, 10.05).

$t = [-\ln(q)] \times \exp(\alpha_0 + \alpha_1 TRT)$

$\hat{t} = [-\ln(q)] \times \exp(2.159 + 1.527(TRT))$

Estimated Survival Times by S(t) Quartiles for TRT = 1 and TRT = 0 (Exponential Model)

S(t) = q	$\hat{t}_{TRT=0}$	$\hat{t}_{TRT=1}$
0.25	12.0	55.3
0.50	6.0	27.6
0.75	2.5	11.5

The parameter estimates can be used to estimate the time $\hat{t}$ to any value of S(t) = q. The table on the left lists the estimated time (in weeks) for the first, second (median), and third quartiles of S(t) using the expression for $\hat{t}$ shown above for both the treated and placebo groups. In this example, survival time is the time to remission for leukemia patients.

$\hat{\gamma} = 4.60$ (for TRT = 1 vs. TRT = 0)

Ratio of survival times:

$$\frac{55.3}{12.0} = \frac{27.6}{6.0} = \frac{11.5}{2.5} = 4.60$$

The ratio of survival times for each row in the table comparing TRT = 1 vs. TRT = 0 is 4.60, which not coincidently is the estimate of the acceleration factor (see left). The estimated acceleration factor suggests that the experimental treatment is effective for delaying remission by stretching survival time by a factor of 4.60. Although the hazard ratio is a more familiar measure of association for health scientists, the acceleration factor has an intuitive appeal, particularly for describing the efficacy of a treatment on survival.

Effect of treatment:

- Stretches survival by a factor of 4.6
- Interpretation of γ has intuitive appeal

HR and γ are reciprocals in exponential models:

$$\widehat{HR}(\text{TRT} = 1 \text{ vs. } 0) = \exp(-1.527) = 0.22$$

$$\hat{\gamma}(\text{TRT} = 1 \text{ vs. } 0) = \exp(1.527) = 4.60$$

In general

$\gamma > 1 \Rightarrow$ exposure benefits survival
HR $> 1 \Rightarrow$ exposure harmful to survival

$\gamma > 1 \Rightarrow$ exposure harmful to survival
HR $< 1 \Rightarrow$ exposure benefits survival

$\gamma = \text{HR} = 1 \Rightarrow$ no effect from exposure

Recall from Section III that the hazard ratio for the effect of treatment was estimated at $\exp(-1.527) = 0.22$ using the PH form of the exponential model. This result illustrates a key property of the exponential model: the corresponding acceleration factor and hazards ratio (e.g., TRT = 1 vs. TRT = 0) are reciprocals of each other. This property is unique to the exponential model. What can be generalized, however, is that **an acceleration factor greater than one for the effect of an exposure implies that being exposed (i.e., TRT = 1) is beneficial to survival whereas a hazard ratio greater than one implies being exposed is harmful to survival (and vice versa)**.

Exponential PH and AFT models:

- Same model
- Different parameterization
- Same estimates for
  - Survival function
  - Hazard function
  - Median survival

Although the exponential PH and AFT models focus on different underlying assumptions, they are in fact **the same model**. The only difference is in their parameterization. The resulting estimates for the survival function, hazard function, and median survival do not differ between these models (see Practice Exercises 6 and 7).

For those experienced with Poisson regression:

Exponential and Poisson models

- Assume a constant rate
- Different data structure
  - Poisson — aggregate counts
  - Exponential — individual level
- Use different outcomes
  - Poisson — number of cases
  - Exponential — time to event
- Yield equivalent parameter estimates
  - With same data and same covariates in the model

For those who have experience with Poisson regression, there is a close connection between the exponential and Poisson models. Both distributions assume an underlying constant rate. In fact, if the data are structured such that all the cases and the total time at risk are aggregated for each pattern of covariates (e.g., TRT = 1 and TRT = 0) and the log of the corresponding person-time at risk is used as an offset, then a Poisson model will yield equivalent parameter estimates as the exponential PH model. The difference is that the random outcome for the Poisson model is the count of events over a fixed amount of time at risk and the random outcome for the exponential model is the time (at risk) to event.

Exponential model is special case of Weibull model

We continue with the remission data example and present the more general Weibull model, which includes the exponential model as a special case. In the next section we also show a graphical approach for evaluating the appropriateness of the Weibull (and thus also the exponential) model.

VI. Weibull Example

Weibull Model:

Hazard function: $h(t) = \lambda p t^{p-1}$ (where $p > 0$ and $\lambda > 0$)

p is a shape parameter

- $p > 1$ hazard increases over time
- $p = 1$ constant hazard (exponential model)
- $p < 1$ hazard decreases over time

Additional shape parameter offers greater flexibility

The **Weibull model** is the most widely used parametric survival model. Its hazard function is $h(t) = \lambda p t^{p-1}$, where p and $\lambda > 0$. As with the exponential model, λ will be reparameterized with regression coefficients. The additional parameter p is called a **shape parameter** and determines the shape of the hazard function. If $p > 1$ then the hazard increases as time increases. If $p = 1$ then the hazard is constant and the Weibull model reduces to the exponential model ($h(t) = \lambda$). If $p < 1$ then the hazard decreases over time. The addition of this shape parameter gives the Weibull model greater flexibility than the exponential model, yet the hazard function remains relatively simple (basically a scaling of t raised to some fixed power).

Unique property for Weibull model
AFT ⇒ PH and PH ⇒ AFT
Holds if p is fixed

The Weibull model has the property that **if the AFT assumption holds then the PH assumption also holds** (and vice versa). This **property is unique** to the Weibull model (Cox and Oakes, 1984) and holds if p does not vary over different levels of covariates.

HR vs. AFT

Hazard ratio ⇒ Comparison of rates

Acceleration factor ⇒ Effect on survival

The PH assumption allows for the estimation of a hazard ratio enabling a comparison of rates among different populations. The AFT assumption allows for the estimation of an acceleration factor, which can describe the direct effect of an exposure on survival time.

Useful Weibull property:

- ln[−ln S(t)] is linear with ln(t)
- Enables graphical evaluation using KM survival estimates

The Weibull model also has another key property: **the log(–log) of S(t) is linear with the log of time**. This allows a graphical evaluation of the appropriateness of a Weibull model by **plotting the log negative log of the Kaplan–Meier survival estimates against the log of time.**

Linearity of ln(t)

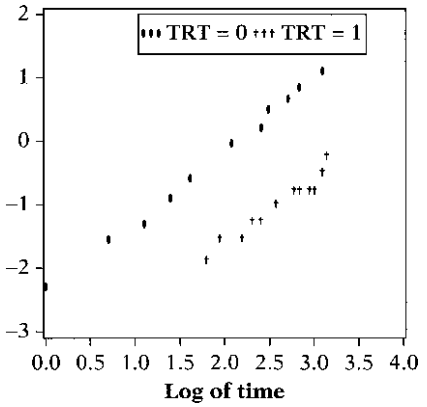
$$S(t) = \exp(-\lambda t^p)$$
$$\Rightarrow \ln[-\ln S(t)] = \ln(\lambda) + p \ln(t)$$

Intercept = $\ln(\lambda)$, Slope = p

To see this linear relationship: start with the Weibull survival function $S(t) = \exp(-\lambda t^p)$, take the log of S(t), multiply by negative one, and take the log again (see left). For the Weibull distribution, the $\ln[-\ln(S(t))]$ is a linear function of ln(t) with slope p and intercept $p \ln(\lambda)$. If the slope equals one then t follows an exponential distribution.

Remission data: evaluate Weibull assumption for TRT = 1 and TRT = 0

$\ln[-\ln \hat{S}(t)]$ plotted against ln(t)



We again return to the remission data and evaluate the appropriateness of the Weibull assumption for the treated (TRT = 1) and placebo (TRT = 0) groups. On the left is the plot of the log negative log Kaplan-Meier survival estimates against the log of time for TRT = 1 and TRT = 0. Both plots look reasonably straight suggesting that the Weibull assumption is reasonable. Furthermore, the lines appear to have the same slope (i.e., are parallel, same p) suggesting that the PH (and thus the AFT) assumptions hold. If this common slope equals one (i.e., $p = 1$), then survival time follows an exponential distribution. The Weibull model output containing the parameter estimates includes a statistical test for the hypothesis $p = 1$ or equivalently for $\ln(p) = 0$ (for testing the exponential assumption). This is examined later in this section.

Summary of possible results for plot of $\ln[-\ln \hat{S}(t)]$ against ln(t)

1. Parallel straight lines ⇒ Weibull, PH, and AFT assumptions hold
2. Parallel straight lines with slope of 1 ⇒ Exponential. PH and AFT
3. Parallel but not straight lines ⇒ PH but not Weibull, not AFT (can use Cox model)
4. Not parallel and not straight ⇒ Not Weibull, PH violated
5. Not parallel but straight lines ⇒ Weibull holds, but PH and AFT violated, different p

On the left is a summary of five possible results from an examination of the log negative log Kaplan-Meier survival estimates plotted against the log of time for two or more levels of covariates. The key points are that **straight lines support the Weibull assumption** and **parallel curves support the PH assumption**. If the plots are parallel but not straight then the PH assumption holds but not the Weibull. Assessing whether the curves are parallel is a familiar approach for evaluating the PH assumption in a Cox model (see Chapter 4 and Computer Appendix). An interesting scenario occurs if the lines are straight but not parallel. In this situation the Weibull assumption is supported but the PH and AFT assumptions are violated. If the lines are not parallel, then p is not constant across levels of covariates. In Section IX of this chapter, we present a method for modeling the shape parameter p as a function of predictor variables, but typically p is assumed fixed.

Previous plot suggests Weibull and PH assumption reasonable for TRT

An examination of the plot on the previous page suggests that the Weibull and PH assumptions are reasonable for treatment (TRT). First the PH form of the model is presented and then the AFT form.

Weibull PH model:

$h(t) = \lambda p t^{p-1}$
where $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$.

The Weibull hazard function is $h(t) = \lambda p t^{p-1}$. A Weibull PH model is defined by reparameterizing lambda λ as $\exp(\beta_0 + \beta_1 \text{ TRT})$. The hazard ratio is obtained by substituting TRT = 1 and TRT = 0 into the hazard functions (see left). After canceling we obtain the familiar result $\exp(\beta_1)$. Note that this result depends on p having the same value for TRT = 1 and TRT = 0, otherwise time (t) would not cancel in the expression for the HR (i.e., PH assumption not satisfied).

Hazard ratio (TRT = 1 vs. TRT = 0)

$$\text{HR} = \frac{\exp(\beta_0 + \beta_1) p t^{p-1}}{\exp(\beta_0) p t^{p-1}} = \exp(\beta_1)$$

Remission Data

Weibull regression log relative-hazard form

_t	Coef.	Std. Err.	z	p > |z|
trt	−1.731	.413	−4.19	0.000
_cons	−3.071	.558	−5.50	0.000
/ln_p	.312	.147	2.12	0.034
p	1.366	.201		
1/p	.732	.109		

On the left is stata output from running the PH form of the Weibull model. There are parameter estimates for the coefficient of TRT, the intercept (called _cons), and for three forms of the shape parameter: p, $1/p$, and $\log(p)$. The estimate for p is 1.366 suggesting an increase in the hazard as survival time increases (because $\hat{p} > 1$). A statistical test for H_0: $\log(p) = 0$ yields a p-value of 0.034. At a significance level of 0.05 we would reject the null and decide p is not equal to 1, suggesting that the exponential model is not appropriate.

Weibull PH

$\widehat{HR}(\text{TRT} = 1 \text{ vs. } 0) = \exp(-1.731) = 0.18$

$95\%\ \text{CI} = \exp[-1.731 \pm 1.96(0.413)] = (0.08, 0.40)$

An estimated hazard ratio of 0.18 is obtained by exponentiating the estimated coefficient (−1.731) of the TRT variable. The 95% confidence interval for this HR is calculated to be (0.08, 0.40) indicating a significant preventive effect of treatment. These results are similar to those obtained from the exponential model in which the estimated hazard ratio was 0.22.

Weibull: $\widehat{HR} = 0.18$
Exponential: $\widehat{HR} = 0.22$
Suggests preventive effect of TRT

Comparing Cox and Weibull PH models

Cox: estimate β_1

$h(t) = h_0(t)\exp(\beta_1 \text{TRT})$

baseline hazard unspecified

It can be instructive to compare the Cox and Weibull PH models. The Cox PH model with treatment as the only predictor is stated as $h_0(t)\exp(\beta_1 \text{TRT})$. There is one parameter to estimate (β_1) and the distribution of the baseline hazard ($h_0(t)$) remains unspecified.

Weibull: estimate β_0, β_1, p

$h(t) = \lambda p t^{p-1}$ where

$\lambda = \exp(\beta_0 + \beta_1 \text{TRT})$.

$h(t) = [\exp(\beta_0) p t^{p-1}]\exp(\beta_1 \text{TRT})$.

baseline hazard specified parametrically

With some manipulation, the Weibull PH model can also be expressed as a product of a baseline hazard and $\exp(\beta_1$ TRT) (see left). There are three parameters to estimate β_0, β_1, and p that fully specify the hazard.

$$S(t) = \exp(-\lambda t^p)$$

solve for t

$$t = [-\ln S(t)]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

$$\text{let } \frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 RX)$$

$$t = [-\ln S(t)]^{1/p} \times \exp(\alpha_0 + \alpha_1 TRT)$$

Scaling of t

An AFT model can also be formulated with the Weibull distribution. We develop the AFT parameterization similarly to that done with the exponential model, by solving for t in terms of a fixed S(t). The Weibull survival function is $S(t) = \exp(-\lambda t^p)$. Taking the natural log, multiplying by negative 1, raising to the power $1/p$, and then multiplying by the reciprocal of $\lambda^{1/p}$, yields the expression for t shown on the left. By reparameterizing $1/\lambda^{1/p} = \exp(\alpha_0 + \alpha_1 TRT)$, it can be shown that α_1, the coefficient of the predictor variable TRT, is used to scale the time to any fixed value of S(t) (see left).

Let $S(t) = q$

$$t = [-\ln(q)]^{1/p} \times \exp(\alpha_0 + \alpha_1 TRT)$$

The expression for t is restated on the left in terms of any fixed probability $S(t) = q$. For example, to find an expression for the median survival time t_m, substitute $q = 0.5$ (see left).

Median survival time ($q = 0.5$)

$$t_m = [-\ln(0.5)]^{1/p} \times \exp(\alpha_0 + \alpha_1 TRT)$$

Acceleration factor, γ (TRT = 1 vs. TRT = 0)

$$\gamma = \frac{[-\ln(q)]^{1/p} \exp(\alpha_0 + \alpha_1)}{[-\ln(q)]^{1/p} \exp(\alpha_0)} = \exp(\alpha_1)$$

The acceleration factor γ is obtained as the ratio of the times to $S(t) = q$ for TRT = 1 and for TRT = 0. After canceling, γ reduces to $\exp(\alpha_1)$, where α_1 is the coefficient of the TRT variable. As with the PH form of the model, this result depends on p not varying by treatment status; otherwise γ would depend on q.

Remission Data

Weibull regression accelerated failure-time form

_t	Coef.	Std. Err.	z	P>|z|
trt	1.267	.311	4.08	0.000
_cons	2.248	.166	13.55	0.000
/ln_p	.312	.147	2.12	0.034
p	1.366	.201		
1/p	.732	.109		

Output from running a Weibull AFT model is shown on the left. The estimates for each form of the shape parameter (p, $1/p$, and $\ln(p)$) are the same as obtained from the previously shown PH form of the model.

The estimated acceleration factor of 3.55 is obtained by exponentiating the estimated coefficient (1.267) of the TRT variable. The 95% confidence interval for γ is calculated to be (1.93, 6.53). There results are shown at the top of the following page.

Weibull AFT:

$\hat{\gamma}(\text{TRT} = 1 \text{ vs. } 0) = \exp(1.267)$
$= 3.55$

95% CI $= \exp[1.267 \pm 1.96(0.311)]$
$= (1.93, 6.53)$

These results suggest that the median (or any other quantile of) survival time is increased by a factor of 3.55 for those receiving the treatment compared to the placebo. Recall that the acceleration factor was estimated at 4.60 using the exponential model. However, the exponential model uses a much stronger assumption: that the hazards are constant.

Weibull: $\hat{\gamma} = 3.55$
Exponential: $\hat{\gamma} = 4.60$ (assumes $h(t) = \lambda$)

Relating Weibull AFT and PH coefficients

Corresponding coefficients obtained from the PH and AFT forms of the Weibull models are related as follows: $\beta_j = -\alpha_j\, p$ for the jth covariate. This can most easily be seen by formulating the parameterization equivalently in terms of $\ln(\lambda)$ for both the PH and AFT forms of the model as shown on the left.

AFT: $\lambda^{1/p} = \exp[-(\alpha_0 + \alpha_1 \text{TRT})]$
$(1/p)\ln \lambda = -(\alpha_0 + \alpha_1 \text{TRT})$
$\ln \lambda = -p\,(\alpha_0 + \alpha_1 \text{TRT})$

PH: $\lambda = \exp(\beta_0 + \beta_1 \text{TRT})$
$\ln \lambda = \beta_0 + \beta_1 \text{TRT}$

Relationship of coefficients:

$\beta_j = -\alpha_j\, p$ so that
$\beta = -\alpha$ for exponential ($p = 1$)

This relationship is illustrated utilizing the coefficient estimates we obtained for TRT: $-1.731 = (-1.267)(1.366)$. Note for the exponential model in which $p = 1$, the PH and AFT coefficients are related, $\beta = -\alpha$.

Relating estimates for TRT (PH vs. AFT)

$-1.731 = (-1.267)(1.366)$

Next: log-logistic model

- Hazard may be nonmonotonic

Weibull model

- Hazard does not change direction

In the next example the log-logistic model is presented. In contrast to the Weibull, the hazard function for the log-logistic distribution allows for some nonmonotonic behavior in the hazard function.

VII. Log-Logistic Example

The log-logistic distribution accommodates an AFT model but not a PH model. Its hazard function is shown on the left. The shape parameter is $p(>0)$.

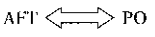
Log-logistic hazard: $h(t) = \dfrac{\lambda p t^{p-1}}{1 + \lambda t^p}$
(where $p > 0$ and $\lambda > 0$)

Shape of hazard function:

$p \leq 1$ hazard decreases over time
$p > 1$ hazard first increases and then decreases over time (unimodal)

If $p \leq 1$ then the hazard decreases over time. If $p > 1$, however, the hazard increases to a maximum point and then decreases over time. In this case ($p > 1$), the hazard function is said to be **unimodal**.

Log-logistic modeling assumptions:



PO: Survival Odds ratio constant over time

Unlike the Weibull model, a log-logistic AFT model is not a PH model. However, the log-logistic AFT model is a proportional odds (PO) model. **A proportional odds survival model is a model in which the survival odds ratio is assumed to remain constant over time**. This is analogous to a proportional hazard model where the hazard ratio is assumed constant over time.

Survival odds

$$\frac{S(t)}{(1-S(t))} = \frac{P(T > t)}{P(T \leq t)}$$

The survival odds is the odds of surviving beyond time t (i.e., S(t)/(1 − S(t)). This is the probability of not getting the event by time t divided by the probability of getting the event by time t.

Failure odds by time t

$$\frac{(1-S(t))}{S(t)} = \frac{P(T \leq t)}{P(T > t)}$$

The failure odds is the odds of getting the event by time t (i.e., (1 − S(t))/S(t)), which is the reciprocal of the survival odds (see left).

Log-logistic survival and failure functions

$$\mathrm{S(t)} = \frac{1}{1+\lambda t^p} \quad 1 - \mathrm{S(t)} = \frac{\lambda t^p}{1+\lambda t^p}$$

The log-logistic survival function (S(t)) and failure function (1 − S(t)) are shown on the left.

Failure odds

$$\frac{1-S(t)}{S(t)} = \frac{\left(\frac{\lambda t^p}{1+\lambda t^p}\right)}{\left(\frac{1}{1+\lambda t^p}\right)} = \lambda t^p$$

The failure odds simplifies in a log-logistic model to λt^p (see left).

Log-logistic PO model:

- Reparameterize λ in terms of Xs and βs

  $$\mathrm{SOR} = \frac{\mathrm{S_1(t)}/(1-\mathrm{S_1(t)})}{\mathrm{S_2(t)}/(1-\mathrm{S_2(t)})}$$

- SOR satisfies PO if SOR constant over time
- SOR constant

  ⇕

FOR = 1/SOR constant

A log-logistic proportional odds model can be formulated by reparameterizing λ in terms of predictor variables and regression parameters. We come back to this point later in this section.

A survival odds ratio (SOR) is defined as the ratio of survival odds for two groups of subjects, as shown on the left.

An SOR satisfies a PO assumption provided the SOR does not depend on time.

Also, if the SOR does not depend on time, then the failure odds ratio (FOR = 1/SOR) also does not depend on time.

Log Odds Is Linear with ln(t)

$$\log(\text{failure odds}) = \ln(\lambda t^p) = \ln(\lambda) + p[\ln(t)]$$

Intercept $= \ln(\lambda)$ slope $= p$

The log of the failure odds is $\ln(\lambda t^p)$, which can be rewritten as $\ln(\lambda) + p[\ln(t)]$. In other words, **the log odds of failure is a linear function of the log of time** with slope p and intercept $\ln(\lambda)$. This is a useful result enabling a graphical evaluation for the appropriateness of the log-logistic distribution.

Evaluate log-logistic assumption graphically

- Plot $\ln\left[\frac{(1-\hat{S}(t))}{\hat{S}(t)}\right]$ against ln(t)
- If log-logistic, then plot is linear with slope $= p$

The log-logistic assumption can be graphically evaluated by **plotting** $\ln(1 - \hat{S}(t))/(\hat{S}(t))$ **against ln(t) where** $\hat{S}(t)$ **are the Kaplan–Meier survival estimates**. If survival time follows a log-logistic distribution, then the resulting plots should be a straight line of slope p.

Alternatively

- Plot $\ln\left(\frac{((\hat{S}(t))}{(1-\hat{S}(t))}\right)$ against ln(t)
- If log-logistic, then plot is linear with slope $= -p$

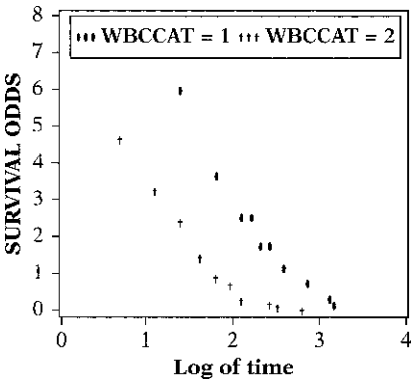
We could alternatively plot the log of the survival odds, $\ln(\hat{S}(t))/(1 - \hat{S}(t))$, against ln(t). If the log-logistic assumption is correct the resulting plots should be a straight line of slope $-p$.

Remission Data

WBCCAT: white blood cell count variable medium = 1 vs. high = 2

$\ln\left[\frac{\hat{S}(t)}{(1-\hat{S}(t))}\right]$ plotted against ln(t).

We next consider a different variable from the remission data: a dichotomous variable for white blood cell count (WBCCAT) coded medium = 1 and high = 2.



On the left is the plot of the log odds of survival (obtained from the Kaplan–Meier survival estimates) against the log of time comparing **medium** (WBCCAT = 1) and **high** (WBCCAT = 2) blood cell counts. The points for WBCCAT = 1 lie above the points for WBCCAT = 2 indicating that the survival odds are higher for those with a medium white blood cell count compared to high. The lines look reasonably straight and parallel, at least until the estimated odds of survival approaches zero.

If we accept the proposition that the lines look straight, then the log-logistic assumption is reasonable. Because the lines look parallel, the proportional odds (PO) assumption is also reasonable. If the PO assumption holds in a log-logistic model then the AFT assumption also holds.

Straight lines $\Longrightarrow$ Log-logistic

Parallel plots $\Longrightarrow$ PO

Log-logistic and PO $\Longrightarrow$ AFT

Log-logistic and Weibull graphical approach analogous

- Check PH for Weibull
- Check PO for log-logistic

AFT log-logistic model

$$S(t) = \frac{1}{1+\lambda t^p} = \frac{1}{1+\left(\lambda^{1/p} t\right)p}$$

solve for t to obtain

$$t = \left[\frac{1}{S(t)} - 1\right]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

$$\text{let } \frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{WBCCAT})$$

$$t = \left[\frac{1}{S(t)} - 1\right]^{1/p} \times \exp(\alpha_0 + \alpha_1 \mathbf{WBCCAT})$$

Scaling of t

Let S(t) − q

$$t = [q^{-1} - 1]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{ WBCCAT})$$

Median survival time (q = 0.5):

$$t_m = [2 - 1]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{ WBCCAT})$$

The key points from above are:

a. **straight lines support the log-logistic assumption**,
b. **parallel curves support the PO assumption**, and
c. **If the log-logistic and PO assumptions hold, then the AFT assumption also holds**.

The graphical evaluation for the log-logistic assumption is analogous to the graphical analysis of the Weibull assumption presented in the last section, except here the PO assumption rather than the PH assumption is evaluated by checking for parallel lines.

Next we consider an AFT log-logistic model with white blood cell count as the only predictor comparing WBCCAT = 2 (high count) and WBCCAT = 1 (medium count).

We develop the AFT parameterization by solving for t in terms of a fixed S(t). Starting with the expression for S(t), taking reciprocals, subtracting 1, raising to the power $1/p$, and then multiplying by the reciprocal of $\lambda^{1/p}$, yields the expression for t shown on the left. By reparameterizing $1/\lambda^{1/p} = \exp(\alpha_0 + \alpha_1 \text{ WBCCAT})$, we allow α_1, the coefficient of the predictor variable WBCCAT, to be used for the multiplicative scaling of time to any fixed value of S(t) (see left).

The expression for t is restated on the left in terms of any fixed probability S(t) = q. For example, to find an expression for the median survival time t_m, substitute q = 0.5 (see left).

Acceleration factor,
γ (WBCCAT = 2 vs. WBCCAT = 1)

$$= \frac{[q^{-1} - 1]^{1/p} \exp(\alpha_0 + 2\alpha_1)}{[q^{-1} - 1]^{1/p} \exp(\alpha_0 + 1\alpha_1)}$$
$$= \exp(\alpha_1)$$

The acceleration factor γ is found by taking the ratio of the times to S(t) = q for WBCCAT = 2 and for WBCCAT = 1. After canceling, γ reduces to $\exp(\alpha_1)$, where α_1 is the coefficient of the WBCCAT variable in the AFT model.

Log-logistic regression accelerated failure-time form

_t	Coef.	Std. Err.	z	P>|z|
wbccat	−.871	.296	−2.94	0.003
_cons	3.495	.498	7.09	0.000
ln_gam	−.779	.164	−4.73	0.000
gamma	.459	0.756		

$$p = 1/(0.459) = 2.18$$

The output from running the AFT log-logistic model is shown on the left. The coefficient estimate for WBCCAT is −0.871, which is statistically significant with a p-value of 0.003 (far right column of output).

Stata provides estimates of the reciprocal of p (gamma = $1/p$) rather than for p. The estimate for gamma is 0.459. Therefore, the estimate for p is 1/(0.459) = 2.18.

WBCCAT = 2 vs. WBCCAT = 1 (log-logistic):

$$\hat{\gamma} = \exp(-0.871) = 0.42$$
$$95\%\ \text{CI for}\ \gamma = \exp[-0.871 \pm 1.96(0.296)]$$
$$= (0.23,\ 0.75)$$

An estimate of the acceleration factor $\hat{\gamma}$ comparing WBCCAT = 2 to WBCCAT = 1 is found by exponentiating the estimate −0.871 of α_1 to obtain 0.42. The 95% confidence interval for γ is calculated to be (0.23, 0.75).

Comparing estimated survival

$$\hat{S}_1(t) = \hat{S}_2(0.42t)$$

Survival function for WBCCAT = 1

Survival function for WBCCAT = 2

These results suggest that the time for going out of remission is "accelerated" for patients with a **high** white blood cell count compared to those with a **medium** count by an estimated factor of 0.42. In terms of the survival functions estimated from this model, $\hat{S}_1(t) = \hat{S}_2(0.42t)$ where $\hat{S}_1(t)$ and $\hat{S}_2(t)$ are the respective survival functions for patients with medium and high blood cell counts.

Failure odds

$$\frac{1 - S(t)}{S(t)} = \frac{\left(\frac{\lambda t^p}{1+\lambda t^p}\right)}{\left(\frac{1}{1+\lambda t^p}\right)} = \lambda t^p$$

where $\lambda = \exp(\beta_0 + \beta_1\ \text{WBCCAT})$

The proportional odds form of the log-logistic model can also be formulated by reparameterizing λ. Recall that the log-logistic failure odds is λt^p.

By setting $\lambda = \exp(\beta_0 + \beta_1\ \text{WBCCAT})$, a (failure) odds ratio comparing WBCCAT = 2 to WBCCAT = 1 can be calculated (see left). After canceling, the odds ratio reduces to $\exp(\beta_1)$.

OR (WBCCAT = 2 vs. WBCCAT = 1)

$$= \frac{t^p \exp(\beta_0 + 2\beta_1)}{t^p \exp(\beta_0 + 1\beta_1)} = \exp(\beta_1)$$

Comparing AFT and PO (log-logistic)
Relationship of coefficients:

$\beta_j = -\alpha_j p$

Since $\hat{\alpha} = -0.871$ and $\hat{p} = 2.18$
Then,

$\hat{\beta}_1 = -(-0.871)(2.18) = 1.90$

and

$\widehat{OR} = \exp(1.90) = 6.69$

The corresponding coefficients for log-logistic PO and AFT models are related by $\beta_j = -\alpha_j p$ for the jth covariate. This result is obtained using a similar argument to that presented for the Weibull example in the previous section.

The estimate for α_1 in the AFT model is -0.871 and the estimate for p is 2.18. Therefore, an estimate for β_1 can be found by multiplying $-(-0.871)$ times 2.18 yielding 1.90. An estimate of the odds ratio is found by exponentiating this estimate, $\exp(1.90) = 6.69$. (Unfortunately, neither Stata nor SAS estimates the proportional odds form of the model.)

VIII. A More General Form of the AFT Model

Exponential: $S(t) = \exp(-\lambda t)$

- AFT Form: $\frac{1}{\lambda} = \exp(\alpha_0 + \alpha_1 \text{ TRT})$
- PH Form: $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$

Weibull: $S(t) = \exp(-\lambda t^p)$

- AFT Form: $\frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{ TRT})$
- PH Form: $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$

Log-logistic: $S(t) = \frac{1}{1 + \lambda t^p}$

- AFT Form: $\frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{ WBCCAT})$
- PO Form: $\lambda = \exp(\beta_0 + \beta_1 \text{ WBCCAT})$

On the left is a summary of the models discussed in the previous sections. These models were formulated by reparameterizing the survival (and hazard) functions in terms of regression parameters and predictor variables.

An advantage for stating the models in this form is that the interpretation and relationships between parameters are specific to each distribution.

However, there are more general ways these models could be stated. The Cox PH model is a more general way of stating the proportional hazards model. In this section we discuss a more general formulation of the AFT model.

General Form of AFT Model (One Predictor)

$$\ln(T) = \alpha_0 + \alpha_1 TRT + \epsilon$$

ϵ: random error

Consider an AFT model with one predictor (TRT) in which T represents a random variable for survival time. The model can be expressed on the log scale as shown on the left, where ϵ is random error following some distribution.

With additional parameter

$$\ln(T) = \alpha_0 + \alpha_1 TRT + \sigma\epsilon$$

σ scales the error

Some distributions have an additional parameter (σ) scaling ϵ. The model including this additional parameter can be expressed as shown on the left, where the random error ϵ is multiplied by a scale parameter σ.

If $\epsilon \sim N(0, 1)$, then

$\ln(T) \sim N(\mu = \alpha_0 + \alpha_1 TRT, sd = \sigma)$

Similar to linear regression (except for inclusion of censorships)

If ϵ follows a standard normal distribution and $\ln(T) = \alpha_0 + \alpha_1 TRT + \sigma\epsilon$, then ln(T) would follow a normal distribution with mean $\mu = \alpha_0 + \alpha_1 TRT$ and standard deviation σ. For this situation, the model would look like a standard linear regression. The **key difference** between fitting this survival model and a standard linear regression **is the inclusion of censored observations** in the data.

In general,

$\mu_{\ln(T)} \neq (\alpha_0 + \alpha_1 TRT)$, $sd \neq \sigma$

Interpretation of parameters depends on distribution

In general, for other then the normal distribution, the mean of ln(T) is not necessarily $\alpha_0 + \alpha_1 TRT$ and its standard deviation is not σ. In other words, it should not be assumed that the mean of ϵ is 0 and the standard deviation is 1. The interpretation of the parameters depends on the underlying distribution.

Let $\sigma = \frac{1}{p}$, then

$$\ln(T) = \alpha_0 + \alpha_1 TRT + \frac{1}{p}\epsilon$$

Sometimes the model is parameterized using $\sigma = 1/p$. The model can then be restated by replacing $\sigma\epsilon$ with $(1/p)\,\epsilon$.

Additive model in terms of ln(T)

but

multiplicative model in terms of T

The AFT model is **additive on the log scale but a multiplicative model with respect to T**.

$$T = \exp\left(\alpha_0 + \alpha_1 TRT + \frac{1}{p}\epsilon\right)$$

$$= \exp[(\alpha_0 + \alpha_1 TRT)] \times \exp\left(\frac{1}{p}\epsilon\right)$$

In particular, the model can be expressed in terms of T by exponentiating ln(T), as shown on the left.

Collapse α_0 into baseline term

$$T_0 = \exp(\alpha_0)\exp\left(\frac{1}{p}\epsilon\right)$$

The model may also be expressed by collapsing the intercept into a baseline random term T_0 (see left). In this setting, T_0 is a random variable representing the survival time of the placebo group (TRT = 0).

so that $T = T_0 \exp(\alpha_1 \text{ TRT})$ where T_0 is a random variable for TRT = 0

Thus, the AFT model has a form analogous to the a Cox PH model, although the baseline term T_0 is a random variable rather than a constant.

AFT model may be expressed in terms of T or ln(T)

Comparing Distributions: T and ln(T)

T	ln(T)
Exponential	Extreme minimum value
Weibull	Extreme minimum value
Log-logistic	Logistic
Lognormal	Normal

In summary, an AFT model may be expressed by reparameterizing a specific distribution, or may be more generally expressed either in terms of a random variable T (for survival time), or ln (T). If T follows a Weibull distribution then ln (T) follows a distribution called the extreme minimum value distribution (see table on left). Similarly, if T follows a log-logistic or lognormal distribution then ln(T) follows a logistic or normal distribution, respectively. The logistic and normal distributions are similarly shaped, and are both symmetric about their mean.

IX. Other Parametric Models

In the previous sections we presented examples of the exponential, Weibull, and log-logistic models. In this section we briefly discuss some other parametric survival models.

Generalized Gamma Model

- Supported by SAS and Stata
- S(t), h(t) expressed in terms of integrals
- Contains three parameters
- Weibull, lognormal are special cases

The **generalized gamma model** is a parametric survival model that is supported by both SAS and Stata software. The hazard and survival function for this model is complicated and can only be expressed in terms of integrals. The generalized gamma distribution has three parameters allowing for great flexibility in its shape. The Weibull and lognormal distributions are special cases of the generalized gamma distribution (see Practice Exercises 12 to 14).

Lognormal Model

Similar to log-logistic
Difference:
Log-logistic: AFT and PO
Lognormal: AFT but not PO

The **lognormal model** also has a relatively complicated hazard and survival function that can only be expressed in terms of integrals. The shape of the lognormal distribution is very similar to the log-logistic distribution and yields similar model results. A difference is that although the lognormal model accommodates an accelerated failure time model, it is not a proportional odds model.

Gompertz Model

- PH model but not AFT
- One predictor (TRT) in model:

$h(t) = [\exp(\gamma t)] \times \exp(\beta_0 + \beta_1 \text{ TRT})$

↗ parametrically specified

$h_0(t) = \exp(\gamma t)$

$\gamma > 0$ hazard exponentially increases with t
$\gamma < 0$ hazard exponentially decreases with t
$\gamma = 0$ constant hazard (exponential model)

Parametric models need not be AFT models. The **Gompertz** model is a parametric proportional hazards model but not an AFT model. The model can be expressed in a form similar to that of a Cox PH model except that the baseline hazard is specified as the hazard of a Gompertz distribution containing a shape parameter γ (see left).

If $\gamma > 0$ then the hazard exponentially increases over time. If $\gamma < 0$ then the hazard exponentially decreases over time. If $\gamma = 0$ then the hazard is constant and reduces to the exponential model.

AFT model: multiplicative

$T = \exp(\alpha_0 + \alpha_1 \text{ TRT} + \epsilon)$
$= \exp(\alpha_0) \times \exp(\alpha_1 \text{ TRT}) \times \exp(\epsilon)$

but
additive on log scale:
$\ln(T) = \alpha_0 + \alpha_1 \text{ TRT} + \epsilon$

The AFT model is a **multiplicative model** (i.e., a multiplicative scaling of failure time). It becomes an additive model on the log scale (see left side).

Additive failure time model
$T = \alpha_0 + \alpha_1 \text{TRT} + \epsilon$

↗ T rather than log(T) is linear with TRT

An alternative parametric model is to define an **additive failure time model** in terms of T, rather than ln(T). Consider the model: $T = \alpha_0 + \alpha_1$ TRT $+ \epsilon$. Now T, rather than ln(T), is expressed as a linear function of the regression parameters. SAS supports such an additive failure time model (see Computer Appendix).

Modeling the Shape Parameter (e.g., Weibull and log-logistic)

Typical Weibull model

$h(t) = \lambda p t^{p-1}$

where $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$
p unaffected by predictors

Many parametric models contain an extra shape (or ancillary) parameter beyond the regression parameters. For example, the Weibull and log-logistic models contain a shape parameter p. Typically, this parameter is considered fixed, unaffected by changes in the values of predictor variables.

Alternative Weibull model
models the ancillary parameter p

$h(t) = \lambda p t^{p-1}$

where $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$
$p = \exp(\delta_0 + \delta_1 \text{ TRT})$

Not a PH or AFT model if $\delta_1 \neq 0$ but still a Weibull model

An alternative approach is to model the shape parameter in terms of predictor variables and regression coefficients. In the Weibull model shown on the left, both λ and p are modeled as functions of treatment status (TRT). If δ_1 is not equal to zero, then the value of p differs by TRT. For that situation, the PH (and thus the AFT) assumption is violated because t^{p-1} will not cancel in the hazard ratio for TRT (see Practice Exercises 15 to 17).

Choosing appropriate model

- Evaluate graphically
  - Exponential
  - Weibull
  - Log-logistic
- Akaike's information criterion
  - Compares model fit
  - Uses −2 log likelihood

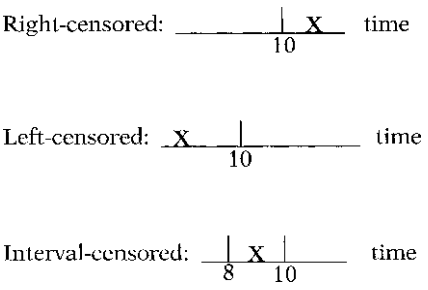
Choosing the most appropriate parametric model can be difficult. We have provided graphical approaches for evaluating the appropriateness of the exponential, Weibull, and log-logistic models. **Akaike's information criterion (AIC)** provides an approach for comparing the fit of models with different underlying distributions, making use of the −2 log likelihood statistic (described in Practice Exercises 11 and 14).

X. The Parametric Likelihood

- Function of observed data and unknown parameters
- Based on outcome distribution f(t)
- Censoring complicates survival data
  - Right-censored
  - Left-censored
  - Interval-censored

The likelihood for any parametric model is a function of the observed data and the model's unknown parameters. The form of the likelihood is based on the probability density function f(t) of the outcome variable. A complication of survival data is the possible inclusion of censored observations (i.e., observations in which the exact time of the outcome is unobserved). We consider three types of censored observations: right-censored, left-censored, and interval-censored.

Examples of Censored Subjects



Right-censored. Suppose a subject is lost to follow-up after 10 years of observation. The time of event is not observed because it happened after the 10th year. This subject is right-censored at 10 years because the event happened to the *right* of 10 on the time line (i.e., $t > 10$).

Left-censored. Suppose a subject had an event before the 10th year but the exact time of the event is unknown. This subject is left-censored at 10 years (i.e., $t < 10$).

Interval-censored. Suppose a subject had an event between the 8th and 10th year (exact time unknown). This subject is interval-censored (i.e., $8 < t < 10$).

Formulating the Likelihood

Barry, Gary, Larry, . . ., Outcome Distribution $f(t)$

Subject	Event Time	Likelihood Contribution
Barry	t = 2	$f(2)$
Gary	$t > 8$ (right-censored)	$\int_8^\infty f(t)dt$
Harry	t = 6	$f(6)$
Carrie	$t < 2$ (left-censored)	$\int_0^2 f(t)dt$
Larry	$4 < t < 8$ (interval-censored)	$\int_4^8 f(t)dt$

The table on the left illustrates how the likelihood is formulated for data on five subjects. We assume a probability density function f(t) for the outcome. Barry gets the event at time t = 2. His contribution to the likelihood is f(2). Gary is right-censored at t = 8. The probability that Gary gets the event after t = 8 is found by integrating f(t) from 8 to infinity. This is Gary's contribution to the likelihood. Harry gets the event at time t = 6. His contribution to the likelihood is f(6). Carrie is left-censored at t = 2. Her contribution to the likelihood is obtained by integrating f(t) from zero to 2. Finally, Larry is interval-censored between t = 4 and t = 8. His contribution to the likelihood is found by integrating f(t) from 4 to 8.

Likelihood (L)

Product of individual contributions

$$L = f(2)x\int_8^\infty f(t)dt \times f(6) \times \int_0^2 f(t)dt \times \int_0^8 f(t)dt$$

(Barry × Gary × Harry × Carrie × Larry)

The full likelihood (L) is found by taking the product of each subject's independent contribution to the likelihood. The likelihood for this example is shown on the left.

Assumptions for formulating L

- No competing risks
  - Competing event does not prohibit event of interest
  - Death of all causes is classic example of no competing risk

The formulation of this likelihood uses the assumption of **no competing risks**. In other words, we assume that no competing event will prohibit any subject from eventually getting the event of interest (see Chapter 9). Death from all causes is the classic example of an outcome that in reality has no competing risk. For other outcomes, the no competing risk assumption is more of a theoretical construct.

- Subjects independent
  - Allows L to be formulated as product of subjects' contributions

Another assumption is that individual contributions to the likelihood are independent. This assumption allows the full likelihood to be formulated as the product of each individual's contribution.

- Follow-up time continuous
  - No gaps in follow-up

A third assumption is that each subjects follow-up time is continuous without gaps (i.e., once subjects are out of the study, they do not return). If gaps are allowed, the likelihood can be modified to accommodate such a scenario.

Revisit example with Barry, Gary, Larry, ...
f(t) is Weibull
SMOKE is only predictor

1 = Smoker
0 = Nonsmoker

In the last example, we did not specify the probability density f(t), nor did we specify any covariates. We revisit this example, assuming f(t) is Weibull with one predictor SMOKE in the model (coded 1 for smokers and 0 for nonsmokers).

Weibull: $h(t) = \lambda pt^{p-1}$,
$S(t) = \exp(-\lambda t^p)$

$f(t) = h(t)S(t)$
$f(t) = \lambda pt\ p-1\ exp(-\lambda t\ p)$

where $\lambda = \exp(\beta 0 + \beta 1\ \text{SMOKE})$
(PH form of the model)

The Weibull hazard and survival functions are shown on the left. The probability density function f(t) is the product of the hazard and survival functions. The parameterization will use the proportional hazards (PH) form of the Weibull model: $\lambda = \beta_0 + \beta_1$ SMOKE.

Data Layout for Right-, Left-, and Interval-Censoring Using SAS

ID	LOWER	UPPER	SMOKE
Barry	2	2	1
Gary	8	-	0
Harry	6	6	0
Carrie	-	2	0
Larry	4	8	1

Right-censored: UPPER missing
Left-censored: LOWER missing
Interval-censored: LOWER < UPPER
Not censored: LOWER = UPPER

On the left is the data layout for running parametric models containing right-, left-, and interval censored data in a form suitable for using the SAS procedure PROC LIFETEST (version 8.2). There are two time variables LOWERand UPPER. Barry got the event at t = 2, so both LOWER and UPPER get the value 2. Gary was right-censored at 8 ($t < 8$) so LOWER gets the value 8 and UPPER is set to missing. Carrie is left-censored at 2 ($t > 2$) so LOWER is set to missing and UPPER gets the value 2. Larry was interval-censored with LOWER = 4 and UPPER = 8. Barry and Larry are smokers whereas Gary, Harry, and Carrie are nonsmokers.

Weibull Likelihood (L)

Product of individual contributions

$$L = f(2) \times \int_{8}^{\infty} f(t)dt \times f(6) \times \int_{0}^{2} f(t)dt \times \int_{4}^{8} f(t)dt$$

$$\begin{aligned} L = {} & \exp(\beta_0 + \beta_1)p(2)^{p-1} \exp(-\exp(\beta_0 + \beta_1)2^p) \\ & \times \int_{8}^{\infty} \exp(\beta_0)p(t)^{p-1} \exp(-\exp(\beta_0)t^p)dt \\ & \times \exp(\beta_0)p(6)^{p-1} \exp(-\exp(\beta_0)6^p) \\ & \times \int_{0}^{2} \exp(\beta_0)p(t)^{p-1} \exp(-\exp(\beta_0)t^p)dt \\ & \times \int_{4}^{8} \exp(\beta_0 + \beta_1)p(t)^{p-1} \\ & \quad \times \exp(-\exp(\beta_0 + \beta_1)t^p)dt \end{aligned}$$

The full likelihood using the Weibull distribution can now be formulated as a product of each individual's contribution (shown on the left). We have used a small dataset (five subjects) for ease of illustration but the process can be generalized for any number of subjects.

Obtaining maximum likelihood estimates

Solve system of equations:

$$\frac{\partial Ln(L)}{\partial \beta_j} = 0 \quad \mathrm{j} = 1, 2, \ldots, \mathrm{N}$$

where N = # of parameters

Once the likelihood is formulated, the question becomes: *which values of the regression parameters would maximize L*? The process of maximizing the likelihood is typically carried out by setting the partial derivative of the natural log of L to zero and then solving the system of equations (called the score equations). The parameter estimates (e.g., $\hat{p}$, $\hat{\beta}_0$, $\hat{\beta}_1$) that maximize L are the maximum likelihood estimates.

XI. Interval-Censored Data

Parametric likelihood

- Handles right-, left-, or interval-censored data

Cox likelihood

- Designed to handle right-censored data.

One advantage of a parametric model compared to a Cox model is that the parametric likelihood easily accommodates right-, left-, or interval-censored data. The Cox likelihood, by contrast, easily handles right-censored data but does not directly accommodate left- or interval-censored data.

Interval-censored study design

- Check for nonsymptomatic outcome once a year
- If outcome newly detected, exact time occurred during previous year
- Left-censoring special case of interval-censoring
  - Zero the lower boundary of the interval

Sometimes the design of a study is such that all the data are interval-censored. For example, consider a study in which healthcare workers examine subjects once a year, checking for a nonsymptomatic outcome. If an event was first detected in the beginning of the third year, then the exact time of the outcome occurred sometime between the second and third years. In this framework left-censoring can be considered a special case of interval-censoring with zero as the lower boundary of the interval.

Parametric model can be fitted

- f(t) specified
- Contribution to likelihood for each subject
  - Integrate f(t) over event interval

A parametric model can easily be fitted using the methods described in the previous section. Once a distribution for the outcome, f(t), is specified, each subject's contribution to the likelihood is obtained by integrating f(t) over the interval in which he or she had the event.

Binary regression

- Alternative approach for interval-censored data
- Outcome coded
  - 0 if subject survives interval
  - 1 if subject gets event during interval
- Useful approach if
  - Ample number of events in each interval
  - Prefer not to specify f(t)

A binary regression (e.g., logistic regression) is an alternative approach that may be considered if all the data are interval-censored. With this method, the outcome variable can be coded zero if the subject survives the interval and coded one if the subject gets the event during the interval. This approach is particularly useful if there are an ample number of events in each interval and the analyst prefers not to specify a distribution f(t) for continuous survival time.

Information on Three Subjects

Subject 1: Gets event in first interval
Subject 2: Survives first interval Survives second interval Gets event in third interval
Subject 3: Survives first interval Gets event in second interval

For illustration, consider a small dataset containing three subjects. Subject 1 gets the event in the first interval of follow-up, subject 2 gets the event in the third interval, and subject 3 gets the event in the second interval of follow-up.

Data Layout for Binary Regression

SUBJECT	EVENT	D_1	D_2	D_3	TRT
1	1	1	0	0	1
2	0	1	0	0	0
2	0	0	1	0	0
2	1	0	0	1	0
3	0	1	0	0	1
3	1	0	1	0	1

The data layout is shown on the left. Each observation represents one interval of follow-up time allowing multiple observations per subject. EVENT is the dichotomous outcome variable. Subject 1 had the event in the first interval (EVENT = 1) and thus has one observation. Subject 2 has three observations because she survived the first two intervals (EVENT = 0) but got the event in the third interval. D_1 is a dummy variable coded 1 if the observation represents the first interval and 0 otherwise. Similarly, D_2 is coded 1 for the second interval and D_3 is coded 1 for the third interval.

EVENT: dichotomous outcome coded 1 if event, 0 for no event during the interval

D_1, D_2, D_3: dummy variables for intervals 1, 2, and 3 coded 1 if in the corresponding interval, 0 otherwise

TRT: Treatment coded 1 for new treatment, 0 for placebo

TRT is the predictor of interest, coded 1 for the new treatment and 0 for the placebo. TRT could be coded as a time-independent or time-dependent variable. In this example, TRT is time-independent because TRT does not change values over different intervals corresponding to the same subject.

Logistic Model

$$\text{Logit } P(Y = 1) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 \text{TRT}$$

A logistic model (shown at left) containing the three dummy variables and TRT can be formulated with the data in this form.

where $P(Y = 1)$ is the probability of event for a given interval conditioned on survival of previous intervals

Interpretation of Parameters

β_1: Log odds of event in 1st interval among TRT = 0

β_2: Log odds of event in 2nd interval given survival of 1st interval among TRT = 0

β_3: Log odds of event in 3rd interval given survival of first two intervals among TRT = 0

β_4: Log odds ratio for TRT

Care must be taken with the interpretation of the parameters: β_1 is the log odds of the event occurring in the first interval among the placebo group; β_2 is the log odds of the event occurring in the second interval conditioned on survival of the first interval among the placebo group; β_3 is the log odds of the event occurring in the third interval conditioned on survival of the first and second intervals among the placebo group; and β_4 is the log odds ratio for TRT.

D_1, D_2, D_3 play similar role as intercept

- Baseline measure when covariates are zero
- 3 parameters rather than 1 intercept
  - Baseline measure may differ for each interval

The dummy variables play a similar role to that of the intercept in a conventional regression, providing a baseline outcome measure for the case in which all predictors are zero (e.g., TRT = 0). In general, the baseline measure may differ for each interval, which is the reason that the model contains 3 dummy variables rather than 1 intercept.

Odds Ratio (TRT = 1 vs. TRT = 0) = $\exp(\beta_4)$

Model uses PO assumption

- OR constant over time
- PO assumption can be tested
  - Include interaction terms with TRT and dummy variables
  - Significant interaction suggests PO violation
  - Need ample data to practically carry out test

The odds ratio comparing TRT = 1 to TRT = 0 is obtained by exponentiating β_4. This model uses the **proportional odds (PO) assumption** in that the odds ratio is assumed constant over time (or at least constant at the end of each interval). This assumption can be tested by including interaction (product) terms with TRT and two of the dummy variables in the model. A statistically significant product term would suggest a violation of the PO assumption. However, if there are sparse data corresponding to particular intervals, it will not be practical to carry out such a test on those intervals.

Alternative Binary Model

$$\log(-\log(1 - P(Y = 1))) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 TRT$$

where $1 - P(Y = 1)$ is the probability of surviving a given interval conditioned on survival of previous intervals

Logistic regression is not the only type of binary regression that may be considered for interval-censored data. An alternative binary model (shown on the left) uses the **complementary log–log link** function rather than the **logit link** function that is used for the more familiar logistic regression.

Complementary log–log link

- Log–log survival modeled as linear function of regression parameters

Logit link

- Log odds of failure modeled as linear function of regression parameters

A model using a complementary log–log link function expresses the log negative log survival probability as a linear function of regression parameters. By contrast, a model using a logit link function expresses the log odds of failure (i.e., getting the event) as a linear function of regression parameters.

Complementary log–log model is PH model

- HR (TRT = 1 vs. TRT = 0) = $\exp(\beta_4)$
- HR constant over time

The complementary log–log binary model is a proportional hazards model. The hazard ratio comparing TRT = 1 to TRT = 0 is obtained by exponentiating β_4.

Log–log survival curves:
parallel $\Rightarrow$ additive effects
$\Rightarrow$ PH

Complementary log–log link:
additive effects on log–log scale
$\Rightarrow$PH

Recall we can use log–log survival curves to evaluate the PH assumption for a Cox model. If the effects are additive (e.g., parallel for TRT = 1 and TRT = 0) then the PH assumption is assumed to hold. The underlying idea is similar for the complementary log–log link function in that additive effects are assumed on the log–log scale (e.g., comparing TRT = 1 to TRT = 0).

In theory
- Survival time is continuous

In practice
- Survival time measured in intervals
  - If event occurred in month 7 then event occurred in an interval of time

In theory, the time-to-event variable in survival analyses is thought of as a continuous variable. In practice, however, the time variable is typically an interval of time. For example, if time is measured in months and an event occurs in month 7 then the event is recorded as having occurred in a specific *interval* lasting a month.

Discrete survival analysis

- Discrete time
- for example, number of menstrual cycles to pregnancy rather than time to pregnancy
  - Fraction of cycle does not make sense

Discrete survival analysis is a survival analysis in which the outcome variable is discrete, both in theory and in practice. For example, consider a study in which women who stop using oral contraception are followed until pregnancy. The outcome is defined as the number of menstrual cycles until pregnancy. The number of cycles rather than the time to pregnancy is used because the cycle length varies among women and a woman ovulates only once per menstrual cycle (i.e., one opportunity per cycle to become pregnant). The number of cycles is a discrete outcome. A fraction of a cycle does not make sense.

Analyzing discrete survival data

- Can use binary regression
- Analogous to interval-censored data
  - Discrete outcome — subjects survive discrete units of time
  - Interval outcomes — subjects survive intervals of time

Binary regression, as described in this section, can be applied for discrete survival outcomes in a similar manner to that described for interval-censored outcomes. With this method, subjects can be conceptualized as surviving discrete units of time analogously as subjects surviving continuous intervals of time.

XII. Frailty Models

Frailty

- Random component
- Accounts for extra variability from unobserved factors

In this section we consider the inclusion of frailty to a survival model. Frailty is a random component designed to account for variability due to unobserved individual-level factors that is otherwise unaccounted for by the other predictors in the model.

Conceptualize S(t) two ways:

- For an individual
- Averaging over a theoretical large population

Consider a survival model with a continuous age variable and dichotomous smoking status variable as the only predictors. Under this model the survival function for a 33-year-old smoker might be conceptualized in different ways. One way is as the survival function for an individual 33-year-old smoker. The second way is as some kind of averaging over a theoretical large population of 33-year-old smokers.

With Frailty Component

Jake and Blake
1. May have different S(t) due to unobserved factors
2. Extra source of variability in outcome (e.g., more variation than expected under Weibull)

Now suppose a "frailty" component is included in the model. Under this model, we can conceptualize survival functions specific to each individual. If Jake and Blake are both 33-year-old smokers, not only might their observed failure times be different, but under this model **their individual survival functions could also be different**. Jake may be more "frail" than Blake due to unobserved factors accounting for individual level differences in his hazard and survival functions. These unobserved factors may contribute an extra layer of heterogeneity, leading to greater variability in survival times than might be expected under the model (e.g., Weibull) without the frailty component.

Without Frailty Component

Jake and Blake
1. Have same S(t)
2. May have different event times because event time is random, following some distribution (e.g., Weibull)

The frailty component α ($\alpha > 0$)

- Unobserved multiplicative effect on hazard
- Follows distribution $g(\alpha)$ with $\mu = E(\alpha) = 1$
- $Var(\alpha) = \theta$, parameter to be estimated

The frailty α is an unobserved multiplicative effect on the hazard function assumed to follow some distribution $g(\alpha)$ with $\alpha > 0$ and the mean of α equal to 1. The variance of α is a parameter θ (theta) that is typically estimated from the data.

Hazard and survival conditioned on frailty

$h(t|\alpha) = \alpha h(t)$

$S(t|\alpha) = S(t)^{\alpha}$

An individual's hazard function conditional on the frailty can be expressed as α **multiplied by h(t)**. Using the relationship between the survival and hazard functions, the corresponding conditional survival function can be expressed as **S(t) raised to the** α **power**.

$\alpha > 1$

- Increased hazard: $\alpha h(t) > h(t)$
- Decreased survival: $S(t)^{\alpha} < S(t)$

$\alpha < 1$

- Decreased hazard: $\alpha h(t) < h(t)$
- Increases survival: $S(t)^{\alpha} > S(t)$

$\alpha = 1$ (average frailty): $\alpha h(t) = h(t)$

Individuals with $\alpha > 1$ have an increased hazard and decreased probability of survival compared to those of average frailty ($\alpha = 1$). Similarly, individuals with $\alpha < 1$ have a decreased hazard and increased probability of survival compared to those of average frailty.

Survival functions (with frailty models)

1. Conditional, $S(t|\alpha)$, individual level
2. Unconditional, $S_U(t)$, population level

With frailty models, we distinguish the individual level or **conditional survival function** $S(t|\alpha)$ discussed above, from the population level or **unconditional survival function** $S_U(t)$, which represents a population average. Once the frailty distribution $g(\alpha)$ is chosen, the unconditional survival function is found by integrating over the conditional survival function $S(t|\alpha)$ times $g(\alpha)$, with respect to α. The corresponding unconditional hazard $h_U(t)$ can then be found using the relationship between the survival and hazard functions (see left).

Unconditional survival function $S_U(t)$

$$S_U(t) = \int_0^{\infty} S(t|\alpha) g\{\alpha\} d\alpha$$

$$h_U(t) = \frac{-d[S_U(t)]/dt}{S_U(t)}$$

Frailty distribution $g(\alpha)$, $\alpha > 0$, $E(\alpha) = 1$

Stata offers choices for $g(\alpha)$
1. Gamma
2. Inverse-Gaussian

Both distributions parameterized in terms of θ, where $\text{Var}(\alpha) = \theta$

Any distribution for $\alpha > 0$ with a mean of 1 can theoretically be used for the distribution of the frailty. Stata supports two distributions: the **gamma distribution** and the **inverse-Gaussian distribution** for the frailty. With the mean fixed at 1, both these distributions are parameterized in terms of the variance θ and typically yield similar results.

EXAMPLE

Vet Lung Cancer Trial
Predictors:
TX (dichotomous: 1 = standard, 2 = test)
PERF (continuous: 0 = worst, 100 = best)
DD (disease duration in months)
AGE (in years)
PRIORTX (dichotomous: 0 = none, 10 = some)

Model 1. No Frailty

Weibull regression (PH form)
Log likelihood = −206.20418

_t	Coef.	Std. Err.	z	p > |z|
tx	.137	.181	0.76	0.450
perf	−.034	.005	−6.43	0.000
dd	.003	.007	0.32	0.746
age	−.001	.009	−0.09	0.927
priortx	−.013	.022	−0.57	0.566
_cons	−2.758	.742	−3.72	0.000
/ln_p	−.018	.065	−0.27	0.786
p	.982	.064		
1/p	1.02	.066		

To illustrate the use of a frailty model, we apply the data from the Veteran's Administration Lung Cancer Trial described in Chapter 5. The exposure of interest is treatment status TX (standard = 1, test = 2). The control variables are performance status (PERF), disease duration (DD), AGE, and prior therapy (PRIORTX), whose coding is shown on the left. The outcome is time to death (in days).

Output from running a Weibull PH model without frailty using Stata software is shown on the left (Model 1). The model can be expressed: $h(t) = \lambda p t^{p-1}$ where

$$\lambda = \exp(\beta_0 + \beta_1 \text{TX} + \beta_2 \text{PERF} + \beta_3 \text{DD} + \beta_4 \text{AGE} + \beta_5 \text{PRIORTX}).$$

The estimate of the hazard ratio comparing TX = 2 vs. TX = 1 is $\exp(0.137) = 1.15$ controlling for performance status, disease duration, age, and prior therapy. The estimate for the shape parameter is 0.982 suggesting a slightly decreasing hazard over time.

EXAMPLE: (continued)

Model 2. With Frailty

Weibull regression (PH form)
Gamma frailty
Log likelihood = −200.11338

_t	Coef.	Std. Err.	z	P>|z|
tx	.105	.291	0.36	0.719
perf	−.061	.012	−5.00	0.000
dd	−.006	.017	−0.44	0.663
age	−.013	.015	−0.87	0.385
priortx	−.006	.035	−0.18	0.859
_cons	−2.256	1.100	−2.05	0.040
/ln_p	.435	.141	3.09	0.002
/ln_the	−.150	.382	−0.39	0.695
p	1.54	.217		
1/p	.647	.091		
theta	.861	.329		

Likelihood ratio test of theta = 0:
chibar2(01) = 12.18
Prob>=chibar2=0.000

Model 2 (output on left) is the same Weibull model as Model 1 except that a frailty component has been included. The frailty in Model 2 is assumed to follow a gamma distribution with mean 1 and variance equal to theta (θ). The estimate of theta is 0.861 (bottom row of output). A variance of zero (theta = 0) would indicate that the frailty component does not contribute to the model. A likelihood ratio test for the hypothesis theta = 0 is shown directly below the parameter estimates and indicates a chi-square value of 12.18 with 1 degree of freedom yielding a highly significant p-value of 0.000 (rounded to three decimals).

Notice how all the **parameter estimates are altered with the inclusion of the frailty**. The estimate for the shape parameter is now 1.54, quite different from the estimate 0.982 obtained from Model 1. The inclusion of frailty not only has an impact on the parameter estimates but also complicates their interpretation.

Comparing Model 2 with Model 1

- There is one additional parameter to estimate in Model 2
- The actual values of individuals' frailty are not estimated in Model 2
- The coefficients for the predictor variables in Models 1 and 2 have different estimates and interpretation
- The estimate of the shape parameter is <1.0 for Model 1 and >1.0 for Model 2

Before discussing in detail how the inclusion of frailty influences the interpretation of the parameters, we overview some of the key points (listed on the left) that differentiate Model 2 (containing the frailty) and Model 1.

Model 2 contains one additional parameter, the variance of the frailty. However, the actual values of each subject's frailty are not estimated. The regression coefficients and Weibull shape parameter also differ in their interpretations for Model 2 compared to Model 1. We now elaborate on these points.

Model 2

Hazard for jth individual:

$h_j(t|\alpha_j) = \alpha_j h(t)$ j = 1, 2,. . ., n

where $h(t) = \lambda p t^{p-1}$
with $\lambda = \exp(\beta_0 + \beta_1 \text{TX} + \beta_2 \text{PERF} + \beta_3 \text{DD} + \beta_4 \text{AGE} + \beta_5 \text{PRIORTX})$
and where $\alpha \sim$ gamma ($\mu = 1$, variance = θ)

For Model 2 we can express the Weibull model with a gamma distributed frailty in terms of the individual level hazard for the jth subject.

If α_j denotes the frailty for the jth subject, then that subject's hazard $\mathbf{h_j}$ **(t|α_j)** can be expressed as α_j **multiplied by h(t)**, where h(t) is the Weibull hazard function parameterized in terms of the predictor variables and their regression coefficients (see left).

α_j not estimable
- An α_j associated with each subject
- Too many parameters

The values for each α_j are not estimable because there is a level of frailty associated with each data point. If we tried to estimate each subject's frailty, then there would be more parameters to estimate than observations in the dataset and the model would be overparameterized.

Rather, var[g(α)] is estimated
- Gamma is 2-parameter distribution
  - Mean set at 1.0
  - $\theta = \text{Var}(\alpha)$ is estimated

Rather, the variance of the frailty is estimated. The gamma distribution is a two-parameter distribution. Because the mean is set at 1, we need only estimate its variance to fully specify the frailty distribution.

Interpreting coefficients in Model 2

$$\widehat{HR} = \exp\left(\hat{\beta}_1\right) = 1.11$$

Estimates HR comparing two individuals
- With same α
- One with TX = 2, other with TX = 1
- With same levels of other predictors

The estimated coefficient for TX using Model 2 is 0.105. By exponentiating, we obtain exp (0.105) = 1.11. This is the **estimated hazard ratio for two individuals having the same frailty** in which one takes the test treatment and the other takes the standard treatment controlling for the other co-variates in the model. Thus, for two individuals with the same frailty, we can use the coefficient estimates from Model 2 to estimate the ratio of conditional hazards.

Recall: $h(t|\alpha) = \alpha h(t)$

$TX = 1$: $h_1(t|\alpha_1) = \alpha_1 h_1(t)$
$TX = 2$: $h_2(t|\alpha_2) = \alpha_2 h_1(t)$

If $\frac{h_2(t)}{h_1(t)} = \exp(\beta_1)$

then $\frac{\alpha_1 h_1(t)}{\alpha_2 h_2(t)} = \exp(\beta_1)$

only if $\alpha_1 = \alpha_2$

To clarify, recall that the individual level or conditional hazard function can be expressed as a multiplied by h(t). Suppose $h_1(t|\alpha_1)$ and $h_2(t|\alpha_2)$ are the conditional hazard functions for individuals who use the standard and test treatments, respectively, at the mean levels of the other covariates. If the ratio of $h_2(t)$ and $h_1(t)$ equals $\exp(\beta_1)$, then the ratio of $h_2(t|\alpha_2)$ and $h_1(t|\alpha_1)$ equals $\exp(\beta_1)$ only if the individuals have the same level of frailty (i.e., $\alpha_1 = \alpha_2$; see left).

Another interpretation for $\exp(\beta_1)$

- Ratio of conditional hazards from the same individual
- Effect for individual taking test rather than standard treatment

Another way to interpret the exponentiated coefficient for TRT, $\exp(\beta_1)$, is as a ratio of conditional hazards from the **same** individual. This measure can be used to estimate an effect for an individual taking the test treatment instead of the standard treatment.

Model 1 ($\hat{p} = 0.982$)
Decreasing hazard for individual and population because ($\hat{p} < 1$)

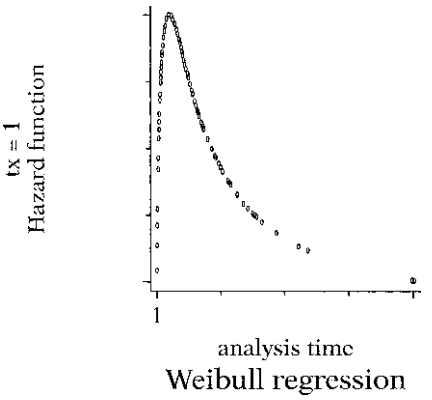
Model 2 ($\hat{p} = 1.54$)
Complication:
Individual level hazard
vs
Population level hazard

A somewhat striking difference in the output from Model 1 and Model 2 is the estimate of the shape parameter. The hazard estimated from Model 1 (without the frailty) is estimated to decrease over time because $\hat{p} < 1$. By contrast, the estimated individual level hazard from Model 2 is estimated to increase over time because $\hat{p} > 1$. However, the interpretation of the shape parameter in Model 2 has an additional complication that should be considered before making direct comparisons with Model 1. For frailty models, we have to distinguish between the **individual level** and **population level** hazards.

For Model 2
Conditional hazard increases
but
unconditional hazard unimodal

Although the estimated individual level or **conditional** hazard from Model 2 is estimated to increase over time, the estimated population level or **unconditional** hazard does not strictly increase. The unconditional hazard first increases but then decreases to zero, resulting in a **unimodal** shape due to the effect of the frailty, as we will now explain.

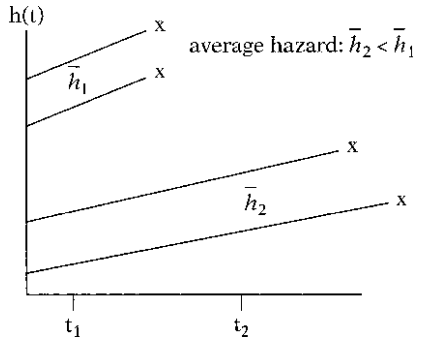
Estimated unconditional hazard Model 2 (TX = 1, mean level for other covariates, $\hat{p} = 1.54$)



On the left is a plot (from Model 2) of the estimated unconditional hazard for those on standard treatment (TX = 1) with mean values for the other covariates. The graph is unimodal, with the hazard first increasing and then decreasing over time. So each individual has an estimated increasing hazard ($\hat{p} = 1.54$), yet the hazard averaged over the population is unimodal, rather than increasing. How can this be?

The answer is that the population is comprised of individuals with different levels of frailty. The more frail individuals ($\alpha > 1$) have a greater hazard and are more likely to get the event earlier. Consequently, over time, the "at risk group" has an increasing proportion of less frail individuals ($\alpha > 1$), decreasing the population average, or unconditional, hazard.

Four increasing individual level hazards, but average hazard decreases from t_1 to t_2



To clarify the above explanation, consider the graph on the left in which the **hazards for four individuals increase** linearly over time until their event occurs. The two individuals with the highest hazards failed between times t_1 and t_2 and the other two failed after t_2. Consequently, the average hazard ($\bar{h}_2$) of the two individuals still at risk at t_2 is less than the average hazard ($\bar{h}_1$) of the four individuals at risk at t_1. Thus the average hazard of the "at risk" population **decreased** from t_1 to t_2 (i.e., $\bar{h}_2 < \bar{h}_1$) because the individuals surviving past t_2 were less frail than the two individuals who failed earlier.

Frailty Effect

$h_U(t)$ eventually decreases
because
"at risk group" becoming less frail over time

This property, in which the unconditional hazard eventually decreases over time because the "at risk group" has an increasing proportion of less frail individuals, is called the **frailty effect**.

Unconditional hazard $h_U(t)$ with gamma frailty

$$h_U(t) = \frac{h(t)}{1 - \theta \ln[S(t)]}$$

If $\theta = 0$ then $h_U(t) = h(t)$ (no frailty)

The unconditional hazard function $h_U(t)$, **with gamma frailty** is shown on the left.

If $\theta = 0$, then $h_U(t)$ reduces to $h(t)$ indicating that there is no frailty.

For Model 2:

- $h(t)$ and $S(t)$ are Weibull
- At $t = 0$
  - $h_U(t) = h(t)$ (increasing)
- As t gets large
  - If $\theta > 0$ then $h_U(t) \to 0$
- So $h_U(t)$ increases and then decreases (unimodal)

An examination of the expression for $h_U(t)$ gives further insight into how we obtained an estimated unconditional hazard of unimodal shape. $S(t)$ and $h(t)$ represent the survival and hazard functions ignoring the frailty, which for Model 2 corresponds to a Weibull distribution. If $t = 0$ then $h_U(t) = h(t)$, which for Model 2 yields an estimated increasing hazard. As t gets larger, and if $\theta > 0$, the denominator gets larger (because $\ln[S(t)]$ is negative) until eventually $h_U(t)$ approaches zero. So $h_U(t)$ is increasing at $t = 0$ but eventually decreases to zero, which means at some point in time, $h_U(t)$ changes direction.

Population level hazards (with gamma frailty)

$$h_{U1}(t) = \frac{h_1(t)}{1 - \theta \ln[S_1(t)]} \quad \text{for TX} = 1$$

$$h_{U2}(t) = \frac{h_2(t)}{1 - \theta \ln[S_2(t)]} \quad \text{for TX} = 2$$

Ratio of unconditional hazards (not PH)

$$\frac{h_{U2}(t)}{h_{U1}(t)} = \frac{h_2(t)}{h_1(t)} \times \frac{1 - \theta \ln[S_1(t)]}{1 - \theta \ln[S_2(t)]}$$

A consequence of the frailty effect is the need to distinguish between the ratio of individual level hazards and the ratio of population level hazards. For the population level hazards, the **PH assumption is violated** when a gamma (or inverse-Gaussian) distributed frailty is added to a PH model. To see this for gamma frailty, let $h_{U1}(t)$ and $h_{U2}(t)$ be the unconditional hazard functions representing the standard and test treatments, respectively, at the mean levels of the other covariates. The ratio of these hazards is shown on the left.

If

$$\frac{h_2(t)}{h_1(t)} = \exp(\beta_1)$$

then

$$\frac{h_{U2}(t)}{h_{U1}(t)} = \exp(\beta_1) \times \frac{1 - \theta \ln[S_1(t)]}{1 - \theta \ln[S_2(t)]}$$

not constant over time, PH violated

If the ratio of $h_2(t)$ and $h_1(t)$ equals $\exp(\beta_1)$, then the ratio of the unconditional hazards equals $\exp(\beta_1)$ times the ratio of $1 - \theta \ln[S_1(t)]$ and $1 - \theta \ln[S_2(t)]$. This latter ratio is a function of time and only cancels when t equals zero. Therefore the ratio of the unconditional hazards is not constant over time, thus violating the PH assumption.

Plots of $\hat{S}(t)$

- Generally averaged over population
  - An important consideration for frailty models

Generally, survival plots are estimated over a population average (e.g., Kaplan–Meier). When considering PH models without frailty, we do not need to distinguish between the conditional and unconditional survival functions. However, this distinction needs to be considered with frailty models.

Suppose $\ln[-\ln \hat{S}(t)]$ curves for TX start parallel but then converge over time:

1. It may be effect of TX weakens over time ⇓ PH model not appropriate
2. It may be effect of TX is constant over time but unobserved heterogeneity is in population ⇓ PH model with frailty is appropriate

Suppose we plot Kaplan–Meier log–log survival estimates evaluating the PH assumption for treatment (TX = 2 vs. TX = 1), and the plots start out parallel but then begin to converge over time. One interpretation is that the effect of the treatment weakens over time. For this interpretation, a PH model is not appropriate. Another interpretation is that the effect of the treatment remains constant over time but the plots converge due to unobserved heterogeneity in the population. For this interpretation, a PH model with frailty would be appropriate.

Model 2 (Weibull with frailty)

- Used PH parameterization
- Can equivalently use AFT parameterization

Recall, from Section VI of this chapter, that a Weibull PH model is also an AFT model. The only difference is in the way the model is parameterized. We next present the AFT form of Model 2.

Unconditional survival function $S_U(t)$ with gamma frailty $g(\alpha)$

$$S_U(t) = \int_0^\infty S(t|\alpha) g(\alpha) d\alpha$$
$$= [[1 - \theta \ln S(t)]^{-1/\theta}$$

Before stating the model, we show the unconditional survival function using gamma frailty. Recall that the unconditional survival function is obtained by integrating over the frailty, as shown on the left.

Model 3 (Weibull AFT with gamma frailty)

$$S_U(t) = [1 - \theta \ln S(t)]^{-1/\theta}$$

where $S(t) = \exp(-\lambda t^p)$ (Weibull) and

$$\frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + a_1 TX + \alpha_2 PERF + \alpha_3 DD + \alpha_4 AGE + \alpha_5 PRIORTX)$$

Model 3 (the AFT form of Model 2) is presented in terms of the unconditional survival function $S_U(t)$. The unconditional survival function is a function of S(t), which represents the Weibull survival function. The Weibull survival function, in turn, is parameterized in terms of the shape parameter p and regression coefficients using AFT parameterization (see left).

Model 3 Output

Weibull regression (**AFT form**)
Gamma frailty
Log likelihood = −200.11338

_t	Coef.	Std. Err.	z	P > |z|
tx	−.068	.190	−0.36	0.721
perf	.040	.005	8.37	0.000
dd	.004	.009	0.44	0.661
age	.008	.009	0.89	0.376
priortx	.004	.023	0.18	0.860
_cons	1.460	.752	1.94	0.052
/ln_p	.435	.141	3.09	0.002
/ln_the	−.150	.382	−0.39	0.695
p	1.54	.217		
1/p	.647	.091		
theta	.861	.329		

Likelihood ratio test of theta = 0:
chibar2(01) = 12.18
Prob>=chibar2 = 0.000

The output for Model 3, shown on the left, is similar to that obtained from Model 2. The estimates for theta and p are identical to those obtained from Model 2. The difference is that the regression coefficients obtained with Model 3 use AFT parameterization, i.e., multiply by $-p = -1.54$ to get the PH coefficient estimates in Model 2.

$$\hat{\gamma}(\text{TX = 2 vs. 1}) = \exp(-0.068)$$
$$= 0.93$$

An estimated acceleration factor of 0.93 **comparing two individuals with the same level of frailty**, for the effect of treatment (TX = 2 vs. TX = 1) and controlling for the other covariates, is obtained by exponentiating the estimated coefficient (−0.068) of the TX variable.

Comparing individuals with same α

Interpreting $\hat{\gamma}$

- Taking test treatment reduces individual's median survival time by factor of 0.93
- Suggests slightly harmful effect
- $\hat{\alpha}_1$ is not significant (p = 0.721)

Another interpretation for this estimate is that an individual taking the test treatment instead of the standard treatment reduces her median survival time (i.e., contracts her individual level survival function) by an estimated factor of 0.93. This estimate suggests a slight harmful effect from the test treatment compared to the standard treatment. However, the estimated coefficient for TX is not significant, with a p-value of 0.721.

PH assumption
Individual level PH $\nRightarrow$ Population level PH

AFT assumption
Individual level AFT $\Rightarrow$ Population level AFT

A key difference between the PH and AFT formulations of this model is that if the AFT assumption holds at the individual level, then it will also hold at the population level using the gamma (or inverse-Gaussian) distributed frailty.

Population level survival (with gamma frailty)

$S_{U1}(t) = [[1 - \theta \text{ In } S_1(t)]^{-1/\theta}$

$S_{U2}(t) = [[1 - \theta \text{ In } S_2(t)]^{-1/\theta}$

To see this for gamma frailty, let $S_{U1}(t)$ and $S_{U2}(t)$ be the unconditional survival functions representing the standard and test treatments respectively, at the mean levels of the other covariates.

Also let γ represent the individual level acceleration factor for treatment; that is, $S_1(t) = S_2(\gamma t)$. Then $S_{U1}(t) = S_{U2}(\gamma t)$ (see left).

If $S_1(t) = S_2(\gamma t)$
then

$$S_{U1}(t) = [[1 - \theta In S_1(t)]^{-1/\theta}$$
$$= [[1 - \theta In S_2(\gamma t)]^{-1/\theta}$$
$$= S_{U2}(\gamma t)$$

Thus,
Individual level AFT
$\Rightarrow$ Population level AFT

Thus, for models with gamma frailty, if the AFT assumption holds at the individual level then it also holds at the population level.

Coefficient estimates from Model 3

- Applies to individual or population
- Interpretation of $\exp(\hat{\alpha}_1) = 0.93$
  - Median survival time for individual reduced by factor of 0.93
  - Median survival time reduced in population by factor of 0.93

The coefficient estimates obtained from Model 3 can therefore be used at the population level as well as the individual level. So another interpretation for the estimated acceleration factor for treatment is that the test treatment reduces the median survival time in the population by an estimated factor of 0.93.

Models 2 and 3:
Same model, different parameterization
Same estimates for
$S(t)$, $S_U(t)$, $h(t)$, $h_U(t)$

Model 2 and Model 3 are the same model but use different parameterizations. The models provide identical estimates for the hazard and survival functions.

Models 2 and 3: Weibull with gamma frailty

- Unimodal unconditional hazard

Log-logistic model

- Accommodates unimodal hazard without a frailty component

Recall that the estimated unconditional hazard function obtained from this frailty model is of unimodal shape. Alternatively, a log-logistic (or lognormal) model, which accommodates a unimodal-shaped hazard function, could have been run without the frailty (see Practice Exercises 8 to 11 for comparison).

Parametric likelihood with frailty

- Uses $f_U(t)$, where $f_U(t) = h_U(t) S_U(t)$
- Formulated similarly to that described in Section X with $f_U(t)$ replacing $f(t)$
- Additional parameter θ

The likelihood for Model 3 can be formulated using the unconditional probability density function $f_U(t)$ which is the product of the unconditional hazard and survival functions. The likelihood is constructed in a similar manner to that described previously in this chapter except that $f_U(t)$ is used for the likelihood rather than $f(t)$ (see Section X). The main difference is that there is one additional parameter to estimate, the variance of the frailty.

Shared Frailty

- Clusters share same frailty
- For example, subjects from same family may share unobserved factors
  - Shared frailty designed to account for such similarities

Another type of frailty model is the **shared frailty model**. With this model, clusters of subjects are assumed to share the same frailty. For example, subjects from the same family may be similar with respect to some unobserved genetic or environmental factors. Allowing family members to share the same frailty is designed to account for such similarities.

Unshared Frailty

- The type of frailty we have described previous to this point
- Frailty distributed independently among subjects

By contrast, the frailty described previous to this point (**unshared frailty**) has been assumed to be distributed independently among subjects.

Shared Frailty Models

- Similar to random effect regression models
- Accounts for within-cluster correlation
- θ is a measure of the degree of correlation

Adding shared frailty to a survival model plays an analogous role to that of adding a random effect to a linear regression as a way to account for correlation between clusters of observations (Kleinbaum and Klein 2010). The estimate for the variance parameter θ in a shared frailty model can be thought of as a measure of the degree of correlation, where $\theta = 0$ indicates no within-cluster correlation.

Hazard conditional on shared frailty (for jth subject in kth cluster)

$h_{jk}(t|\alpha_k) = \alpha_k h_{jk}(t)$
where
$$h_{jk}(t) = h(t|\mathbf{X}_{jk})$$
for $j = 1, 2, \ldots, n_k$
and total n_k subjects in k^{th} cluster

If family is the cluster variable, then
subjects of same family have same α_k

For a shared frailty model, the conditional hazard function for the jth subject from the kth cluster can be expressed as α_k **multiplied by $\mathbf{h_{jk}(t)}$** where $h_{jk}(t)$ depends on the subject's covariates $\mathbf{X}_{jk}$. Notice that the frailty α_k is subscripted by k, but not by j. This indicates that subjects from the same cluster share the same frailty. If, for example, subjects are clustered by family, then subjects from the same family are assumed to have the same frailty.

Shared and unshared frailty

- Fundamentally the same
  - Accounts for variation due to unobservable factors
- Difference in data to which they are applied
  - Affects interpretation and methods of estimation

The frailty in a shared frailty model or unshared frailty model **is fundamentally the same**, a random effect to account for a source of variation due to unobservable, or latent, factors. **However, the data to which the shared and unshared frailty is applied are different, affecting differences in interpretation and methods of estimation**.

Unshared frailty models

- Subjects assumed independent

Shared frailty models

- Accounts for dependence among subjects who share frailty

For unshared frailty models, a subject's survival is assumed to be independent of the survival of other subjects in the study population. For shared frailty models, however, the frailty accounts for dependence among subjects who share the same frailty. Shared frailty provides an approach to account for correlation in the data due to unobservable factors common within clusters of subjects.

Likelihood for shared frailty models

- More complicated than for unshared frailty models
- Unconditional contribution of each cluster formulated separately by integrating out $g(\alpha)$
- Full likelihood formed as product of unconditional contribution from each cluster

The formulation of the likelihood is more complicated for shared frailty models than it is for unshared frailty models. To construct the shared frailty likelihood, the unconditional contribution for each cluster of subjects is formulated separately by integrating out the frailty from the product of each subject's conditional contribution. The full likelihood is then formulated as the product of the contributions from each cluster (see Gutierrez 2002 for details).

Shared frailty in Cox model

- Provided by Stata
  - Only gamma distributed shared frailty available
- Accounts for within-group correlation

Up to this point we have discussed frailty in terms of parametric models. Stata (version 8) allows shared frailty to be included in a Cox model in order to account for within-group correlation. The conditional hazard function for the jth subject from the kth cluster can be expressed as α_k multiplied by the baseline hazard $h_0(t)$ multiplied by $\exp(\beta \mathbf{X}_{jk})$. The frailty component is assumed to follow some distribution even though the distribution is unspecified for the rest of the model. Stata only allows a gamma distribution for the frailty to be included with a Cox model.

Cox shared frailty model

$h_{ij}(t|\alpha_j) = \alpha_k h_0(t) \exp(\beta X_{jk})$
for $j = 1, 2, \ldots, n_k$
total of n_k subjects in kth cluster

PH violation of $h_U(t)$ in Cox model

- if gamma-distributed frailty included
- Interpreting coefficient estimates
  - Only used for HR estimates among those who share same α

If a gamma-distributed frailty component is added to the Cox model, then the PH assumption is not satisfied for the unconditional hazards. In this framework, the frailty in a Cox model can be thought of as a source of random error that causes violation of the PH assumption at the population level. Consequently, care must be taken in the interpretation of the coefficient estimates. They can only be used to obtain estimates for hazard ratios conditioned on the same level of frailty.

Recurrent events

- Multiple events from same subject
- Events from same subject may be correlated
- Clusters are formed representing each subject
  - Different subjects do not share frailty
  - Observations from same subject share frailty

Shared frailty models can also be applied to recurrent event data. It is reasonable to expect that multiple events occurring over follow-up from the same individual would be correlated. To handle within-subject correlation, clusters are formed, each containing observations from the same subject. In this setting, it is not the case that different subjects share the same frailty. Rather, multiple observations representing the same subject share the same frailty.

Recurrent events:

- Topic of next chapter (Chapter 8)

Survival analyses on recurrent events are the focus of the next chapter (Chapter 8) of this text. An example of a Weibull model with shared frailty applied to recurrent event data is presented in the next chapter.

XIII. Summary

Parametric Models

- Assume distribution for survival time
- Distribution specified in terms of parameters
- Parameters estimated from data

In this chapter we presented parametric survival models as an alternative to the Cox model. They are called parametric models because the distribution of the time-to-event variable is specified in terms of unknown parameters, which are estimated from the data. Distributions that are commonly utilized are the exponential, the Weibull, the log-logistic, the lognormal, and the generalized gamma.

$f(t)$ specified $\Rightarrow$ corresponding $S(t)$, $h(t)$ also determined

Moreover,
Specifying one of f(t), S(t), or h(t) determines all three functions

More precisely, for parametric survival models, it is the probability density function f(t) of the distribution that is specified in terms of the parameters. Once f(t) is specified, the corresponding survival and hazard functions S(t) and h(t) can also be determined. Moreover, specifying any one of the probability density function, survival function, or hazard function allows the other two functions to be determined.

Parametric models

- Need not be PH models
- Many are AFT models

The proportional hazards (PH) assumption is the underlying assumption for a Cox PH model. However, parametric survival models need not be proportional hazards models. Many parametric models are acceleration failure time (AFT) models rather than proportional hazards models.

Acceleration factor (γ)

- Key measure of association in AFT models
- Describes stretching or contraction of S(t)

The acceleration factor (γ) is the key measure of association obtained in an AFT model. It describes the "stretching out" or contraction of survival functions when comparing one group to another. If $S_1(t)$ and $S_2(t)$ are the survival functions for Group 1 and Group 2, respectively, then the AFT assumption can be expressed as $S_2(t) = S_1(\gamma t)$.

AFT assumption

$$S_2(t) = S_1(\gamma t)$$
$$\uparrow \quad \uparrow$$
Group 2 Group 1

Detailed examples presented:

- Exponential model
- Weibull model
- Log-logistic model

We presented detailed examples of the exponential, Weibull, and log-logistic model using the remission dataset.

Exponential Model

- h(t) = λ (constant hazard)
- Special case of Weibull model

The underlying assumption for an exponential model, a special case of the Weibull model, is that the hazard function is constant over time (i.e., h(t) = λ). The Weibull model is unique in that if the PH assumption holds then the AFT assumption also holds (and vice versa). The log-logistic model does not satisfy the PH assumption. However, if the AFT assumption holds in a log-logistic model, then the proportional odds (PO) assumption also holds (and vice versa).

Weibull Model

- AFT ⇔ PH

Log-logistic Model

- Not a PH model
- AFT ⇔ PO

PO assumption

$$\text{OR} = \frac{S(t,x^*)/[1-S(t,x^*)]}{S(t,x)/[1-S(t,x)]}$$

OR is constant over time

The idea underlying the proportional odds assumption is that the survival (or failure) odds ratio comparing two specifications of covariates remains constant over time.

Graphical Evaluation

Weibull and Exponential

- Plot ln[−ln $\hat{S}$(t)] against ln(t)

Log-logistic:

- Plot $\ln\left[\frac{\hat{S}(t)}{(1-\hat{S}(t))}\right]$ against ln(t).

Check for linearity

We presented graphical approaches for evaluating the appropriateness of the exponential, Weibull, and log-logistic model by plotting a function of the Kaplan–Meier survival estimates $\hat{S}(t)$ against the log of time and then checking for linearity.

For evaluation of the exponential and Weibull assumptions, the ln[−ln $\hat{S}(t)$] is plotted against ln(t) and for evaluation of the log-logistic assumption the log odds of $\hat{S}(t)$ is plotted against ln(t).

Presented other parametric models

- Generalized gamma model
- Lognormal model
- Gompertz model

We briefly discussed other parametric models such as the generalized gamma, lognormal, and Gompertz models and showed additional parametric approaches such as modeling ancillary (shape) parameters as a function of predictor variables.

Contributions to Likelihood

If event at t, contributes f(t)

If censored, integrate over f(t)

$$\int_0^{t_1} f(t)dt : \text{left} - \text{censored at } t_1$$

$$\int_{t_1}^{\infty} f(t)dt : \text{right} - \text{censored at } t_1$$

$$\int_{t_1}^{t_2} f(t)dt : \text{interval} - \text{censored from } t_1 \text{ to } t_2$$

The parametric likelihood was developed and includes a discussion of left-, right-, and interval-censored data. If a subject has an event at time t, then that subject's contribution to the likelihood is f(t). On the other hand, if a subject is censored (i.e., exact time of event unknown), then the subject's contribution to the likelihood is found by integrating over f(t). The integration limits are determined by the time and type of censorship (see left).

Full likelihood (L)

$$L = \prod_{j=1}^{N} L_j \quad \text{j} = 1, 2, \ldots, \text{N}$$

where L_j is the contribution from jth subject and N = # of subject's

Assuming independence among subjects, the full likelihood can be formulated as a product of each subject's contribution.

Binary regression for interval-censored data

- Follow-up divided into intervals
  - Allows for multiple observations per subject
- Binary outcome variable defined
  - Indicates survival or failure over each interval

We showed how binary regression could be applied to interval-censored data by defining a dichotomous outcome variable indicating subjects' survival or failure over each interval of their follow-up. The data layout for this type of analysis allows multiple observations per subject, representing intervals of survival prior to failure (or censorship).

Binary regression for discrete survival analysis

- Analogous to interval-censored data
  - Discrete outcome–subjects survive discrete units of time
  - Interval outcomes–subjects survive intervals of time

Binary regression can also be used for discrete survival analysis in which the "time-to-event" variable is considered discrete rather than continuous. The data layout is similar to that for interval-censored data except subjects are conceptualized as surviving discrete units of time rather than continuous intervals of time.

Frailty, α

$$h(t|\alpha) = \alpha h(t)$$

multiplicative from effect on h(t)
mean = 1, varience = θ

θ estimated from data

We concluded with a discussion of frailty models. The frailty α is a multiplicative random effect on the hazard designed to account for individual-level unobserved factors that add an extra layer of variability beyond what has already been specified in the model. The frailty is generally assumed to follow a distribution with mean equal to 1 and is typically parameterized in terms of the variance θ which is estimated from the data.

Chapters

1. Introduction to Survival Analysis
2. Kaplan–Meier Curves and the Log-Rank Test
3. The Cox Proportional Hazard Model
4. Evaluating the Proportional Hazards Assumption
5. The Stratified Cox Procedure
6. Extension of the Cox Proportional Hazards Model for Time-Dependent Covariates
- ✓7. Parametric Survival Models

Next:

8. Recurrent Event Survival Analysis

The presentation is now complete. The reader can review the detailed outline that follows and then answer the practice exercises and test.

In the next chapter (8) entitled "Recurrent Event Survival Analysis," we consider approaches for analyzing data in which individuals may have more than one event over the course of their follow-up.

Detailed Outline

I. Overview (pages 292–294)

A. Parametric Survival Models

i. Outcome assumed to follow specified distribution

ii. Weibull, exponential (a special case of the Weibull), log-logistic, lognormal, and generalized gamma are supported with popular software (SAS and Stata)

iii. Contrasts with Cox model in which baseline hazard and survival functions are not specified

II. Probability Density Function in Relation to the Hazard and Survival Function (pages 294–295)

A. If any one of the hazard h(t), survival S(t), or probability density f(t) functions is known then the other two functions can be determined.

B. If f(t) is specified, then $S(t) = \int_t^\infty f(u)du$

C. If S(t) is specified, then $h(t) = (-d[S(t)]/dt)/S(t)$ and $f(t) = (-d[S(t)])/dt$

D. If h(t) is specified, then $S(t) = \exp\left(-\int_0^t h(u)du\right)$ and $f(t) = h(t)S(t)$

III. Exponential Example (pages 295–297)

A. Hazard is constant (i.e., not a function of time) in an exponential model

i. Stronger assumption than the PH assumption that the HR is constant

B. Exponential PH model (one predictor X_1)

i. In terms of the hazard: $h(t) = \lambda$ where $\lambda = \exp(\beta_0 + \beta_1 X_1)$

ii. Hazard ratio: HR ($X_1 = 1$ vs. $X_1 = 0$) = $\exp(\beta_1)$

IV. Accelerated Failure Time Assumption (pages 298–300)

A. Underlying assumptions

i. AFT — effect of covariates is multiplicative with respect to survival time

ii. PH — effect of covariates is multiplicative with respect to the hazard

B. The acceleration factor (γ) is the key measure of association in an AFT
   i. Acceleration factor is a ratio of survival times corresponding to any fixed value of S(t); that is, t_A/t_B where A and B denote two individuals for which $S(t_A) = S(t_B)$
   ii. $S_2(t) = S_1(\gamma t)$, survival function for Group 1, $S_1(t)$ is stretched (or contracted) by a factor of γ compared to survival function for Group 2, $S_2(t)$

C. AFT illustration
   i. Dogs are said to grow older 7 times faster than humans, $S_D(t) = S_H(7t)$

V. Exponential Example Revisited (pages 300–304)

A. Exponential AFT model (one predictor X_1)
   i. In terms of survival: $S(t) = \exp(-\lambda t)$ where $\lambda = \exp[-(\alpha_0 + \alpha_1 X_1)]$
   ii. In terms of time: $t = [-\ln(S(t)] \times \exp(\alpha_0 + \alpha_1 X_1)$
   iii. Acceleration factor ($X_1 = 1$ vs. $X_1 = 0$), $\gamma = \exp(\alpha_1)$

B. An exponential PH model is an exponential AFT model (but uses different parameterization)
   i. $\beta_j = -\alpha_j$, where β_j and $-\alpha_j$ are PH and AFT parameterization for the jth covariate
   ii. $\alpha > 1$ for ($X_1 = 1$ vs. $X_1 = 0$) implies effect of $X_1 = 1$ is beneficial to survival
   iii. HR > 1 for ($X_1 = 1$ vs. $X_1 = 0$) implies effect of $X_1 = 1$ is harmful to survival

C. Exponential model is a special case of a Weibull model
   i. Graphical approach for evaluating appropriateness of exponential model is described in the section on the Weibull example

VI. Weibull Example (pages 304–309)

A. PH form of the Weibull model (one predictor X_1)
   i. In terms of the hazard: $h(t) = \lambda p t^{p-1}$ where $\lambda = \exp(\beta_0 + \beta_1 X_1)$
   ii. Hazard ratio: HR ($X_1 = 1$ vs. $X_1 = 0$) = $\exp(\beta_1)$

iii. Weibull hazard is monotonic with its direction determined by the value of the shape parameter p
   a. $p > 1$ hazard increases over time
   b. $p = 1$ constant hazard (exponential model)
   c. $p < 1$ hazard decreases over time

A. Graphical approach for evaluating appropriateness of Weibull model
   i. Plot the log negative log of the Kaplan–Meier survival estimates against the log of time for each pattern of covariates
      a. If Weibull assumption is correct then plots should be straight lines of slope p
      b. If exponential assumption is correct then plots should be straight lines with slope equal to one ($p = 1$)
      c. If plots are parallel straight lines then Weibull PH and AFT assumptions are reasonable

B. AFT form of the Weibull model (one predictor X_1)
   i. In terms of survival:
   $S(t) = \exp(-\lambda t^p) = \exp[-(\lambda^{1/p}t)^p]$ where $\lambda^{1/p} = \exp[-(\alpha_0 + \alpha_1 X_1)]$
   ii. In terms of time:
   $t = [-\ln(S(t)]^{1/p} \times \exp(\alpha_0 + \alpha_1 X_1)$
   iii. Acceleration factor ($X_1 = 1$ vs. $X_1 = 0$), $\gamma = \exp(\alpha_1)$

C. A Weibull PH model is a Weibull AFT model (but uses different parameterization)
   i. Unique property of Weibull model (exponential is special case, $p = 1$)
   ii. $\beta_j = -\alpha_j p$ where β_j and α_j are PH and AFT parameterization, respectively, for the jth covariate

VII. Log-Logistic Example (pages 309–314)

A. Log-logistic hazard function:
$h(t) = \lambda p t^{p-1}/(1 + \lambda t^p)$.
   i. $p \leq 1$ hazard decreases over time
   ii. $p > 1$ hazard first increases and then decreases over time (unimodal)

B. Graphical approach for evaluating appropriateness of log-logistic model
   i. Plot the log of the survival odds (using KM estimates) against the log of time for each pattern of covariates
      a. if log-logistic assumption is correct then plots should be straight line of slope $-p$
      b. If plots are parallel straight lines then log-logistic proportional odds (PO) and AFT assumptions are reasonable

C. Log-logistic AFT model (one predictor X_1):
   i. In terms of survival:
   $S(t) = 1/(1+\lambda t^P) = 1/(1+(\lambda t^{1/P} t)^P)$ where $\lambda^{1/p} = \exp(-(\alpha_0 + \alpha_1 X_1))$
   ii. In terms of time:

$$t = \left[\frac{1}{S(t)} - 1\right]^{1/p} \times \exp(\alpha_0 + \alpha_1 X_1)$$

   iii. Acceleration factor ($X_1 = 1$ vs. $X_1 = 0$), $\gamma = \exp(\alpha_1)$

D. Log-logistic proportional odds (PO) model (one predictor X_1)
   i. In terms of survival: $S(t) = 1/(1 + \lambda t^p)$ where $\lambda = \exp(\beta_0 + \beta_1 X_1)$
   ii. Odds of an event (failure odds) by time t: $(1 - S(t))/S(t) = \lambda t^p$
   iii. Odds of surviving event (survival odds) beyond t: $S(t)/(1 - S(t)) = 1/\lambda t^p$
   iv. Failure odds ratio:
   HR ($X_1 = 1$ vs. $X_1 = 0$) $= \exp(\beta_1)$
      a. PO assumption is that the odds ratio is constant over time
   v. Survival odds ratio:
   HR ($X_1 = 1$ vs. $X_1 = 0$) $= \exp(-\beta_1)$
      a. Survival odds ratio is reciprocal of failure odds ratio

E. A log-logistic AFT model is a log-logistic PO model (but uses different parameterization)
   i. Log-logistic model is not a proportional hazards (PH) model
   ii. $\beta_j = -\alpha_j p$ where β_j and α_j are PO and AFT parameterization for the jth covariate
      a. Shape parameter with Stata is parameterized as gamma $= 1/p$

VIII. A More General Form of the AFT Model (pages 314–316)

A. General form with one predictor (X_1):
$\ln(T) = \alpha_0 + \alpha_1 X_1 + \epsilon$

B. Include additional parameter, σ:
$\ln(T) = \alpha_0 + \alpha_1 X_1 + \sigma\epsilon$

C. Let $\sigma = 1/p \Rightarrow \ln(T) = \alpha_0 + \alpha_1 X_1 + (1/p)\,\epsilon$

D. Additive in terms of ln(T) but multiplicative in terms of T:

$$T = \exp\left(\alpha_0 + \alpha_1 X_1 + \frac{1}{p}\epsilon\right)$$
$$= \exp[\alpha_0 + \alpha_1 X_1] \times \exp\left(\frac{1}{p}\epsilon\right)$$

E. Collapse α_0 into baseline term,let
$T_0 = \exp(\alpha_0)\exp\left(\frac{1}{p} \in\right)$:
so $T = \exp(\alpha_1 X_1) \times T_0$

IX. Other Parametric Models (pages 316–318)

A. Generalized gamma model
   i. Additional shape parameters give flexibility in distribution shape
   ii. Weibull and lognormal are special cases

B. Lognormal model
   i. ln(T) follows a normal distribution
   ii. Accommodates AFT model

C. Gompertz model
   i. PH model, not AFT model

D. Modeling failure time as an additive model
   i. Additive model with one predictor:
   $T = \alpha_0 + \alpha_1 TRT + \epsilon$ (no log link)

E. Modeling ancillary parameters
   i. Typically shape parameter p is considered a fixed constant
   ii. Can reparameterize shape parameter in terms of predictor variables and regression coefficients

X. The Parametric Likelihood (pages 318–321)

A. Product of each subject contribution (assuming independence)

B. Subject's contribution uses probability density function f(t)
   i. Subject contributes f(t) if event is observed at time t

ii. Integrate over f(t) if subject is censored
   a. Integrate from 0 to t if subject is left-censored at t
   b. Integrate from t to infinity if subject is right-censored at t
   c. Integrate over interval of censorship if subject is interval-censored

XI. Interval-Censored Data (pages 321–326)

A. Binary regression is alternative approach if data are interval-censored

B. Binary outcome variable represents survival or failure over each subinterval of subject's follow-up

C. Specify a link function when using binary regression
   i. Logit link for logistic regression
   ii. Complementary log–log link is an alternative to logistic regression

D. Discrete survival analysis
   i. Time-to-event variable is discrete
   ii. Binary regression can be applied in a similar manner to that of interval-censored data

XII. Frailty Models (pages 326–340)

A. The frailty α is an unobserved multiplicative effect on the hazard function
   i. Hazard, conditioned on the frailty, $h(t|\alpha) = \alpha h(t)$
   ii. Survival, conditioned on the frailty, $S(t|\alpha) = S(t)^{\alpha}$

B. Frailty assumed to follow some distribution $g(\alpha)$ of mean 1 and variance θ
   i. The variance θ is a parameter estimated by the data
   ii. Gamma distribution offered by Stata and R software

C. Designed to account for unobserved individual-level factors that influence survival
   i. Distinction is made between the individual-level and population-level hazards. PH assumption may hold on individual level but not on population level

D. Shared frailty models allow individuals to share the same frailty
   i. Play similar role as adding a random effect to a linear regression
   ii. Can account for within-group correlation.

XIII. Summary (pages 341–344)

Practice Exercises

Answer questions 1 to 5 as true or false (circle T or F)

T F 1. The acceleration factor comparing exposed and unexposed subjects, (E = 1 vs. E = 0), is a ratio of their median survival times (time to S(t) = 0.5), or more generally the ratio of their times to any fixed value of S(t) = q.

T F 2. Let $S_0(t)$ be the survival function for unexposed subjects (E = 0) and let $S_1(t)$ be the survival function for exposed subjects (E = 1). If $S_0(t) = S_1(3t)$ then the median survival time for the unexposed subjects is 3 times longer than the median survival time for the exposed subjects.

T F 3. The Cox proportional hazards model is a parametric model.

T F 4. if the acceleration failure time (AFT) assumption holds in a Weibull model then the proportional hazards assumption also holds.

T F 5. The hazard is assumed constant in a log-logistic model.

Questions 6 and 7 make use of the output (copied below) presented in Sections III and V containing an example of the exponential model. This example used the remission data with treatment status (coded TRT = 1 for the experimental treatment and TRT = 0 for the placebo). The exponential survival and hazard functions are, respectively, $S(t) = \exp(-\lambda t)$ and $h(t) = \lambda$ where $\lambda = \exp[-(\alpha_0 + \alpha_1 TRT)]$ for the AFT parameterization and $\lambda = \exp(\beta_0 + \beta_1 TRT)$ for the PH parameterization. The output for both the AFT and PH forms of the model are presented.

Exponential regression accelerated failure-time form $\lambda = \exp[-(\alpha_0 + \alpha_1 TRT)]$

_t	Coef.	Std. Err.	z	p>|z|
trt	1.527	.398	3.83	0.00
_cons	2.159	.218	9.90	0.00

Exponential regression log relative-hazard form $\lambda = \exp(\beta_0 + \beta_1 TRT)$

_t	Coef.	Std. Err.	z	p>|z|
trt	−1.527	.398	3.83	0.00
_cons	−2.159	.218	−9.90	0.00

6. In this chapter it was shown in an exponential model that the time to event is $t = [-\log(S(t)] \times (1/\lambda)$ given a fixed value of S(t). Use the output from the AFT form of the model to estimate the median survival time (in weeks) for the treated group (TRT = 1) and the placebo group (TRT = 0).

7. Use the output from the PH form of the model to estimate the median survival time for the treated group (TRT = 1) and the placebo group (TRT = 0). Notice the answers from Questions 6 and 7 are the same, illustrating that the AFT and PH forms of the exponential model are just different parameterizations of the same model.

Questions 8 to 11 refer to a log-logistic AFT model using the data from the Veteran's Administration Lung Cancer Trial. The exposure of interest is treatment status TX (standard = 1, test = 2). The control variables are performance status (PERF), disease duration (DD), AGE, and prior therapy (PRIORTX). These predictors are used in the section on frailty models. The outcome is time to death (in days). The output is shown below.

Log-logistic regression — accelerated failure-time form

Log likelihood = −200.196		LR chi2(5) = 61.31 Prob > chi2 = 0.0000		
_t	Coef.	Std. Err.	z	p>|z|
tx	−.054087	.1863349	−0.29	0.772
perf	.0401825	.0046188	8.70	0.000
dd	.0042271	.0095831	0.44	0.659
age	.0086776	.0092693	0.94	0.349
priortx	.0032806	.0225789	0.15	0.884
_cons	1.347464	.6964462	1.93	0.053
/ln_gam	−.4831864	.0743015	−6.50	0.000
gamma	.6168149	.0458303		

8. State the AFT log-logistic model in terms of S(t) (note gamma = $1/p$).

9. Estimate the acceleration factor γ with a 95% confidence interval comparing the test and standard treatment (TX = 2 vs. TX = 1). Interpret your answer.

10. The AFT log-logistic model is also a proportional odds model. Use the output to estimate the odds ratio (odds of death) comparing the test and standard treatment. Also estimate the survival odds ratio comparing the test and standard treatment.

11. The Akaike Information Criterion (AIC) is a method designed to compare the fit of different models. For this question, three models are compared using the same 5 predictors:
    1. A Weibull model without frailty (presented as Model 1 in the section on frailty models);
    2. A Weibull model containing a frailty component (presented as Model 2 in the section on frailty models); and
    3. The log-logistic model presented above.

Below is a table containing the log likelihood statistic for each model.

Model	Frailty	Number of parameters	Log likelihood
1. Weibull	No	7	−206.204
2. Weibull	Yes	8	−200.193
3. Log-logistic	No	7	−200.196

The goal for this question is to calculate the AIC statistic for each model and select the model based on this criterion. **The AIC statistic is calculated as: −2 log likelihood + 2p** (where p is the number of parameters in the model). A smaller AIC statistic suggests a better fit. The addition of 2 times p can be thought of as a penalty if nonpredictive parameters are added to the model. Each model contains the 5 predictors, an intercept, and a shape parameter. Model 2 contains an additional variance parameter (theta) because a frailty component is included in the model. The log likelihood was unchanged when a frailty component was added to the log-logistic model (not shown in table).

Note that if we are just comparing Models 1 and 2 we could use the likelihood ratio test because Model 1 is nested (contained) in Model 2. The likelihood ratio test is considered a superior method to the AIC for comparing models but cannot be used to compare the log-logistic model to the other two, because that model uses a different distribution.

Which of the three models should be selected based on the AIC?

Questions 12 to 14 refer to a generalized gamma model using the Veterans' data with the same five predictor variables that were used in the model for Questions 8 to 10. The generalized gamma distribution contains two shape parameters (kappa and sigma) that allow great flexibility in the shape of the hazard. If kappa = 1, the model reduces to a Weibull distribution with p = 1/sigma. If kappa = 0 the model reduces to a lognormal distribution. The output is shown below.

Gamma regression — accelerated failure-time form

Log likelihood = −200.626		LR chi2(5) = 52.86 Prob > chi2 = 0.0000		
_t	Coef.	Std. Err.	z	p>|z|
tx	−.131	.1908	−0.69	0.491
perf	.039	.0051	7.77	0.000
dd	.0004	.0097	0.04	0.965
age	.008	.0095	0.89	0.376
priortx	.004	.0229	0.17	0.864
_cons	1.665	.7725	2.16	0.031
/ln_sig	.0859	.0654	1.31	0.189
/kappa	.2376	.2193	1.08	0.279
sigma	1.0898	.0714		

12. Estimate the acceleration factor γ with a 95% confidence interval comparing the test and standard treatment (TX = 2 vs. TX = 1).

13. Use the output to test the null hypothesis that a lognormal distribution is appropriate for this model.

14. A lognormal model was run with the same five predictors (output not shown) and yielded very similar parameter estimates to those obtained from the generalized gamma model shown above. The value of the log likelihood for the lognormal model was −201.210. Compare the AIC of the generalized gamma model, the lognormal model, and the log-logistic model from Question 11 and select a model based on that criterion. Note: each model contains an intercept and five predictors. The generalized gamma distribution contains two additional shape parameters and the log-logistic and lognormal distributions each contain one additional shape parameter (see Question 11 for further details on the AIC).

Questions 15 to 17 refer to a Weibull model using the remission data with treatment as the only predictor (coded TRT = 1 for the test treatment and TRT = 0 for the placebo). In this model both λ and p are modeled as functions of the predictor TRT. The model can be stated in terms of the hazard function: $h(t) = \lambda p t^{p-1}$ where $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT})$ and $p = \exp(\delta_0 + \delta_1 \text{ TRT})$. Typically, the shape parameter in a Weibull model is assumed constant (i.e., $\delta_1 = 0$) across levels of covariates. This model is discussed in the section of this chapter called "Other Parametric Models." The output obtained using Stata is shown below.

Weibull regression — log relative-hazard form

			LR chi2(1) = 1.69		
Log likelihood = −47.063396			Prob > chi2 = 0.1941		
	_t	Coef.	Std. Err.	z	p>|z|
_t					
	trt	−1.682	1.374	−1.22	0.221
	_cons	−3.083	.646	−4.77	0.000
ln_p					
	trt	−.012	.328	−0.04	0.970
	_cons	.315	.174	1.82	0.069

15. Even though λ is parameterized similarly to that in a PH Weibull model, this model is not a PH model because the shape parameter p varies across treatment groups. Show the PH assumption is violated in this model by estimating the hazard ratios for TRT = 0 vs. TRT = 1 after 10 weeks and after 20 weeks of follow-up.
16. Perform a statistical test on the hypothesis $\delta_1 = 0$ (the coefficient for the treatment term for $\ln(p)$). Note: if we assume $\delta_1 = 0$, then the model reduces to the example of the Weibull PH model presented in Section VI of this chapter.
17. Consider the plot of the log negative log of the Kaplan–Meier survival estimates against the log of time for TRT = 1 and TRT = 0. How should the graph look if $\delta_1 = 0$?

Test

Answer the following true or false questions (circle T or F).

T F 1. The accelerated failure time model and proportional hazards model are both additive models.

T F 2. If the survival function is known then the hazard function can be ascertained (and vice versa).

T F 3. If survival time follows a Weibull distribution then a plot of the ln[−ln S(t)] against ln(t) should be a straight line.

T F 4. If the acceleration failure time (AFT) assumption holds in a log-logistic model then the proportional hazards assumption also holds.

T F 5. If the acceleration factor for the effect of an exposure (exposed vs. unexposed) is greater than one, then the exposure is harmful to survival.

T F 6. Let $S_0(t)$ be the survival function for unexposed subjects (E = 0) and let $S_1(t)$ be the survival function for exposed subjects (E = 1). If γ is the acceleration factor comparing E = 1 vs. E = 0 then $S_0(t) = S_1(\gamma t)$.

T F 7. Frailty models are designed to provide an approach to account for unobserved individual-level characteristics.

T F 8. If you include a gamma distributed frailty component to the model, then you will see an additional parameter estimate for the variance of the frailty in the model output.

T F 9. If survival time T follows a Weibull distribution, then ln(T) also follows a Weibull distribution.

T F 10. If a subject is lost to follow-up after 5 years, then the subject is left-censored.

Questions 11 to 17 refer to a Weibull model run with the "addicts" dataset. The predictor of interest is CLINIC (coded 1 or 2) for two methadone clinics for heroin addicts. Covariates include DOSE (continuous) for methadone dose (mg/day), PRISON (coded 1 if patient has a prison record and 0 if not), and a prison–dose product term (called PRISDOSE). The outcome is time (in days) until the person dropped out of the clinic or was censored. The Weibull survival and hazard functions are, respectively, $S(t) = \exp(-\lambda t^p)$ and $h(t) = \lambda p t^{p-1}$ where $\lambda^{1/p} = \exp[-(\alpha_0 + \alpha_1 \text{CLINIC} + \alpha_2 \text{PRISON} + \alpha_3 \text{DOSE} + \alpha_4 \text{PRISDOSE})]$ for the AFT parameterization and $\lambda = \exp[\beta_0 + \beta_1 \text{CLINIC} + \beta_2 \text{PRISON} + \beta_3 \text{DOSE} + \beta_4 \text{PRISDOSE}]$ for the PH parameterization. The Stata output for both the AFT and PH forms of the model are presented as follows:

Weibull regression
accelerated failure-time form

Log likelihood = −260.74854

_t	Coef.	Std. Err.	z	P > |z|
clinic	.698	.158	4.42	0.000
prison	.145	.558	0.26	0.795
dose	.027	.006	4.60	0.000
prisdose	−.006	.009	−0.69	0.492
_cons	3.977	.376	10.58	0.000
/ln_p	.315	.068	4.67	0.000
p	1.370467			
1/p	.729678			

Weibull regression
log relative-hazard form

Log likelihood = −260.74854

_t	Coef.	Std. Err.	z	P > |z|
clinic	−.957	.213	−4.49	0.000
prison	−.198	.765	−0.26	0.795
dose	−.037	.008	−4.63	0.000
prisdose	.009	.013	0.69	0.491
_cons	−5.450	.702	−7.76	0.000
/ln.p	.315	.068	4.67	0.000
p	1.370467			
1/p	.729678			

11. Estimate the acceleration factor with a 95% confidence interval comparing CLINIC = 2 vs. CLINIC = 1. Interpret this result.
12. Estimate the hazard ratio with a 95% confidence interval comparing CLINIC = 2 vs. CLINIC = 1. Interpret this result.
13. Estimate the coefficient for CLINIC in the PH Weibull model using the results reported in the output from the AFT form of the model. Hint: the coefficients for a Weibull PH and AFT model are related $\beta_j = -\alpha_j p$ for the jth covariate.
14. Is the product term PRISDOSE included in the model to account for potential interaction or potential confounding of the effect of CLINIC on survival?
15. Use the output to estimate the median survival time for a patient from CLINIC = 2 who has a prison record and receives a methadone dose of 50 mg/day. Hint: use the relationship that $t = [-\ln S(t)]^{1/p} \times (1/\lambda^{1/p})$ for a Weibull model.
16. Use the output to estimate the median survival time for a patient from CLINIC = 1 who has a prison record and receives a methadone dose of 50 mg/day.
17. What is the ratio of your answers from Questions 15 and 16 and how does this ratio relate to the acceleration factor?

Questions 18 and 19 refer to the Weibull model (in AFT form) that was used for the previous set of questions (Questions 11 to 17). The only difference is that a frailty component is now included in the model. A gamma distribution of mean 1 and variance theta is assumed for the frailty. The output shown on in the following contains one additional parameter estimate (for theta).

Weibull regression
accelerated failure-time form
Gamma frailty

Log likelihood = −260.74854

_t	Coef.	Std. Err.	z	P > |z|
clinic	.698	.158	4.42	0.000
prison	.145	.558	0.26	0.795
dose	.027	.006	4.60	0.000
prisdose	−.006	.009	−0.69	0.492
_cons	3.977	.376	10.58	0.000
/ln_p	.315	.068	4.67	0.000
p	1.370467			
1/p	.729678			
theta	.00000002		.0000262	

Likelihood ratio test of theta=0:
chibar2(01) = 0.00
Prob>=chibar2 = 1.000

18. Did the addition of the frailty component change any of the other parameter estimates (besides theta)? Did it change the log likelihood?
19. A likelihood ratio test for the hypothesis H_0: theta = 0 yields a p-value of 1.0 (bottom of the output). The parameter estimate for theta is essentially zero. What does it mean if theta = 0?

Answers to Practice Exercises

1. T
2. F: The median survival time for the unexposed is 1/3 of the median survival time for the exposed.
3. F: The Cox model is a semiparametric model. The distribution of survival time is unspecified in a Cox model.
4. T
5. F: The hazard is assumed constant in an exponential model.
6. $t = [-\log(S(t)] \times (1/\lambda)$, where $S(t) = 0.5$, and $1/\lambda = \exp(\alpha_0 + \alpha_1 \text{ TRT})$.
   For TRT = 0: estimated median survival = $[-\ln(0.5)] \exp(2.159) = 6.0$ weeks.
   For TRT = 1: estimated median survival = $[-\ln(0.5)] \exp(2.159 + 1.527) = 27.6$ weeks.
7. $t = [-\log(S(t)] (1/\lambda)$, where $S(t) = 0.5$, and $\lambda = \exp(\beta_0 + \beta_1 \text{ TRT}) \Rightarrow 1/\lambda = \exp[-(\beta_0 + \beta_1 \text{TRT})]$.
   For TRT = 0: estimated median survival = $[-\ln(0.5)] \exp[-(-2.159)] = 6.0$ weeks.
   For TRT = 1: estimated median survival = $[-\ln(0.5)] \exp[-(-2.159 - 1.527)] = 27.6$ weeks.
8. $S(t) = 1/(1 + \lambda t^p)$ where $\lambda^{1/p} = \exp[-(\alpha_0 + \alpha_1 \text{ TX} + \alpha_2 \text{ PERF} + \alpha_3 \text{DD} + \alpha_4 \text{AGE} + \alpha_5 \text{PRIORTX})]$.
9. $$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2 PERF + \alpha_3 DD + \alpha_4 AGE + \alpha_5 PRIORTX]}{\exp[\ \alpha_0 + \ \alpha_1(1) + \ \alpha_2 PERF + \ \alpha_3 DD + \ \alpha_4 AGE + \ \alpha_5 PRIORTX]}$$
   $$= \exp(\ \alpha_1)$$
   $$\hat{\gamma} = \exp(-0.054087) = 0.95$$
   $$95\%\ \text{CI} = \exp[-0.054087 \pm 1.96(0.1863349)] = (0.66,\ 1.36)$$
   The point estimate along with the 95% CI suggests a null result.
10. The coefficients for a log-logistic proportional odds (PO) and AFT model are related $\beta_1 = -\alpha_1\ p = -\beta_1/\text{gamma}$, where β_1 is the coefficient for TX in a PO model.
    OR = $\exp(-\alpha_1/\text{gamma})$
    estimated OR = $\exp(-0.054087/0.6168149) = 0.92$
    estimated survival OR = $1/[\exp(-0.054087/0.6168149)] = 1.09$.

11. The AIC statistic is calculated as -2 log likelihood $+2$p (where p is the number of parameters in the model). A smaller AIC statistic suggests a better fit. The AIC statistic is shown below for each of the three models.

Model	Frailty	Number of parameters	Log likelihood	AIC
1. Weibull	No	7	−206.204	426.408
2. Weibull	Yes	8	−200.193	416.386
3. Log-logistic	No	7	−200.196	414.392

Based on the AIC, the log-logistic model is selected yielding the smallest AIC statistic at 414.392.

12. $$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2 PERF + \alpha_3 DD + \alpha_4 AGE + \alpha_5 PRIORTX]}{\exp[\ \alpha_0 + \ \alpha_1(1) + \ \alpha_2 PERF + \ \alpha_3 DD + \ \alpha_4 AGE + \ \alpha_5 PRIORTX]}$$
$$= \exp(\ \alpha_1)$$
$$\hat{\gamma} = \exp(-0.131) = 0.88$$
$$95\%\ \text{CI} = \exp[(-0.131 \pm 1.96(0.1908)] = (0.60,\ 1.28)$$

13. The generalized gamma distribution reduces to a lognormal distribution if kappa $= 0$.

$H_0:$ kappa $= 0$

Wald test statistic: $z = \dfrac{0.2376}{0.2193} = 1.08$ (from output)

p-value: 0.279 (from output)

Conclusion: p-value not significant at a significance level of 0.05. Not enough evidence to reject H_0. The lognormal distribution may be appropriate.

14. The AIC statistic is shown below for the generalized gamma, lognormal, and log-logistic models.

Model	Number of parameters	Log likelihood	AIC
Generalized Gamma	8	−200.626	417.252
Lognormal	7	−201.210	416.420
Log-logistic	7	200.196	414.392

As in Question 11, the log-logistic model is selected yielding the smallest AIC at 414.392.

15. $h(t) = \lambda p t^{p-1}$ where $\lambda = \exp(\beta_0 + \beta_1 TRT)$ and $p = \exp(\delta_0 + \delta_1 TRT)$
let $\lambda_0 = \exp[\beta_0 + \beta_1(0)]$, $\lambda_1 = \exp[\beta_0 + \beta_1(1)]$let $p_0 = \exp[\delta_0 + \delta_1(0)]$, $p_1 = \exp[\delta_0 + \delta_1(1)]\hat{\lambda}_0 = 0.0458$, $\hat{\lambda}_1 = 0.0085$, $\hat{p}_0 = 1.3703$, $\hat{p}_1 = 1.3539$ (calculated using output)

$$\text{HR (TRT} = 0 \text{ vs. TRT} = 1) = \frac{\lambda_0 p_0 t^{p_0 - 1}}{\lambda_1 p_1 t^{p_1 - 1}}$$

$$\widehat{HR}(\text{as a function of t}) = \frac{(0.0458)(1.3703)t^{0.3703}}{(0.0085)(1.3539)t^{0.3539}}$$

$$\widehat{HR}(\text{t} = 10) = \frac{(0.0458)(1.3703)(10^{0.3703})}{(0.0085)(1.3539)(10^{0.3539})} = 5.66$$

$$\widehat{HR}(\text{t} = 20) = \frac{(0.0458)(1.3703)(20^{0.3703})}{(0.0085)(1.3539)(20^{0.3539})} = 5.73$$

The estimated hazard ratios for RX at 10 weeks and at 20 weeks are different, demonstrating that the hazards are not constrained to be proportional in this model. However, the estimated hazard ratios are just slightly different, suggesting that the PH assumption is probably reasonable.

16. H_0: $\delta_1 = 0$

$$\text{Wald test statistic : } z = \frac{-0.0123083}{0.328174} = -0.04 \text{ (from output)}$$

p-value: 0.970 (from output)

Conclusion: p-value is not significant. No evidence to reject H_0. The PH assumption is reasonable.

17. If the Weibull assumption is met, then the plots should be straight lines with slope p. If $\delta_1 = 0$, then the slope p is the same for TRT = 1 and TRT = 0 and the lines are parallel.

8 Recurrent Event Survival Analysis

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_8,
© Springer Science+Business Media, LLC 2012

Introduction

This chapter considers outcome events that may occur more than once over the follow-up time for a given subject. Such events are called "recurrent events." Modeling this type of data can be carried out using a Cox PH model with the data layout constructed so that each subject has a line of data corresponding to each recurrent event. A variation of this approach uses a stratified Cox PH model, which stratifies on the order in which recurrent events occur. Regardless of which approach is used, the researcher should consider adjusting the variances of estimated model coefficients for the likely correlation among recurrent events on the same subject. Such adjusted variance estimates are called "robust variance estimates." A parametric approach for analyzing recurrent event data that includes a frailty component (introduced in Chapter 7) is also described and illustrated.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Overview** (page 366)
II. **Examples of Recurrent Event Data** (pages 366–368)
III. **Counting Process Example** (pages 368–369)
IV. **General Data Layout: Counting Process Approach** (pages 370–371)
V. **The Counting Process Model and Method** (pages 372–376)
VI. **Robust Estimation** (pages 376–378)
VII. **Results for CP Example** (pages 378–379)
VIII. **Other Approaches Stratified Cox** (pages 379–385)
IX. **Bladder Cancer Study Example (Continued)** (pages 385–389)
X. **A Parametric Approach Using Shared Frailty** (pages 389–391)
XI. **A Second Example** (pages 391–395)
XII. **Survival Curves with Recurrent Events** (pages 395–398)
XIII. **Summary** (pages 398–401)

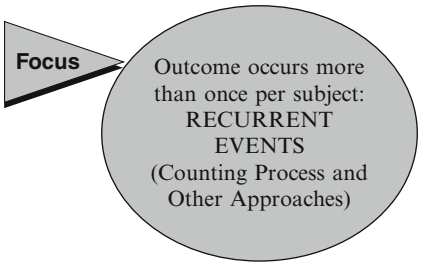
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize examples of recurrent event data.
2. State or recognize the form of the data layout used for the counting process approach for analyzing correlated data.
3. Given recurrent event data, outline the steps needed to analyze such data using the counting process approach.
4. State or recognize the form of the data layout used for the Stratified Cox (SC) approaches for analyzing correlated data.
5. Given recurrent event data, outline the steps needed to analyze such data using the SC approaches.

Presentation

I. Overview



In this chapter we consider outcome events that may occur more than once over the follow-up time for a given subject. Such events are called "recurrent events." We focus on the **Counting Process (CP)** approach for analysis of such data that uses the Cox PH model, but we also describe alternative approaches that use a Stratified Cox (SC) PH model and a frailty model.

II. Examples of Recurrent Event Data

1. Multiple relapses from remission – leukemia patients
2. Repeated heart attacks – coronary patients
3. Recurrence of tumors – bladder cancer patients
4. Deteriorating episodes of visual acuity – macular degeneration patients

Up to this point, we have assumed that the event of interest can occur only once for a given subject. However, in many research scenarios in which the event of interest is not death, a subject may experience an event several times over follow-up. Examples of recurrent event data include:

1. Multiple episodes of relapses from remission comparing different treatments for leukemia patients.
2. Recurrent heart attacks of coronary patients being treated for heart disease.
3. Recurrence of bladder cancer tumors in a cohort of patients randomized to one of two treatment groups.
4. Multiple events of deteriorating visual acuity in patients with baseline macular degeneration, where each recurrent event is considered a more clinically advanced stage of a previous event.

Objective

Assess relationship of predictors to rate of occurrence, allowing for multiple events per subject

For each of the above examples, the event of interest differs, but nevertheless may occur more than once per subject. A logical objective for such data is to assess the relationship of relevant predictors to the rate in which events are occurring, allowing for multiple events per subject.

LEUKEMIA EXAMPLE

Do treatment groups differ in rates of relapse from remission?

In the leukemia example above, we might ask whether persons in one treatment group are experiencing relapse episodes at a higher rate than persons in a different treatment group.

HEART ATTACK EXAMPLE

Do smokers have a higher heart attack rate than nonsmokers?

If the recurrent event is a heart attack, we might ask, for example, whether smokers are experiencing heart attack episodes at a higher rate than nonsmokers.

LEUKEMIA AND HEART ATTACK EXAMPLES

All events are of the same type
The order of events is not important
Heart attacks: Treat as identical events;
Don't distinguish among 1st, 2nd, 3rd, etc. attack

For either of the above two examples, we are treating all events as if they were the same type. That is, the occurrence of an event on a given subject identifies the same disease without considering more specific qualifiers such as severity or stage of disease. We also are not taking into account the order in which the events occurred.

For example, we may wish to treat all heart attacks, whether on the same or different subjects, as identical types of events, and we don't wish to identify whether a given heart attack episode was the first, or the second, or the third event that occurred on a given subject.

BLADDER CANCER EXAMPLE

Compare overall tumor recurrence rate without considering order or type of tumor

The third example, which considers recurrence of bladder cancer tumors, can be considered similarly. That is, we may be interested in assessing the "overall" tumor recurrence rate without distinguishing either the order or type of tumor.

MACULAR DEGENERATION OF VISUAL ACUITY EXAMPLE

A second or higher event is more severe than its preceding event

Order of event is important

The fourth example, dealing with macular degeneration events, however, differs from the other examples. The recurrent events on the same subject differ in that a second or higher event indicates a more severe degenerative condition than its preceding event.

Consequently, the investigator in this scenario may wish to do separate analyses for each ordered event in addition to or instead of treating all recurrent events as identical.

Use a different analysis depending on whether

a. recurrent events are treated as identical
b. recurrent events involve different disease categories and/or the order of events is important

We have thus made an important distinction to be considered in the analysis of recurrent event data. If all recurrent events on the same subject are treated as identical, then the analysis required of such data is different than what is required if either recurrent events involve different disease categories and/or the order that events reoccur is considered important.

Recurrent events identical
⇓
Counting Process Approach (Andersen et al., 1993)

The approach to analysis typically used when recurrent events are treated as identical is called the **Counting Process Approach** (Andersen et al., 1993).

Recurrent events: different disease categories or event order important
⇓
Stratified Cox (SC) Model Approaches

When recurrent events involve different disease categories and/or the order of events is considered important, a number of alternative approaches to analysis have been proposed that involve using stratified Cox (SC) models.

In this chapter, we focus on the **Counting Process (CP)** approach, but also describe the other stratified Cox approaches (in a later section).

III. Counting Process Example

To illustrate the counting process approach, we consider data on two hypothetical subjects (Table 8.1), Al and Hal, from a randomized trial that compares two treatments for bladder cancer tumors.

Table 8.1. 2 Hypothetical Subjects Bladder Cancer Tumor Events

	Time interval	Event indicator	Treatment group
Al	0 to 3	1	1
	3 to 9	1	1
	9 to 21	1	1
	21 to 23	0	1
Hal	0 to 3	1	0
	3 to 15	1	0
	15 to 25	1	0

Al gets recurrent bladder cancer tumors at months 3,9, and 21, and is without a bladder cancer tumor at month 23, after which he is no longer followed. Al received the treatment coded as 1.

Hal gets recurrent bladder cancer tumors at months 3, 15, and 25, after which he is no longer followed. Hal received the treatment coded as 0.

	Al	Hal
No. recurrent events	3	3
Follow-up time	23 months	25 months
Event times from start of follow-up	3, 9, 21	3, 15, 25
Additional months of follow-up after last event	2 months	0 months

Al has experienced 3 events of the same type (i.e., recurrent bladder tumors) over a follow-up period of 23 months. Hal has also experienced 3 events of the same type over a follow-up period of 25 months.

The three events experienced by Al occurred at different survival times (from the start of initial follow-up) from the three events experienced by Hal.

Also, Al had an additional 2 months of follow-up after his last recurrent event during which time no additional event occurred. In contrast, Hal had no additional event-free follow-up time after his last recurrent event.

Table 8.2. Example of Data Layout for Counting Process Approach

Subj	Interval Number	Time Start	Time Stop	Event Status	Treatment Group
Al	1	0	3	1	1
Al	2	3	9	1	1
Al	3	9	21	1	1
Al	4	21	23	0	1
Hal	1	0	3	1	0
Hal	2	3	15	1	0
Hal	3	15	25	1	0

In Table 8.2, we show for these 2 subjects, how the data would be set up for computer analyses using the counting process approach. Each subject contributes a line of data for each time interval corresponding to each recurrent event and any additional event-free follow-up interval. We previously introduced this format as the Counting Process (CP) data layout in section VI of Chapter 1.

Counting process: **Start** and **Stop** times
Standard layout: only **Stop (survival)** times (no recurrent events)

A distinguishing feature of the data layout for the counting process approach is that each line of data for a given subject lists the **start time** and **stop time** for each interval of follow-up. This contrasts with the standard layout for data with no recurrent events, which lists only the stop (**survival**) time.

Subj	Interval Number	Time Start	Time Stop	Event Status	Treatment Group
Sal	1	0	17	1	0
Mal	1	0	12	0	1

Note that if a third subject, Sal, failed without further events or follow-up occurring, then Sal contributes only one line of data, as shown at the left. Similarly, only one line of data is contributed by a (fourth) subject, Mal, who was censored without having failed at any time during follow-up.

IV. General Data Layout: Counting Process Approach

- N subjects
- r_i time intervals for subject i
- d_{ij} event staus (0 or 1) for subject i in interval j
- t_{ij0} start time for subject i in interval j
- t_{ij1} stop time for subject i in interval j
- X_{ijk} value of kth predictor for subject i in interval j
- $i = 1, 2, \ldots, N;\ j = 1, 2, \ldots, n_i;\ k = 1, 2, \ldots, p$

The general data layout for the counting process approach is presented in Table 8.3 for a dataset involving N subjects.

The ith subject has r_i recurrent events. d_{ij} denotes the event status (1 = failure, 0 = censored) for the ith subject in the jth time interval. t_{ij0} and t_{ij1} denote the start and stop times, respectively, for the ith subject in the jth interval. X_{ijk} denotes the value of the kth predictor for the ith subject in the jth interval.

Table 8.3. General Data Layout: CP Approach

Subject	Interval	Status	Start	Stop	Predictors
i	**j**	$\mathbf{d_{ij}}$	$\mathbf{t_{ij0}}$	$\mathbf{t_{ij1}}$	$\mathbf{X_{ij1} \ldots X_{ijp}}$
1	1	d_{11}	t_{110}	t_{111}	$X_{111} \ldots X_{11p}$
1	2	d_{12}	t_{120}	t_{121}	$X_{121} \ldots X_{12p}$
.	.	.	.	.	. .
.	.	.	.	.	. .
.	.	.	.	.	. .
1	r_1	d_{1r_1}	t_{1r_10}	t_{1r_11}	$X_{1r_11} \ldots X_{1r_1p}$
.	.	.	.	.	. .
i	1	d_{i1}	t_{i10}	t_{i11}	$X_{i11} \ldots X_{i1p}$
i	2	d_{i2}	t_{i20}	t_{i21}	$X_{i21} \ldots X_{i2p}$
.	.	.	.	.	. .
.	.	.	.	.	. .
.	.	.	.	.	. .
i	r_i	d_{ir_i}	t_{ir_i0}	t_{ir_i1}	$X_{ir_i1} \ldots X_{ir_ip}$
.	.	.	.	.	. .
N	1	d_{N1}	t_{N10}	t_{N11}	$X_{N11} \ldots X_{N1p}$
N	2	d_{N2}	t_{N20}	t_{N21}	$X_{N21} \ldots X_{N2p}$
.	.	.	.	.	. .
.	.	.	.	.	. .
.	.	.	.	.	. .
N	r_N	d_{Nr_N}	t_{Nr_N0}	t_{Nr_N1}	$X_{Nr_N1} \ldots X_{Nr_Np}$

Subjects are not restricted to have the same number of time intervals (e.g., r_1 does not have to equal r_2) or the same number of recurrent events. If the last time interval for a given subject ends in censorship ($d_{ij} = 0$), then the number of recurrent events for this subject is $r_i - 1$; previous time intervals, however, usually end with a failure ($d_{ij} = 1$).

Also, start and stop times may be different for different subjects. (See the previous section's example involving two subjects.)

As with any survival data, the covariates (i.e., Xs) may be time-independent or time-dependent for a given subject. For example, if one of the Xs is "gender" (1 = female, 0 = male), the values of this variable will be all 1s or all 0s over all time intervals observed for a given subject. If another X variable is, say, a measure of daily stress level, the values of this variable are likely to vary over the time intervals for a given subject.

The second column ("Interval j") in the data layout is not needed for the **CP** analysis, but is required for other approaches described later.

Table 8.4 First 26 Subjects: Bladder Cancer Study

id	int	event	start	stop	tx	num	size
1	1	0	0	0	0	1	1
2	1	0	0	1	0	1	3
3	1	0	0	4	0	2	1
4	1	0	0	7	0	1	1
5	1	0	0	10	0	5	1
6	1	1	0	6	0	4	1
6	2	0	6	10	0	4	1
7	1	0	0	14	0	1	1
8	1	0	0	18	0	1	1
9	1	1	0	5	0	1	3
9	2	0	5	18	0	1	3
10	1	1	0	12	0	1	1
10	2	1	12	16	0	1	1
10	3	0	16	18	0	1	1
11	1	0	0	23	0	3	3
12	1	1	0	10	0	1	3
12	2	1	10	15	0	1	3
12	3	0	15	23	0	1	3
13	1	1	0	3	0	1	1
13	2	1	3	16	0	1	1
13	3	1	16	23	0	1	1
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1
15	1	1	0	7	0	2	3
15	2	1	7	10	0	2	3
15	3	1	10	16	0	2	3
15	4	1	16	24	0	2	3
16	1	1	0	3	0	1	1
16	2	1	3	15	0	1	1
16	3	1	15	25	0	1	1
17	1	0	0	26	0	1	2
18	1	1	0	1	0	8	1
18	2	0	1	26	0	8	1
19	1	1	0	2	0	1	4
19	2	1	2	26	0	1	4
20	1	1	0	25	0	1	2
20	2	0	25	28	0	1	2
21	1	0	0	29	0	1	4
22	1	0	0	29	0	1	2
23	1	0	0	29	0	4	1
24	1	1	0	28	0	1	6
24	2	1	28	30	0	1	6
25	1	1	0	2	0	1	5
25	2	1	2	17	0	1	5
25	3	1	17	22	0	1	5
25	4	0	22	30	0	1	5
26	1	1	0	3	0	2	1
26	2	1	3	6	0	2	1
26	3	1	6	8	0	2	1
26	4	1	8	12	0	2	1
26	5	0	12	30	0	2	1

To illustrate the above general data layout, we present in Table 8.4 the data for the first 26 subjects from a study of recurrent bladder cancer tumors (Byar, 1980 and Wei, Lin, and Weissfeld, 1989). The entire dataset contained 86 patients, each followed for a variable amount of time up to 64 months.

The repeated event being analyzed is the recurrence of bladder cancer tumors after transurethral surgical excision. Each recurrence of new tumors was treated by removal at each examination.

About 25% of the 86 subjects experienced four events.

The exposure variable of interest is drug treatment status (**tx**, 0 = placebo, 1 = treatment with thiotepa). The covariates listed here are initial number of tumors (**num**) and initial size of tumors (**size**) in centimeters. The paper by Wei, Lin, and Weissfeld actually focuses on a different method of analysis (called "marginal"), which requires a different data layout than shown here. We later describe the "marginal" approach and its corresponding layout.

In these data, it can be seen that 16 of these subjects (id #s 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 17, 18, 20, 21, 22, 23) had no recurrent events, 4 subjects had 2 recurrent events (id #s 10, 12, 19, 24), 4 subjects (id #s 13, 14, 16, 25) had 3 recurrent events, and 2 subjects (id #s 15, 26) had 4 recurrent events.

Moreover, 9 subjects (id #s 6, 9, 10, 12, 14, 18, 20, 25, 26) were observed for an additional event-free time interval after their last event. Of these, 4 subjects (id #s 6, 9, 18, 20) experienced only one event (i.e., no recurrent events).

V. The Counting Process Model and Method

The model typically used to carry out the **Counting Process** approach is the standard Cox PH model, once again shown here at the left.

Cox PH Model

$h(t, \mathbf{X}) = h_0(t)\exp[\Sigma\beta_i X_i]$

Need to

Assess PH assumption for X_i
Consider stratified Cox or extended Cox if PH assumption not satisfied
Use extended Cox for time-dependent variables

As usual, the PH assumption needs to be evaluated for any time-independent variable. A stratified Cox model or an extended Cox model would need to be used if one or more time-independent variables did not satisfy the PH assumption. Also, an extended Cox model would be required if inherently time-dependent variables were considered.

The primary difference in the way the Cox model is used for analyzing recurrent event data versus nonrecurrent (one time interval per subject) data is the way several time intervals on the same subject are treated in the formation of the likelihood function maximized for the Cox model used.

Recurrent event data (Likelihood function formed differently)	Nonrecurrent event data
Subjects with > 1 time interval remain in the risk set until last interval is completed	Subjects removed from risk set at time of failure or censorship
Different lines of data are treated as **independent** even though several outcomes on the **same** subject	Different lines of data are treated as **independent** because they come from **different** subjects

To keep things simple, we assume that the data involve only time-independent variables satisfying the PH assumption. For recurrent survival data, a subject with more than one time interval remains in the risk set until his or her last interval, after which the subject is removed from the risk set. In contrast, for nonrecurrent event data, each subject is removed from the risk set at the time of failure or censorship.

Nevertheless, for subjects with two or more intervals, the different lines of data contributed by the same subject are treated in the analysis as if they were independent contributions from different subjects, even though there are several outcomes on the **same** subject.

In contrast, for the standard Cox PH model approach for nonrecurrent survival data, different lines of data are treated as independent because they come from **different** subjects.

Cox PH Model for CP Approach: Bladder Cancer Study

$$h(t, \mathbf{X}) = h_0(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{num} + \gamma_2\ \mathbf{size}]$$

where

tx = 1 if thiotepa, 0 if placebo
num = initial # of tumors
size = initial size of tumors

No-interaction Model

Interaction model would involve product terms

tx × **num** and/or **tx** × **size**

For the bladder cancer study described in Table 8.4, the basic Cox PH model fit to these data takes the form shown at the left.

The primary (exposure) variable of interest in this model is the treatment variable **tx**. The variables **num** and **size** are considered as potential con-founders. All three variables are time-independent variables.

This is a no-interaction model because it does not contain product terms of the form **tx** × **num** or **tx** × **size**. An interaction model involving such product terms could also be considered, but we only present the no-interaction model for illustrative purposes.

Table 8.5. Ordered Failure Time and Risk Set Information for First 26 Subjects in Bladder Cancer Study

Ordered failure times $t_{(f)}$	# in risk set n_f	# failed m_f	# censored in $[t_{(f)}, t_{(f+1)})$	Subject ID #s for outcomes in $[t_{(f)}, t_{(f+1)})$
0	**26**	–	1	1
1	25	1	1	2, 18
2	24	2	0	19, 25
3	24	4	1	3, 13, 14, 16, 26
5	23	1	0	9
6	23	2	0	6, 26
7	23	1	1	4, 15
8	22	1	0	26
9	22	1	0	14
10	22	2	2	5, 6, 12, 15
12	20	2	1	7, 10, 26
15	19	2	0	12, 16
16	19	3	0	10, 13, 15
17	19	1	3	8, 9, 10, 25
21	16	1	0	14
22	16	1	0	25
23	16	1	3	11, 12, 13, 14
24	12	1	0	15
25	11	2	0	16, 20
26	10	1	2	17, 18, 19
28	7	1	4	20, 21, 22, 23, 24
30	3	1	2	24, 25, 26
		32	21	

Table 8.5 at the left provides ordered failure times and corresponding risk set information that would result if the first 26 subjects that we described in Table 8.4 comprised the entire dataset. (Recall that there are 86 subjects in the complete dataset.)

Because we consider 26 subjects, the number in the risk set at ordered failure time $t_{(0)}$ is $n_0 = 26$. As these subjects fail (i.e., develop a bladder cancer tumor) or are censored over follow-up, the number in the risk set will decrease from the fth to the f + 1th ordered failure time provided that no subject who fails at time $t_{(f)}$ either has a recurrent event at a later time or has additional follow-up time until later censorship. In other words, *a subject who has additional follow-up time after having failed at* $t_{(f)}$ *does not drop out of the risk set after* $t_{(f)}$.

Table 8.6. Focus on Subject #s **19** and **25** from Table 8.5

$t_{(f)}$	$n_{(f)}$	$m_{(f)}$	$q_{(f)}$	Subject ID #s
0	26	–	1	1
1	25	1	1	2, 18
2	**24**	2	0	**19, 25**
3	**24**	4	1	3, 13, 14, 16, 26
			•	
			•	
			•	
17	19	1	3	8, 9, 10, **25**
21	16	1	0	14
22	16	1	0	**25**
23	16	1	3	11, 12, 13, 14
24	12	1	0	15
25	11	2	0	16, 20
26	10	1	2	17, 18, **19**
28	7	1	4	20, 21, 22, 23, 24
30	3	1	2	24, **25**, 26

For example, at month $t_{(f)} = 2$, subject #s **19** and **25** fail, but the number in the risk set at that time ($n_f = 24$) does not decrease (by 2) going into the next failure time because each of these subjects has later recurrent events. In particular, subject #**19** has a recurrent event at month $t_{(f)} = 26$ and subject #**25** has two recurrent events at months $t_{(f)} = 17$ and $t_{(f)} = 22$ and has additional follow-up time until censored month 30.

Table 8.7. Focus on Subject #s **3, 13, 14, 16, 26** from Table 8.5

$t_{(f)}$	$n_{(f)}$	$m_{(f)}$	$q_{(f)}$	Subject ID #s
0	26	–	1	1
1	25	1	1	2, 18
2	24	2	0	19, 25
3	**24**	4	1	**3, 13, 14, 16, 26**
5	**23**	1	0	9
6	23	2	0	6, **26**
7	23	1	1	4, 15
8	22	1	0	**26**
9	22	1	0	**14**
10	22	2	2	5, 6, 12, 15
12	20	2	1	7, 10, **26**
15	19	2	0	12, **16**
16	19	3	0	10, **13**, 15
17	19	1	3	8, 9, 10, 25
21	16	1	0	14
22	16	1	0	25
23	16	1	3	11, 12, **13, 14**
24	12	1	0	15
25	11	2	0	**16**, 20
26	10	1	2	17, 18, 19
28	7	1	4	20, 21, 22, 23, 24
30	3	1	2	24, 25, **26**

As another example from Table 8.5, subject #s **3, 13, 14, 16, 26** contribute information at ordered failure time $t_{(f)} = 3$, but the number in the risk set only drops from 24 to 23 even though the last four of these subjects all fail at $t_{(f)} = 3$. Subject #**3** is censored at month 4 (see Table 8.4), so this subject is removed from the risk set after failure time $t_{(f)} = 3$. However, subjects **13, 14, 16,** and **26** all have recurrent events after $t_{(f)} = 3$, so they are not removed from the risk set after $t_{(f)} = 3$.

Subject #26 appears in the last column 5 times. This subject contributes 5 (start, stop) time intervals, fails at months 3, 6, 8, and 12, and is also followed until month 30, when he is censored.

"Gaps" in follow-up time:

0	10	**gap**	25	50
	lost		re-enter	

Another situation, which is not illustrated in these data, involves "gaps" in a subject's follow-up time. A subject may leave the risk set (e.g., lost to follow-up) at, say, time = 10 and then re-enter the risk set again and be followed from, say, time = 25 to time = 50. This subject has a follow-up gap during the period from time = 10 to time = 25.

No Interaction Cox PH Model

$$h(t,\mathbf{X}) = h_0(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{num} + \gamma_2\ \mathbf{size}]$$

The (partial) likelihood function (**L**) used to fit the no-interaction Cox PH model is expressed in typical fashion as the product of individual likelihoods contributed by each ordered failure time and corresponding risk set information in Table 8.5. There are 22 such terms in this product because there are 22 ordered failure times listed in Table 8.5.

Partial likelihood function:

$$\mathbf{L} = \mathbf{L_1} \times L_2 \times \cdots \times \mathbf{L_{22}}$$

$\mathbf{L}_f$ = individual likelihood at $t_{(j)}$
= Pr[failing at $t_{(f)}$ | survival up to $t_{(f)}$]
$f = 1, 2, \ldots, 22$

Each individual likelihood L_f essentially gives the conditional probability of failing at time $t_{(f)}$, given survival (i.e., remaining in the risk set) at $t_{(f)}$.

$$\mathbf{L}_f = \frac{\mathbf{exp}\left(\beta tx_{(f)} + \gamma_1 num_{(f)} + \gamma_2 size_{(f)}\right)}{\sum_{s\ \text{in}\ R(t_{(f)})} \mathbf{exp}\left(\beta tx_{s(f)} + \gamma_1 num_{s(f)} + \gamma_2 size_{s(f)}\right)}$$

If there is only one failure at the jth ordered failure time, L_f is expressed as shown at the left for the above no-interaction model. In this expression $tx_{(f)}$, $num_{(f)}$, and $size_{(f)}$ denote the values of the variables **tx, num,** and **size** for the subject failing at month $t_{(f)}$.

$tx_{(f)}$, $num_{(f)}$, and $size_{(f)}$ values of **tx, num**, and **size** at $t_{(f)}$

The terms $tx_{s(f)}$, $num_{s(f)}$, and $size_{s(f)}$ denote the values of the variables **tx, num,** and **size** for the subject s in the risk set $R(t_{(f)})$. Recall that $R(t_{(f)})$ consists of all subjects remaining at risk at failure time $t_{(f)}$.

$tx_{s(f)}$, $num_{s(f)}$, and $size_{s(f)}$ values of **tx, num**, and **size** for subject s in $R(t_{(f)})$

Data for Subject #25

id	int	event	start	stop	tx	num	size
25	1	1	0	2	0	1	5
25	2	1	2	17	0	1	5
25	3	1	17	**22**	0	1	5
25	4	0	22	30	0	1	5

For example, subject #25 from Table 8.4 failed for the third time at month 22, which is the f = 15th ordered failure time listed in Table 8.5. It can be seen that $n_f = 16$ of the initial 26 subjects are still at risk at the beginning of month 22. The risk set at this time includes subject #25 and several other subjects (#s 12, 13, 14, 15, 16, 18, 19, 26) who already had at least one failure prior to month 22.

f = 15th ordered failure time
$n_{15} = 16$ subjects in risk set at $t_{(15)} = 22$:

$R(t_{(15)} = 22)$ = {subject #s 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26}

$$\mathbf{L}_{15} = \frac{\exp(\beta(0) + \gamma_1(1) + \gamma_2(5))}{\sum_{s \text{ in } R(t_{(15)})} \exp(\beta tx_{s(15)} + \gamma_1 num_{s(15)} + \gamma_1 size_{s(15)})}$$

The corresponding likelihood $\mathbf{L}_{15}$ at $t_{(15)} = 22$ is shown at the left. Subject #25's values $tx_{25(15)} = 0$, $num_{25(15)} = 1$, and $size_{25(15)} = 5$, have been inserted into the numerator of the formula. The denominator will contain a sum of 16 terms, one for each subject in the risk set at $t_{(15)} = 22$.

Computer program formulates partial likelihood **L**
(See Computer Appendix)

The overall partial likelihood **L** will be formulated internally by the computer program once the data layout is in the correct form and the program code used involves the (start, stop) formulation.

VI. Robust Estimation

As illustrated for subject #14 at the left, each subject contributes a line of data for each time interval corresponding to each recurrent event and any additional event-free follow-up interval.

Data for Subject #14

id	int	event	start	stop	tx	num	size
14	1	1	0	3	0	3	1
14	2	1	3	9	0	3	1
14	3	1	9	21	0	3	1
14	4	0	21	23	0	3	1

Up to this point:
the 4 lines of data for subject #14 are treated as independent observations

We have also pointed out that the Cox model analysis described up to this point treats different lines of data contributed by the same subject as if they were independent contributions from different subjects.

Nevertheless,

- Observations of the same subject are correlated
- Makes sense to adjust for such correlation in the analysis

Nevertheless, it makes sense to view the different intervals contributed by a given subject as representing correlated observations on the same subject that must be accounted for in the analysis.

Robust (Empirical) Estimation

- Adjusts
  $\widehat{\mathbf{Var}}\left(\hat{\beta}_k\right)$
  where
  $\hat{\beta}_k$
  is an estimated regression coefficient
- accounts for misspecification of assumed correlation structure

A widely used technique for adjusting for the correlation among outcomes on the same subject is called **robust estimation** (also referred to as **empirical estimation**). This technique essentially involves adjusting the estimated variances of regression coefficients obtained for a fitted model to account for misspecification of the correlation structure assumed (see Zeger and Liang, 1986 and Kleinbaum and Klein, 2010).

CP approach: assumes independence

Goal of robust estimation: adjust for correlation within subjects

In the CP approach, the assumed correlation structure is independence; that is, the Cox PH model that is fit assumes that different outcomes on the same subject are independent. Therefore the goal of robust estimation for the CP approach is to obtain variance estimators that adjust for correlation within subjects when previously no such correlation was assumed.

Same goal for other approaches for analyzing recurrent event data

This is the same goal for other approaches for analyzing recurrent event data that we describe later in this chapter.

Do not adjust

$\hat{\beta}_k$

Only adjust

$\widehat{\mathbf{Var}}\left(\hat{\beta}_k\right)$

Note that the estimated regression coefficients themselves are not adjusted; only the estimated variances of these coefficients are adjusted.

Robust (Empirical) Variance

allows
tests of hypotheses and confidence intervals
that account for correlated data

The robust (i.e., empirical) estimator of the variance of an estimated regression coefficient therefore allows tests of hypotheses and confidence intervals about model parameters that account for correlation within subjects.

Matrix formula:

derived from ML estimation

We briefly describe the formula for the robust variance estimator below. This formula is in matrix form and involves terms that derive from the set of "score" equations that are used to solve for ML estimates of the regression coefficients. This information may be of interest to the more mathematically inclined reader with some background in methods for the analysis of correlated data (Kleinbaum and Klein, 2010).

Formula not essential for using computer packages

However, the information below is not essential for an understanding of how to obtain robust estimators using computer packages. (See Computer Appendix.)

Extension (Lin and Wei, 1989) of **information sandwich estimator** (Zeger and Liang, 1986)

The robust estimator for recurrent event data was derived by Lin and Wei (1989) as an extension similar to the "information sandwich estimator" proposed by Zeger and Liang (1986) for generalized linear models. SAS and Stata use variations of this estimator that give slightly different estimates.

Matrix formula

$$\hat{\mathbf{R}}\left(\hat{\beta}\right) = \widehat{\mathbf{Var}}\left(\hat{\beta}\right)\left[\hat{\mathbf{R}}_{\mathbf{S}}'\hat{\mathbf{R}}_{\mathbf{S}}\right]\widehat{\mathbf{Var}}\left(\hat{\beta}\right)$$

where

$\widehat{\mathbf{Var}}\left(\hat{\beta}\right)$ is the **information matrix**, and

$\hat{\mathbf{R}}_{\mathbf{S}}$ is matrix of **score residuals**

The general form of this estimator can be most conveniently written in matrix notation as shown at the left. In this formula, the variance expression denotes the **information matrix** form of estimated variances and covariances obtained from (partial) ML estimation of the Cox model being fit. The $\hat{\mathbf{R}}_{\mathbf{S}}$ expression in the middle of the formula denotes the matrix of **score residuals** obtained from ML estimation.

Formula applies to other approaches for analyzing recurrent event data

The robust estimation formula described above applies to the **CP** approach as well as other approaches for analyzing recurrent event data described later in this chapter.

VII. Results for CP Example

We now describe the results from using the **CP** approach on the Bladder Cancer Study data involving all 85 subjects.

Table 8.8. Edited SAS Output from CP Approach on Bladder Cancer Data (N = 85 Subjects) Without Robust Variances

Var	DF	Parameter Estimate	Std Error	Chisq	P	HR
tx	**1**	**−0.4071**	**0.2001**	**4.140**	**0.042**	**0.667**
num	1	0.1607	0.0480	11.198	0.001	1.174
size	1	−0.0401	0.0703	0.326	0.568	0.961

−2 LOG L = 920.159

Table 8.8 gives edited output from fitting the no-interaction Cox PH model involving the three predictors **tx, num,** and **size**. A likelihood ratio chunk test for interaction terms **tx** × **num** and **tx** × **size** was nonsignificant, thus supporting the no-interaction model shown here. The PH assumption was assumed satisfied for all three variables.

Table 8.9. Robust Covariance Matrix, CP Approach on Bladder Cancer Data

	tx	**num**	**size**
tx	**0.05848**	−0.00270	−0.00051
num	−0.00270	0.00324	0.00124
size	−0.00051	0.00124	0.00522

Table 8.9 provides the covariance matrix obtained from robust estimation of the variances of the estimated regression coefficients of **tx, num,** and **size**. The values on the diagonal of this matrix give robust estimates of these variances and the off-diagonal values give covariances.

Robust standard error for **tx** = square-root (**.05848**) = **0.2418**

Non**robust standard error for tx** = **0.2001**

Because the exposure variable of interest in this study is **tx,** the most important value in this matrix is **0.05848**. The square root of this value is **0.2418,** which gives the robust standard error of the estimated coefficient of the **tx** variable. Notice that this robust estimator is similar but somewhat different from the nonrobust estimator of **0.2001** shown in Table 8.8.

Summary of Results from CP Approach

Hazard Ratio **tx**: exp(−0.407) = 0.667 (= 1/1.5)

	robust	nonrobust
Wald Chi-Square **tx**:	2.83	4.14
P-value **tx**:	.09	.04

(H_0: no effect of **tx,** H_A: two sided)

95% CI for HR **tx** (robust): (0.414, 1.069)

H_A: one-sided, both p-values < .05

We now summarize the **CP** results for the effect of the exposure variable **tx** on recurrent event survival controlling for **num** and **size**. The hazard ratio estimate of 0.667 indicates that the hazard for the placebo is 1.5 times the hazard for the treatment.

Using robust estimation, the Wald statistic for this hazard ratio is borderline nonsignificant (P = .09). Using the nonrobust estimator, the Wald statistic is borderline significant (P = .04). Both these P-values, however, are for a two-sided alternative. For a one-sided alternative, both P-values would be significant at the .05 level. The 95% confidence interval using the robust variance estimator is quite wide in any case.

We return to the analysis of these data when we discuss other approaches for analysis of recurrent event data.

VIII. Other Approaches Stratified Cox

3 stratified Cox (SC) approaches:

Stratified CP (Prentice, Williams and
Gap Time Peterson, 1981)
Marginal (Wei, Lin, and Weissfeld, 1989)

Goal: distinguish order of recurrent events

We now describe three other approaches for analyzing recurrent event data, each of which uses a Stratified Cox (SC) PH model. They are called **Stratified CP, Gap Time,** and **Marginal**. These approaches are often used to distinguish the order in which recurrent events occur.

Strata variable: time interval # treated as categorical

The "strata" variable for each approach treats the time interval number as a categorical variable.

Example:
maximum of **4** failures per subject

$$\text{Strata variable} = \Downarrow \begin{cases} 1 \text{ for time interval \# 1} \\ 2 \text{ for time interval \# 2} \\ 3 \text{ for time interval \# 3} \\ 4 \text{ for time interval \# 4} \end{cases}$$

For example, if the maximum number of failures that occur on any given subject in the dataset is, say, 4, then time interval #1 is assigned to stratum 1, time interval #2 to stratum 2, and so on.

Time **between** two events:

Stratified CP $\underset{\text{entry}}{0 \quad 50 \rightarrow 80}$

Gap Time $\underset{\text{ev1} \quad \text{ev2}}{0 \rightarrow 30}$

Both **Stratified CP** and **Gap Time** approaches focus on survival time **between** two events. However, **Stratified CP** uses the actual times of the two events from study entry, whereas **Gap Time** starts survival time at 0 for the earlier event and stops at the later event.

Marginal

- Total survival time from study entry until kth event
- Recurrent events of different types

The **Marginal** approach, in contrast to each conditional approach, focuses on **total** survival time from study entry until the occurrence of a specific (e.g., kth) event; this approach is suggested when recurrent events are viewed to be of different types.

Stratified CP for Subject 10

id	int	event	start	stop	tx	num	size
10	1	1	0	12	0	1	1
10	2	1	12	16	0	1	1
10	3	0	16	18	0	1	1

The **stratified CP** approach uses the exact same (start, stop) data layout format used for the **CP** approach, except that for **Stratified CP,** an SC model is used rather than a standard (unstratified) PH model. The strata variable here is **int** in this listing.

Gap Time for Subject 10

(stop = Interval Length Since Previous Event)							
id	int	event	start	stop	tx	num	size
10	1	1	0	12	0	1	1
10	2	1	0	4	0	1	1
10	3	0	0	2	0	1	1

The **Gap Time** approach also uses a (start, stop) data layout, but the start value is always 0 and the stop value is the time interval length since the previous event. The model here is also a SC model.

Marginal approach
Standard (nonrecurrent event) layout, i.e., without (start, stop) columns

The **Marginal** approach uses the standard (nonrecurrent event) data layout instead of the (start, stop) layout, as illustrated below.

Marginal Approach for Subject 10

id	int	event	stime	tx	num	size
10	1	1	12	0	1	1
10	2	1	16	0	1	1
10	3	0	18	0	1	1
10	4	0	18	0	1	1

The **Marginal** approach layout, shown at the left, contains four lines of data in contrast to the three lines of data that would appear for subject #10 using the **CP, Stratified CP,** and **Gap Time** approaches

Marginal approach
Each subject at risk for all failures that might occur

The reason why there are 4 lines of data here is that, for the **Marginal** approach, each subject is considered to be at risk for all failures that might occur, regardless of the number of events a subject actually experienced.

actual failures $\leq$ # possible failures

Bladder cancer data:

Maximum # (possible) failures = 4

So, subject 10 (as well as all other subjects) gets 4 lines of data

Because the maximum number of failures being considered in the bladder cancer data is 4 (e.g., for subject #s 15 and 26), subject #10, who failed only twice, will have two additional lines of data corresponding to the two additional failures that could have possibly occurred for this subject.

Fundamental Difference Among the 3 SC Approaches

Risk set differs for strata after first event

The three alternative SC approaches (**Stratified CP, Gap Time,** and **Marginal**) fundamentally differ in the way the risk set is determined for strata corresponding to events after the first event.

Gap Time: time until 1st event does not influence risk set for later events (i.e., clock reset to 0 after event occurs)

With **Gap Time,** the time until the first event does not influence the composition of the risk set for a second or later event. In other words, the clock for determining who is at risk gets reset to 0 after each event occurs.

Stratified CP: time until 1st event influences risk set for later events

In contrast, with **Stratified CP,** the time until the first event affects the composition of the risk set for later events.

Marginal: risk set determined from time since study entry

With the **Marginal** approach, the risk set for the kth event (k = 1, 2, ...) identifies those at risk for the kth event since entry into the study.

EXAMPLE

			Days		
ID	**Status**	**Stratum**	**Start**	**Stop**	**tx**
M	1	1	0	100	1
M	1	2	100	105	1
H	1	1	0	30	0
H	1	2	30	50	0
P	1	1	0	20	0
P	1	2	20	60	0
P	1	3	60	85	0

Suppose, for example, that Molly (M), Holly (H), and Polly (P) are the only three subjects in the dataset shown at the left. Molly receives the treatment (tx = 1) whereas Holly and Polly receive the placebo (tx = 0). All three have recurrent events at different times. Also, Polly experiences three events whereas Molly and Holly experience two.

Stratified CP

Stratum 1			Stratum 2		
$t_{(f)}$	n_f	$R(t_{(f)})$	$t_{(f)}$	n_f	$R(t_{(f)})$
0	3	{M, H, P}	20	1	{P}
20	3	{M, H, P}	30	2	{H, P}
30	2	{M, H}	50	2	{H, P}
100	1	{M}	60	1	{P}
			105	1	{M}

The table at the left shows how the risk set changes over time for strata 1 and 2 if the **Stratified CP** approach is used. For **stratum 2**, there are no subjects in the risk set until t = 20, when Polly gets the earliest first event and so becomes at risk for a second event. Holly enters the risk set at t = 30. So at t = 50, when the earliest second event occurs, the risk set contains Holly and Polly. Molly is not at risk for getting her second event until t = 100. The risk set at t = 60 contains only Polly because Holly has already had her second event at t = 50. And the risk set at t = 105 contains only Molly because both Holly and Polly have already had their second event by t = 105.

Gap Time

Stratum 1			Stratum 2		
$t_{(f)}$	n_f	$R(t_{(f)})$	$t_{(f)}$	n_f	$R(t_{(f)})$
0	3	{M, H, P}	0	3	{M, H, P}
20	3	{M, H, P}	5	3	{M, H, P}
30	2	{M, H}	20	2	{H, P}
100	1	{M}	40	1	{P}

The next table shows how the risk set changes over time if the **Gap Time** approach is used. Notice that the data for **stratum 1** are identical to those for **Stratified CP**. For **stratum 2,** however, all three subjects are at risk for the second event at t = 0 and at t = 5, when Molly gets her second event 5 days after the first occurs. The risk set at t = 20 contains Holly and Polly because Molly has already had her second event by t = 20. And the risk set at t = 40 contains only Polly because both Molly and Holly have already had their second event by t = 40.

Marginal

Stratum 1			Stratum 2		
$t_{(f)}$	n_f	$R(t_{(f)})$	$t_{(f)}$	n_f	$R(t_{(f)})$
0	3	{M, H, P}	0	3	{M, H, P}
20	3	{M, H, P}	50	3	{M, H, P}
30	2	{M, H}	60	2	{M, P}
100	3	{M}	105	1	{M}

We next consider the **Marginal** approach. For **stratum 1**, the data are identical again to those for **Stratified CP**. For **stratum 2,** however, all three subjects are at risk for the second event at t = 0 and at t = 50, when Holly gets her second event. The risk set at t = 60 contains Molly and Polly because Holly has already had her second event at t = 50. And the risk set at t = 105 contains only Molly because both Holly and Polly have already had their second event by t = 60.

Stratum 3 for Marginal approach follows

Because Polly experienced three events, there is also a third stratum for this example, which we describe for the marginal approach only.

Marginal
Stratum 3

$t_{(f)}$	n_f	$R(t_{(f)})$
0	3	{M, H, P}
85	2	{M, P}

Note: H censored by t = 85

Using the marginal approach, all three subjects are considered at risk for the third event when they enter the study (t = 0), even though Molly and Holly actually experience only two events. At t = 85, when Polly has her third event, Holly, whose follow-up ended at t = 50, is no longer in the risk set; which still includes Molly because Molly's follow-up continues until t = 105.

Basic idea (**Marginal** approach):

Each failure considered a separate process

Allows stratifying on

- Failure order
- Different failure type (e.g., stage 1 vs. stage 2 cancer)

The basic idea behind the **Marginal** approach is that it allows each failure to be considered as a separate process. Consequently, the **Marginal** approach not only allows the investigator to consider the ordering of failures as separate events (i.e., strata) of interest, but also allows the different failures to represent different types of events that may occur on the same subject.

Stratified Cox PH (SC) Model for all 3 alternative approaches

Use standard computer program for SC (e.g., SAS's PHREG, Stata's stcox, SPSS's coxreg, R's Coxph)

All three alternative approaches, although differing in the form of data layout and the way the risk set is determined, nevertheless use a stratified Cox PH model to carry out the analysis. This allows a standard program that fits a SC model (e.g., SAS's PHREG) to perform the analysis.

No-interaction SC model for bladder cancer data

$$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{num} + \gamma_2\ \mathbf{size}]$$

where g = 1, 2, 3, 4

The models used for the three alternative SC approaches are therefore of the same form. For example, we show on the left the no-interaction SC model appropriate for the bladder cancer data we have been illustrating.

Two types of SC models:

No-interaction versus interaction model
Typically compared using LR test

As described previously in Chapter 5 on the stratified Cox procedure, a no-interaction stratified Cox model is not appropriate if there is interaction between the stratified variables and the predictor variables put into the model. Thus, it is necessary to assess whether an interaction version of the SC model is more appropriate, as typically carried out using a likelihood ratio test.

Version 1: Interaction SC Model

$$h_g(t,\mathbf{X}) = h_{0g}(t)\ \mathbf{exp}[\beta_g\ \mathbf{tx} + \gamma_{1g}\ \mathbf{num} + \gamma_{2g}\ \mathbf{size}]$$
$$g = 1, 2, 3, 4$$

For the bladder cancer data, we show at the left two equivalent versions of the SC interaction model. The first version separates the data into 4 separate models, one for each stratum.

Version 2: Interaction SC Model

$$\begin{aligned} h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[&\beta\ \mathbf{tx} + \gamma_1\ \mathbf{num} \\ &+ \gamma_2\ \mathbf{size} + \delta_{11}(Z_1^* \times \mathbf{tx}) \\ &+ \delta_{12}(Z_2^* \times \mathbf{tx}) + \delta_{13}(Z_3^* \times \mathbf{tx}) \\ &+ \delta_{21}(Z_1^* \times \mathbf{num}) + \delta_{22}(Z_2^* \times \mathbf{num}) \\ &+ \delta_{23}(Z_3^* \times \mathbf{num}) + \delta_{31}(Z_1^* \times \mathbf{size}) \\ &+ \delta_{32}(Z_2^* \times \mathbf{size}) + \delta_{33}(Z_3^* \times \mathbf{size})] \end{aligned}$$

where Z_1^*, Z_2^*, and Z_3^* are 3 dummy variables for the 4 strata.

The second version contains product terms involving the stratified variable with each of the 3 predictors in the model. Because there are 4 strata, the stratified variable is defined using 3 dummy variables Z_1^*, Z_2^*, and Z_3^*.

H_0 (Version 1)

$$\begin{aligned} \beta_1 &= \beta_2 = \beta_3 = \beta_4 \equiv \beta, \\ \gamma_{11} &= \gamma_{12} = \gamma_{13} = \gamma_{14} \equiv \gamma_1, \\ \gamma_{21} &= \gamma_{22} = \gamma_{23} = \gamma_{24} \equiv \gamma_2 \end{aligned}$$

H_0 (Version 2)

$$\begin{aligned} \delta_{11} &= \delta_{12} = \delta_{13} = \delta_{21} = \delta_{22} \\ &= \delta_{23} = \delta_{31} = \delta_{32} = \delta_{33} \\ &= 0 \end{aligned}$$

The null hypotheses for the LR test that compares the interaction with the no-interaction SC model is shown at the left for each version. The df for the LR test is 9.

Interaction SC model may be used regardless of LR test result

- Allows separate HRs for **tx** for each stratum
- if no-interaction SC, then only an overall effect of **tx** can be estimated

Even if the no-interaction SC model is found more appropriate from the likelihood ratio test, the investigator may still wish to use the interaction SC model in order to obtain and evaluate different hazard ratios for each stratum. In other words, if the no-interaction model is used, it is not possible to separate out the effects of predictors (e.g., **tx**) within each stratum, and only an overall effect of a predictor on survival can be estimated.

Recommend using

robust estimation

$$\hat{\mathbf{R}}\left(\hat{\beta}\right) = \widehat{\mathbf{Var}}\left(\hat{\beta}\right)\left[\hat{\mathbf{R}}'_{\mathbf{S}}\hat{\mathbf{R}}_{\mathbf{S}}\right]\widehat{\mathbf{Var}}\left(\hat{\beta}\right)$$

to adjust for correlation of observations on the same subject

For each of the SC alternative approaches, as for the **CP** approach, it is recommended to use **robust estimation** to adjust the variances of the estimated regression coefficients for the correlation of observations on the same subject. The general form for the robust estimator is the same as in the **CP** approach, but will give different numerical results because of the different data layouts used in each method.

IX. Bladder Cancer Study Example (Continued)

We now present and compare SAS results from using all four methods we have described – **CP, Stratified CP, Gap Time,** and **Marginal** – for analyzing the recurrent event data from the bladder cancer study.

Table 8.10. Estimated βs and HRs for **tx** from Bladder Cancer Data

Model	$\hat{\beta}$	$\hat{\mathrm{HR}} = \exp(\hat{\beta})$
CP	−0.407	0.666 (=1/1.50)
SCP	−0.334	0.716 (=1/1.40)
GT	−0.270	0.763 (=1/1.31)
M	−0.580	0.560 (=1/1.79)

CP = Counting Process,
SCP = Stratified CP
GT = Gap Time, M = Marginal

Table 8.10 gives the regression coefficients for the **tx** variable and their corresponding hazard ratios (i.e., **exp**$(\hat{\beta})$ for the no-interaction Cox PH models using these four approaches). The model used for the **CP** approach is a standard Cox PH model whereas the other three models are SC models that stratify on the event order.

HR for **M: 0.560** (=1/1.79)
differs from
HRs for **CP: 0.666** (=1/1.50),
SCP: 0.716 (=1/1.40),
GT: 0.763 (=1/1.31)

From this table, we can see that the hazard ratio for the effect of the exposure variable **tx** differs somewhat for each of the four approaches, with the **Marginal** model giving a much different result from that obtained from the other three approaches.

Table 8.11 Estimated βs, SE(β)s, and P-Values for **tx** from No-Interaction Model for Bladder Cancer Data

Model	$\hat{\beta}$	SE(NR)	SE(R)	P(NR)	P(R)
CP	−0.407	0.200	0.242	.042	.092
SCP	−0.334	0.216	0.197	.122	.090
GT	−0.270	0.208	00.208	.195	.194
M	−0.580	0.201	0.303	.004	.056

CP = Counting Process, SCP = Stratified CP,
GT = Gap Time, M = Marginal,
NR = Nonrobust, R = Robust, P = Wald P-value

Table 8.11 provides, again for the exposure variable **tx** only, the regression coefficients, robust standard errors, nonrobust standard errors, and corresponding Wald Statistic P-values obtained from using the no-interaction model with each approach.

SE(NR) differs from **SE(R)**
P(NR) differs from **P(R)**
but no clear pattern

for example,
CP: P(NR) = .042 < P(R) = .092
SCP: P(NR) = .122 > P(R) = .090
GT: P(NR) = .195 = P(R) = .194

The nonrobust and robust standard errors and P-values differ to some extent for each of the different approaches. There is also no clear pattern to suggest that the nonrobust results will always be either higher or lower than the corresponding robust results.

Wald test statistic(s):

$$Z = \hat{\beta}/SE(\hat{\beta}) \Leftrightarrow Z^2 = [\hat{\beta}/SE(\hat{\beta})]^2 \sim N(0,1) \textbf{ under } \mathbf{H_0}\textbf{: } \boldsymbol{\beta} = \mathbf{0} \sim \chi^2_{1\ df}$$

The P-values shown in Table 8.11 are computed using the standard Wald test Z or chi-square statistic, the latter having a chi-square distribution with 1 df under the null hypothesis that there is no effect of **tx**.

Table 8.12 Estimated βs and Robust SE(β)s for tx from Interaction SC Model for Bladder Cancer Data

	Interaction SC Model				
Model	Str1 $\hat{\beta}_1$ (SE)	Str2 $\hat{\beta}_2$ (SE)	Str3 $\hat{\beta}_3$ (SE)	Str4 $\hat{\beta}_4$ (SE)	No Interaction $\hat{\beta}$ (SE)
CP	—	—	—	—	−.407 (.242)
SCP	−.518 (.308)	−.459 (.441)	.117 (.466)	−.041 (.515)	−.334 (.197)
GT	−.518 (.308)	−.259 (.402)	.221 (.620)	−.195 (.628)	−.270 (.208)
M	−.518 (.308)	−.619 (.364)	−.700 (.415)	−.651 (.490)	−.580 (.303)

CP = Counting Process, SCP = Stratified CP
GT = Gap Time, M = Marginal

Table 8.12 presents, again for the exposure variable **tx** only, the estimated regression coefficients and robust standard errors for both the interaction and the no-interaction SC models for the three approaches (other than the CP approach) that use a SC model.

Notice that for each of the three SC modeling approaches, the estimated βs and corresponding standard errors are different over the four strata as well as for the no-interaction model. For example, using the **Stratified CP** approach, the estimated coefficients are −0.518, − **0.459,** − **0.117,** − **0.041,** and − **0.334** for strata 1 through 4 and the no-interaction model, respectively.

Version 1: Interaction SC Model

$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_g\mathbf{tx} + \gamma_{1g}\,\mathbf{num} + \gamma_{2g}\,\mathbf{size}]$

$g = 1, 2, 3, 4$

Note: subscript g allows for different regression coefficients for each stratum

Such differing results over the different strata should be expected because they result from fitting an interaction SC model, which by definition allows for different regression coefficients over the strata.

Stratified CP for Subject 10

id	int	event	start	stop	tx	num	size
10	1	1	0	12	0	1	1

Gap Time for Subject 10

id	int	event	start	stop	tx	num	size
10	1	1	0	12	0	1	1

Marginal Approach for Subject 10

id	int	event	stime	tx	num	size
10	1	1	12	0	1	1

Note: int = stratum #

Notice also that for stratum 1, the estimated β and its standard error are identical (−**0.518** and **0.308**, resp.) for the **Stratified CP, Gap Time,** and **Marginal** approaches. This is as expected because, as illustrated for subject #10 at the left, the survival time information for first stratum is the same for stratum 1 for the three SC approaches and does not start to differ until stratum 2.

Marginal approach

start time = 0 always
stop time = **stime**

Subject # 10: (start, stop) = (0, 12)

Although the data layout for the **marginal** approach does not require (start,stop) columns, the start time for the first stratum (and all other strata) is 0 and the stop time is given in the **stime** column. In other words, for stratum 1 on subject #10, the stop time is 0 and the start time is 12, which is the same as for the **Stratified CP** and **Gap Time** data for this subject.

Bladder Cancer Study

1. Which approach is best?
2. Conclusion about **tx**?

So, based on all the information we have provided above concerning the analysis of the bladder cancer study,

1. Which of the four recurrent event analysis approaches is best?
2. What do we conclude about the estimated effect of **tx** controlling for num and size?

Which of the 4 approaches is best?
It depends!

CP: Don't want to distinguish recurrent event order
Want overall effect

The answer to question 1 is probably best phrased as, "**It depends!**" Nevertheless, if the investigator does not want to distinguish between recurrent events on the same subject and wishes an overall conclusion about the effect of **tx,** then the **CP** approach seems quite appropriate, as for this study.

If event order important:

Choose from the 3 SC approaches.

If, however, the investigator wants to distinguish the effects of **tx** according to the order that the event occurs (i.e., by stratum #), then one of the three SC approaches should be preferred. So, which one?

Stratified CP: time of recurrent event from entry into the study

The **Stratified CP** approach is preferred if the study goal is to use time of occurrence of each recurrent event from entry into the study to assess a subject's risk for an event of a specific order (i.e., as defined by a stratum #) to occur.

Gap Time: Use time from previous event to next recurrent event

The **Gap Time** approach would be preferred if the time interval of interest is the time (reset from 0) from the previous event to the next recurrent event rather than time from study entry until each recurrent event.

Marginal: Consider strata as representing different event types

Finally, the **Marginal** approach is recommended if the investigator wants to consider the events occurring at different orders as different types of events, for example different disease conditions.

Stratified CP versus **Marginal** (subtle choice)

Recommend: Choose Stratified CP unless strata represent different event types

We (the authors) consider the choice between the **Stratified CP** and **Marginal** approaches to be quite subtle. We prefer **Stratified CP,** provided the different strata do not clearly represent different event types. If, however, the strata clearly indicate separate event processes, we would recommend the **Marginal** approach.

What do we conclude about **tx**?

Conclusions based on results from **CP** and **Stratified CP** approaches

Overall, based on the above discussion, we think that the **CP** approach is an acceptable method to use for analyzing the bladder cancer data. If we had to choose one of the three SC approaches as an alternative, we would choose the **Stratified CP** approach, particularly because the order of recurrent events that define the strata doesn't clearly distinguish separate disease processes.

Table 8.13. Comparison of Results Obtained from No-Interaction Models Across Two Methods for Bladder Cancer Data

	Counting process	Stratified CP
Parameter estimate	−0.407	−0.334
Robust standard error	0.2418	0.1971
Wald chi-square	2.8338	2.8777
p-value	0.0923	0.0898
Hazard ratio	0.667	0.716
95% confidence interval	(0.414, 1.069)	(0.486, 1.053)

Table 8.13 summarizes the results for the **CP** and **Stratified CP** approaches with regard to the effect of the treatment variable (**tx**), adjusted for the control variables **num** and **size**. We report results only for the no-interaction models, because the interaction SC model for the **Stratified CP** approach was found (using LR test) to be not significantly different from the no-interaction model.

The results are quite similar for the two different approaches. There appears to be a small effect of **tx** on survival from bladder cancer: $\widehat{\text{HR}}(\textbf{CP}) = 0.667 = 1/1.50$, $\widehat{\text{HR}}(\textbf{C1}) = 0.716 = 1/1.40$. This effect is borderline nonsignificant (2-sided tests): P(**CP**) = .09 = P(**SCP**). 95% confidence intervals around the hazard ratios are quite wide, indicating an imprecise estimate of effect.

Overall, therefore, these results indicate that there is no strong evidence that **tx** is effective (after controlling for **num** and **size**) based on recurrent event survival analyses of the bladder cancer data.

X. A Parametric Approach Using Shared Frailty

Compared 4 approaches in previous section

- Each used a Cox model
- Robust standard errors
  - Adjusts for correlation from same subject

In the previous section we compared results obtained from using four analytic approaches on the recurrent event data from the bladder cancer study. Each of these approaches used a Cox model. Robust standard errors were included to adjust for the correlation among outcomes from the same subject.

We now present a parametric approach

- Weibull PH model
- Gamma shared frailty component
- Bladder Cancer dataset
  - Data layout for the counting process approach

In this section we present a parametric approach for analyzing recurrent event data that includes a frailty component. Specifically, a Weibull PH model with a gamma distributed shared frailty component is shown using the Bladder Cancer dataset. The data layout is the same as described for the counting process approach. It is recommended that the reader first review Chapter 7, particularly the sections on Weibull models (Section VI) and frailty models (Section XII).

Can review Chapter 7
Weibull model (Section VI)
Frailty models (Section XII)

Hazard conditioned on frailty α_k

$h_i(t|\alpha,\mathbf{X}_i) = \alpha_i h(t|\mathbf{X}_i)$

where $\alpha \sim$ gamma($\mu = 1$, var $= \theta$) and where $h(t|\mathbf{X}_i) = \lambda_i p t^{p-1}$ (Weibull) with $\lambda_{fk} = \exp(\beta_0 + \beta_1 \mathbf{tx}_i + \beta_2 \mathbf{num}_i + \beta_3 \mathbf{size}_i)$

We define the model in terms of the hazard of any (recurrent) outcome on the ith subject conditional on his or her frailty α_i. The frailty is a multiplicative random effect on the hazard function $h(t|\mathbf{X}_i)$, assumed to follow a gamma distribution of mean 1 and variance theta θ. We assume $h(t|\mathbf{X}_i)$ follows a Weibull distribution (shown at left).

Including shared frailty

- Accounts for unobserved factors
  - Subject specific
  - Source of correlation
  - Observations clustered by subject

The frailty is included in the model to account for variability due to unobserved subject-specific factors that are otherwise unaccounted for by the other predictors in the model. These unobserved subject-specific factors can be a source of within-subject correlation. We use the term **shared frailty** to indicate that observations are clustered by subject and each cluster (i.e., subject) shares the same level of frailty.

Robust standard errors

- Adjusts standard errors
- Does not affect coefficient estimates

In the previous sections, we have used robust variance estimators to **adjust** the standard errors of the coefficient estimates to account for within-subject correlation. Shared frailty is not only an adjustment, but also is built into the model and can have an impact on the estimated coefficients as well as their standard errors.

Shared frailty

- Built into model
- Can affect coefficient estimates and their standard errors

The model output (obtained using Stata version 10) is shown on the left. The inclusion of frailty in a model (shared or unshared) leads to one additional parameter estimate in the output (theta, the variance of the frailty). A likelihood ratio test for theta = 0 yields a statistically significant p-value of 0.003 (bottom of output) suggesting that the frailty component contributes to the model and that there is within-subject correlation.

Weibull regression (PH form)
Gamma shared frailty
Log likelihood = −184.73658

_t	Coef.	Std. Err.	z	P > |z|
tx	−.458	.268	−1.71	0.011
num	.184	.072	2.55	0.327
size	−.031	.091	−0.34	0.730
_cons	−2.952	.417	−7.07	0.000
/ln_p	−.119	.090	−1.33	0.184
/ln_the	−.725	.516	−1.40	0.160
p	.888	.080		
1/p	1.13	.101		
theta	.484	.250		

Likelihood ratio test of theta = 0:
chibar(01) = 7.34
Prob > = chibar2 = 0.003

The estimate for the Weibull shape parameter p is 0.888 suggesting a slightly decreasing hazard over time because $\hat{p} < 1$. However, the Wald test for $\ln(p) = 0$ (or equivalently $p = 1$) yields a non-significant p-value of 0.184.

Comparing Hazard Ratios

Weibull with frailty model

$$\widehat{HR}(\mathbf{tx}) = \exp(-0.458) = 0.633$$

$$95\%\,\text{CI} = \exp[-0.458 \pm 1.96(0.268)] = (0.374,\ 1.070)$$

Counting processes approach with Cox model

$$\widehat{HR}(\mathbf{tx}):\ \exp(-0.407) = 0.667$$

95% CI for HR **tx** (robust): (0.414, 1.069)

An estimated hazard ratio of 0.633 for the effect of treatment comparing two individuals with the same level of frailty and controlling for the other covariates is obtained by exponentiating the estimated coefficient (−0.458) for **tx**. The estimated hazard ratio and 95% confidence intervals are similar to the corresponding results obtained using a counting processes approach with a Cox model and robust standard errors (see left).

Interpretations of HR from frailty model

- Compares 2 individuals with same α
- Compares individual with himself
  - What is effect if individual had used treatment rather than placebo?

Another interpretation for the estimated hazard ratio from the frailty model involves the comparison of an individual to himself. In other words, this hazard ratio describes the effect on an individual's hazard (i.e., conditional hazard) if that individual had used the treatment rather than the placebo.

XI. A Second Example

Age-Related Eye Disease Study (AREDS)

Outcome

Age-related macular degeneration (AMD)

Clinical trial
Evaluate effect of treatment with high doses of antioxidants and zinc on progression of AMD

n = 43 (subset of data analyzed here)

We now illustrate the analysis of recurrent event survival data using a new example. We consider a subset of data from the Age-Related Eye Disease Study (AREDS), a long-term multicenter, prospective study sponsored by the U.S. National Eye Institute of the clinical course of age-related macular degeneration (AMD) (see AREDS Research Group, 2003).

In addition to collecting natural history data, AREDS included a clinical trial to evaluate the effect of high doses of antioxidants and zinc on the progression of AMD. The data subset we consider consists of 43 patients who experienced ocular events while followed for their baseline condition, macular degeneration.

Exposure

tx = 1 if treatment, 0 if placebo

8 years of follow-up

The exposure variable of interest was treatment group (**tx**), which was coded as 1 for patients randomly allocated to an oral combination of antioxidants, zinc, and vitamin C versus 0 for patients allocated to a placebo. Patients were followed for 8 years.

Two possible events

First event: visual acuity score <50 (i.e., poor vision)

Each patient could possibly experience two events. The first event was defined as the sudden decrease in visual acuity score below 50 measured at scheduled appointment times. Visual acuity score was defined as the number of letters read on a standardized visual acuity chart at a distance of 4 m, where the higher the score, the better the vision.

Second event: clinically advanced severe stage of macular degeneration

The second event was considered a successive stage of the first event and defined as a clinically advanced and severe stage of macular degeneration. Thus, the subject had to experience the first event before he or she could experience the second event.

4 approaches for analyzing recurrent event survival data carried out on macular degeneration data

Each model contains **tx, age,** and **sex**.

We now describe the results of using the four approaches for analyzing recurrent event survival with these data. In each analysis, two covariates **age** and **sex** were controlled, so that each model contained the variables **tx, age,** and **sex**.

CP model

$$h(t,\mathbf{X}) = h_0(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{age} + \gamma_2\ \mathbf{sex}]$$

No-interaction SC model

$$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{age} + \gamma_2\ \mathbf{sex}]$$

where g = 1, 2

Interaction SC model:

$$h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_g\ \mathbf{tx} + \gamma_{1g}\ \mathbf{age} + \gamma_{2g}\ \mathbf{sex}]$$

where g = 1, 2

The counting process **(CP)** model is shown here at the left together with both the no-interaction and interaction SC models used for the three stratified Cox (SC) approaches.

Table 8.14 Comparison of Parameter Estimates and Robust Standard Errors for Treatment Variable (tx) Controlling for Age and Sex (Macular Degeneration Data)

	"Interaction" Cox stratified model		"No-interaction" SC model
Model	Stratum 1	Stratum 2	
	$\hat{\beta}_1$ (SE)	$\hat{\beta}_2$ (SE)	$\hat{\beta}_3$ (SE)
Counting process	n/a	n/a	−0.174 (0.104)
SCP	−0.055 (0.286)	−0.955 (0.443)	−0.306 (0.253)
GT	−0.055 (0.286)	−1.185 (0.555)	−0.339 (0.245)
Marginal	−0.055 (0.286)	−0.861 (0.465)	−0.299 (0.290)

Interaction SC models are preferred (based on LR test results) to use of no-interaction SC model

Table 8.15. Comparison of Results for the Treatment Variable (tx) Obtained for **Stratified CP** and Marginal Approaches (Macular Degeneration Data)

		Stratified CP	Marginal
Estimate	$\hat{\beta}_1$	−0.0555	−0.0555
	$\hat{\beta}_2$	−0.9551	−0.8615
	$\hat{\beta}$	−0.306	−0.2989
Robust std. error	SE($\hat{\beta}_1$)	0.2857	0.2857
	SE($\hat{\beta}_2$)	0.4434	0.4653
	SE($\hat{\beta}$)	0.2534	0.2902
Wald chi-square	H_0:$\beta_1 = 0$	0.0378	0.0378
	H_0:$\beta_2 = 0$	**4.6395**	**3.4281**
	H_0:$\beta = 0$	1.4569	1.0609
P-value	H_0:$\beta_1 = 0$	0.8458	0.8478
	H_0:$\beta_2 = 0$	**0.0312**	**0.0641**
	H_0:$\beta = 0$	0.2274	0.3030
Hazard ratio	exp($\hat{\beta}_1$)	0.946	0.946
	exp($\hat{\beta}_2$)	**0.385**	**0.423**
	exp($\hat{\beta}$)	0.736	0.742
95% Conf. interval	exp($\hat{\beta}_1$)	(0.540, 1.656)	(0.540, 1.656)
	exp($\hat{\beta}_2$)	**(0.161, 0.918)**	**(0.170, 1.052)**
	exp($\hat{\beta}$)	(0.448, 1.210)	(0.420, 1.310)

In Table 8.14, we compare the coefficient estimates and their robust standard errors for the treatment variable (**tx**) from all four approaches. This table shows results for both the "interaction" and "nointeraction" stratified Cox models for the three approaches other than the counting process approach.

Notice that the estimated coefficients for β_1 and their corresponding standard errors are identical for the three SC approaches. This will always be the case for the first stratum regardless of the data set being considered.

The estimated coefficients for β_2 are, as expected, somewhat different for the three SC approaches. We return to these results shortly.

LR tests for comparing the "no-interaction" with the "interaction" SC models were significant ($P < .0001$) for all three SC approaches (details not shown), indicating that an interaction model was more appropriate than a no-interaction model for each approach.

In Table 8.15, we summarize the statistical inference results for the effect of the treatment variable (**tx**) for the **Stratified CP** and **Marginal** approaches only.

We have not included the **CP** results here because the two events being considered are of very different types, particularly regarding severity of illness, whereas the CP approach treats both events as identical replications. We have not considered the **Gap Time** approach because the investigators were more likely interested in survival time from baseline entry into the study than the survival time "gap" from the first to second event.

Because we previously pointed out that the interaction SC model was found to be significant when compared to the corresponding no-interaction SC model, we focus here on the treatment (**tx**) effect for each stratum (i.e., event) separately.

First event:

	SCP	Marginal
$\widehat{HR}$	0.946	0.946
p-value	0.85	0.85

Based on the Wald statistics and corresponding P-values for testing the effect of the treatment on survival to the *first event* (i.e., H_0: $\beta_1 = 0$), both the **Stratified CP** and **Marginal** approaches give the identical result that the estimated treatment effect ($\widehat{HR} = 0.946 = 1/1.06$) is neither meaningful nor significant ($P = 0.85$).

Second event:

	SCP	Marginal
$\widehat{HR}$	0.385	0.423
p-value	0.03	0.06
95% CI	(0.16, 0.92)	(0.17, 1.05)

For the *second event*, indicating a clinically severe stage of macular degeneration, the Wald statistic P-value for the **Stratified CP** approach is 0.03, which is significant at the .05 level, whereas the corresponding P-value for the **Marginal** approach is 0.06, borderline nonsignificant at the .05 level.

The estimated HR for the effect of the treatment is ($\widehat{HR} = 0.385 = 1/2.60$) using the **Stratified CP** approach and its 95% confidence interval is quite wide but does not contain the null value of 1. For the **Marginal** approach, the estimated HR is $\widehat{HR} = 0.423 = 1/2.36$, also with a wide confidence interval, but includes 1.

Conclusions regarding 1st event:

- No treatment effect
- Same for **Stratified CP** and **Marginal** approaches

Thus, based on the above results, there appears to be no effect of treating patients with high doses of antioxidants and zinc on reducing visual acuity score below 50 (i.e., the first event) based on either **Stratified CP** or **Marginal** approaches to the analysis.

Conclusions regarding 2nd event:

- Clinically moderate and statistically significant treatment effect
- Similar for **Stratified CP** and **Marginal** approaches, but more support from **Stratified CP** approach

However, there is evidence of a clinically moderate and statistically significant effect of the treatment on protection (i.e., not failing) from the second more severe event of macular degeneration. This conclusion is more supported from the **Stratified CP** analysis than from the **Marginal** analysis.

Comparison of **Stratified CP** with **Marginal** Approach

What if results had been different?

Despite similar conclusions from both approaches, it still remains to compare the two approaches for these data. In fact, if the results from each approach had been very different, it would be important to make a choice between these approaches.

Recommend **Stratified CP** if

Can assume 2nd event cannot occur without 1st event previously occurring
⇓
Should consider survival time to 2nd event **conditional on** experiencing 1st event

Recommend **Marginal** if

Can assume each subject at risk for 2nd event whether or not 1st event previously occurred
⇓
2nd event considered a separate event, that is, **unconditional** of the 1st event
⇓
Should consider survival times to 2nd event for **all** subjects

Macular degeneration data: recommend **Marginal** approach

In general: carefully consider interpretation of each approach

Nevertheless, we authors find it difficult to make such a decision, even for this example. The **Stratified CP** approach would seem appropriate if the investigators assumed that the second event cannot occur without the first event previously occurring. If so, it would be important to consider survival time to the second event only for (i.e., **conditional on**) those subjects who experience a first event.

On the other hand, the **Marginal** approach would seem appropriate if each subject is considered to be at risk for the second event whether or not the subject experiences the first event. The second event is therefore considered separate from (i.e., **unconditional of**) the first event, so that survival times to the second event need to be included for all subjects, as in the **Marginal** approach.

For the macular degeneration data example, we find the **Marginal** approach persuasive. However, **in general**, the choice among all four approaches is not often clear-cut and requires careful consideration of the different interpretations that can be drawn from each approach.

XII. Survival Curves with Recurrent Events

Goal: Plot and Interpret Survival Curves

Types of survival curves:

KM (empirical): Chapter 2
Adjusted (Cox PH): Chapters 3 and 4

Previously: 1 (nonrecurrent) event
Now:
Survival plots with recurrent events?

An important goal of most survival analyses, whether or not a regression model (e.g., Cox PH) is involved, is to plot and interpret/compare survival curves for different groups. We have previously described the Kaplan–Meier (KM) approach for plotting empirical survival curves (Chapter 2) and we have also described how to obtain adjusted survival curves for Cox PH models (Chapters 3 and 4).

This previous discussion only considered survival data for the occurrence of one (nonrecurrent) event. So, how does one obtain survival plots when there are recurrent events?

Focus on one ordered event at a time

$S_1(t)$: 1st event
$S_2(t)$: 2nd event
...
$S_k(t)$: kth event

The answer is that survival plots with recurrent events only make sense when the focus is on one ordered event at a time. That is, we can plot a survival curve for **survival to a first event, survival to a second event,** and so on.

Survival to a 1st event

$S_1(t) = \Pr(T_1 > t)$

where
T_1 = survival time up to occurrence of 1st event
(ignores later recurrent events)

For **survival to a first event,** the survival curve describes the probability that a subject's time to occurrence of a first event will exceed a specified time. Such a plot essentially ignores any recurrent events that a subject may have after a first event.

Survival to a 2nd event

$S_2(t) = \Pr(T_2 > t)$

where
T_2 = survival time up to occurrence of 2nd event

For **survival to a second event**, the survival curve describes the probability that a subject's time to occurrence of a second event will exceed a specified time.

Two versions

Stratified:
T_{2c} = time from 1st event to 2nd event, restricting data to 1st event subjects
Marginal:
T_{2m} = time from study entry to 2nd event, ignoring 1st event

There are two possible versions for such a plot.

Stratified: use survival time from time of first event until occurrence of second event, thus restricting the dataset to only those subjects who experienced a first event.
Marginal: use survival time from study entry to occurrence of second event, ignoring whether a first event occurred.

Survival to a kth event ($k \geq 2$)

$S_k(t) = \Pr(T_k > t)$

where
T_k = survival time up to occurrence of kth event

Similarly, for **survival to the kth event**, the survival curve describes the probability that a subject's time to occurrence of the kth event will exceed a specified time.

Two versions

Stratified:
T_{kc} = time from the $k-1$st to kth event, restricting data to subjects with $k-1$ events
Marginal:
T_{km} = time from study entry to kth event, ignoring previous events

As with survival to the second event, there are two possible versions, **Stratified** or **Marginal,** for such a plot, as stated on the left.

EXAMPLE

ID	Status	Stratum	Days Start	Days Stop	tx
M	1	1	0	100	1
M	1	2	100	105	1
H	1	1	0	30	0
H	1	2	30	50	0
P	1	1	0	20	0
P	1	2	20	60	0
P	1	3	60	85	0

We now illustrate such survival plots for recurrent event data by returning to the small dataset previously described for three subjects Molly (M), Holly (H), and Polly (P), shown again on the left.

Deriving $S_1(t)$: **Stratum 1**

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_1(t_{(f)})$
0	3	0	0	{M, H, P}	1.00
20	3	1	0	{M, H, P}	0.67
30	2	1	0	{M, H}	0.33
100	1	1	0	{M}	0.00

Deriving $S_{2c}(t)$: **Stratum 2 (Stratified GT)**

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)}\}$	$S_{2c}(t_{(f)})$
0	3	0	0	{M, H, P}	1.00
5	3	1	0	{M, H, P}	0.67
20	2	1	0	{M, P}	0.33
450	1	1	0	{M}	0.00

Deriving $S_{2m}(t)$: **Stratum 2 (Marginal)**

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)}\}$	$S_{2m}(t_{(f)})$
0	3	0	0	{M, H, P}	1.00
20	3	1	0	{M, H, P}	0.67
30	2	1	0	{H, P}	0.33
100	1	1	0	{P}	0.00

The survival plot for survival to the first event $\mathbf{S_1(t)}$ is derived from the stratum 1 data layout for any of the three alternative SC analysis approaches. Recall that m_f and q_f denote the number of failures and censored observations at time $t_{(f)}$. The survival probabilities in the last column use the KM product limit formula.

The **Stratified** survival plot for survival to the second event is derived from the stratum 2 data layout for the **Gap Time** approach. We denote this survival curve as $\mathbf{S_{2c}(t)}$. Notice that the survival probabilities here are identical to those in the previous table; however, the failure times $t_{(f)}$ in each table are different.

The **Marginal** survival plot for survival to the second event is derived from the stratum 2 data layout for the **Marginal** approach. We denote this survival curve as $\mathbf{S_{2m}(t)}$. Again, the last column here is identical to those in the previous two tables, but, once again, the failure times $t_{(f)}$ in each table are different.

Survival Plots for Molly, Holly and Polly Recurrent Event Data (n = 3)

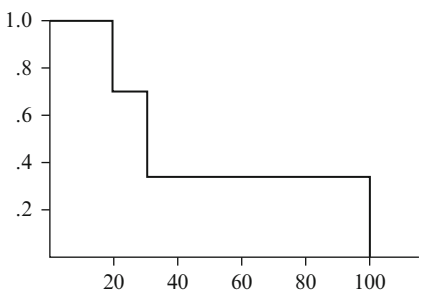


Figure 8.1. $\mathbf{S_1(t)}$: Survival to 1st Event

The survival plots that correspond to the above three data layouts are shown in Figures 8.1 to 8.3.

Figure 8.1 shows survival probabilities for the first event, ignoring later events. The risk set at time zero contains all three subjects. The plot drops from $S_1(t) = 1$ to $S_1(t) = 0.67$ at $t = 20$, drops again to $S_1(t) = 0.33$ at $t = 30$ and falls to $S_1(t) = 0$ at $t = 100$ when the latest first event occurs.

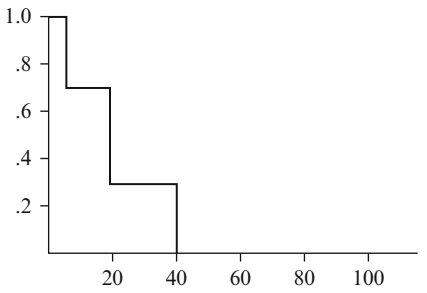


Figure 8.2. $S_{2c}(t)$: Survival to 2nd Event (Stratified GT)

Figure 8.2 shows **Stratified GT** survival probabilities for the second event using survival time **from the first event to the second event**. Because all three subjects had a first event, the risk set at time zero once again contains all three subjects. Also, the survival probabilities of 1, 0.67, 0.33, and 0 are the same as in Figure 8.1. Nevertheless, this plot differs from the previous plot because the survival probabilities are plotted at different survival times (t = 5, 20, 40 in Figure 8.2 instead of t = 20, 30, 100 in Figure 8.1)

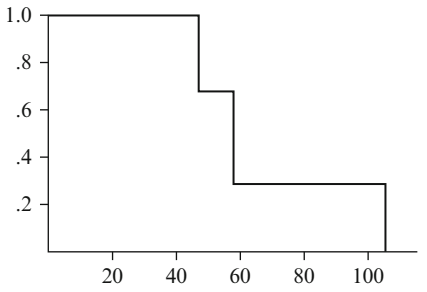


Figure 8.3. $S_{2m}(t)$: Survival to 2nd Event (Marginal)

Figure 8.3 shows **Marginal** survival probabilities for the second event using survival time **from study entry to the second event, ignoring the first event**. The survival probabilities of 1, 0.67, 0.33, and 0 are once again the same as in Figures 8.1 and 8.2. Nevertheless, this plot differs from the previous two plots because the survival probabilities are plotted at different survival times (t = 50, 60, 105 in Figure 8.3).

XIII. Summary

4 approaches for recurrent event data
Counting process (CP), Stratified CP, Gap Time, Marginal

We have described four approaches for analyzing recurrent event survival data.

The 4 approaches

- Differ in how risk set is determined
- Differ in data layout
- All involve standard Cox model program
- Latter three approaches use a SC model

These approaches differ in how the risk set is determined and in data layout. All four approaches involve using a standard computer program that fits a Cox PH model, with the latter three approaches requiring a stratified Cox model, stratified by the different events that occur.

Identical recurrent events
⇓
CP approach

The approach to analysis typically used when recurrent events are treated as identical is called the **CP Approach**.

Recurrent events: different disease categories or event order important

⇓

Stratified Cox (SC) approaches

When recurrent events involve different disease categories and/or the order of events is considered important, the analysis requires choosing among the three alternative SC approaches.

CP approach: **Start** and **Stop** times

Standard layout: only **Stop** (survival) times (no recurrent events)

The data layout for the counting process approach requires each subject to have a line of data for each recurrent event and lists the **start time** and **stop time** of the interval of follow-up. This contrasts with the standard layout for data with no recurrent events, which lists only the stop (**survival**) time on a single line of data for each subject.

Stratified CP: same **Start** and **Stop** Times as **CP**, but uses **SC** model

The **Stratified CP** approach uses the exact same (start, stop) data layout format used for the **CP** approach, except that for **Stratified CP,** the model used is a SC PH model rather than an unstratified PH model.

Gap Time: Start and **Stop** Times
Start = 0 always
Stop = time since previous event
SC model

The **Gap Time** approach also uses a (start, stop) data layout, but the start value is always 0 and the stop value is the time interval length since the previous event. The model here is also a SC model.

Marginal approach:

Standard layout (nonrecurrent event), that is, without (**Start, Stop**) columns
Each failure is a separate process

The **Marginal** approach uses the standard (nonrecurrent event) data layout instead of the (start, stop) layout. The basic idea behind the **Marginal** approach is that it allows each failure to be considered as a separate process.

Recommend using **robust estimation** to adjust for correlation of observations on the same subject.

For each of the SC alternative approaches, as for the **CP** approach, it is recommended to use **robust estimation** to adjust the variances of the estimated regression coefficients for the correlation of observations on the same subject.

Application 1: Bladder Cancer study
n = 86
64 months of follow-up

We considered two applications of the different approaches described above. First, we compared results from using all four methods to analyze data from a study of bladder cancer involving 86 patients, each followed for a variable time up to 64 months.

Repeated event: recurrence of bladder cancer tumor; up to 4 events

The repeated event analyzed was the recurrence of a bladder cancer tumor after transurethral surgical excision. Each recurrence of new tumors was treated by removal at each examination. About 25% of the 86 subjects experienced four events.

tx = 1 if thiotepa, 0 if placebo
num = initial # of tumors
size = initial size of tumors

The exposure variable of interest was drug treatment status (**tx,** 0 = placebo, 1 = treatment with thiotepa), There were two covariates: initial number of tumors (**num**) and initial size of tumors (**size**).

CP results: no strong evidence for **tx** ($\widehat{HR} = 0.67$, P = .09, 95% CI: 0.414, 1.069)

Results for the **CP** approach, which was considered appropriate for these data, indicated that there was no strong evidence that **tx** is effective after controlling for **num** and **size**.

Alternative parametric approach

- Weibull PH model
- Gamma shared frailty component
- Bladder cancer dataset
- Similar HR and confidence interval as for counting process approach

An alternative approach for analyzing recurrent event data was also described using a parametric model containing a frailty component (see Chapter 7). Specifically, a Weibull PH model with a gamma distributed frailty was fit using the bladder cancer dataset. The resulting estimated HR and confidence interval were quite similar to the counting process results.

Application 2: Clinical trial

n = 43
8 years of follow-up
High doses of antioxidants and zinc
Age-related macular degeneration

The second application considered a subset of data (n = 43) from a clinical trial to evaluate the effect of high doses of antioxidants and zinc on the progression of age-related macular degeneration (AMD). Patients were followed for 8 years.

Exposure: **tx** = 1 if treatment, 0 if placebo
Covariates: **age, sex**

The exposure variable of interest was treatment group (**tx**). Covariates considered were **age** and **sex**.

Two possible events:

1st event: visual acuity score <50 (i.e., poor vision)
2nd event: clinically advanced severe stage of macular degeneration

Each patient could possibly experience two events. The first event was defined as the sudden decrease in visual acuity score below 50. The second event was considered a successive stage of the first event and defined as a clinically advanced and severe stage of macular degeneration.

Focus on **Stratified CP** vs. **Marginal** (events were of different types)

Because the two events were of very different types and because survival from baseline was of primary interest, we focused on the results for the **Stratified CP** and **Marginal** approaches only.

Interaction SC model ✓
No-interaction SC model ×

An interaction SC model was more appropriate than a no-interaction model for each approach, thus requiring separate results for the two events under study.

Conclusions regarding 1st event

- No treatment effect
- Same for **Stratified CP** and **Marginal** approaches

The results for the first event indicated no effect of the treatment on reducing visual acuity score below 50 (i.e., the first event) from either **Stratified CP** or **Marginal** approaches to the analysis.

Conclusions regarding 2nd event

- Clinically moderate and statistically significant treatment effect

However, there was evidence of a clinically moderate and statistically significant effect of the treatment on the second more severe event of macular degeneration.

Macular degeneration data: prefer **Marginal** approach (but not clear-cut)

The choice between the Stratified CP and marginal approaches for these data was not clear-cut, although the **Marginal** approach was perhaps more appropriate because the two events were of very different types.

In general: carefully consider interpretation of each approach

In general, however, the choice among all four approaches requires careful consideration of the different interpretations that can be drawn from each approach.

Survival plots: one ordered event at a time Two versions for survival to kth event:
Stratified: only subjects with $k - 1$ events
Marginal: ignores previous events

Survival plots with recurrent events are derived one ordered event at a time. For plotting survival to a kth event where $k \geq 2$, one can use either a **Stratified** or **Marginal** plot, which typically differ.

Detailed Outline

I. Overview (page 366)

A. Focus: outcome events that may occur more than once over the follow-up time for a given subject, that is, "recurrent events."

B. **Counting Process (CP)** approach uses the Cox PH model.

C. Alternative approaches that use a Stratified Cox (SC) PH model and a frailty model.

II. Examples of Recurrent Event Data (pages 366–368)

A. 1. Multiple relapses from remission: leukemia patients.

2. Repeated heart attacks: coronary patients.

3. Recurrence of tumors: bladder cancer patients.

4. Deteriorating episodes of visual acuity: macular degeneration patients.

B. Objective of each example: to assess relationship of predictors to rate of occurrence, allowing for multiple events per subject.

C. Different analysis required depending on whether:

1. Recurrent events are treated as identical (**counting process approach**), or

2. Recurrent events involve different disease categories and/or the order of events is important (**stratified Cox** approaches).

III. Counting Process Example (pages 368–369)

A. Data on two hypothetical subjects from a randomized trial that compares two treatments for bladder cancer tumors.

B. Data set-up for **Counting Process (CP)** approach:

1. Each subject contributes a line of data for each time interval corresponding to each recurrent event and any additional event-free follow-up interval.

2. Each line of data for a given subject lists the **start time** and **stop time** for each interval of follow-up.

IV. General Data Layout: Counting Process Approach (pages 370–371)

A. r_i time intervals for subject i.
d_{ij} event status (0 or 1) for subject i in interval j.
t_{ij0} start time for subject i in interval j.
t_{ij1} stop time for subject i in interval j.
X_{ijk} value of kth predictor for subject i in interval j.
$i = 1, 2, \ldots, N$; $j = 1, 2, \ldots, r_i$; $k = 1, 2, \ldots, p$.

B. Layout for subject i:

i	j	d_{ij}	t_{ij0}	t_{ij1}	X_{ij1}	X_{ijp}
i	1	d_{i1}	t_{i10}	t_{i11}	X_{111}	X_{i1p}
i	2	d_{i2}	t_{i20}	t_{i21}	X_{121}	X_{i2p}
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
i	r_i	d_{ir_i}	t_{ir_i0}	t_{ir_i1}	X_{ir_i1}	X_{ir_ip}

C. Bladder Cancer Study example:
1. Data layout provided for the first 26 subjects (86 subjects total) from a 64-month study of recurrent bladder cancer tumors.
2. The exposure variable: drug treatment status (**tx,** 0 = placebo, 1 = treatment with thiotepa).
3. Covariates: initial number of tumors (**num**) and initial size of tumors (**size**).
4. Up to 4 events per subject.

V. The Counting Process Model and Method (pages 372–376)

A. The model typically used to carry out the **Counting Process (CP)** approach is the standard Cox PH model: $h(t, \mathbf{X}) = h_0(t) \exp[\Sigma \beta_i X_i]$.

B. For recurrent event survival data, the (partial) likelihood function is formed differently than for nonrecurrent event survival data:
1. A subject who continues to be followed after having failed at $t_{(f)}$ does not drop out of the risk set after $t_{(f)}$ and remains in the risk set until his or her last interval of follow-up, after which the subject is removed from the risk set.
2. Different lines of data contributed by the same subject are treated in the analysis as if they were independent contributions from different subjects.

C. For the bladder cancer data, the Cox PH Model for **CP** approach is given by

$$h(t, \mathbf{X}) = h_0(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma_1\ \mathbf{num} + \gamma_2\ \mathbf{size}].$$

D. The overall partial likelihood **L** from using the **CP** approach will be automatically determined by the computer program used once the data layout is in the correct **CP** form and the program code used involves the (start, stop) formulation.

VI. Robust Estimation (pages 376–378)

A. In the **CP** approach, the different intervals contributed by a given subject represent correlated observations on the same subject that must be accounted for in the analysis.

B. A widely used technique for adjusting for the correlation among outcomes on the same subject is called **robust estimation**.

C. The goal of **robust estimation** for the CP approach is to obtain variance estimators that adjust for correlation within subjects when previously no such correlation was assumed.

D. The **robust estimator** of the variance of an estimated regression coefficient allows tests of hypotheses and confidence interval estimation about model parameters to account for correlation within subjects.

E. The general form of the **robust estimator** can be most conveniently written in matrix notation; this formula is incorporated into the computer program and is automatically calculated by the program with appropriate coding.

VII. Results for CP Example (pages 378–379)

A. Edited output is provided from fitting the no-interaction Cox PH model involving the three predictors **tx, num**, and **size**.

B. A likelihood ratio chunk test for interaction terms **tx** × **num** and **tx** × **size** was nonsignificant.

C. The PH assumption was assumed satisfied for all three variables.

D. The robust estimator of 0.2418 for the standard deviation of **tx** was similar though somewhat different from the corresponding nonrobust estimator of 0.2001.

E. There was not strong evidence that **tx** is effective after controlling for **num** and **size** ($\widehat{HR}$ = 0.67, two-sided P = .09, 95% CI: 0.414, 1.069).

F. However, for a one-sided alternative, the p-values using both robust and nonrobust standard errors were significant at the .05 level.

G. The 95% confidence interval using the robust variance estimator is quite wide.

VIII. Other Approaches Stratified Cox
(pages 379–385)

A. The "strata" variable for each of the three SC approaches treats the time interval number for each event occurring on a given subject as a stratified variable.

B. Three alternative approaches involving SC models need to be considered if the investigator wants to distinguish the order in which recurrent events occur.

C. These approaches all differ from what is called **competing risk** survival analysis in that the latter allows each subject to experience only one of several different types of events over follow-up.

D. **Stratified CP** approach:
   1. Same **Start** and **Stop** Times as **CP**.
   2. **SC** model.

E. **Gap Time** approach:
   1. **Start** and **Stop** Times, but **Start** = 0 always and **Stop** = time since previous event.
   2. **SC** model.

F. **Marginal** approach:
   1. Uses standard layout (nonrecurrent event); no (**Start, Stop**) columns.
   2. Treats each failure is a separate process.
   3. Each subject at risk for all failures that might occur, so that # actual failures < # possible failures.
   4. **SC** model.

G. Must decide between two types of **SC** models:
   1. No-interaction **SC** versus interaction **SC**.
   2. Bladder cancer example:
      No-interaction model: $h_g(t, \mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta\,\mathbf{tx} + \gamma_1\,\mathbf{num} + \gamma_2\,\mathbf{size}]$ where g = 1, 2, 3, 4.
      Interaction model: $h_g(t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_g\mathbf{tx} + \gamma_{1g}\mathbf{num} + \gamma_{2g}\mathbf{size}]$. where g = 1, 2, 3, 4.

H. Recommend using **robust estimation** to adjust for correlation of observations on the same subject.

IX. Bladder Cancer Study Example (Continued) (pages 385–389)

A. Results from using all four methods **– CP, Stratified CP, Gap Time**, and **Marginal –** on the bladder cancer data were compared.

B. The hazard ratio for the effect of **tx** based on a no-interaction model differed somewhat for each of the four approaches, with the marginal model being most different:

M: 0.560 **CP:** 0.666 **SCP:** 0.716 **GT:** 0.763

C. The nonrobust and robust standard errors and P-values differed to some extent for each of the different approaches.

D. Using an interaction SC model, the estimated βs and corresponding standard errors are different over the four strata (i.e., four events) for each model separately.

E. The estimated β's and corresponding standard errors for the three alternative SC models are identical, as expected (always for first events).

F. Which of the four recurrent event analysis approaches is best?

1. Recommend **CP** approach if do not want to distinguish between recurrent events on the same subject and desire overall conclusion about the effect of **tx**.
2. Recommend one of the three SC approaches if want to distinguish the effect of **tx** according to the order in which the event occurs.
3. The choice between the **Stratified CP** and **Marginal** is difficult, but prefer **Stratified CP** because the strata do not clearly represent different event types.

G. Overall, regardless of the approach used, there was no strong evidence that **tx** is effective after controlling for **num** and **size**.

X. A Parametric Approach Using Shared Frailty (pages 389–391)

A. Alternative approach using a parametric model containing a frailty component (see Chapter 7).

B. Weibull PH model with a gamma distributed frailty was fit using the bladder cancer dataset.

C. Estimated HR and confidence interval were quite similar to the counting process results.

D. Estimated frailty component was significant ($P = 0.003$).

XI. A Second Example (pages 391–395)

A. Clinical trial (n = 43, 8-year study) on effect of using high doses of antioxidants and zinc (i.e., **tx** = 1 if yes, 0 if no) to prevent age-related macular degeneration.

B. Covariates: **age** and **sex**.

C. Two possible events:

1. First event: visual acuity score <50 (i.e., poor vision).
2. Second event: clinically advanced stage of macular degeneration.

D. Focus on **Stratified CP** vs. **Marginal** because events are of different types.

E. Interaction SC model significant when compared to no-interaction SC model.

F. Conclusions regarding 1st event:

1. No treatment effect (HR = 0.946, P = 0.85).
2. Same for **Stratified CP** and **Marginal** approaches.

G. Conclusions regarding 2nd event.

1. **Stratified CP:** $\widehat{\text{HR}} = 0.385 = 1/2.60$, two-sided P = 0.03.
2. **Marginal:** $\widehat{\text{HR}} = 0.423 = 1/2.36$, two-sided P = 0.06).
3. Overall, clinically moderate and statistically significant treatment effect.

H. **Marginal** approach preferred because 1st and 2nd events are different types.

XII. Survival Curves with Recurrent Events (pages 395–398)

A. Survival plots with recurrent events only make sense when the focus is on one ordered event at a time.

B. For survival from a 1st event, the survival curve is given by $S_1(t) = \Pr(T_1 > t)$ where T_1 = survival time up to occurrence of the 1st event (ignores later recurrent events).

C. For survival from the kth event, the survival curve is given by $S_k(t) = \Pr(T_k > t)$ where T_k = survival time up to occurrence of the kth event).

D. Two versions for $S_k(t)$:
   i. **$S_{kc}(t)$ Stratified:** T_{kc} = time from the (k-1)st to kth event, restricting data to subjects with k-1 events.
   ii. **$S_{km}(t)$ Marginal:** T_{km} = time from study entry to kth event, ignoring previous events.

E. Illustration of survival plots for recurrent event data using a small dataset involving three subjects Molly (M), Holly (H), and Polly (P).

XIII. Summary (pages 398–401)

A. Four approaches for analyzing recurrent event survival data: the **counting process (CP), Stratified CP, Gap Time**, and **Marginal** approaches.

B. Data layouts differ for each approach.

C. **CP** approach uses Cox PH model; other approaches use Cox SC model.

D. Choice of approach depends in general on carefully considering the interpretation of each approach.

E. Should use **robust estimation** to adjust for correlation of observations on the same subject.

Practice Exercises

Answer questions 1 to 15 as true or false (circle T or F).

T F 1. A recurrent event is an event (i.e., failure) that can occur more than once over the follow-up on a given subject.

T F 2. The **Counting Process (CP)** approach is appropriate if a given subject can experience more than one different type of event over follow-up.

T F 3. In the data layout for the **CP** approach, a subject who has additional follow-up time after having failed at time $t_{(f)}$ does not drop out of the risk set after $t_{(f)}$.

T F 4. The **CP** approach requires the use of a stratified Cox (SC) PH model.

T F 5. Using the **CP** approach, if exactly two subjects fail at month $t_{(f)} = 10$, but both these subjects have later recurrent events, then the number in the risk set at the next ordered failure time does not decrease because of these two failures.

T F 6. The goal of **robust estimation** for the **CP** approach is to adjust estimated regression coefficients to account for the correlation of observations within subjects when previously no such correlation was assumed.

T F 7. **Robust estimation** is recommended for the **CP** approach but not for the alternative SC approaches for analyzing recurrent event survival data.

T F 8. The p-value obtained from using a robust standard error will always be larger than the corresponding p-value from using a nonrobust standard error.

T F 9. The **Marginal** approach uses the exact same (start, stop) data layout format used for the **CP** approach, except that for the **Marginal** approach, the model used is a stratified Cox PH model variable rather than a standard (unstratified) PH model.

T F 10. Suppose the maximum number of failures occurring for a given subject is **five** in a dataset to be analyzed using the **Marginal** approach. Then a subject who failed only twice will contribute five lines of data corresponding to his or her two failures and the three additional failures that could have possibly occurred for this subject.

T F 11. Suppose the maximum number of failures occurring for a given subject is **five** in a dataset to be analyzed using the **Stratified CP** approach. Then an interaction SC model used to carry out this analysis will have the following general model form: $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g}X_1 + \beta_{2g}X_2 + \cdots + \beta_{pg}X_p]$, $g = 1,2,3,4,5$.

T F 12. Suppose a no-interaction SC model using the **Stratified CP** approach is found (using a likelihood ratio test) not statistically different from a corresponding interaction SC model. Then if the no-interaction model is used, it will not be possible to separate out the effects of predictors within each stratum representing the recurring events on a given subject.

T F 13. In choosing between the **Stratified CP** and the **Marginal** approaches, the **Marginal** approach would be preferred provided the different strata clearly represent different event types.

T F 14. When using an interaction SC model to analyze recurrent event data, the estimated regression coefficients and corresponding standard errors for the first stratum always will be identical for the **Stratified CP, Gap Time,** and **Marginal** approaches.

T F 15. The choice among the **CP, Stratified CP, Gap Time**, and **Marginal** approaches depends upon whether a no-interaction SC or an interaction SC model is more appropriate for one's data.

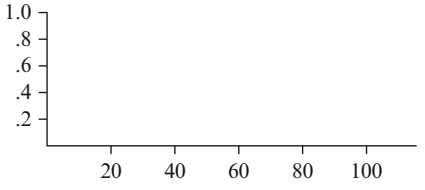
16. Suppose that Allie (A), Sally (S), and Callie (C) are the only three subjects in the dataset shown below. All three subjects have two recurrent events that occur at different times.

ID	Status	Stratum	Start	Stop	tx
A	1	1	0	70	1
A	1	2	70	90	1
S	1	1	0	20	0
S	1	2	20	30	0
C	1	1	0	10	1
C	1	2	10	40	1

Fill in the following data layout describing survival (in weeks) **to the first event (stratum 1)**. Recall that m_f and q_f denote the number of failures and censored observations at time $t_{(f)}$. The survival probabilities in the last column use the KM product limit formula.

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_1(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

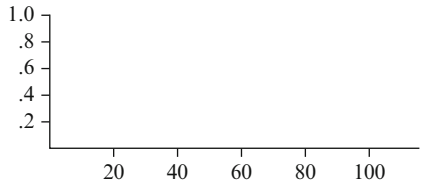
17. Plot the survival curve that corresponds to the data layout obtained for Question 16.



18. Fill in the following data layout describing survival (in weeks) **from the first to second event** using the **Gap Time** approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_2(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

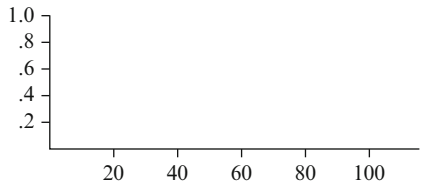
19. Plot the survival curve that corresponds to the data layout obtained for Question 18.



20. Fill in the following data layout describing survival (in weeks) **to the second event** using the **Marginal** approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_2(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
30	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

21. Plot the survival curve that corresponds to the data layout obtained for Question 20.



22. To what extent do the three plots obtained in Questions 17, 19, and 21 differ? Explain briefly.

Test

1. Suppose that Bonnie (B) and Lonnie (L) are th only two subjects in the dataset shown below, wher both subjects have two recurrent events that occur ar different times.

ID	Status	Stratum	Start	Stop
B	1	1	0	12
B	1	2	12	16
L	1	1	0	20
L	1	2	20	23

   a. Fill in the empty cells in the following data layout describing survival time (say, in weeks) to the first event (stratum 1):

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	2	0	0	{B, L}
12				
20				

   b. Why will the layout given in part a be the same regardless of whether rhe analysis approach is the Counting Process (CP), Stratified CP, Gap Time, or Marginal approaches?

   c. Fill in the empty cells in the following data layout describing survival time (say, in weeks) from the first to the second event (stratum 2) using the Stratified CP approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	0	0	0	-
16				
23				

   d. Fill in the empty cells in the following data layout describing survival time (say, in weeks) from the first to the second event (stratum 2) using the Gap Time approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	2	0	0	{B, L}
3				
4				

e. Fill in the empty cells in the following data layout describiing survival time (say, in weeks) from the first to the second event (stratum 2) using the Marginal approach:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	2	0	0	{B, L}
16				
23				

f. For the **Stratified CP** approach described in part c, determine which of the following choices is correct. Circle the number corresponding to the one and only one correct choice.
   i. Lonnie is in the risk set when Bonnie gets her second event.
   ii. Bonnie is in the risk set when Lonnie gets her second event.
   iii. Neither is in the risk set for the other's second event.

g. For the **Gap Time** approach described in part d, determine which of the following choices is correct. Circle the number corresponding to the one and only one correct choice.
   i. Lonnie is in the risk set when Bonnie gets her second event.
   ii. Bonnie is in the risk set when Lonnie gets her second event.
   ii. Neither is in the risk set for the other's second event.

h. For the **Marginal** approach described in part e, determine which of the following choices is correct. Circle the number corresponding to the one and only one correct choice.
   i. Lonnie is in the risk set when Bonnie gets her second event.
   ii. Bonnie is in the risk set when Lonnie gets her second event.
   iii. Neither is in the risk set for the other's second event.

2. The dataset shown below in the counting process layout comes from a clinical trial involving 36 heart attack patients between 40 and 50 years of age with implanted defibrillators who were randomized to one of two treatment groups (**tx**, = 1 if treatment A, = 0 if treatment B) to reduce their risk for future heart attacks over a 4-month period. The event of interest was experiencing a "high energy shock" from the defibrillator. The outcome is time (in days) until an event occurs. The covariate of interest was Smoking History (1 = ever smoked, 0 = never smoked). Questions about the analysis of this dataset follow.

Col 1 = id, Col 2 = event, Col 3 = start, Col 4 = stop, Col 5 = tx, Col 6 = smoking

01	1	0	39	0	0	12	1	0	39	0	1
01	1	39	66	0	0	12	1	39	80	0	1
01	1	66	97	0	0	12	0	80	107	0	1
02	1	0	34	0	1	13	1	0	36	0	1
02	1	34	65	0	1	13	1	36	64	0	1
02	1	65	100	0	1	13	1	64	95	0	1
03	1	0	36	0	0	14	1	0	46	0	1
03	1	36	67	0	0	14	1	46	77	0	1
03	1	67	96	0	0	14	0	77	111	0	1
04	1	0	40	0	0	15	1	0	61	0	1
04	1	40	80	0	0	15	1	61	79	0	1
04	0	80	111	0	0	15	0	79	111	0	1
05	1	0	45	0	0	16	1	0	57	0	1
05	1	45	68	0	0	16	0	57	79	0	1
05	.	68	.	0	0	16	.	79	.	0	1
06	1	0	33	0	1	17	1	0	37	0	1
06	1	33	66	0	1	17	1	37	76	0	1
06	1	66	96	0	1	17	0	76	113	0	1
07	1	0	34	0	1	18	1	0	58	0	1
07	1	34	67	0	1	18	1	58	67	0	1
07	1	67	93	0	1	18	0	67	109	0	1
08	1	0	39	0	1	19	1	0	58	1	1
08	1	39	72	0	1	19	1	58	63	1	1
08	1	72	102	0	1	19	1	63	106	1	1
09	1	0	39	0	1	20	1	0	45	1	0
09	1	39	79	0	1	20	1	45	72	1	0
09	0	79	109	0	1	20	1	72	106	1	0
10	1	0	36	0	0	21	1	0	48	1	0
10	1	36	65	0	0	21	1	48	81	1	0
10	1	65	96	0	0	21	1	81	112	1	0
11	1	0	39	0	0	22	1	0	38	1	1
11	1	39	78	0	0	22	1	38	64	1	1
11	1	78	108	0	0	22	1	64	97	1	1

(*Continued on next page*)

23	1	0	51	1	1	30	1	0	57	1	0
23	1	51	69	1	1	30	1	57	78	1	0
23	0	69	98	1	1	30	1	78	99	1	0
24	1	0	43	1	1	31	1	0	44	1	1
24	1	43	67	1	1	31	1	44	74	1	1
24	0	67	111	1	1	31	1	74	96	1	1
25	1	0	46	1	0	32	1	0	38	1	1
25	1	46	66	1	0	32	1	38	71	1	1
25	1	66	110	1	0	32	1	71	105	1	1
26	1	0	33	1	1	33	1	0	38	1	1
26	1	33	68	1	1	33	1	38	64	1	1
26	1	68	96	1	1	33	1	64	97	1	1
27	1	0	51	1	1	34	1	0	38	1	1
27	1	51	97	1	1	34	1	38	63	1	1
27	0	97	115	1	1	34	1	63	99	1	1
28	1	0	37	1	0	35	1	0	49	1	1
28	1	37	79	1	0	35	1	49	70	1	1
28	1	79	93	1	0	35	0	70	107	1	1
29	1	0	41	1	1	36	1	0	34	1	1
29	1	41	73	1	1	36	1	34	81	1	1
29	0	73	111	1	1	36	1	81	97	1	1

Table T.1 below provides the results for the treatment variable (**tx**) from no-interaction models over all four recurrent event analysis approaches. Each model was fit using either a Cox PH model (**CP** approach) or a Stratified Cox (SC) PH model (**Stratified CP, Gap Time, Marginal** approaches) that controlled for the covariate **smoking**.

Table T.1. Comparison of Results for the Treatment Variable (**tx**) Obtained from No-Interaction Models[a] Across Four Methods (Defibrillator Study)

Model	**CP**	**Stratified CP**	**Gap Time**	**Marginal**
Parameter estimate[b]	0.0839	0.0046	−0.0018	−0.0043
Robust standard error	0.1036	0.2548	0.1775	0.2579
Chi-square	0.6555	0.0003	0.0001	0.0003
p-value	0.4182	0.9856	0.9918	0.9866
Hazard ratio	1.087	1.005	0.998	0.996
95% confidence interval	(0.888, 1.332)	(0.610, 1.655)	(0.705, 1.413)	(0.601, 1.651)

[a] No-interaction SC model fitted with PROC PHREG for the **Stratified CP, Gap Time** and **Marginal** methods; no-interaction standard Cox PH model fitted for **CP** approach.
[b] Estimated coefficient of **tx** variable.

2. a. State the hazard function formula for the no-interaction model used to fit the **CP** approach.
   b. Based on the **CP** approach, what do you conclude about the effect of treatment (tx)? Explain briefly using the results in Table T.1.
   c. State the hazard function formulas for the no-interaction and interaction SC models corresponding to the use of the **Marginal** approach for fitting these data.
   d. Table T.1 gives results for "no-interaction" SC models because likelihood ratio (LR) tests comparing a "no-interaction" with an "interaction" SC model were not significant. Describe the (LR) test used for the marginal model (full and reduced models, null hypothesis, test statistic, distribution of test statistic under the null).
   e. How can you criticize the use of a no-interaction SC model for any of the SC approaches, despite the finding that the above likelihood ratio test was not significant?
   f. Based on the study description given earlier, why does it make sense to recommend the **CP** approach over the other alternative approaches?
   g. Under what circumstances/assumptions would you recommend using the **Marginal** approach instead of the **CP** approach?

Table T.2 below provides ordered failure times and corresponding risk set information that result for the 36 subjects in the above Defibrillator Study dataset using the Counting Process (**CP**) data layout format.

Table T.2. Ordered Failure Times and Risk Set Information for Defibrillator Study (CP)

Ordered failure times $t_{(f)}$	# in risk set n_f	# failed m_f	# censored in $[t_{(f)}, t_{(f+1)})$	Subject ID #s for outcomes in $[t_{(f)}, t_{(f+1)})$
0	36	0	0	—
33	36	2	0	6, 26
34	36	3	0	2, 7, 36
36	36	3	0	3, 10, 13
37	36	2	0	17, 28
38	36	4	0	22, 32, 33, 34
39	36	5	0	1, 8, 9, 11, 12
40	36	1	0	4
41	36	1	0	29
43	36	1	0	24
44	36	1	0	31

(*Continued on next page*)

Table T.2. (*Continued*)

Ordered failure times $t_{(f)}$	# in risk set n_f	# failed m_f	# censored in $[t_{(f)}, t_{(f+1)})$	Subject ID #s for outcomes in $[t_{(f)}, t_{(f+1)})$
45	36	2	0	5, 20
46	36	2	0	14, 25
48	36	1	0	21
49	36	1	0	35
51	36	2	0	23, 27
57	36	2	0	16, 30
58	36	2	0	18, 19
61	36	1	0	15
63	36	2	0	19, 34
64	36	3	0	13, 22, 33
65	36	2	0	2, 10
66	36	3	0	1, 6, 25
67	36	4	0	3, 7, 18, 24
68	36	2	0	5, 26
69	35	1	0	23
70	35	1	0	35
71	35	1	0	32
72	35	2	0	8, 20
73	35	1	0	29
74	35	1	0	31
76	35	1	0	17
77	35	1	0	14
78	35	2	0	11, 30
79	35	3	1	9, 15, 16, 28
80	34	2	0	4, 12
81	34	2	0	21, 36
93	34	2	0	7, 28
95	32	1	0	13
96	31	5	0	3, 6, 10, 26, 31
97	26	5	0	1, 22, 27, 33, 36
98	22	0	1	23
99	21	2	0	30, 34
100	19	1	0	2
102	18	1	0	8
105	17	1	0	32
106	16	2	0	19, 20
107	14	1	1	12, 35
108	12	1	0	11
109	11	0	2	9, 18
110	9	1	0	25
111	8	0	5	4, 14, 15, 24, 29
112	3	1	0	21
113	2	0	1	17
115	1	0	1	27

h. In Table T.2, why does the number in the risk set ($\mathbf{n}_f$) remain unchanged through failure time (i.e., day) 68, even though 50 events occur up to that time?

i. Why does the number in the risk set change from 31 to 26 when going from time 96 to 97?

j. Why is the number of failures ($\mathbf{m}_f$) equal to 3 and the number of censored subjects equal to 1 in the interval between failure times 79 and 80?

k. What 5 subjects were censored in the interval between failure times 111 and 112?

l. Describe the event history for subject #5, including his or her effect on changes in the risk set.

Based on the **CP** data layout of Table T.2, the following table (T.3) of survival probabilities has been calculated.

Table T.3. Survival Probabilities for Defibrillator Study Data Based on CP Layout

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)})\Pr(T > t_{(f)} \mid T \geq t_{(f)})$
0	36	0	0	1.0
33	36	2	0	$1 \times 34/36 = .94$
34	36	3	0	$.94 \times 33/36 = .87$
36	36	3	0	$.87 \times 33/36 = .79$
37	36	2	0	$.79 \times 34/36 = .75$
38	36	4	0	$.75 \times 32/36 = .67$
39	36	5	0	$.67 \times 31/36 = .57$
40	36	1	0	$.57 \times 35/36 = .56$
41	36	1	0	$.56 \times 35/36 = .54$
43	36	1	0	$.54 \times 35/36 = .53$
44	36	1	0	$.53 \times 35/36 = .51$
45	36	2	0	$.51 \times 34/36 = .48$
46	36	2	0	$.48 \times 34/36 = .46$
48	36	1	0	$.46 \times 35/36 = .44$
49	36	1	0	$.44 \times 35/36 = .43$
51	36	2	0	$.43 \times 34/36 = .41$
57	36	2	0	$.41 \times 34/36 = .39$
58	36	2	0	$.39 \times 34/36 = .36$
61	36	1	0	$.36 \times 35/36 = .35$
63	36	2	0	$.35 \times 34/36 = .33$
64	36	3	0	$.33 \times 33/36 = .31$
65	36	2	0	$.31 \times 34/36 = .29$
66	36	3	0	$.29 \times 33/36 = .27$
67	36	4	0	$.27 \times 32/36 = .24$
68	36	2	0	$.24 \times 34/36 = .22$
69	35	1	0	$.22 \times 34/35 = .22$
70	35	1	0	$.22 \times 34/35 = .21$
71	35	1	0	$.21 \times 34/35 = .20$
72	35	2	0	$.20 \times 33/35 = .19$
73	35	1	0	$.19 \times 34/35 = .19$
74	35	1	0	$.19 \times 34/35 = .18$
76	35	1	0	$.18 \times 34/35 = .18$
77	35	1	0	$.18 \times 34/35 = .17$
78	35	2	0	$.17 \times 33/35 = .16$

(*Continued on next page*)

Table T.3. (*Continued*)

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)})Pr(T > t_{(f)} \mid T \geq t_{(f)})$
79	35	3	1	$.16 \times 31/35 = .14$
80	34	2	0	$.14 \times 32/34 = .13$
81	34	2	0	$.13 \times 32/34 = .13$
95	32	1	0	$.13 \times 31/32 = .12$
96	31	5	0	$.12 \times 26/31 = .10$
97	26	5	0	$.10 \times 21/26 = .08$
98	22	0	1	$.08 \times 22/22 = .08$
99	21	2	0	$.08 \times 19/21 = .07$
100	19	1	0	$.07 \times 18/19 = .07$
102	18	1	0	$.07 \times 17/18 = .06$
105	17	1	0	$.06 \times 16/17 = .06$
106	16	2	0	$.06 \times 14/16 = .05$
107	14	1	1	$.05 \times 13/14 = .05$
108	12	1	0	$.05 \times 21/26 = .05$
109	11	0	2	$.05 \times 11/11 = .05$
110	9	1	0	$.05 \times 8/9 = .04$
111	8	0	5	$.04 \times 8/8 = .04$
112	3	1	0	$.04 \times 2/3 = .03$
113	2	0	1	$.03 \times 2/2 = .03$
115	1	0	1	$.03 \times 1/1 = .03$

Suppose the survival probabilities shown in Table T.3 are plotted on the y-axis versus corresponding ordered failure times on the x-axis.

m. What is being plotted by such a curve? (Circle one or more choices.)
   i. $Pr(T_1 > t)$ where T_1 = time to first event from study entry.
   ii. $Pr(T > t)$ where T = time from any event to the next recurrent event.
   iii. $Pr(T > t)$ where T = time to any event from study entry.
   iv. Pr(not failing prior to time t).
   v. None of the above.

n. Can you criticize the use of the product limit formula for $S(t_{(f)})$ in Table T.3? Explain briefly.

o. Use Table T.2 to complete the data layouts for plotting the following survival curves.

i. $S_1(t) = Pr(T_1 > t)$ where T_1 = time to first event from study entry

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)}) \times Pr(T_1 > t \mid T_1 \geq t)$
0	36	0	0	1.00
33	36	2	0	0.94
34	34	3	0	0.86
36	31	3	0	0.78
37	28	2	0	0.72
38	26	4	0	0.61
39	22	5	0	0.47
40	17	1	0	0.44
41	16	1	0	0.42
43	15	1	0	0.39
44	14	1	0	0.36
45	13	2	0	0.31
46	11	2	0	0.25
48	9	1	0	0.22
49	8	1	0	0.19
51	-	-	-	-
57	-	-	-	-
58	-	-	-	-
61	-	-	-	-

ii. **Gap Time** $S_{2c}(t) = Pr(T_{2c} > t)$ where T_{2c} = time to second event from first event.

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)}) \times Pr(T_1 > t \mid T_1 \geq t)$
0	36	0	0	1.00
5	36	1	0	0.97
9	35	1	0	0.94
18	34	2	0	0.89
20	32	1	0	0.86
21	31	2	1	0.81
23	28	1	0	0.78
24	27	1	0	0.75
25	26	1	0	0.72
26	25	2	0	0.66
27	23	2	0	0.60
28	21	1	0	0.58
29	20	1	0	0.55
30	19	1	0	0.52

(*Continued on next page*)

(Continued)

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)}) \times Pr(T_1 > t \mid T_1 \geq t)$
31	18	3	0	0.43
32	15	1	0	0.40
33	14	5	0	0.26
35	9	1	0	0.23
39	8	2	0	0.17
40	-	-	-	-
41	-	-	-	-
42	-	-	-	-
46	-	-	-	-
47	-	-	-	-

iii. **Marginal** $S_{2m}(t) = Pr(T_{2m} > t)$ where T_{2m} = time to second event from study entry.

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S(t_{(f-1)}) \times Pr(T_1 > t \mid T_1 \geq t)$
0	36	0	0	1.00
63	36	2	0	0.94
64	34	3	0	0.86
65	31	2	0	0.81
66	29	3	0	0.72
67	26	4	0	0.61
68	22	2	0	0.56
69	20	1	0	0.53
70	19	1	0	0.50
71	18	1	0	0.47
72	17	2	0	0.42
73	15	1	0	0.39
74	14	1	0	0.36
76	13	1	0	0.33
77	12	1	0	0.31
78	11	2	0	0.25
79	-	-	-	-
80	-	-	-	-
81	-	-	-	-
97	-	-	-	-

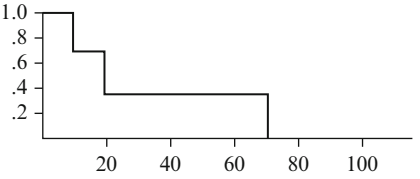
p. The survival curves corresponding to each of the data layouts (a, b, c) described in Question 14 will be different. Why?

Answers to Practice Exercises

1. T
2. F: The **Marginal** approach is appropriate if events are of different types.
3. T
4. F: The **Marginal, Stratified CP**, and **Gap Time** approaches all require a SC model, whereas the **CP** approach requires a standard PH model.
5. T
6. F: Robust estimation adjusts the **standard errors** of regression coefficients.
7. F: Robust estimation is recommended for all four approaches, not just the **CP** approach.
8. F: The P-value from robust estimation may be either larger or smaller than the corresponding P-value from nonrobust estimation.
9. F: Replace the word **Marginal** with **Stratified CP** or **Gap Time**. The Marginal approach does not use (Start, Stop) columns in its layout.
10. T
11. T
12. T
13. T
14. T
15. F: The choice among the **CP, Stratified CP, Gap Time**, and **Marginal** approaches depends on carefully considering the interpretation of each approach.
16.

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$	$S_1(t_{(f)})$
0	3	0	0	{A, S, C}	1.00
10	3	1	0	{A, S, C}	0.67
20	2	1	0	{A, S}	0.33
70	1	1	0	{A}	0.00

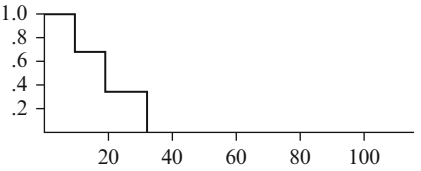
17. $\mathbf{S_1(t)}$



18.

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)}\}$	$S_2(t_{(f)})$ Gap Time
0	3	0	0	{A, S, C}	1.00
10	3	1	0	{A, S, C}	0.67
20	2	1	0	{A, C}	0.33
30	1	1	0	{C}	0.00

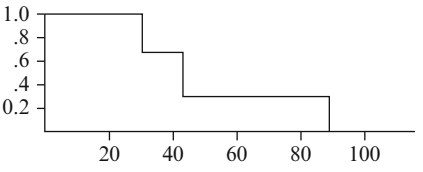
19. $\mathbf{S_{2c}(t)}$ **Gap Time**



20.

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)}\}$	$S_2(t_{(f)})$ Marginal
0	3	0	0	{A, S, C}	1.00
30	3	1	0	{A, S, C}	0.67
40	2	1	0	{A, C}	0.33
90	1	1	0	{A}	0.00

21. $\mathbf{S_{2m}(t)}$ **Marginal**



22. All three plots differ because the risk sets for each plot are defined differently inasmuch as the failure times are different for each plot.

9 Competing Risks Survival Analysis

D.G. Kleinbaum and M.Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_9,
© Springer Science+Business Media, LLC 2012

Introduction

This chapter considers survival data in which each subject can experience only one of several different types of events over follow-up. This situation contrasts with the topic of the preceding chapter in which subjects could experience more than one event of a given type. When only one of several different types of event can occur, we refer to the probabilities of these events as "competing risks," which explains the title of this chapter.

Modeling competing risks survival data can be carried out using a Cox model, a parametric survival model, or models that use the *cumulative incidence* (rather than survival). In this chapter, we mainly consider the Cox model because of its wide popularity and also because of the availability of computer programs that use the Cox model for analysis of competing risks.

The typical ("cause-specific") approach for analyzing competing risks data is to perform a survival analysis for each event type separately, where the other (competing) event types are treated as censored categories. There are two primary drawbacks to the above method. One problem is that the above method requires the assumption that competing risks are independent. The second problem is that the generalized Kaplan–Meier (KM)-based product-limit survival curve obtained from fitting separate Cox models for each event type has questionable interpretation when there are competing risks.

Unfortunately, if the independence assumption is incorrect, there is no direct methodology available for analyzing competing risks simultaneously. The only "indirect" method for addressing this problem involves carrying out a "sensitivity analysis" that treats subjects with events from competing risks as all being event-free or as all experiencing the event of interest. An example of this "sensitivity" approach is provided.

The primary alternative summary curve to the KM-based survival curve is the "cumulative incidence curve (CIC)," which estimates the "marginal probability" of an event (both terms are defined in this chapter). This CIC is not estimated using a product-limit formulation, and its computation is not included in mainstream statistical packages. Moreover, the independence of competing risks is still required when a proportional hazard model is used to obtain hazard ratio estimates for individual competing risks as an intermediate step in the computation of a CIC. Nevertheless, the CIC has a meaningful interpretation in terms of treatment utility regardless of whether competing risks are independent. A variation of the CIC, called the "conditional probability curve (CPC)," provides a risk probability conditional on an individual not experiencing any of the other competing risks by time t.

An equivalent approach to the cause-specific method for analyzing competing risks is called the Lunn–McNeil (LM) approach. The LM approach allows only one model to be fit rather than separate models for each event type and, moreover, allows flexibility to perform statistical inferences about simpler versions of the LM model. This approach has added appeal in that competing events are not considered as simply being censored. Nevertheless, as with the cause-specific approach, the LM method assumes the independence of competing risks.

Abbreviated Outline

The outline below gives the user a preview of the material covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Overview** (page 430)
II. **Examples of Competing Risks Data** (pages 430–432)
III. **Byar Data** (pages 433–434)
IV. **Method 1: Separate Models for Different Event Types** (pages 434–437)
V. **The Independence Assumption** (pages 437–443)
VI. **Cumulative Incidence Curves (CIC)** (pages 444–453)
VII. **Conditional Probability Curves (CPC)** (pages 453–455)
VIII. **Method 2—The Lunn-McNeil (LM) Approach** (pages 455–461)
IX. **Method 2a: Alternative Lunn–McNeil (LM_{alt}) Approach** (pages 461–464)
X. **Method 1 (Separate Models) versus Method 2 (LM Approach)** (pages 465–468)
XI. **Summary** (pages 468–473)

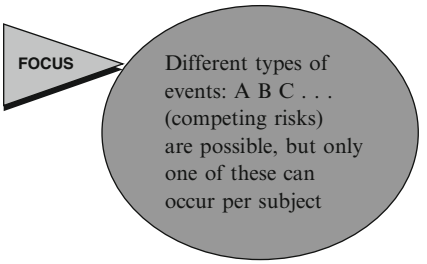
Objectives

Upon completing this chapter, the learner should be able to:

1. State or recognize examples of competing risks survival data.
2. Given competing risks data, outline the steps needed to analyze such data using separate Cox models.
3. Given computer output from the analysis of competing risk data, carry out an analysis to assess the effects of explanatory variables on one or more of the competing risks.
4. State or describe the independence assumption typically required in the analysis of competing risks data.
5. Describe how to carry out and/or interpret a "sensitivity analysis" to assess the independence assumption about competing risks.
6. State why a survival function obtained from competing risk data using the Cox model has a questionable interpretation.
7. State or describe the "cumulative incidence" approach for analyzing competing risks data.
8. Given competing risk data, describe how to calculate a CIC and/or a CPC curve.
9. Given competing risks data, outline the steps needed to analyze such data using the Lunn–McNeil method.
10. Given computer output from fitting either a LM or LM_{alt} model, carry out an analysis to assess the effect of explanatory variables on one or more of the competing risks.

Presentation

I. Overview



In this chapter, we consider survival data in which each subject can experience only one of different types of events over follow-up. The probabilities of these events are typically referred to as **competing risks**. We describe how to use the Cox PH model to analyze such data, the drawbacks to this approach, and some approaches for addressing these drawbacks.

II. Examples of Competing Risks Data

1. Dying from either lung cancer or stroke
2. Advanced cancer patients either dying from surgery or getting hospital infection
3. Soldiers dying in accident or in combat
4. Limb sarcoma patients developing local recurrence, lung metastasis, or other metastasis over follow-up

Competing risks occur when there are at least two possible ways that a person can fail, but only one such failure type can actually occur. For example,

1. A person can die from lung cancer or from a stroke, but not from both (although he can have both lung cancer and atherosclerosis before he dies);
2. Patients with advanced-stage cancer may die after surgery before their hospital stay is long enough for them to get a hospital infection;
3. Soldiers in war may die during combat or may die by (e.g., traffic) accident;
4. In a clinical trial, patients with nonmetastatic limb sarcoma undergoing chemotherapy and surgery might develop a local recurrence, lung metastasis, or other metastasis after follow-up.

Each example above allows only **one** event out of several possible events to occur per subject

If event not death, then recurrent events are possible

Competing risks + recurrent events beyond scope of this chapter

For each of the above examples, the possible events of interest differ, but only one such event can occur per subject. Note, however, if at least one of the possible event types does not involve death, it is also possible that such events can recur over follow-up. Thus, although the analysis of recurrent events that also involves competing risks may be required, this more complex topic is beyond the scope of this chapter (see Tai et al., 2001).

Objective: assess

$X_1, X_2, \ldots, X_p \Rightarrow$ Failure rate (survival probability)

for any one event allowing for competing risks from other possible events

A logical objective for competing risks data is to assess the relationship of relevant predictors to the failure rate or corresponding survival probability of any **one** of the possible events allowing for the competing risks of the other ways to fail.

Another objective

Compare hazard rate for event A with hazard rate for event B

We might also want to compare the failure rates (e.g., using a hazard ratio) for two or more possible events, controlling for relevant predictors.

Lung Cancer vs. Stroke (1)

HR_{LC}(E vs. not E) = 1? (allowing for competing risk from stroke)

In the lung cancer versus stroke example above, we might ask whether the lung cancer death rate in "exposed" persons is different from the lung cancer rate in "unexposed" persons, allowing for the possibility that subjects could have died from a stroke instead.

HR(LC vs. Stroke) = 1? (controlling for predictors)

We might also want to know if the lung cancer death rate differs from the stroke death rate controlling for predictors of interest.

Surgery Death vs. Hospital Infection (2)

$HR_{HOSPINF}$(E vs. not E) = 1? (allowing for competing risk from surgery)

Note: death from surgery reduces number of hospital infections to be treated

In the second example, the competing risks are death from surgery versus development of a hospital infection. For infection control investigators, the hospital infection event is of primary interest. Nevertheless, the occurrence of death from surgery reduces the burden of hospital infection control required. Thus, the estimation of hospital infection rates are complicated by the competing risk of death from surgery.

Accidental Death vs. Combat Death (3)

HR_{COMBAT} (E vs. not E) (allowing competing risk of accidental death)

Suppose entire company dies at accident time t before entering combat

$\Downarrow$

$S_{COMBAT}(t) = P(T_{COMBAT} > t) = 1$

where T_{COMBAT} = time to combat death

The third example involves competing risks of death from either combat or accident in a company of soldiers. Here, primary interest concerns the hazard ratio for combat death comparing two exposure groups. Suppose the entire company dies at time t in a helicopter accident on their way to a combat area. Because no one died in combat by time t, the survival probability of not dying in combat is one, even though no combat took place.

However,

T_{C+A} = combat or accidental death

$\Downarrow$

"event free" $S_{C+A}(t) = P(T_{C+A} > t) = 0$

However, if we define the outcome of interest as death from either combat or accident, the "event free" survival probability is zero after the accident occurred (at time t).

Moreover,

$$S_{KM}(T_{COMBAT} > t)$$

is undefined because the risk set is empty at time t

Moreover, the KM survival probability for combat death at time t is undefined because no one was at risk for a combat death at time t.

Competing Risks Data Survival Curve Interpretation?

This example points out that when there are competing risks, the interpretation of a survival curve may be difficult or questionable (more on this issue later).

Limb sarcoma patients (4)

In the fourth example involving limb sarcoma patients, the competing risks are the three failure types shown at the left.

Competing risks
1 = local recurrence, 2 = lung metastasis, or 3 = other metastasis

In this study, the investigators wanted hazard ratios for each failure type, allowing for competing risks from the other two failure types.

$HR_{\mathbf{c}}$(E vs. not E), $\mathbf{c} = 1, 2, 3$ (allowing for competing risk from other two failure types)

HR(Lung Metastasis vs. Local Recurrence)? Controlling for Predictors

It was also of interest to compare the failure rates for lung metastasis versus local recurrence (or any other two of the three failure types), controlling for relevant predictors.

No failure types involve death

$\Downarrow$

Recurrent events possible

Because none of the failure types involves death, recurrent events are possible for any of the three failure types. If, however, the information on only the first failure is targeted, the classical competing risk methodology described in this chapter can be applied.

But can use classical competing risk methods if focus on **only first failure**

III. Byar Data

- Randomized clinical trial
- Compare treatments for Stage III and IV prostate cancer
- Rx status: placebo or one of 3 dose levels of DES

We now introduce an example of competing risks survival analysis of data from a randomized clinical trial (Byar and Green, 1980) comparing treatments for prostate cancer. We henceforth refer to this as the **Byar data**. Patients with Stages III (local extension beyond the prostate gland) and IV (distant metastases, elevated acid phosphatase, or both) prostate cancer were randomized to receive either a placebo or one of three dose levels of the active treatment diethylstilbestrol (DES).

Competing risks: deaths from

Cancer (main focus)
CVD
Other

In this study, patients could die from prostate cancer, cardiovascular disease, or other causes. Covariate information was also collected to account for the possible influence of predictors on survival. These data have been analyzed extensively (Byar and Corle, 1977, Kay, 1986, and Lunn and McNeil, 1995). Some grouping of the predictors was considered to be clinically meaningful.

Covariate information collected

Some predictors grouped

Predictors	Value	Category
Treatment (Rx)	0	Placebo, 0.2 mg DES
	1	1.0,5 mg DES
Age at diagnosis Diagnosis (Age)	0	$\leq$74 years
	1	75–79 years
	2	$\geq$80 years
Standardized[a] weight (Wt)	0	$\geq$100
	1	80–99
	2	$>$80
Performance status (PF)	0	Normal
	1	Limitation of activity
History of CVD (Hx)	0	No
	1	Yes
Hemoglobin (Hg)	0	$\geq$ 12 g/100 ml
	1	9.0–11.9 g/100 ml
	2	$<$9 g/100 ml
Size of the primary lesion (SZ)	0	$<$30 cm^2
	1	$\geq$30 cm^2
Gleeson Score[+](SG)	0	$\leq$10
	1	$>$10

[a] weight (kg) – height (cm) + 200
[+] index of tumor invasiveness/ aggressiveness

Key risk factors related to the primary outcome of interest (cancer deaths) and the appropriate grouping is shown at the left.

Primary interest was to assess the effect of treatment (Rx) adjusted for relevant risk factors in the presence of the competing risks. Notice from the table that the Rx variable is grouped into a binary variable by coding subjects receiving the placebo or 0.2 mg of DES as 0 and coding subjects receiving 1.0 or 5.0 mg of DES as 1.

Independence assumption (discussed later)

Next

Analysis of competing risks survival data

Assume independent competing risks

From a clinical perspective, these three competing risks can be considered to be **independent** (e.g., failure from heart disease and/or other causes of death is unrelated to risk of failure from prostate cancer). We discuss this "independence assumption" in more detail in a subsequent section of this chapter.

We now describe the approach typically used to analyze competing risks data. This approach assumes that competing risks are independent. We illustrate this approach using the Byar data.

IV. Method 1: Separate Models for Different Event Types

- Use Cox (PH) model
- Estimate separate hazards or HRs for each failure type
- Other (competing) failure types are treated as censored
- Persons lost to follow-up or withdrawal are also censored

The typical approach for analyzing competing risks data uses the Cox (PH) model to separately estimate hazards and corresponding hazard ratios for each failure type, treating the other (competing) failure types as censored in addition to those who are censored from loss to follow-up or withdrawal. We refer to this approach as **Method 1** because we later describe an alternative approach (Method 2) that requires only a single model to be fit.

If only one failure type of interest

⇓

Estimate only one hazard or HR

If only one failure type is of primary interest, then the analysis might be restricted to estimating hazards or hazard ratios for that type only (but still treating the competing failure types as censored).

Cause-specific hazard function

$$h_c(t) = \lim_{\Delta t \to 0} P(t \le T_{\mathbf{c}} < t + \Delta t | T_{\mathbf{c}} \ge t)/\Delta t$$

where T_c = time-to-failure from event **c**

c = 1, 2, …, C (# of event types)

To describe this method mathematically, we define the **cause-specific hazard function** shown at the left. The random variable $T_{\mathbf{c}}$ denotes the time-to-failure from event type **c**. Thus, $h_{\mathbf{c}}(t)$ gives the instantaneous failure rate at time t for event type **c**, given not failing from event **c** by time t.

Cox PH cause-specific model (event-type **c**):

$$h_{\mathbf{c}}(t, \mathbf{X}) = h_{0\mathbf{c}}(t) \exp\left[\sum_{i=1}^{p} \beta_{ic} X_i\right], \quad \mathbf{c} = 1, \ldots, C$$

β_{ic} allows effect of X_i to differ by event-type

Using a Cox PH model that considers predictors $\mathbf{X} = (X_1, X_2, \ldots, X_p)$, the **cause-specific hazard model** for event type **c** has the form shown at the left. Note that $\beta_{\mathbf{ic}}$, the regression coefficient for the ith predictor, is subscripted by **c** to indicate that the effects of the predictors may be different for different event types.

Byar Data Example

Competing Risks: **Cancer, CVD, Other**

Cause-specific model: **Cancer**
No-interaction model:

$$h_{\mathbf{Ca}}(t, \mathbf{X}) = h_{0\mathbf{Ca}}(t)\exp[\beta_{1\mathbf{Ca}}Rx + \beta_{2\mathbf{Ca}}Age + \beta_{3\mathbf{Ca}}Wt + \beta_{4\mathbf{Ca}}PF + \beta_{5\mathbf{Ca}}Hx + \beta_{6\mathbf{Ca}}HG + \beta_{7\mathbf{Ca}}SZ + \beta_{8\mathbf{Ca}}SG]$$

$$HR_{\mathbf{Ca}}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1\mathbf{Ca}}]$$

CVD and **Other** deaths are censored

Cause-specific model: **CVD**

$$h_{\mathbf{CVD}}(t,\mathbf{X}) = h_{0\mathbf{CVD}}(t)\exp[\beta_{1\mathbf{CVD}}Rx + \beta_{2\mathbf{CVD}}Age + \beta_{3\mathbf{CVD}}Wt + \beta_{4\mathbf{CVD}}PF + \beta_{5\mathbf{CVD}}Hx + \beta_{6\mathbf{CVD}}HG + \beta_{7\mathbf{CVD}}SZ + \beta_{8\mathbf{CVD}}SG]$$

$$HR_{\mathbf{CVD}}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1\mathbf{CVD}}]$$

Cancer and **Other** are censored

Cause-specific model: **Other**

$$H_{\mathbf{OTH}}(t,\mathbf{X}) = h_{0\mathbf{OTH}}(t)\exp[\beta_{1\mathbf{OTH}}Rx + \beta_{2\mathbf{OTH}}Age + \beta_{3\mathbf{OTH}}Wt + \beta_{4\mathbf{OTH}}PF + \beta_{5\mathbf{OTH}}Hx + \beta_{6\mathbf{OTH}}HG + \beta_{7\mathbf{OTH}}SZ + \beta_{8\mathbf{OTH}}SG]$$

Cancer and **CVD** are censored

We illustrate the above model using the Byar data involving the three competing risks and the eight predictors.

A no-interaction cause-specific model for **Cancer** death (**Ca**) is shown at the left. From this model, the hazard ratio for the effect of Rx controlling for the other variables is $\exp[\beta_{1\mathbf{Ca}}]$.

Because **Cancer** is the event-type of interest, the two competing event-types, **CVD** and **Other**, need to be treated as censored in addition to usual censored observations (i.e., for persons who are either lost to follow-up or withdraw from the study).

Similarly, if **CVD** is the event-type of interest, the cause-specific no-interaction hazard model and the hazard ratio formula for the effect of treatment is shown at the left, and the event types **Cancer** and **Other** would be treated as censored.

And finally, if **Other** is the event-type of interest, the cause-specific no-interaction hazard model and the hazard ratio formula for the effect of treatment is shown at the left, and the event types **Cancer** and **CVD** would be treated as censored.

Edited output for each of the above three cause-specific models is now presented.

First, we show the results for the event type **Cancer**, treating **CVD** and **Other** as censored

Table 9.1. Edited Output for Cancer with **CVD** and **Other** Censored

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Rx	1	−0.550	0.170	0.001	0.577
Age	1	0.005	0.142	0.970	1.005
Wt	1	0.187	0.138	0.173	1.206
PF	1	0.253	0.262	0.334	1.288
Hx	1	−0.094	0.179	0.599	0.910
HG	1	0.467	0.177	0.008	1.596
SZ	1	1.154	0.203	0.000	3.170
SG	1	1.343	0.202	0.000	3.830

Log likelihood = − 771.174

$\widehat{HR}_{\mathbf{Ca}}(\text{RX} = 1 \text{ vs. RX} = 0)$
$= \exp(-0.550) = 0.577$
Wald ChiSq $= (-.550/.170)^2$
$= 10.345$ (P $= 0.001$)
Signif. below .01 level

From this output, the adjusted $\widehat{HR}$ for the effect of Rx is 0.577 (=1/1.73).

The P-value for a two-tailed Wald test is 0.001; thus Rx has a significant positive effect on survival for **Cancer** death with competing risks from **CVD** and **Other** deaths.

95% CI for $\exp[\beta_{1\mathbf{Ca}}]$:
$\exp[-0.550 \pm 1.96(0.170)]$
$= (0.413, 0.807)$

Also, the 95% confidence interval for this HR is (0.413, 0.807) = (1/2.43, 1/1.24).

Table 9.2. Edited Output for **CVD** with **Cancer** and **Other** Censored

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Rx	1	0.354	0.174	0.042	1.425
Age	1	0.337	0.134	0.012	1.401
Wt	1	0.041	0.150	0.783	1.042
PF	1	0.475	0.270	0.079	1.608
Hx	1	1.141	0.187	0.000	3.131
HG	1	0.018	0.202	0.929	1.018
SZ	1	−0.222	0.364	0.542	0.801
SG	1	−0.023	0.186	0.900	0.977

Log likelihood = −763.001

We next provide edited output when the event-type is **CVD**, treating **Cancer** and **Other** as censored.

$\widehat{HR}_{\mathbf{CVD}}(\text{RX} = 1 \text{ vs. RX} = 0)$
$= \exp(0.354) = 1.425$
Wald ChiSq $= (.354/.174)^2$
$= 4.220$ (P $= 0.042$)
Signif. at .05 level

Here, the adjusted $\widehat{HR}$ for the effect of Rx is 1.425.

The P-value for a two-tailed Wald test is 0.042; thus, Rx has a significant (P < .05) but negative effect on survival for **CVD** death with competing risks from **Cancer** and **Other** deaths.

95% CI for $\exp[\beta_{1\mathbf{CVD}}]$:
exp.$[0.354 \pm 1.96(0.174)]$
$= (1.013, 2.004)$

The 95% confidence interval for this HR is (1.013, 2.004).

Table 9.3. Edited Output for **Other** with **Cancer** and **CVD** Censored

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Rx	1	−0.578	0.279	0.038	0.561
Age	1	0.770	0.204	0.000	2.159
Wt	1	0.532	0.227	0.019	1.702
PF	1	0.541	0.422	0.200	1.718
Hx	1	0.023	0.285	0.935	1.023
HG	1	0.357	0.296	0.228	1.428
SZ	1	0.715	0.423	0.091	2.045
SG	1	−0.454	0.298	0.127	0.635

Log likelihood = −297.741

Last, we provide edited output when the event-type is **Other**, treating **Cancer** and **CVD** as censored.

$\hat{HR}_{\mathbf{OTH}}(RX = 1 \text{ vs. } RX = 0) = \exp(-0.580) = 0.560$

Here. the adjusted $\hat{HR}$ for the effect of Rx is 0.561 (= 1/ 1.78).

Wald ChiSq = $(-.578/.279)^2$ =4.29 (P=0.038)
Signif. at .05 level

The P-value for a two-tailed Wald test is .038; thus, Rx has a significant ($P < .05$) protective effect on survival for **Other** deaths with competing risks from **Cancer** and **CVD** deaths.

95% CI for $\exp[\beta_{1\mathbf{OTH}})$:
$\exp[-0.578 +- 1.96(0.279)]$
=(0.325, 0.969)

The 95% confidence interval for this HR is (0.325,0.969), which is somewhat imprecise.

Not assessed in the above analysis:

PH assumption

Interaction of Rx with control variables

We have thus completed a competing risk analysis af the Byar data assuming that a no-interaction Cox PH model is appropriate. We haven't actually checked the PH assumption for any of the variables in the model nor have we assessed whether there is significant interaction between Rx and the other variables being controlled. Typically, these situations should be explored to ensure a more appropriate analysis.

V. The Independence Assumption

Censoring: a major concern in survival analysis

At the beginning of this text in Chapter 1, we introduced the concept of censoring as a major concern for the analysis of survival data. We distinguished between right- and left-censoring and indicated our focus in the text would be on right-censoring, which occurs more often.

Right-censoring vs. left-censoring
↓
- more often
- our focus

Typical assumption:
Censoring is Independent

- Required for all approaches/ models described to this point
- Relevant for competing risks

We also introduced in Chapter 1 an important assumption about censoring that is typically assumed for all approaches/models for analyzing survival data described up to this point, including data with competing risks. This assumption is often stated as follows: **censoring is independent**.

Two other (different) assumptions
Random censoring
Non-informative censoring

In addition, we distinguished (in Chapter 1) "independent" censoring from two other assumptions: "random" censoring and "non-informative" censoring, but emphasized the importance of the independence assumption.

Independent censoring:

Chapter 1 context: no competing risks

$h(t \mid G, Ce) = h(t \mid G, NCe)$ where

G denotes any subgroup of subjects at risk at time t
$h(t \mid G, Ce)$ denotes hazard for censored subjects in subgroup G
$h(t \mid G, NCe)$ denotes hazard for noncensored subjects in subgroup G

Bias possible:
$\hat{S}(t)$ may over-estimate $S(t)$
if
large proportion of subjects censored at time t actually fail after time t

In Chapter 1, we defined **independent censoring** in a context that assumed the absence of competing risks as follows:

Within any subgroup of interest, subjects who are censored at time t should be representative of all subjects in that subgroup who remained at risk at time t with respect to their survival experience.

Non-independent censoring unfortunately can lead to **biased results** in a survival analysis. A bias can result if people who get censored are more likely to fail than those not censored. Thus, the estimated survival probability at any time t may over-estimate the true survival probability at time t if a large proportion of those with unknown status (i.e., censored) actually failed.

Independent censoring with competing risks

⇓

Different types of censoring:

- failure from competing risks
- lost-to-follow-up
- withdrawal
- end of study

When the survival analysis problem involves competing risks, the requirement of **independent censoring** has the additional complication that there are different types of censoring that are possible. That is, when focusing on the cause-specific hazard for event-type **A**, say, competing risks other than **A** are also considered as censored in addition to standard censorship from lost-to-follow-up, withdrawal or ending of the study.

EXAMPLE (Byar data)

3 competing risks:
Cancer, CVD, or Other deaths
Independent censoring?

Suppose censoring is **independent** and Harry censored at time t

⇓

$h_{\mathbf{Ca}}(t \mid G, Ce) = h_{\mathbf{Ca}}(t \mid G, NCe)$

Suppose censoring is **not independent** and Harry died from **CVD** or **Other** Cause at time t

⇓

$h_{\mathbf{Ca}}(t \mid G, Ce) \neq h_{\mathbf{Ca}}(t \mid G, NCe)$

For example, in the Byar data set, there were three competing risks of interest, **Cancer, CVD,** or **Other** deaths. What, then, must we assume if censoring in this study were independent?

Suppose **censoring is independent** and we focus on cause-specific deaths for **Cancer**, then any subject (e.g., Harry) in the risk set at time t with a given set (G) of covariates who is censored at time t is presumed to have the same failure rate as any noncensored subject in the risk set with the same set of covariates regardless of whether the reason for censoring is either a **CVD** or **Other** death, withdrawal from study, or lost-to-follow-up.

On the other hand, suppose **censoring were not independent,** then if Harry was censored at time t because he died from **CVD** or **Other** causes at time t, Harry's (unknown) failure rate at time t for dying of **Cancer** would differ from the Cancer failure rate for noncensored subjects at time t (who didn't previously die of **Cancer, CVD**, or **Other** causes prior to time t).

Important assumption for competing risks

Censoring is independent regardless of different types of censoring possible

Synonym: **Competing risks are independent**

The important message at this point when analyzing competing risks survival data is that it is typically assumed that **censoring is independent** regardless of the different ways that censoring can occur, including failure from competing risks other than the cause-specific event-type of interest. A synonymous expression is to say that **competing risks are independent**, which we henceforth adopt in our remaining discussion of this topic.

Questions about independence assumption

1. How can we determine whether this assumption is satisfied?
2. How can we proceed with the analysis to consider the possibility that the assumption is not satisfied?

So, if we typically require that competing risks are independent, (1) how can we determine whether this assumption is satisfied and (2) how can we proceed with the analysis to consider the possibility that the assumption is not satisfied?

Answer to 1:
We can never explicitly prove the assumption is satisfied for given data.

For example, Byar data: **Cancer** death
Then can't determine would have died from **Cancer** if hadn't died from **CVD**.

Unfortunately, the answer to the first question is that we can never explicitly prove that competing risks are or are not independent for a given dataset. For example, in the Byar dataset, we cannot determine for certain whether a subject who died from, say, **CVD** at time t would have died from **Cancer** if he hadn't died from **CVD**.

CVD death
⇓
Cancer death unobservable

In general
Failure from competing risk A
⇓
Failure from competing risk B unobservable

In other words, dying from **Cancer** at time t is an unobservable outcome for a subject who died from **CVD** at or before time t. More generally, failure from a competing risk at time t is unobservable for a subject who has already failed from a different competing risk up to time t.

Answer to 2:
Alternative strategies available but no strategy is always best

Because we can never fully determine whether competing risks are independent, how can we proceed with the analysis of competing risks survival data? The answer is that there are several alternative strategies, but no one strategy that is always best.

Strategy 1
Decide assumption satisfied on clinical/biological/other grounds

One strategy is to decide on clinical/biological/other grounds without any data analysis that the independence assumption is satisfied and then carry out the analysis assuming independence.

EXAMPLE OF STRATEGY 1—CANCER VS. CVD

Decide independence if subjects who were censored because of **CVD** death were no more or less likely to have died from **Cancer**.

For example, suppose the two competing risks are **Cancer** deaths and **CVD** deaths. Then you may decide that the assumption of independent competing risks is reasonable if at any time t, subjects who were censored because of **CVD** death were no more or less likely to have died from **Cancer**.

Strategy 2
Include common risk factors for competing risks in survival model

A second strategy is to measure those variables that are common risk factors for competing risks being considered and then include those variables in the survival model. For example, with **Cancer** and **CVD**, perhaps including age and smoking status in the survival model might remove common effects on competing risks.

EXAMPLE OF STRATEGY 2—CANCER VS. CVD

Include age smoking in model to remove the common effects of these variables on competing risks.

Criticism of Strategies 1 and 2
Assumptions cannot be verified with observed data

A criticism of each of the above strategies is that they both rely on assumptions that cannot be verified with the observed data.

Strategy 3
Use a sensitivity analysis

- Considers "worst-case" violations of the independence assumption

Another strategy (3) that can be used is a sensitivity analysis. As with Strategies 1 and 2, a sensitivity analysis cannot explicitly demonstrate whether the independence assumption is satisfied. However, this strategy allows the estimation of parameters by considering "worst-case" violations of the independence assumption.

Sensitivity analysis

- Determines extreme ranges for estimated parameters of one's model

Thus, using a sensitivity analysis, the investigator can determine extreme ranges for the estimated parameters in one's model under violation of the independence assumption.

If "worst-case" not meaningfully different from independence
then
at most a small bias when assuming independence

If such "worst-case" results do not meaningfully differ from results obtained under the independence assumption, then the investigator may conclude that at most a small bias can result from an analysis that assumes independence.

If "worst-case" meaningfully
different from independence
then
only extreme of bias but not
actual bias is determined

If, on the other hand, the sensitivity analysis provides results that meaningfully differ from results obtained under the independence assumption, the investigator learns only the extremes to which the results could be biased without adjusting for the actual bias.

EXAMPLE BYAR DATA

Cause-specific focus: **Cancer**
Censored: **CVD** deaths. **Other** deaths, usual censoring

Worst-case situations:

1. **CVD** or **Other** deaths are assumed to die of cancer instead
2. **CVD** or **Other** deaths assumed to survive as long as the largest survival time observed in the study

We now illustrate how a sensitivity analysis can be carried out using the Byar data, where we focus on the cause-specific survival for **Cancer** deaths, treating **CVD** and **Other** deaths as censored in addition to usual censoring.

The following two worst-case situations are considered. (1) All subjects that are censored because of **CVD** or **Other** deaths are assumed to die of cancer instead. (2) All subjects that are censored because of **CVD** or **Other** deaths survive as long as the largest survival time observed in the study.

Table 9.4. Edited Output for Cancer Worst-Case (1)

var	DF	Coef	Std.Err.	p > |z|	Haz. Ratio
Rx	1	−0.185	0.110	0.092	0.831
Age	1	0.286	0.087	0.001	1.332
Wt	1	0.198	0.093	0.032	1.219
PF	1	0.402	0.170	0.018	1.495
Hx	1	0.437	0.112	0.000	1.548
HG	1	0.292	0.120	0.015	1.339
SZ	1	0.672	0.159	0.000	1.958
SG	1	0.399	0.115	0.001	1.491

Log likelihood = −1892.091

Table 9.5. Edited Output for Cancer Worst-Case (2)

Var	DF	Coef	Std.Err.	p > |z|	Haz. Ratio
Rx	1	−0.411	0.169	0.015	0.663
Age	1	−0.118	0.139	0.394	0.888
Wt	1	0.086	0.138	0.532	1.090
PF	1	0.125	0.254	0.622	1.133
Hx	1	−0.266	0.179	0.138	0.767
HG	1	0.314	0.169	0.063	1.369
SZ	1	0.825	0.197	0.000	2.282
SG	1	1.293	0.201	0.000	3.644

Log likelihood = −839.631

Table 9.4 and **Table 9.5** give edited output for the above two scenarios followed by a repeat of the output previously shown (**Table 9.1**) under the independence assumption.

To carry out worst-case scenario (1), the Status variable (indicating whether a subject failed or was censored) was changed in the dataset from 0 to 1 for each subject that had a **CVD** or **Other** death.

For worst-case scenario (2), the longest survival time observed in the study was 76 weeks. Thus, the survival time for each subject that had a **CVD** or **Other** death was changed in the dataset from the actual time of death to 76 weeks.

To evaluate the results of the sensitivity analysis, we need to compare the output in **Table 9.1**, which assumes that competing risks are independent, with output for worst-case situations provided in **Table 9.4** and **Table 9.5**. We focus this comparison on the estimated coefficient of the Rx variable.

Table 9.1 (Repeated). Edited Output for Cancer with **CVD** and Other censored (Assumes Competing Risks Independent)

Var	DF	Coef	Std.Err.	P>|z|	Haz. Ratio
Rx	1	−0.550	0.170	0.001	0.577
Age	1	0.005	0.142	0.970	1.005
Wt	1	0.187	0.138	0.173	1.206
PF	1	0.253	0.262	0.334	1.288
Hx	1	−0.094	0.179	0.599	0.910
HG	1	0.467	0.177	0.008	1.596
SZ	1	1.154	0.203	0.000	3.170
SG	1	1.343	0.202	0.000	3.830

Log likelihood = −771.174

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Worst-Case (1):					
Rx	1	−0.185	0.110	0.092	0.831
Worst-Case (2):					
Rx	1	−0.411	0.169	0.015	0.663
Independent competing risks:					
Rx	1	−0.550	0.171	0.001	0.577

The first line of output corresponding to the Rx variable is shown at the left for both worst-case scenarios together with the output obtained from assuming independent competing risks.

	WC(1)	WC(2)	Independent
$\widehat{HRs}$	0.831	0.663	0.577
P-values	0.092	0.015	0.001
	(N.S.)	(<.05)	(< <.01)

These results for the RX variable show considerable differences among all three scenarios. In particular, the three estimated hazard ratios are 0.831 (= 1/1.20), 0.663 (=1/1.51), and .577 (=1/1.73). Also, the P-values for the significance of the effect of Rx (0.092, 0.015, .001) lead to different conclusions about the effect of Rx.

Independence Nonindependence
_____x_____ [_____] ___
.577 .663 .831

Note that the HR obtained from assuming independence does not lie between the HRs from the two worst-case scenarios. This should not be surprising because both worst-case scenarios assume non-independence.

If
competing risks not independent
then
conclusions about the effect of Rx could be very different

These results suggest that if the competing risks were not independent, then the conclusions about the effect of Rx could be somewhat different.

But,

- Have not demonstrated whether independence assumption satisfied
- Have not obtained correct results under violation of independence assumption

However, these results do not demonstrate whether the independence assumption is satisfied, nor do they provide estimates of the unbiased hazard ratios and corresponding Wald tests under violation of the independent assumption.

Worst-case (1)
More departure from independence
More realistic
More emphasis
than
Worst-case (2)

Worst-case (1) gives more departure from independence than worst-case (2). It can also be argued that worst-case (1) is more realistic and thus should be emphasized more than worst-case (2), because subjects who were censored because of **CVD** or **Other** deaths would not be expected to survive the entire study if they hadn't died.

Sensitivity analysis: approaches can vary for example,

- Randomly select subset of 50% (or 25%) of subjects censored with **CVD** or **Other** deaths
- Assume everyone in subset dies of **Cancer**

The previous observation suggests that the investigator can vary the approach used to either carry out or interpret such a sensitivity analysis. For example, an alternative approach would be to modify worst-case (1) by randomly selecting a subset of 50% (or 25%) of subjects censored with **CVD** or **Other** deaths and then assuming that everyone in this subset dies of **Cancer** instead.

Main point:

Sensitivity analysis is one of several strategies to address concern about independence assumption

Evaluates how badly biased the results can get if independence not satisfied

In any case, the main point here is that a sensitivity analysis of the type we have illustrated is one of several strategies that can be used to address concern about the independence assumption. Such a sensitivity analysis allows the investigator to evaluate how badly biased the results could get if the independence assumption is not satisfied.

Nevertheless

- No method to directly assess independence assumption
- Typical analysis assumes independence assumption is satisfied

Nevertheless, as mentioned earlier, there is no method currently available that can directly assess the independence assumption nor guarantee correct estimates when the independence assumption is violated. Consequently, the typical survival analysis assumes that the independence assumption is satisfied when there are competing risks, even if this is not the case.

VI. Cumulative Incidence Curves (CIC)

Survival curves S(t):
provide summary information over time of survival experience

KM: empirical approach for estimating survival curves

Adjusted survival curves: generalized version of KM using a regression model to adjust for covariates

Up to now: One event-type of interest (no competing risks)

We have previously discussed (Chapter I and beyond) the use of survival curves to provide summary information over time of the survival experience of (suh) groups of interest. The Kaplan-Meier (KM) approach (Chapter 2), also called the product-limit approach, is a widely used empirical method for estimating survival culves. A generalized version of KM can be used with a regression (e.g., Cox) model to estimate adjusted survival curves (Chapter 3) that account for covariates. Up to now, such survival curves have been described only for the situation when there is only one event- type of interest.

Competing risks: **KM** may not be as informative as when only one risk

When competing risks are being considered, the KM survival curve may not be as informative as with only one risk.

Hypothetical Study

- n = 100 subjects
- All subjects with prostate cancer

Consider the following hypothetical scenario: a 5-month follow-up of 100 subjects with (say, prostate) cancer. Suppose that at 3 months from start of follow-up, 99 of the 100 subjects die from CVD. And at 5 months, the 1 remaining subject dies from prostate cancer.

Survt (months)	# Died	Cause
3	99	CVD
5	1	Cancer

Study goal: cause-specific cancer survival
Censored: CVD deaths

The goal of the study is to determine the cause-specific survival experience for cancer mortality, where a CVD death is considered as censored.

Table 9.6. Hypothetical Survival Data

f	t_f	n_f	m_f	q_f	$S_{Ca}(t_f) \leftrightarrow$ KM
0	0	100	0	0	1
1	3	100	0	99	1
2	**5**	1	1	–	**0**

Table 9.6 summarizes the survival experience in this hypothetical study. The first five columns of this table show the ordered failure-time interval number (f), the time of failure (t_f), the number in the risk set (n_f), the number who fail (m_f), and the number who are censored at each failure time (q_f), assuming that a subject who died of CVD at a given time is censored at that time. The last column shows the KM survival probabilities $S_{Ca}(t_f)$ for cause-specific cancer at each failure time.

Risk set at $t_f = 5$: 1 subject

$\Pr(T > 5 \mid T \geq 5) = (1 - 1)/2 = 0$

$\mathbf{KM_{Ca}}$: $\mathbf{S_{Ca}}\,(t = 5)$
$= S(t = 4) \times \Pr(T > 5 | T \geq 5)$
$= \quad 1 \quad \times \quad 0$
$= \quad \mathbf{0}$

From this table, we see that there is only one subject in the risk set at 5 months, and that this subject fails at month 5. The conditional probability of surviving past 5 months given survival up to 5 months is $(1 - 1)/1 = 0$, so that the KM survival probability at 5 months is 0.

$\mathbf{KM_{Ca}} = 0 \Rightarrow \mathrm{Risk}_{\mathbf{Ca}}\,(T \leq 5)$
$= 1 - 0 = \mathbf{1}$

Thus, use of the $\mathrm{KM}_{\mathbf{Ca}}$ curve in the presence of competing risks (for CVD), suggests that the 5-month risk for cancer death is 1; that is, $1 - \mathbf{S_{Ca}}\,(t = 5)$. Nevertheless, because 99 patients died of CVD instead of cancer, the proportion of the initial 100 subjects who died of cancer is .01, a very small "risk" in contrast to the KM-based "risk" of 1.

Nevertheless,

$$\frac{1 \text{ cancer death}}{100 \text{ initial subjects}} = 0.01 \quad \text{(small)}$$

Question:
How many of the 99 CVD deaths would have died of cancer at $t = 5$ if they hadn't died of CVD at $t = 3$?

A natural question at this point is, how many of the 99 patients who died of CVD at 3 months would have died of cancer by 5 months instead if they hadn't died of CVD?

Cannot answer: unobservable

Unfortunately, we cannot ever answer this question because those dying of CVD cannot be observed further once they have died.

Table 9.7. Hypothetical Survival Data Sensitivity Analysis A (99 **CVD** Deaths of Cancer at $t = 5$)

f	t_f	n_f	m_f	q_f	$S_{Ca}(t_f) \leftrightarrow KM$
0	0	100	0	0	1
1	3	100	0	0	1
2	5	100	100	0	**0**

But we can consider a sensitivity-type of analysis to see what might happen under certain alternative scenarios. Suppose, for example, that all 99 subjects who died of CVD at 3 months would have died of cancer at 5 months if they hadn't died of CVD. Also assume as before that the 100th subject survived up to 5 months but then immediately died. The survival experience for this situation is shown in **Table 9.7**. Notice that the KM survival probability at month 5 is 0, which is the same value as obtained in the original dataset.

KM method assumes non-informative (i.e., independent) censoring

⇓

$\Pr(T > 5 | \text{censored at month 3})$

=

$\Pr(T > 5 | \text{survived to month 5}) = 0$

⇓

99 CVDs deaths would have been cancer deaths at month 5

The reason why **Tables 9.6** and **9.7** give the same 5-month survival probability (=0) is that the KM method assumes independent censoring. For the original data (**Table 9.6**), independent censoring requires that those who were censored at month 3 were as likely to have died from cancer at month 5 as those who were in the risk set at month 5. Because the one person in the risk set at month 5 actually died from cancer, then the KM method assumes that all 99 CVD deaths being viewed as censored would have been cancer deaths at month 5, which is what is represented in **Table 9.7**.

Table 9.8. Hypothetical Survival Data Sensitivity Analysis B (99 **CVD** Deaths of survive past t = 5)

f	t_f	n_f	m_f	q_f	$S_{Ca}(t_f) \leftrightarrow KM$
0	0	100	0	0	1
1	3	100	0	0	1
2	5	100	1	99	**0.99**

Now let's consider a different version (B) of a sensitivity analysis. Suppose that all 99 subjects who died of CVD at 3 months would **not** have died of cancer at 5 months if they hadn't died of CVD. Also assume as before that the 100th subject survived up to 5 months but then immediately died. The survival experience for this situation is shown in **Table 9.8**.

Table 9.8: $S_{\mathbf{Ca}}(t = 5) = 0.99$ different from
Table 9.6: $S_{\mathbf{Ca}}(t = 5) = 0$

The KM survival probability at month 5 is 0.99 (i.e., close to 1), which is very different from the value of 0 obtained in the original dataset (**Table 9.6**).

Focus on 1 − S(t) = Risk:
$\text{Risk}_{\mathbf{Ca}}(T \leq 5) = 1 - 0.99 = \mathbf{0.01}$

If we then focus on 1 − S(t) instead of S(t), sensitivity analysis B suggests that the 5-month risk for cancer death is 0.01 (i.e., 1 − 0.99).

Table 9.6: $\text{Risk}_{\mathbf{Ca}}(T \leq 5) = \mathbf{1}$ derived from the data

Table 9.8: $\text{Risk}_{\mathbf{Ca}}(T \leq 5) = \mathbf{0.01}$ derived from sensitivity analysis

but also derived directly from data as a **marginal probability**

We thus see that the KM-based risk of 1 computed from the actual data (**Table 9.6**) is quite different from the KM-based risk of .01 computed in **Table 9.8**, where the latter derives from a sensitivity analysis that does not use the actual data. Note, however, that a "risk" of .01 for cancer death can be derived directly from the actual data by treating the CVD deaths as cancer survivors. That is, .01 is the proportion of all subjects who actually developed cancer regardless of whether they died from CVD. This proportion is an example of what is called a **marginal probability**.

Which is more informative,

$\text{Risk}_{\mathbf{Ca}}(T \leq 5) = \mathbf{1}$ or $\mathbf{0.01}$?

So which of these two "risk" estimates (1 vs. 01) is more informative? Actually, they are both informative in different ways.

Answer: both informative

"Risk" of .01 considers treatment utility

for example, proportion of cancer patients needing treatment

The "risk" of .01 is informative from the standpoint of treatment utility for cancer because in these data, the proportion of cancer patients needing treatment is quite small when allowing for competing risks.

"Risk" of 1 considers etiology, providing competing risks are independent

for example, cancer survival is unlikely after 5 months

On the other hand, the "risk" of 1, corresponding to the survival probability of 0, is informative from an etiologic standpoint providing competing risks are independent; for example, cancer patients who don't die of CVD would be expected to die from their cancer by 5 months; that is, cancer survival is unlikely after 5 months.

Main point

KM survival curve may not be very informative

- Requires independence assumption about competing risks
- Independence assumption cannot be verified

The main point of the above illustration is that when there are competing risks, the KM survival curve may not be very informative because it is based on an independence assumption about competing risks that cannot be verified.

Alternative to KM: **Cumulative Incidence Curve (CIC)** uses marginal probabilities

This has led to alternative approaches to KM for competing risk data. One such alternative, called the **Cumulative Incidence Curve (CIC)**, involves the use of marginal probabilities as introduced above. (Kalbfleisch and Prentice, 1980).

Only one risk: $\mathbf{CIC} = \mathbf{1} - \mathbf{KM}$

CIC with competing risk

- Derived from cause-specific hazard function
- Estimates **marginal probability** when competing risks are present
- Does not require independence assumption

In the simplest case, if there is only one risk, the **CIC** is (1 - KM). With competing risks, however, the **CIC** is derived from a cause-specific hazard function, provides estimates of the "marginal probability" of an event in the presence of competing events, and does not require the assumption that competing risks are independent.

Marginal probabilities:

- useful to assess treatment utility in cost-effectiveness analyses
- example: 0.01 = 5-month marginal probability for Cancer (**Table 9.6**)

Such marginal probabilities are relevant to clinicians in cost-effectiveness analyses in which risk probabilities are used to assess treatment utility. For example, the 0.01 (5-month) marginal probability for cancer derived from the hypothetical data in **Table 9.6** illustrates small treatment utility for cancer.

Steps to construct **CIC**:

1. Estimate hazard at ordered failure times t_f for event-type (**c**) of interest:

$$\hat{h}_c(t_f) = m_{cf}/n_f$$

where
m_{cf} = # of events for event-type **c** at time t_f
n_f = # of subjects at risk at time t_j

How does one construct a **CIC**? We first estimate the hazard at ordered time points t_f when the event of interest occurs. This hazard estimate is simply the number of events that occur at t_f divided by the number at risk at t_f (analogous to the KM estimate). We can write this as $\hat{h}_{\mathbf{c}}(t_f) = m_{\mathbf{cf}}/n_f$ where the $m_{\mathbf{cf}}$ denotes the number of events for risk **c** at time t_f and n_f is the number of subjects at that time. Thus, at any particular time, $m_{\mathbf{cf}}/n_f$ is the estimated proportion of subjects failing from risk **c**.

2. Estimate
$S(t_{f-1})$ = **overall** survival probability of surviving previous time (t_{f-1})
overall ⇒ subject survives all other competing events

To be able to fail at time t_f, the subject needs to be "around to fail", i.e., he must have survived the previous time when a failure occurred. The probability of surviving the previous time t_{f-1} is denoted $S(t_{f-1})$, where S(t) denotes the **overall survival curve** rather than the cause-specific survival curve $S_{\mathbf{c}}(t)$. We must consider "overall" survival here, since the subject must have survived all other competing events.

3. Compute estimated incidence of failing from event-type **c** at time t_f:

$$\hat{I}_c(t_f) = \hat{S}(t_{f-1}) \times \hat{h}_c(t_f)$$

The probability (i.e., incidence) of failing from event-type **c** at time t_f is then simply the probability of surviving the previous time period multiplied by $\hat{h}_{\mathbf{c}}(t_f)$.

4.

$$\mathbf{CIC_c}(t_f) = \sum_{f'=1}^{f} \hat{I}_c(t_{f'}) = \sum_{f'=1}^{f} \hat{S}(t_{f'-1})\hat{h}_c(t_{f'})$$

The **cumulative incidence** ($\mathbf{CIC}_c$) for event-type **c** at time t_f is then the *cumulative* sum up to time t_f (i.e., from f′=1 to f′=f) of these incidence values over all event-type **c** failure times.

CIC = 1 − **KM** ↔ no competing risks
but
$\mathbf{CIC_c}(t_f) \neq 1 - \mathbf{KM_c}$ ↔ competing risks

Although, as previously mentioned, the **CIC** is equal to 1 − **KM** when there are no competing risks, formula 4 for $\mathbf{CIC}_c(t_f)$ differs from 1 - $\mathbf{KM}_c$ when there are competing risks. In particular, the **CIC** formula (4) uses the overall survival function S(t) that counts events from competing risks in addition to the event-type of interest as failures. In contrast, the formula for 1 − $\mathbf{KM}_c$ uses the event-type-specific survival function $S_{\mathbf{c}}(t)$, which treats failures from competing risks as censored observations; this formula has been called the *censor method* (Arriagada et al.,1992).

since

$$1 - \mathbf{KM_c} = \sum_{f'=1}^{f} \hat{S}_c(t_{f'-1})\hat{h}_c(t_{f'})$$

(censor method)

Example of **CIC** calculation

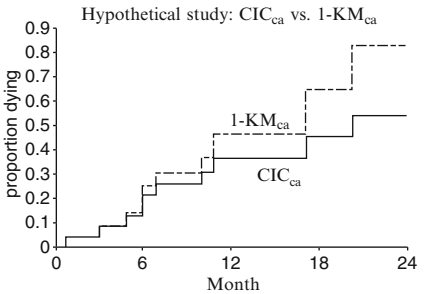
- n = 24 subjects
- all subjects receive treatment XRT for head and neck cancer

Survival time in (months)

Died of disease: 0.7, 3, 4.9, 6, 6, 6.9, 10, 10.8, 17.1, 20.3
Died of other causes: 1.5, 2.8, 3.8, 4.7, 7, 10, 10, 11.2
Censored: 3.2, 7.6, 10, 11, 15, 24.4

Table 9.9. CIC calculation Hypothetical data

t_f	n_f	m_f	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$CIC_{ca}(t_f)$
0	24	0	0	-	0	0
0.7	24	1	0.042	1.000	0.042	0.042
1.5	23	0	0	0.958	0	0.042
2.8	22	0	0	0.916	0	0.042
3.0	21	1	0.048	0.875	0.042	0.083
3.2	20	0	0	0.833	0	0.083
3.8	19	0	0	0.833	0	0.083
4.7	18	0	0	0.789	0	0.083
4.9	17	1	0.059	0.745	0.044	0.127
6	16	2	0.125	0.702	0.088	0.215
6.9	14	1	0.071	0.614	0.044	0.259
7.0	13	0	0	0.570	0	0.259
7.6	12	0	0	0.526	0	0.259
10	11	1	0.091	0.526	0.048	0.307
10.8	7	1	0.143	0.383	0.055	0.361
11.0	6	0	0	0.328	0	0.361
11.2	5	0	0	0.328	0	0.361
15	4	0	0	0.262	0	0.361
17.1	3	1	0.333	0.262	0.087	0.449
20.3	2	1	0.5	0.175	0.087	**0.536**
24.4	1	0	0	0.087	0	0.536



We illustrate the calculation of a **CIC** through an example.

Consider another hypothetical study involving 24 individuals receiving radiotherapy (XRT) for the treatment of head and neck cancer. Patients may either die of the disease (cancer), other causes or still be alive at the time of analysis.

The data are shown on the left.

The calculations required for the cumulative incidence curve $\mathbf{CIC_{ca}}$ for the event-type "death from cancer (i.e., **ca**)" are shown in **Table 9.9**.

From the table, we can see that the highest **CIC** probability of 0.536 is reached when t= 20.3 weeks when the last observed event occurred. Thus, the cumulative risk (i.e., marginal probability) for a Cancer death by week 20 is about 53.6% when allowing for the presence of competing risks for CVD and Other Deaths.

Because the **CIC** curve describes "cumulative incidence", a plot of the curve starts at 0 when t=0 and is a nondecreasing function up until the latest time of individual follow-up (t=24.4). We show on the left the graphs of both the $\mathbf{CIC_{ca}}$ and $1 - \mathbf{KM_{ca}}$. Notice that $1 - \mathbf{KM_{ca}}$ overestimates the probability of failure for the event-type "death from cancer (**ca**)."

CIC Summary

- Gives marginal probability.
- Does not use product limit formulation
- Not included in mainstream commercially avilable statistical packages (e.g., SAS, STATA, SPSS)

Thus, as the example illustrates, the "marginal probability" estimated by the **CIC** does not use a product-limit (i.e., KM) formulation. Moreover, the computation of a **CIC** is currently not included in mainstream commercially available statistical packages.

Independence of competing risks not required for **CIC** approach.

As mentioned earlier, the assumption of independent competing risks is not required for the calculation of the **CIC**, in contrast to the KM survival curve, which requires this assumption.

Nevertheless, **CIC** requires

$h(t) = h_{\mathbf{c1}}(t) + h_{\mathbf{c2}}(t) + \ldots + h_{\mathbf{ck}}(t)$

where

$h(t)$ = overall hazard

$h_{\mathbf{c}}(t)$ = hazard for event type **c**

Nevertheless, the **CIC** does require that the overall hazard is the sum of the individual hazards for all the risk types (Kalbfleisch and Prentice, 1980). The latter assumption will be satisfied, however, whenever competing risks are mutually exclusive and events are nonrecurrent, i.e., one and only one event can occur at any one time and only once over time.

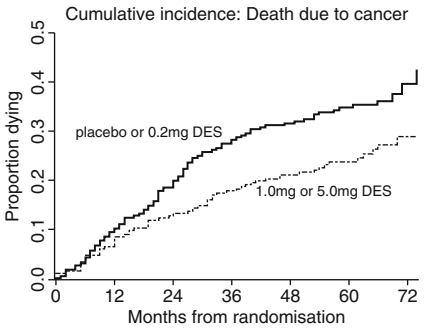
Note: Satisfied if

- Mutually exclusive event types
- Nonrecurrent events

Comparing **CIC**'s for two or more groups:

- Statistical test available (Gray, 1989)
- Analogous to log-rank test
- No independence assumption
- Does not adjust for covariates

Gray (1988) developed a test to compare two or more **CIC**s. This test is analogous to the log-rank test. The independence assumption is not required. However, this test does not adjust for covariates.



The plot shown on the left gives the **CIC**s for the two treatments for the Byar data that we originally introduced in Section III of this chapter.

Gray's test results: $\chi^2 = 6.6$, df = 1
P-value: 0.01

Using Gray's test to compare the two **CIC**s shown in the plot, we find the two curves to be significantly different (P=0.01).

PH model used to obtain **CIC**
⇓
Independence of competing risks required
(but **CIC** meaningful for treatment utility)

So far, we have described the **CIC** without considering (Cox PH) models that account for covariates. However, **when a PH model is used** to obtain hazard ratio estimates for individual competing risks as an intermediate step in the computation of a **CIC, the independence of competing risks is required.** In any case, the **CIC** has a meaningful interpretation in terms of treatment utility regardless of whether competing risks are independent.

Modeling **CIC** with covariates using PH model: Fine and Gray (1999)

(**CIC** also called **subdistribution** function)

Software available (Accord, 1997)

Fine and Gray (1999) provide methodology for modeling the **CIC** with covariates using a proportional hazards assumption. They refer to the **CIC** curves as **subdistribution** functions. Software is available that allows for such models to be fitted (Accord, 1997).

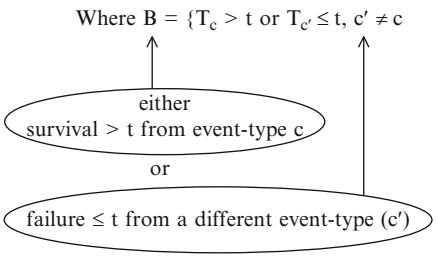
Fine and Gray model analogous to Cox PH model:

- use a hazard function defined from a CIC
- effects of predictors (e.g., HRs) have similar interpretation

The **CIC** models developed by Fine and Gray are analogous to the Cox PH model but, for any failure type, they model a hazard function (also called a **sub-distribution hazard**) derived from a **CIC**. The results from fitting these models have a similar interpretation regarding the effects of predictors in the model as can be derived from the (standard) Cox PH model approach for competing risks data.

Sub-distribution hazard function (for event-type c):

$$h_{\mathbf{c},CIC}(t) = \lim_{\Delta \to 0} \frac{\Pr(t < T_{\mathbf{c}} < t + \Delta t | B)}{\Delta t}$$



The (sub-distribution) hazard function for event-type **c** used in the Fine and Gray method is defined by $h_{\mathbf{c},CIC}(t)$ on the left. This function, based on the CIC, gives the hazard rate for a cause-specific event at time t based on the risk set that remains at time t after accounting for all previously occurring event-types (i.e., including competing risks).

In the above hazard formula, the expression defined by B to the right of the conditioning symbol (|) in the formula conditions on a subject's surviving from event-type **c** past time t or on a subject's failing from a different event-type (e.g., **c′**) at or before time t; this accounts for the occurrence of all event-types prior to time t.

Equivalent formula for $h_{\mathbf{c},CIC}(t)$:

$$h_{\mathbf{c},CIC}(t) = \frac{dCIC_{\mathbf{c}}(t)/dt}{1 - CIC_{\mathbf{c}}(t)}$$

analogous to

Hazard when no competing risks:

$$h(t) = \frac{f(t)}{S(t)}$$

where f(t) = probability density function

$$= \lim_{\Delta t \to 0} \frac{\Pr[t \leq T < t + \Delta t]}{\Delta t}$$

An equivalent mathematical expression for $h_{\mathbf{c},CIC}(t)$ is shown on the left, where $dCIC_{\mathbf{c}}(t)/dt$ is the derivative of the CIC function for event-type **c** at time t.

This expression for $h_{\mathbf{c},CIC}(t)$ is analogous to a similar expression shown on the left for the hazard function h(t) when there are no competing risks (see Chapter 7, Section II). In the latter case, h(t) is equal to the probability density function f(t) divided by the survival function S(t) shown on the left.

The analogy:

Competing Risks	No Competing Risks
$h_{\mathbf{c},CIC}(t)$	h(t)
$dCIC_{\mathbf{c}}(t)/dt$	f(t) = dF(t)/dt where F(t) = Pr(T ≤ t)
$1-CIC_{\mathbf{c}}(t)$	S(t) = 1 − F(t)

Note: $1 - CIC_{\mathbf{c}}(t)$ is not strictly a survival curve since the formula for $CIC_c(t)$ uses overall survival S(t) in its calculation rather than survival from event-type c, i.e., $S_{\mathbf{c}}(t)$

To clarify the analogy further, the term in $dCIC_{\mathbf{c}}(t)/dt$ in the numerator of the formula for $h_{\mathbf{c},CIC}(t)$ corresponds to the density function f(t) in the numerator of h(t). Note that f(t) is equal to dF(t)/dt, where F(t) is cumulative distribution function F(t); in other words, the derivative of $CIC_{\mathbf{c}}(t)$ in $h_{\mathbf{c},CIC}(t)$ corresponds to the derivative of F(t), which is f(t) in the expression for h(t).

Also, when there are no competing risks, S(t) equals 1 − F(t), which is analogous to the denominator $1 - CIC_{\mathbf{c}}(t)$ of $h_{\mathbf{c},CIC}(t)$. Strictly speaking, however, $1 - CIC_{\mathbf{c}}(t)$ is not directly equivalent to a survival curve for event-type c, since the formula $CIC_{\mathbf{c}}(t)$ treats the occurrence of events from other event-types as failures rather than as censored observations.

Fine and Gray's CIC model:

$$h_{\mathbf{c},CIC}(t) = h_{0\mathbf{c},CIC}(\exp)[\sum_{i=1}^{p} \gamma_i X_i]$$

The CIC model developed by Fine and Gray is shown on the left. This model is analogous to a Cox model, except that the model considers the subdistribution hazard function $h_{\mathbf{c},CIC}(t)$ instead of the hazard function h(t).

PH assumption satisfied in above model:

$$HR_{\mathbf{c},CIC}(\mathbf{X}^*, \mathbf{X}) = \exp[\sum_{i=1}^{p} \gamma_i (X_i^* - X_i)$$

analogous to

$$HR_{CoxPH}(\mathbf{X}^*, \mathbf{X}) = \exp[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)$$

The model shown here satisfies the PH assumption for the subpopulation hazard being modeled, i.e., the general HR formula is essentially the same as for the Cox model, except that the β's in the Cox PH model are now replaced by γ's in the Fine and Gray model.

Can use "extended" Fine and Gray model to account for time-dependent variables.

The Fine and Gray model can be extended to allow for variables not satisfying the PH, including time-dependent variables.

Table 9.10. Edited Output for Cancer with CVD and Other censored Byar data (Fine and Gray CIC approach)

Var	DF	Coef	Std.Err.	P>|z|	Haz. Ratio
Rx	1	−0.414	0.171	0.008	0.661
Age	1	−0.112	0.145	0.221	0.894
Wt	1	0.088	0.146	0.274	1.092
PF	1	0.126	0.260	0.313	1.135
Hx	1	−0.256	0.182	0.080	0.774
HG	1	0.321	0.191	0.046	1.379
SZ	1	0.841	0.207	0.001	2.318
SG	1	1.299	0.198	0.001	3.665

−2 LOG L = 1662.766546

For the Byar data, the fitted Fine and Gray **CIC** model that focuses on cancer deaths as the event type of interest is shown in **Table 9.10** below which we have repeated in **Table 9.1**, which uses the standard competing risks Cox PH model approach.

Table 9.1. (Repeated). Edited Output for Cancer with **CVD** and Other Censored (Standard Cox PH approach)

Var	DF	Coef	Std.Err.	p > |z|	Haz. Ratio
Rx	1	−0.550	0.170	0.001	0.577
Age	1	0.005	0.142	0.970	1.005
Wt	1	0.187	0.138	0.173	1.206
PF	1	0.253	0.262	0.334	1.288
Hx	1	−0.094	0.179	0.599	0.910
HG	1	0.467	0.177	0.008	1.596
SZ	1	1.154	0.203	0.000	3.170
SG	1	1.343	0.202	0.000	3.830

Log likelihood = −771.174

Although corresponding coefficient estimates and standard errors are different in the two outputs, both outputs are reasonably similar.

Fine and Gray **CIC** (**Table 9.10**)	Standard Cox PH (**Table 9.1**)
$\hat{\beta}_{Rx}$: −0.414	−0.550
$\hat{HR}_{Rx}$: 0.661 $= (\frac{1}{1.513})$	0.577 $= (\frac{1}{1.733})$
P-value : 0.008	0.001

For example, the estimated coefficient of Rx is −0.414 in **Table 9.10** vs. −0.550 in **Table 9.1**. The corresponding hazard ratio estimates ($e^{\hat{\beta}}$) are 0.661 (=1/1.513) and 0.577 (=1/1.733), respectively, so that the strength of the association is slightly weaker using the Fine and Gray approach for these data, although both hazard ratios are highly significant.

VII. Conditional Probability Curves (CPC)

A third measure of failure risk: **CPC** (Other measures: **1-KM** and **CIC**)

$\mathbf{CPC_c} = \Pr(T_c \leq t \mid T \geq t)$

where T_c = time until event c occurs
T = time until any competing risk event occurs

Another approach to competing risks is called the **cumulative conditional probability** or **CPC**. CPCs provide a third summary measure, in addition to (1 - KM) and CIC, of the risk of failure of an event in the presence of competing risks. Put simply, the $\mathbf{CPC_c}$ is the probability of experiencing an event **c** by time t, *given that an individual has not experienced any of the other competing risks by time t*.

$\mathbf{CPC_c} = CIC_c/(1\text{-}CIC_{c'})$

where $\mathbf{CIC_{c'}}$ = CIC from risks other than **c**

Graphs of CPC's obtained from CIC's

Tests to compare CPC's:
Pepe and Mori (1993) −2 curves
Lunn (1998) −g curves

Example of **CPC** calculation

- n = 24 subjects
- all subjects receive treatment XRT for head and neck cancer
- CIC_{ca} previously calculated (Table 9.9)

For event-type **c**, the **CPC** is defined by $\mathbf{CIC_c}$ divided by (1 - $\mathbf{CIC_{c'}}$), where $\mathbf{CIC_{c'}}$ is the cumulative incidence of failure from risks other than risk **c** (i.e., all other risks considered together).

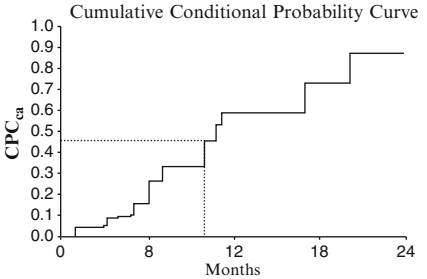
Graphs of **CPC** curves can be obtained from CIC curves and have been studied by Pepe-Mori (1993) and Lunn (1998). Pepe-Mori provide a test to compare two **CPC** curves. Lunn (1998) extended this test to g-groups and allow for strata.

We illustrate the calculation of a **CPC** using the previously described hypothetical study data involving 24 individuals receiving radiotherapy (XRT) for the treatment of head and neck cancer (**ca**). For these data, the calculations for the $\mathbf{CIC_{ca}}$ were given in previous **Table 9.9**, below which was shown the graph of the $\mathbf{CIC_{ca}}$ curve.

Table 9.11 CPC calculation—Hypothetical data

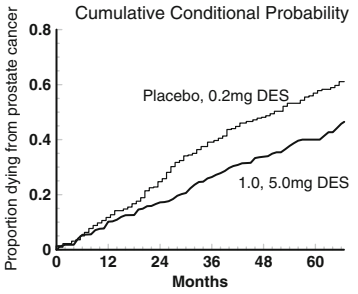
t_f	n_f	$CIC_{ca}(t_f)$	$CIC_{oth}(t_f)$	$CPC_c = CIC_c/(1-CIC_{c'})$
0	24	0	0	0
0.7	24	0.042	0	0.042
1.5	23	0.042	0.042	0.043
2.8	22	0.042	0.083	0.045
3.0	21	0.083	0.083	0.091
3.2	20	0.083	0.083	0.091
3.8	19	0.083	0.127	0.095
4.7	18	0.083	0.171	0.101
4.9	17	0.127	0.171	0.153
6	16	0.215	0.171	0.259
6.9	14	0.259	0.171	0.312
7.0	13	0.259	0.215	0.330
7.6	12	0.259	0.215	0.330
10	11	0.307	0.311	**0.445**
10.8	7	0.361	0.311	0.524
11.0	6	0.361	0.311	0.524
11.2	5	0.361	0.376	0.579
15	4	0.361	0.376	0.579
17.1	3	0.449	0.376	0.719
20.3	2	0.536	0.376	**0.860**
24.4	1	0.536	0.376	0.860

Table 9.11 on the left gives the calculation of the **CPC** for the event-type "death from cancer". A graph of the $\mathbf{CPC_{ca}}$ curve is shown below the table.



From the $\mathbf{CPC_{ca}}$ curve, we can see, for example, that at 10 months, the probability of dying from cancer given that no other type of event has occurred is 0.455. Similarly, at 20.3 months, the probability of dying of cancer given that no other type of event has occurred is 0.860.

Example: Byar Data



Test for equality: p-value = .01 (Pepe-Mori)

Returning to the Byar dataset originally introduced in Section III of this chapter, the plot shown here gives the **CPC** curves comparing the two DES treatments. These curves give the probability of an event (death) from prostate cancer at any particular time *given* that the patient is alive at this time to experience the event.

(Note: the Fine and Gray approach has not been extended to model the **CPC**s in a regression framework.)

The Pepe-Mori test shows a significant difference between these two CPC curves.

VIII. Method 2—The Lunn-McNeil (LM) Approach

Method I: separate estimates for each failure type, treating the competing failure types as censored

Method 2: **LM** Approach

- Uses a single Cox (PH) model
- Gives identical results as obtained from Method 1
- Allows flexibility to perform statistical inferences not available from Method 1

We have previously (Section IV) described an approach (called Method 1) for analyzing competing risks data that uses the Cox (PH) model to separatelyestimate hazards and corresponding hazard ratios for each failure type, treating the other(competing) failure types as censored in addition?to those not failing from any event-type.

We now describe Method 2, called the **Lunn-McNeil (LM)** approach, that allows only one Cox PH model to be fit rather than separate models for each event-type (i,e., Method I above). This approach, depending on the variables put in the model, can give identical results to those obtained from separate models. Moreover, the **LM** approach allows flexihility to perform statistical inferences about various features of the competing risk models that can not be conveniently assessed using Method 1.

Table 9.12. Augmented Data for ith Subject at Time t_i Using LM Approach

Subj	Stime	Status	D_1	D_2	$D_3 \ldots D_C$	$X_1 \ldots X_p$
i	t_i	e_1	1	0	0 . . . 0	$X_{i1} \ldots X_{ip}$
i	t_i	e_2	0	1	0 . . . 0	$X_{i1} \ldots X_{ip}$
i	t_i	e_3	0	0	1 . . . 0	$X_{i1} \ldots X_{ip}$
.	.	.	.	.	. . .	. . .
.	.	.	.	.	. . .	. . .
.	.	.	.	.	. . .	. . .
i	t_i	e_C	0	0	0 . . . 1	$X_{i1} \ldots X_{ip}$

D_1, D_2, D_3,. . ., D_C: indicators for event-types

To carry out the LM approach, the data layout must be augmented. If there are C competing risks, the original data must be duplicated C times, one row for each failure type as shown in **Table 9.12** for the ith subject with survival time t_i in the table. Also, C dummy variables D_1, D_2, D_3,. . ., D_C are created as shown in the table. The value of the status variable e_c, with **c** going from 1 to C, equals 1 if event type **c** occurs at time **c**, and equals 0 if otherwise. The Xs in the table denote the predictors of interest and, as shown in the table, are identical in each row of the table.

The dummy variables D_1, D_2, D_3,. . .,D_C are indicators that distinguish the C competing risks (i.e., event-types).

Definition

D_c equals 1 for event-type **c** and 0 otherwise, $c = 1, 2, \ldots, C$

Thus, the dummy variable D_c equals 1 for event-type **c** and 0 otherwise.

for example,

Event-type 1: $D_1 = 1$, $D_2 = 0$, $D_3 = 0, \ldots, D_C = 0$
Event-type 2: $D_1 = 0$, $D_2 = 1$, $D_3 = 0, \ldots, D_C = 0$
Event-type 3: $D_1 = 0$, $D_2 = 0$, $D_3 = 1, \ldots, D_C = 0$

For example, for event type 1, the Ds are $D_1 = 1$, $D_2 = 0$, $D_3 = 0, \ldots, D_C = 0$; for event-type 2, the Ds are $D_1 = 0$, $D_2 = 1$, $D_3 = 0, \ldots, D_C = 0$; and for event-type 3, the Ds are $D_1 = 0$, $D_2 = 0$, $D_3 = 1, \ldots, D_C = 0$.

Table 9.13. Augmented Data for Subjects 1, 14, 16, and 503 from Byar Data Using LM Approach

Subj	Stime	Status	**CA**	**CVD**	**OTH**	**Rx**	**Age**	**Wt**
1	72	0	1	0	0	0	1	2
1	72	0	0	1	0	0	1	2
1	72	0	0	0	1	0	1	2
14	49	1	1	0	0	0	0	0
14	49	0	0	1	0	0	0	0
14	49	0	0	0	1	0	0	0
16	3	0	1	0	0	1	2	1
16	3	1	0	1	0	1	2	1
16	3	0	0	0	1	1	2	1
503	41	0	1	0	0	0	1	0
503	41	0	0	1	0	0	1	0
503	41	1	0	0	1	0	1	0

Table 9.13 shows observations for subject #s 1, 14, 16, and 503 from the Byar dataset. The **CA, CVD**, and **OTH** columns denote the C = 3 dummy variables D_1, D_2, and D_3, respectively. The last three columns, labeled **Rx, Age**, and **Wt** give values for three of the eight predictors.

In this table, there are three lines of data for each subject, which correspond to the three competing risks, **Cancer** death, **CVD** death, and **Other** death, respectively. The survival time (**Stime**) for subject 1 is 72, for subject 14 is 49, for subject 16 is 3, and for subject 503 is 41.

Subject 1: Censored
Subject 14: died of Cancer
Subject 16: died of CVD
Subject 503: died from OTH

From the Status and Event (i.e., **CA, CVD, OTH**) columns, we can see that subject 1 was censored, subject 14 died of **Cancer**, subject 16 died of **CVD**, and subject 503 died from **Other** causes.

	Rx	Age	Wt
Subject 1	0	1	2
Subject 16	1	2	1

For subject 1, the values for the predictors Rx, Age, and Wt, were 0, 1, and 2, respectively. These values appear identically in the last three columns of the three rows for this subject. Similarly, for subject 16, the predictor values for Rx, Age, and Wt, are 1, 2, and 1, respectively.

General Stratified Cox LM Model

$$g = 1, 2, \ldots, C$$

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp\left[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \delta_{21} D_2 X_1 + \delta_{22} D_2 X_2 + \cdots + \delta_{2p} D_2 X_p + \delta_{31} D_3 X_1 + \delta_{32} D_3 X_2 + \cdots + \delta_{3p} D_3 X_p + \cdots + \delta_{C1} D_C X_1 + \delta_{C2} D_C X_2 + \cdots + \delta_{Cp} D_C X_p\right]$$

1st row: predictors $X_1, X_2, \ldots, X_p$
2nd row: product terms
$D_2X_1, D_2X_2, \ldots, D_2X_p$

Cth row: product terms
$D_CX_1, D_CX_2, \ldots, D_CX_p$

To use the **LM** approach with augmented data to obtain identical results from fitting separate models (Method 1), an interaction version of a stratified Cox PH model is required. A general form for this model based on the notation used for the column heading variables in **Table 9.12** is shown at the left.

Recall that the $X_1, X_2, \ldots, X_p$ denote the p predictors of interest. $D_2, D_3, \ldots, D_C$ are $C - 1$ dummy variables that distinguish the C event-types. Note that event-type 1 ($g = 1$) is the referent group, so variable D_1 is omitted from the model. Thus, the first row in the exponential formula contains the Xs, the second row contains product terms involving D_2 with each of the Xs, and so on, with the last (Cth) row containing product terms of D_C with each of the Xs. The strata ($g = 1, \ldots, C$) are the C event-types.

LM Hazard Model for Event–Type 1

$$h_1^*(t, \mathbf{X}) = h_{01}^*(t) \times \exp\left[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p\right] \quad (D_2 = D_3 = \cdots = D_C = 0)$$

For event-type 1 ($g = 1$), the above stratified Cox model simplifies to the expression shown at the left. Note that because $g = 1$, the values of the dummy variables $D_2, D_3, \ldots, D_C$ are all 0.

No product terms in model:

$$HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\beta_1]$$

Product terms X_jX_1 in model:

$$HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\beta_1 + \Sigma\beta_jX_j]$$

Thus, for g = 1, if X_1 is a (0,1) variable, the other Xs are covariates, and there are no product terms X_jX_1 in the model, the formula for the HR for the effect of X_1 adjusted for the covariates is $\exp[\beta_1]$. The more general exponential formula described in Chapter 3 would need to be used instead to obtain adjusted HRs if there are interaction terms in the model of the form X_jX_1.

LM Hazard Model for Event-Type g (> 1)

$$\begin{aligned} h^*_g(t, \mathbf{X}) &= h^*_{0g}(t) \\ &\quad \times \exp\Big[\beta_1X_1 + \beta_2X_2 + \cdots + \beta_pX_p \\ &\qquad + \delta_{g1}X_1 + \delta_{g2}X_2 + \cdots + \delta_{gp}X_p\Big] \\ &= h^*_{0g}(t)\exp\big[(\beta_1 + \delta_{g1})X_1 + (\beta_2 + \delta_{g2})X_2 \\ &\qquad + \cdots + \big(\beta_p + \delta_{gp}\big)X_p\big] \end{aligned}$$

For any g greater than 1, the general hazard model simplifies to a hazard function formula that contains only those product terms involving the subscript g, because $D_g = 1$ and $D_{g'} = 0$ for g′ not equal to g.

With a little algebra we can combine coefficients of the same predictor to rewrite this hazard model as shown here.

No product terms X_jX_1 in the model and g > 1:

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[(\beta_1 + \delta_{g1})]$$

Thus, for g > 1, if X_1 is a (0,1) variable, the other Xs are covariates, and there are no product terms X_jX_1 in the model, the formula for the HR for the effect of X_1 adjusted for the covariates is **exp**$[\beta_1 + \delta_{g1}]$. This HR expression would again need to be modified if the model contains product terms of the form X_jX_1.

Product terms X_jX_1 in the model and g > 1:

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[(\beta_1 + \delta_{g1}) + \Sigma(\beta_j + \delta_{gj}X_j)]$$

EXAMPLE OF LM MODEL FOR BYAR DATA

We now illustrate the above general **LM** model formation using the Byar data.

Separate models approach (Method 1): Cause-specific model: **Cancer** **CVD** and **Other** deaths censored

Recall that using Method 1, the separate models approach, the Cox hazard formula used to fit a separate model for **Cancer** deaths, treating **CVD** and **Other** deaths as censored is repeated here.

No-interaction model

$$\begin{aligned} h_{\mathbf{Ca}}(t,\mathbf{X}) &= h_{0\mathbf{Ca}}(t)\exp[\beta_{1\mathbf{Ca}}Rx \\ &\quad + \beta_{2\mathbf{Ca}}Age + \beta_{3\mathbf{Ca}}Wt + \\ &\quad \beta_{4\mathbf{Ca}}PF + \beta_{5\mathbf{Ca}}Hx + \beta_{6\mathbf{Ca}}HG \\ &\quad + \beta_{7\mathbf{Ca}}SZ + \beta_{8\mathbf{Ca}}SG] \end{aligned}$$

$$HR_{\mathbf{Ca}}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1\mathbf{Ca}}]$$

Also shown is the formula for the hazard ratio for the effect of the Rx variable, adjusted for other variables in the model.

LM SC Model for Byar Data

$g = 1, 2, 3$

$$h_g^*(t, \mathbf{X}) = h_{0g}(t) \times \exp[\beta_1 RX + \beta_2 Age + \cdots + \beta_8 SG + \delta_{21} D_2 Rx + \delta_{22} D_2 Age + \cdots + \delta_{28} D_2 SG + \delta_{31} D_3 Rx + \delta_{32} D_3 Age + \cdots + \delta_{38} D_3 SG]$$

1st row: predictors
Rx, Age, Wt, PF,. . ., SG
2nd row: products
D_2Rx, D_2Age, . . ., D_2SG
3rd row: products
D_3Rx, D_2Age, . . ., D_3SG
D_2 = CVD and D_3 = OTH are (0,1) dummy variables that distinguish the 3 event-types

Using the general **LM** data layout given in Table 9.11, the stratified Cox **LM** model for the Byar data that incorporates C = 3 event-types is shown at the left. The strata, denoted by g = 1, 2, 3, identify the three event-types as **Cancer, CVD**, and **Other**, respectively.

Notice that in the exponential part of the model, there are 3 rows of terms that correspond to the 3 event-types of interest. The first row contains p = 8 predictors Rx, Age, Wt, PF, HX, HG, SZ, SG. The second row contains product terms of the dummy variable D_2 (the **CVD** indicator) with each of the 8 predictors. Similarly, the third row contains product terms of D_3 (the **Other** indicator) with each of the predictors.

$$HR_{\mathbf{Ca}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1]$$
$$HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[(\beta_1 + \delta_{21})]$$
$$HR_{\mathbf{OTH}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[(\beta_1 + \delta_{31})]$$

From the above model, it follows that the hazard ratio formulas for the effects of Rx corresponding to each event-type are as shown at the left. Notice that for **CVD** and **Other** deaths, the coefficient δ_{g1} of the product term D_gRx, g = 2, 3, is added to the coefficient β_1 of Rx in the exponential term.

Table 9.14. Edited Output for LM Model (Interaction SC)-Byar Data

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Rx	1	−0.550	0.170	0.001	0.577
Age	1	0.005	0.142	0.970	1.005
Wt	1	0.187	0.138	0.173	1.206
PF	1	0.253	0.262	0.334	1.288
Hx	1	−0.094	0.179	0.599	0.910
HG	1	0.467	0.177	0.008	1.596
SZ	1	1.154	0.203	0.000	3.170
SG	1	1.343	0.202	0.000	3.830
RxCVD	1	0.905	0.244	0.000	2.471
AgeCVD	1	0.332	0.196	0.089	1.394
WtCVD	1	−0.146	0.203	0.472	0.864
PFCVD	1	0.222	0.377	0.556	1.248
HxCVD	1	1.236	0.259	0.000	3.441
HGCVD	1	−0.449	0.268	0.094	0.638
SZCVD	1	−1.375	0.417	0.001	0.253
SGCVD	1	−1.366	0.275	0.000	0.255
RxOth	1	−0.028	0.327	0.932	0.972
AgeOth	1	0.764	0.248	0.002	2.147
WtOth	1	0.344	0.265	0.194	1.411
PFOth	1	0.288	0.497	0.562	1.334
HxOth	1	0.117	0.337	0.727	1.125
HGOth	1	−0.111	0.345	0.748	0.895
SZOth	1	−0.439	0.470	0.350	0.645
SGOth	1	−1.797	0.360	0.000	0.166

log likelihood = −1831.92

Table 9.14 shows edited output obtained from fitting the above **LM** model.

The first eight rows of output in this table are identical to the corresponding eight rows of output in the previously shown **Table 9.1** obtained from Method 1, which fits a separate model for **Cancer** deaths only. This equivalence results because the first eight rows of the **LM** output correspond to the reduced version of the **LM** model when $D_2 = D_3 = 0$, which identifies **Cancer** as the event of interest.

However, the remaining 16 rows of **LM** output are **not** identical to the corresponding 8 rows of **Table 9.2** (for **CVD**) and 8 rows of **Table 9.3** (for **Other**). Note that the remaining 16 coefficients in the **LM** output identify the δ_{gj} coefficients in the **LM** model rather than the sum $(\beta_1 + \delta_{gj})$ required for computing the HR when g = 2 and 3.

$\widehat{HR}_{\mathbf{Ca}}(\text{Rx} = 1 \text{ vs. Rx} = 0)$
$= \exp[-0.550] = 0.577$
$= (1/1.733)$
Wald ChiSq $= (-.550/.171)^2$
$= 10.345(\text{P} = 0.001)$

95% CI for $\exp[\beta_{1\mathbf{Ca}}]$:
$\exp[-0.550 \pm 1.96(0.171)]$
$= (0.413, 0.807)$

From **Table 9.14**, the adjusted $\widehat{HR}$ for the effect of Rx when the event-type is **Cancer** can be read directly off the output as 0.577. Also, the Wald statistic for testing H_0: $\beta_1 = 0$ is highly significant (P = .001). The corresponding 95% confidence interval for this HR has the limits (0.413, 0.807).

LM results for **Cancer** identical to Method 1 results for **Cancer**

These results are identical to those obtained for the adjusted $\widehat{HR}$, the Wald test, and interval estimate obtained in **Table 9.1** using Method 1 to assess the effect of Rx on survival for cancer death.

$\widehat{HR}_{CVD}(\text{Rx} = 1 \text{ vs. Rx} = 0)$
$= \exp\left(\hat{\beta}_1 + \hat{\delta}_{11}\right)$
$= \exp(-0.550 + 0.905)$
$= \exp(0.355) = 1.426$
$\widehat{HR}_{OTH}(\text{Rx} = 1 \text{ vs. Rx} = 0)$
$= \exp\left(\hat{\beta}_1 + \hat{\delta}_{21}\right)$
$= \exp(-0.550 - 0.028)$
$= \exp(-0.578) = 0.561$

Using **Table 9.14** to obtain adjusted $\widehat{HR}$ for the Rx effect when the event-type is **CVD** or **Other**, we must exponentiate the sum $(\hat{\beta}_1 + \hat{\delta}_{g1})$ for g = 2 and 3, respectively.

LM results for **CVD** and **Other** identical to Method 1 results for **CVD** and **Other**

These results are shown at the left, and they are identical to those obtained in **Tables 9.2 and 9.3** using Method 1.

Wald test statistics for **CVD** and **Other**

$$\text{Wald}_{\text{CVD}} = \left[\frac{\hat{\beta}_1 + \hat{\delta}_{11}}{\text{SE}_{\hat{\beta}_1 + \hat{\delta}_{21}}}\right]^2$$

$$\text{Wald}_{\text{OTH}} = \left[\frac{\hat{\beta}_1 + \hat{\delta}_{11}}{\text{SE}_{\hat{\beta}_1 + \hat{\delta}_{21}}}\right]^2$$

Note, however, that using the **LM** model to obtain Wald test statistics and 95% confidence intervals for the HRs for **CVD** and **Other**, the mathematical formulas (shown at left for the Wald tests) require obtaining **standard errors of the sums** $(\hat{\beta}_1 + \hat{\delta}_{g1})$ for g = 2 and 3, whereas the output in Table 9.14 gives only **individual standard errors** of $\hat{\beta}_1, \hat{\delta}_{11}$ and $\hat{\delta}_{21}$.

Computer packages provide for computation of the above formulas

SAS: **Contrast** statement

STATA: **lincom** command

SAS ans STATA provide special syntax to specify the computer code for such computations: SAS's PHREG allows a "Contrast" statement; STATA allows a "lincom" command.

Alternative **LM** formulation (**LM_{alt}** model)

Output identical to Method 1 (**Tables 9.1, 9.2, 9.3**)

Nevertheless, there is an alternative version of the **LM** model that avoids the need for special syntax. This alternative formulation, which we call the **LM_{alt}** model, results in output that is identical to the output from the separate models (Method 1) approach for analyzing competing risk data as given in **Tables 9.1** through **9.3**.

IX. Method 2a: Alternative Lunn–McNeil (LM_{alt}) Approach

- Uses same data layout as **Table 9.12**
- Column headings:
  - Dummy variables $D_1, D_2, \ldots, D_C$
  - Predictor variables $X_1, X_2, \ldots, X_p$
- Above variables are transformed into product terms

The data layout required to fit the **LM_{alt}** model is the same as shown earlier in **Table 9.12**. However, the variables listed in the columns of this table, namely, the dummy variables $D_1, D_2, \ldots, D_C$ and the predictor variables $X_1, X_2, \ldots, X_p$, serve as basic variables that are transformed into product terms that define the **LM_{alt}** model.

1st row of **LM_{alt}** model:
product terms
$D_1 X_1, D_1 X_2, \ldots, D_1 X_p$
coefficients $\delta'_{11}, \ldots, \delta'_{1p}$

1st row of **LM** model
predictors $X_1, X_2, \ldots, X_p$
coefficients $\beta_1, \ldots, \beta_p$

The primary difference in the two formulas is that the first row of the exponential term in the **LM_{alt}** model contains product terms $D_1X_1, D_1X_2, \ldots, D_1X_p$ with coefficients denoted $\delta'_{11}, \ldots, \delta'_{1p}$ whereas the first row in the **LM** model contains the predictors $X_1, X_2, \ldots, X_p$ without product terms and coefficients denoted $\beta_1, \ldots, \beta_p$.

General Stratified Cox **LM_{alt}** *Model*

$$g = 1, \ldots, C$$

$$\begin{aligned} h'_g(t, \mathbf{X}) &= h'_{0g}(t) \\ &\times \exp\left[\delta'_{11} D_1 X_1 + \delta'_{12} D_1 X_2 + \cdots + \delta'_{1p} D_1 X_p \right. \\ &+ \delta'_{21} D_2 X_1 + \delta'_{22} D_2 X_2 + \cdots + \delta'_{2p} D_2 X_p \\ &+ \delta'_{31} D_3 X_1 + \delta'_{32} D_3 X_2 + \cdots + \delta'_{3p} D_3 X_p \\ &+ \cdots \\ &\left. + \delta'_{C1} D_C X_1 + \delta'_{C2} D_C X_2 + \cdots + \delta'_{Cp} D_C X_p \right] \end{aligned}$$

The general form of the **LM_{alt}** model is shown at the left. We have used a superscript prime ($'$) to distinguish the hazard model formula for the **LM_{alt}** model from the corresponding formula for the **LM** model given earlier.

- $\mathbf{LM}_{\text{alt}}$ and **LM** models are different
- Estimated regression coefficients will not be identical
- Estimated HRs, test statistics, and interval estimates are identical
- Computational formulas are different

Because the $\mathbf{LM}_{\text{alt}}$ model and the **LM** model are different hazard models, their estimated regression coefficients will not be identical. Nevertheless, when used on the same dataset, the estimated HRs of interest and their corresponding test statistics and interval estimates are identical even though the formulas used to compute these statistics are different for the two models.

$\mathbf{LM}_{\text{alt}}$ *Hazard Model for Event-Type 1*

$$h'_1(t,\mathbf{X}) = h'_{01}(t) \times \exp\left[\delta'_{11}X_1 + \delta'_{12}X_2 + \cdots + \delta'_{1p}X_p\right]$$

$$(D_1 = 1,\ D_2 = D_3 = \cdots = D_C = 0)$$

$$HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp\left[\delta'_{11}\right]$$

(no products $X_j X_i$ in model)

For g = 1 (i.e., event-type 1), the $\mathbf{LM}_{\text{alt}}$ model simplifies to the expression shown at the left. Note that because g = 1, the values of the dummy variables are $D_1 = 1$, and $D_2 = D_3 = \ldots = D_C = 0$.

Thus, for g = 1, if X_1 is a (0,1) variable, the other Xs are covariates, and there are no product terms of the form X_jX_1 in the model, the formula for the HR for the effect of X_1 adjusted for the covariates is $\exp[\delta'_{11}]$.

LM $HR = \exp[\beta_1]$

Recall that for the **LM** model, the corresponding HR formula also involved the coefficient of the X_1 variable, denoted as β_1.

$\mathbf{LM}_{\text{alt}}$ *Hazard Model for Event-Type g (> 1)*

$$h'_1(t,\mathbf{X}) = h'_{0g}(t) \times \exp\left[\delta'_{g1}X_1 + \delta'_{g2}X_2 + \cdots + \delta'_{gp}X_p\right]$$

$$\left(D_g = 1 \text{ and } D_{g'} = 0 \text{ for g' } \neq g\right)$$

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp\left[\delta'_{g1}\right]$$

(no products X_jX_1 in model)

For any g greater than 1, the general hazard model simplifies to a hazard function formula that contains only those product terms involving the subscript g, because $D_g = 1$ and $D_{g'} = 0$ for $g' \neq g$.

Thus, for g > 1, if X_1 is a (0,1) variable, the other Xs are covariates, and there are no products $X_j X_1$ in model, the formula for the HR for the effect of X_1 adjusted for the covariates is $\exp[\delta'_{g1}]$.

LM $HR = \exp[\beta_1 + \delta_{g1}]$

Recall that for the **LM** model, the exponential in the HR formula involved the sum $(\beta_1 + \delta_{g1})$.

Statistical inferences (i.e., Wald test, 95% CI)

LM_{alt} model: need standard error for $\hat{\delta}'_{g1}$ (directly provided by output)

LM model: standard error of $(\hat{\beta}_1 + \hat{\delta}_{g1})$. (more complicated computation)

Next: Byar data example of **LM_{alt}** model

Thus for $g > 1$, statistical inferences about HRs using the **LM_{alt}** model only require use of the standard error for $\hat{\delta}'_{g1}$ that is directly provided in the output.

In contrast, the **LM** model requires computing the more complicated standard error of the sum $(\hat{\beta}_1 + \hat{\delta}_{g1})$.

We now illustrate the above general **LM_{alt}** model formation using the Byar data.

LM_{alt} *SC Model for Byar Data*

$g = 1, 2, 3$

$$h'_g(t,\mathbf{X}) = h'_{0g}(t) \times \exp\left[\delta'_{11}D_1Rx + \cdots + \delta'_{18}D_1SG + \delta'_{21}D_2Rx + \cdots + \delta'_{28}D_2SG + \delta'_{31}D_3Rx + \cdots + \delta'_{38}D_3SG\right]$$

The stratified Cox (SC) **LM_{alt}** model that incorporates the $C = 3$ event-types is shown at the left. The strata, denoted by $g = 1, 2, 3$, identify the three event-types, **Cancer, CVD**, and **Other**.

$D_1 =$ **CA**, $D_2 =$ **CVD**, and $D_3 =$ **OTH** are (0,1) dummy variables for the 3 event-types

1st row: products
$D_1Rx, D_1Age, \ldots, D_1SG$
(**LM** predictors, Rx, Age,. . ., SG)
2nd row: products
$D_2Rx, D_2Age, \ldots, D_2SG$
3rd row: products
$D_3Rx, D_3Age, \ldots, D_3SG$

Notice that in the exponential part of the model, the first row contains product terms of the dummy variable D_1 (the **CA** indicator) with each of the 8 predictors Rx, Age, Wt, PF, HX, HG, SZ, SG. Recall that in the **LM** version of this model, the first row contained main effects of the predictors instead of product terms.

The second and third rows, as in the **LM** model, contain product terms of the dummy variable D_2 (the **CVD** indicator) and D_3 (the **OTH** indicator), respectively, with each of the 8 - predictors.

$$HR_{\mathbf{Ca}}(Rx = 1 \text{ vs. } Rx = 0) = \exp\left[\delta'_{11}\right]$$
$$HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0) = \exp\left[\delta'_{21}\right]$$
$$HR_{\mathbf{OTH}}(Rx = 1 \text{ vs. } Rx = 0) = \exp\left[\delta'_{31}\right]$$

From the above model, it follows that the HR formulas for the effects of Rx corresponding to each event-type are of the form $\exp(\delta'_{g1})$, where δ'_{g1} is the coefficient of the product term D_gRx in the **LM_{alt}** model.

$$\text{Wald}_g = \left[\frac{\hat{\delta}'_{g1}}{SE_{\hat{\delta}'_{g1}}}\right]^2$$

g = 1 (**CA**), 2 (**CVD**), 3 (**OTH**)

Consequently, Wald test statistics (shown at the left) and confidence intervals for these HRs use standard errors that are directly obtained from the standard error column from the output obtained for the $\mathbf{LM_{alt}}$ model.

Statistical inference information

$\mathbf{LM_{alt}}$ model: directly provided by output

LM model: not directly provided by output (requires additional computer code)

Thus, the $\mathbf{LM_{alt}}$ model allows the user to perform statistical inference procedures using the information directly provided in the computer output, whereas the **LM** model requires additional computer code to carry out more complicated computations.

Table 9.15. Edited Output for SC $\mathbf{LM_{alt}}$ Model—Byar Data

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
RxCa	1	−0.550	0.170	0.001	0.577
AgeCa	1	0.005	0.142	0.970	1.005
WtCa	1	0.187	0.138	0.173	1.206
PFCa	1	0.253	0.262	0.334	1.288
HxCa	1	−0.094	0.179	0.599	0.910
HGCa	1	0.467	0.177	0.008	1.596
SZCa	1	1.154	0.203	0.000	3.170
SGCa	1	1.343	0.202	0.000	3.830
RxCVD	1	0.354	0.174	0.042	1.429
AgeCVD	1	0.337	0.134	0.012	1.401
WtCVD	1	0.041	0.150	0.783	1.042
PFCVD	1	0.475	0.270	0.079	1.608
HxCVD	1	1.141	0.187	0.000	3.131
HGCVD	1	0.018	0.202	0.929	1.018
SZCVD	1	−0.222	0.364	0.542	0.801
SGCVD	1	−0.023	0.186	0.900	0.977
RxOth	1	−0.578	0.279	0.038	0.561
AgeOth	1	0.770	0.204	0.000	2.159
WtOth	1	0.532	0.227	0.019	1.702
PFOth	1	0.541	0.422	0.200	1.718
HxOth	1	0.023	0.285	0.935	1.023
HGOth	1	0.357	0.296	0.228	1.428
SZOth	1	0.715	0.423	0.091	2.045
SGOth	1	−0.454	0.298	0.127	0.635

log likelihood = − 1831.916

Table 9.15 shows edited output obtained from fitting the above $\mathbf{LM_{alt}}$ model.

The first eight rows of output in this table are identical to the eight rows of output in the previously shown **Table 9.1** obtained from Method 1, which fits a separate model for **Cancer** deaths only, treating **CVD** and Other deaths as censored.

The next eight rows in the table are identical to the eight rows of output in the previous **Table 9.2**, which fits a separate model for **CVD** deaths only, treating **Cancer** and **Other** deaths as censored.

The last eight rows in the table are identical to the eight rows of output in the previous **Table 9.3**, which fits a separate model for **Other** deaths only, treating **Cancer** and **CVD** deaths as censored.

Table 9.15 ($\mathbf{LM_{alt}}$) output identical to **Tables 9.1, 9.2, 9.3 (Method 1)** output combined

Thus, the output in Table 9.15 using the single $\mathbf{LM_{alt}}$ model gives identical results to what is obtained from fitting 3 separate models in **Tables 9.1, 9.2**, and **9.3**.

X. Method 1 (Separate Models) versus Method 2 (LM Approach)

Why bother with **LM** or $\mathbf{LM_{alt}}$ models when you can simply fit 3 separate models?

Answer: Can perform statistical inferences that cannot be done when fitting 3 separate models

The reader may have the following question at this point: Why bother with the **LM** or $\mathbf{LM_{alt}}$ models as long as you can get the same results from fitting three separate models using Method 1? The answer is that the **LM** or $\mathbf{LM_{alt}}$ model formulation allows for performing statistical inferences about various features of the competing risk models that cannot be conveniently assessed when fitting three separate models using Method 1.

LM *Model for Byar Data*

$$
\begin{aligned}
&g = 1, 2, 3\\
&h^*_g(t, \mathbf{X}) = h^*_{0g}(t)\\
&\quad \times \exp[\beta_1 RX + \beta_2 Age + \cdots + \beta_8 SG\\
&\quad + \delta_{21} D_2 Rx + \delta_{22} D_2 Age + \cdots + \delta_{28} D_2 SG\\
&\quad + \delta_{31} D_3 Rx + \delta_{32} D_3 Age + \cdots + \delta_{38} D_3 SG]
\end{aligned}
$$

We illustrate such "extra" inference-making using the **LM** model previously described for the Byar data example. This model is shown again at the left. Equivalent inferences can be made using the $\mathbf{LM_{alt}}$ model (see Exercises at end of this chapter).

Inference question: Byar data

No-interaction SC **LM** model
versus
interaction SC **LM** model

One inference question to consider for the Byar data is whether a no-interaction SC **LM** model is more appropriate than the interaction SC **LM** model defined above.

No-interaction SC model

$$
\begin{aligned}
&g = 1, 2, 3\\
&h^*_g(t, \mathbf{X}) = h^*_{0g}(t)\\
&\quad \times \exp[\beta_1 RX + \beta_2 Age + \cdots + \beta_8 SG]
\end{aligned}
$$

The no-interaction SC model is shown here at the left.

Assumes

$$
\begin{aligned}
HR_{\mathbf{CA}}(X_i) &= HR_{\mathbf{CVD}}(X_i)\\
&= HR_{\mathbf{OTH}}(X_i)\\
&\equiv HR(X_i) \text{ for any } X_i \text{ variable}
\end{aligned}
$$

This model assumes that the hazard ratio for the effect of a single predictor (say, binary) X_i adjusted for the other variables in the model is the same for each event-type of interest.

for example, Rx = 0 vs Rx = 1:

$$
\begin{aligned}
HR_{\mathbf{CA}}(Rx) &= HR_{\mathbf{CVD}}(Rx)\\
&= HR_{\mathbf{OTH}}(Rx)\\
&= \exp[\beta_1]
\end{aligned}
$$

For example, in the above no-interaction SC **LM** model the hazard ratio for the effect of Rx is $\exp[\beta_1]$ for each g, where β_1 is the coefficient of Rx.

H_0: all $\delta_{gj} = 0$,
$g = 2, 3; j = 1, 2, \ldots, 8$
where δ_{gj} is coefficient of $D_g X_j$ in the interaction SC **LM** model

To carry out the comparison of the interaction with the no-interaction SC **LM** models, the null hypothesis being tested is that the coefficients of the 16 product terms (δ_{gj}) in the interaction SC model are equal to zero.

Likelihood Ratio Test

$$LR = -2\log L_R - (-2\log L_F)$$
$$\text{approx } \chi^2_{16} \text{ under } H_0$$
$$R = \text{no interaction SC (reduced) model}$$
$$F = \text{interaction SC (full) model}$$

This null hypothesis is conveniently tested using the **LM** model with a likelihood ratio test statistic obtained by subtracting −2 log L statistics from the two models being compared. The degrees of freedom being tested is 16, the number of δ_{gj} coefficients being set equal to zero under H_0.

Table 9.16. Edited Output – No-Interaction SC **LM** Model–Byar Data

Var	DF	Coef	Std. Err.	p > |z|	Haz. Ratio
Rx	1	−0.185	0.110	0.092	0.831
Age	1	0.287	0.087	0.001	1.332
Wt	1	0.198	0.093	0.032	1.219
PF	1	0.402	0.170	0.018	1.495
Hx	1	0.437	0.112	0.000	1.548
HG	1	0.292	0.120	0.015	1.339
SZ	1	0.672	0.159	0.000	1.958
SG	1	0.399	0.115	0.001	1.491

log likelihood = −1892.091

Table 9.16 gives the output resulting from the no-interaction SC **LM** model for the Byar dataset. In this table, there is one coefficient corresponding to each of the eight predictors in the model, as should be the case for a no-interaction SC model. Nevertheless, baseline hazard functions $h^*_{0g}(t)$ are allowed to be different for different g even if the coefficients are the same for different g.

Table 9.16: Log likelihood$_R$ = −1892.091
Table 9.14: Log likelihood$_F$ = −1831.916

From **Table 9.16**, we find that the log-likelihood statistic for the reduced (no-interaction SC) model is −1892.091. From **Table 9.14** (or **9.15**), the log-likelihood statistic for the full (interaction SC) model is −1831.916.

$$LR = -2\log L_R - (-2\log L_F)$$
$$= -2(-1892.091) - (-2(-1831.916))$$
$$= 120.35 \text{ approx } \chi^2_{16} \text{ under } H_0$$
$$(P < 0.001)$$

The likelihood ratio test statistic (LR) is then calculated to be 120.35, as shown at the left. This statistic has an approximate chi-square distribution with 16 degrees of freedom under H_0.

Reject H_0: interaction SC model more appropriate than no-interaction SC model

The P-value is less than .001, which indicates a highly significant test result, thus supporting use of the full-interaction SC model.

Cancer and **CVD** very different clinically

$$\Downarrow$$

$$HR_{\mathbf{Ca}}(Rx = 1 \text{ vs. } 0) \neq HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } 0)$$

For the Byar dataset, the decision to reject the no-interaction SC model makes sense when considering that two of the competing risks are **Cancer** deaths and **CVD** deaths. Because **Cancer** and **CVD** are clinically very different diseases, one would expect the effect of any of the predictors, particularly Rx, on time to failure to be different for these different disease entities.

DIFFERENT STUDY EXAMPLE

Competing risks: Stage 1 vs. Stage 2 Breast Cancer

$$\Downarrow$$

$$HR_{stg1}(Rx = 0 \text{ vs. } 1) = HR_{stg2}(Rx = 0 \text{ vs. } 1)$$

$$\Downarrow$$

No-interaction SC Cox reasonable depending on similarity of competing risks

Suppose, however, the competing risks for a different study had been, say, two stages of breast cancer. Then it is plausible that the effect from comparing two treatment regimens might be the same for each stage. That is, a no-interaction SC **LM** model may be (clinically) reasonable depending on the (clinical) similarity between competing risks.

Unstratified **LM** model ($\mathbf{LM}_U$):

$$\begin{aligned} h^*(t, \mathbf{X}) &= h_0^*(t) \\ &\times \exp[\gamma_1 CVD + \gamma_2 OTH \\ &+ \beta_1^\bullet Rx + \beta_2^\bullet Age + \cdots + \beta_8^\bullet SG \\ &+ \delta_{21}^\bullet D_2 Rx + \delta_{22}^\bullet D_2 Age + \cdots + \delta_{28}^\bullet D_2 SG \\ &+ \delta_{31}^\bullet D_3 Rx + \delta_{32}^\bullet D_3 Age + \cdots + \delta_{38}^\bullet D_3 SG] \end{aligned}$$

Returning again to the Byar data example, another variation of the **LM** model is shown at the left and denoted $\mathbf{LM}_U$. This is a Cox PH model applied to the augmented data of **Table 9.12** that is not stratified on the competing risks (i.e., there is no subscript g in the model definition). We have used a superscript bullet ($\bullet$) to distinguish the $\mathbf{LM}_U$ model from the **LM** and $\mathbf{LM_{alt}}$ models.

$\mathbf{LM}_U$ model: **CVD** and **OTH** included in model

LM model: **CVD** and **OTH** not included in model

(Both $\mathbf{LM}_U$ and **LM** models use augmented dataset)

The $\mathbf{LM}_U$ model includes the two event-type dummy variables **CVD** and **OTH** in the model, rather than stratifying on these variables. As for the **LM** model, the fit of the $\mathbf{LM}_U$ model is based on the augmented dataset given in **Table 9.12**.

$\mathbf{LM}_U$ model: need to check PH assumption (Chapter 4)

PH assumption not satisfied

$$\Downarrow$$

Use **LM** instead of $\mathbf{LM}_U$ model

Because the $\mathbf{LM}_U$ model is an unstratified Cox PH model, we would want to use the methods of Chapter 4 to assess whether the PH assumption is satisfied for the **CVD** and **OTH** variables (as well as the other variables). If the PH assumption is found wanting, then the (stratified Cox) **LM** model should be used instead.

PH assumption satisfied
⇓
Determine HRs using exponential formula (Chapter 3)

If the PH assumption is satisfied, hazard ratios for the effects of various predictors in the $\mathbf{LM}_U$ model can be determined using the standard exponential formula described in Chapter 3 for the Cox PH model.

Cancer survival (**CVD** = **OTH** = 0):

$$\mathrm{HR}_{\mathbf{Ca}}(\mathrm{Rx}=1 \text{ vs. } \mathrm{Rx}=0) = \exp\left[\beta_1^{\bullet}\right]$$

CVD survival (**CVD** = 1, **OTH** = 0):

$$\mathrm{HR}_{\mathbf{CVD}}(\mathrm{Rx}=1 \text{ vs. } \mathrm{Rx}=0) = \exp\left[\gamma_1+\beta_1^{\bullet}+\delta_{21}^{\bullet}\right]$$

Other survival (**CVD** = 0, **OTH** = 1):

$$\mathrm{HR}_{\mathbf{OTH}}(\mathrm{Rx}=1 \text{ vs. } \mathrm{Rx}=0) = \exp\left[\gamma_2+\beta_1^{\bullet}+\delta_{31}^{\bullet}\right]$$

In particular, to obtain the hazard ratio for the effect of Rx on **Cancer** survival, we would specify **CVD** = **OTH** = 0 in the model and then exponentiate the coefficient of the Rx variable in the model, as shown at the left.

Similar HR expressions (but involving γ_1 and γ_2 also) are obtained for the effect of Rx when **CVD** deaths and **Other** deaths are the event-types of interest.

Essential point

Use of single **LM**-type model offers greater flexibility for the analysis than allowed using Method 1

At this point, we omit further description of results from fitting the $\mathbf{LM}_U$ model to the Byar dataset. The essential point here is that the use of a single **LM**-type model with augmented data allows greater flexibility for the analysis than can be achieved when using Method 1 to fit separate hazard models for each event-type of interest.

XI. Summary

Competing Risks

Each subject can experience only one of several different types of events over follow-up

This chapter has considered survival data in which each subject can experience only one of several different types of events over follow-up. The different events are called **competing risks**.

Typical approach

- Cox PH model
- Separate model for each event-type
- Other (competing) event-types treated as censored

We have described how to model competing risks survival data using a Cox PH model. The typical approach for analyzing competing risks data is to perform a survival analysis for each event-type separately, where the other (competing) event-types are treated as censored categories.

Drawbacks

1. Require independent competing risks that is, censored subjects have same risk as non-censored subjects in risk set

There are two primary drawbacks to the above method. One problem is the requirement that competing risks be independent. This assumption will not be satisfied if those subjects censored from competing risks do not have the same risk for failing as subjects who are not censored from the cause-specific event of interest at that same time.

2. Product-limit (e.g., KM) curve has questionable interpretation

A second drawback is that the estimated product-limit survival curve obtained from fitting separate Cox models for each event-type has questionable interpretation when there are competing risks.

Several alternative strategies regarding independence assumption: No single strategy is always best

Regarding the independence assumption, several alternative strategies for addressing this issue are described, although no single strategy is always best.

Sensitivity analysis: worst-case violations of independence assumption

For example, subjects censored from competing risks treated in analysis as if

- All event-free
- All experience event of interest

A popular strategy is a sensitivity analysis, which allows the estimation of parameters by considering worst-case violations of the independence assumption. For example, subjects censored from competing risks might be treated in the analysis as either all being event-free or all experiencing the event of interest.

- Independence assumption not easily verifiable
- Typical analysis assumes independence assumption is satisfied

Unfortunately, the independence assumption is not easily verifiable. Consequently, the typical competing risks analysis assumes that the independence assumption is satisfied even if this is not the case.

CIC *Alternative to KM*

- Derived from cause-specific hazard function
- Estimates **marginal probability** when competing risks are present
- Does not require independence assumption
- Useful to assess treatment utility in cost-effectiveness analyses

To avoid a questionable interpretation of the KM survival curve, the primary alternative to using KM is the **Cumulative Incidence Curve** (**CIC**), which estimates the **marginal probability** of an event. Marginal probabilities are relevant for assessing treatment utility whether competing risks are independent.

$$\mathrm{CIC}(t_{(f)}) = \sum_{f'=1}^{f} \hat{I}_{\mathbf{c}}(t_{f'}) = \sum_{f'=1}^{f} \hat{S}(t_{f'-1})\hat{h}_{\mathbf{c}}(t_{f'})$$

The formula for the calculating the **CIC** is shown at the left. The $h_{\mathbf{c}}(t_f)$ in the formula is the estimated hazard at survival time t_f for the event-type (**c**) of interest. The term $S(t_{f-1})$ denotes the **overall** survival probability of previous time (t_{f-1}), where "overall survival" indicates a subject that survives all competing events.

$\hat{h}_{\mathbf{c}}(t_f)$ = estimated hazard at ordered failure time t_f for the event-type (**c**)

$S(t_{f-1})$ = **overall** survival probability of previous time (t_{f-1})

CIC

- Does not use product limit formulation
- Not included in mainstream commercially available statistical packages (e.g., SAS, STATA, SPSS, R)

As the formula indicates, the **CIC** is not estimated using a product-limit formulation. Also, its computation is not included in mainstream commercially available standard statistical packages.

PH model used to obtain **CIC**

⇓

Independence of competing risks required

If a proportional hazard model is used to obtain hazard ratio estimates for individual competing risks as an intermediate step in the computation of a **CIC**, the assumption of independent competing risks is still required.

Modeling **CIC** with covariates using PH model: Fine and Gray (1999)

Software available (Gebski, 1997) Fine and Gray model analogous to Cox PH model

Recent work of Fine and Gray (1999) provides methodology for modeling the **CIC** with covariates using a proportional hazards assumption. Software is available for this method (Gebski, 1997, Tai et al., 2001), although not in standard commercial packages.

Alternative to **CIC**

$\mathbf{CPC_c} = \Pr(T_{\mathbf{c}} \leq t \mid T \geq t)$

where $T_{\mathbf{c}}$ = time until event c occurs

T = time until any competing risk event occurs

An alternative to the **CIC** is the Conditional Probability Curve (**CPC**). For risk type **c**, $\mathbf{CPC_c}$ is the probability of experiencing an event **c** by time t, *given that an individual has not experienced any of the other competing risks by time t*.

$\mathbf{CPC_c} = \mathrm{CIC_c}/(1 - \mathrm{CIC_{c'}})$

where $\mathrm{CIC_{c'}}$ = CIC from risks other than **c**

The **CPC** can be computed from the **CIC** through the formula $\mathbf{CPC_c} = \mathrm{CIC_c}/(1 - \mathrm{CIC_{c'}})$, where $\mathrm{CIC_{c'}}$ is the cumulative incidence of failure from risks other than risk **c** (i.e., all other risks considered together).

Tests to compare CPCs:

Pepe and Mori (1993) – 2 curves
Lunn (1998) – g curves

Pepe–Mori provide a test to compare two **CPC** curves. Lunn (1998) extended this test to g-groups and allows for strata.

Method 2: **LM** Approach

- Uses a single Cox (PH) model
- Gives identical results as obtained from Method 1
- Allows flexibility to perform statistical inferences not available from Method 1

We have also described an alternative approach, called the **Lunn–McNeil** (**LM**) approach, for analyzing competing risks data. The **LM** approach allows only one model to be fit rather than separate models for each event-type (Method 1). This method is equivalent to using Method 1. The **LM** model also allows the flexibility to perform statistical inferences to determine whether a simpler version of an initial **LM** model is more appropriate.

Augmented Data for ith Subject at Time t Using **LM** Approach

Subj	Stime	Status	D_1	D_2	$D_3 \ldots D_C$	$X_1 \ldots X_p$
i	t_i	e_1	1	0	0 … 0	$X_{i1} \ldots X_{ip}$
i	t_i	e_2	0	1	0 … 0	$X_{i1} \ldots X_{ip}$
i	t_i	e_3	0	0	1 … 0	$X_{i1} \ldots X_{ip}$
i	t_i	e_C	0	0	0 … 1	$X_{i1} \ldots X_{ip}$

To carry out the **LM** approach, the data layout must be augmented. If there are C competing risks, the original data must be duplicated C times, one row for each failure type.

$g = 1, 2, \ldots, C$

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp\Big[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \delta_{21} D_2 X_1 + \delta_{22} D_2 X_2 + \cdots + \delta_{2p} D_2 X_p + \delta_{31} D_3 X_1 + \delta_{32} D_3 X_2 + \cdots + \delta_{3p} D_3 X_p + \cdots + \delta_{C1} D_C X_1 + \delta_{C2} D_C X_2 + \cdots + \delta_{Cp} D_C X_p\Big]$$

To use the **LM** approach with augmented data to obtain identical results from fitting separate models (Method 1), an interaction version of a stratified Cox (SC) PH model is required. A general form for this model is shown at the left.

LM model: need standard error of $\left(\hat{\beta}_1 + \hat{\delta}_{g1}\right)$

(special syntax required for computation)

The **LM** model can be used to obtain Wald test statistics and 95% confidence intervals for HRs separately for each competing risk. These statistics require obtaining standard errors of the sums of estimated regression coefficients (e.g., $\hat{\beta}_1 + \hat{\delta}_{g1}$). Such computations require special syntax available in standard computer packages such as SAS, STATA, SPSS, and R.

Alternative **LM** *formulation*: $\mathbf{LM_{alt}}$ model

Nevertheless, there is an alternative formation of the **LM** model that avoids the need for special syntax. This alternative formulation, called the $\mathbf{LM_{alt}}$ model, yields output that is identical to the output from the separate models (Method 1) approach for analyzing competing risk data.

$\mathbf{LM_{alt}}$ yields output identical to Method 1

1st row of $\mathbf{LM_{alt}}$ model
$\Downarrow$
Product terms D_1X_1, D_1X_2, ..., D_1X_p Coefficients $\delta'_{11}, \ldots, \delta'_{1p}$

1st row of **LM** model
$\Downarrow$
Predictors $X_1, X_2, \ldots, X_p$
Coefficients $\beta_1, \ldots, \beta_p$

The primary difference in the two formulas is that the first row of the exponential term in the $\mathbf{LM_{alt}}$ model contains product terms D_1X_1, D_1X_2, ..., D_1X_{p1} with coefficients denoted $\delta'_{11}, \ldots, \delta'_{1p}$ whereas the first row in the **LM** model contains the predictors X_1, X_2, ..., X_p without product terms and coefficients denoted $\beta_1, \ldots, \beta_p$.

$\mathbf{LM_{alt}}$ model: $\text{Wald}_g = \left[\dfrac{\delta'_{g1}}{\text{SE}_{\hat{\delta}'_{g1}}}\right]^2$ directly obtained from output

Using the $\mathbf{LM_{alt}}$ model, Wald test statistics (shown at the left) and confidence intervals use standard errors that are directly obtained from the standard error column from the output obtained for the $\mathbf{LM_{alt}}$ model.

Statistical inference information

$\mathbf{LM_{alt}}$ model: directly provided by output
LM model: not directly provided by output (requires additional computer code)

Thus, the $\mathbf{LM_{alt}}$ model allows the user to perform statistical inference procedures using the information directly provided in the computer output, whereas the **LM** model requires additional computer code to carry out more complicated computations.

Advantage of **LM** (Method 2) over method 1:
LM offers flexibility for statistical inferences to consider simpler models

An advantage of using either the **LM** or $\mathbf{LM_{alt}}$ approach instead of fitting separate models (Method 1) is the flexibility to perform statistical inferences that consider simpler versions of an interaction SC **LM** model.

For example,
No-interaction SC **LM** model
versus
interaction SC **LM** model

Unstratified **LM** model
versus
SC **LM** model

For example, one inference question to consider is whether a no-interaction SC **LM** model is more appropriate than an interaction SC model. A different question is whether an unstratified **LM** model is more appropriate than a stratified **LM** model. These questions can be conveniently addressed using a single (i.e., **LM**) model instead of fitting separate models (Method 1).

Overall,

- Can use standard computer packages
- Independence assumption required

Overall, in this chapter, we have shown that competing risks data can be analyzed using standard computer packages provided it can be assumed that competing risks are independent.

Detailed Outline

I. **Overview** (page 430)
   A. Focus: **competing risks** – analysis of survival data in which each subject can experience only one of different types of events over follow-up.
   B. Analysis using Cox PH model.
   C. Drawbacks to typical approach that uses Cox model.
   D. Alternative approaches for analysis.

II. **Examples of Competing Risks Data** (pages 430–432)
   A. Dying from either lung cancer or stroke.
      1. Assess whether lung cancer death rate in exposed" persons is different from lung cancer death rate in "unexposed," allowing for competing risks.
      2. Also, compare lung cancer with stroke death rates controlling for predictors.
   B. Advanced cancer patients either dying from surgery or getting hospital infection.
      1. If focus on hospital infection failure, then death from surgery reduces burden of hospital infection control required.
   C. Soldiers dying in accident or in combat.
      1. Focus on combat deaths.
      2. If entire company dies from accident on way to combat, then KM survival probability for combat death is undefined.
      3. Example illustrates that interpretation of KM curve may be questionable when there are competing risks.
   D. Limb sarcoma patients developing local recurrence, lung metastasis, or other metastasis.
      1. None of failure types involves death, so recurrent events are possible.
      2. Can avoid problem of recurrent events if focus only on time to first failure.
      3. Analysis of recurrent events and competing risks in same data not addressed.

III. **Byar Data** (pages 433–434)
   A. Randomized clinical trial comparing treatments for prostate cancer.
   B. Three competing risks: deaths from prostate cancer; CVD, or other causes.

C. Covariates other than treatment are Age, Weight (Wt), Performance Status (PF), History of CVD (Hx), Hemoglobin (Hg), Lesion size (SZ), and Gleeson score (SG).

D. Competing risks considered independent, for example, death from CVD independent of death from death from cancer.

IV. Method 1: Separate Models for Different Event Types (pages 434–437)

A. Use Cox (PH) model to estimate separate hazards and HRs for each failure type, where other competing risks are treated as censored in addition to usual reasons for censoring: loss to follow-up, withdrawal from study, or end of study.

B. Cause-specific hazard function:
$h_{\mathbf{c}}(t) = \lim_{\Delta t \to 0} P(t \le T_{\mathbf{c}} < t + \Delta t | T_{\mathbf{c}} \ge t)/\Delta t$
where $T_{\mathbf{c}}$ = time-to-failure from event **c,** $\mathbf{c} = 1, 2, \ldots, C$ (# of event types).

C. Cox PH cause-specific model (event-type **c**):

$$h_{\mathbf{c}}(t, \mathbf{X}) = h_{0\mathbf{c}}(t) \exp\left[\sum_{i=1}^{p} \beta_{i\mathbf{c}} X_i\right]$$

where $\mathbf{c} = 1, \ldots, C$, and $\beta_{i\mathbf{c}}$ allows effect of X_i to differ by event-type.

D. Byar data example: Cancer, CVD, Other Deaths are C = 3 competing risks.

1. Cause-specific (no-interaction) model for **Cancer:**

$$\begin{aligned} h_{\mathbf{Ca}}(t, \mathbf{X}) = h_{0\mathbf{Ca}}(t) \exp[&\beta_{1\mathbf{Ca}} Rx + \beta_{2\mathbf{Ca}} Age \\ &+ \beta_{3\mathbf{Ca}} Wt + \beta_{4\mathbf{Ca}} PF + \beta_{5\mathbf{Ca}} Hx \\ &+ \beta_{6\mathbf{Ca}} HG + \beta_{7\mathbf{Ca}} SZ + \beta_{8\mathbf{Ca}} SG] \end{aligned}$$

where **CVD** and **Other** deaths treated as censored observations

$$HR_{\mathbf{Ca}}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1\mathbf{Ca}}]$$

2. Separate cause-specific (no-interaction) models for **CVD** and **Other**.

3. Edited output presented for each cause-specific model:

a. Cause-specific **Cancer** results for RX (with **CVD** and **Other** censored):

$$\widehat{HR}_{\mathbf{Ca}}(RX = 1 \text{ vs. } Rx = 0) = 0.575 \; (P = 0.001)$$

b. Cause-specific **CVD** results for RX (with **Cancer** and **Other** censored):

$$\widehat{HR}_{\mathbf{CVD}}(\mathrm{RX}=1 \text{ vs. } \mathrm{Rx}=0)=1.429\ (\mathrm{P}=0.040)$$

c. Cause-specific **Other** results for RX (with **Cancer** and **CVD** censored):

$$\widehat{HR}_{\mathbf{OTH}}(\mathrm{RX}=1 \text{ vs. } \mathrm{Rx}=0)=0.560\ (\mathrm{P}=0.038)$$

V. The Independence Assumption (pages 437–443)

A. **Independent** censoring: h(t) for censored subjects at time t is the same as for non-censored subjects in the same subgroup at time t

1. Typical (chapter 1) context: no competing risks;
2. Informative censoring can lead to bias results.

B. **(Independent)** censoring with **competing risks.** Censored subjects in the risk set at time t with a given set of covariates have the same failure rate as non-censored subjects in the risk set at time t with the same set of covariates regardless of whether the reason for censoring is a competing risk, withdrawal from study, or loss to follow-up.

1. Non-independent censoring: Subjects in the risk set at time t who are censored from a competing risk do not have the same failure rate as non-censored subjects in the risk set at time t.
2. Synonym: **Competing risks are independent**.

C. **Assessing the independence assumption**.

1. No method available to directly assess the independence assumption nor guarantee unbiased estimates if independence assumption is violated.
2. Consequently, the typical analysis of competing risks assumes that the independence assumption is satisfied, even if not.
3. Strategies for considering independence assumption
   a. Decide that assumption holds on clinical/biological/other grounds:
   b. Include in your model variables that are common risk factors for competing risks.

c. Use a frailty model containing a random effect that accounts for competing risks.
d. Perform a sensitivity analysis by considering worst-case violations of independence assumption.
e. All of above strategies rely on assumptions that cannot be verified from observed data.

4. Example of sensitivity analysis using Byar data.
a. Treat all subjects that die of competing risks CVD and Other as Cancer deaths.
b. Treat all subjects that die of competing risks CVD and Other as surviving as long as the largest survival time in the study.
c. Results suggest that if competing risks are not independent, then conclusions about the effect of Rx could be very different.
d. Alternative sensitivity approach: randomly select a subset (e.g., 50%) of subjects who have CVD or Other deaths and assume everyone in subset dies of Cancer.

VI. Cumulative Incidence Curves (CIC)
(pages 444–453)

A. Hypothetical study: n = 100 subjects, all subjects with prostate cancer

Survt (months)	# Died	Cause
3	99	CVD
5	1	Cancer

Study goal: cause-specific Cancer survival
Censored: CVD deaths
$KM_{\mathbf{Ca}}$: $S_{\mathbf{Ca}}(t = 5) = 0$ and $Risk_{\mathbf{Ca}}(T \leq 5) = 1$

B. How many of 99 deaths from CVD would have died from Cancer if not dying from CVD?
1. No answer is possible because those with CVD deaths cannot be observed further.
2. Sensitivity analysis A: 99 CVD deaths die of Cancer at t = 5.
a. $KM_{\mathbf{Ca}}$: $S_{\mathbf{Ca}}(t = 5) = 0$ and $Risk_{\mathbf{Ca}}(T \leq 5) = 1$ because KM assumes independent censoring; that is, those censored at t = 3 were as likely to die from cancer at t = 5 as those who were in the risk set at t = 5.
b. Same KM result as obtained for actual data.

3. Sensitivity analysis B: 99 CVD deaths survive past $t = 5$.
   a. $KM_{\mathbf{Ca}}$: $S_{\mathbf{Ca}}(t = 5) = 0.99$ and $Risk_{\mathbf{Ca}}(T \leq 5) = 0.01$.
   b. Different KM result than actual data.
   c. Can be derived directly from actual data as a **marginal probability**.
4. Main point: KM survival curve may not be very informative.

C. **Cumulative Incidence Curve (CIC)**: alternative to KM for competing risks.
1. Derived from cause-specific hazard function.
2. Estimates marginal probability.
3. Does not require independence assumption.
4. Has a meaningful interpretation in terms of treatment utility.
5. CIC formula:

$$\mathbf{CIC}(t_f) = \sum_{f'=1}^{f} \hat{I}_c(t_{f'}) = \sum_{f'=1}^{f} \hat{S}(t_{f'-1})\hat{h}_{\mathbf{c}}(t_{f'})$$

6. Calculation of CIC for another hypothetical dataset.
7. Tests have been developed (Pepe and Mori, 1992) for comparing the equality of **CIC**s for two or more groups: analogous to log rank test.
8. When a PH model is used to obtain hazard ratio estimates in the computation of a **CIC**, the independence of competing risks is required.
9. Fine and Gray (1999) provide methodology for modeling the **CIC** (also called **subdistribution function**) with covariates using a proportional hazards assumption: analogous to fitting Cox PH model.
10. Example of Fine and Gray output compared with Cox PH output for Byar data.

VII. Conditional Probability Curves (CPC) (pages 453–455)

A. $\mathbf{CPC_c} = Pr(T_{\mathbf{c}} \leq t | T \geq t)$ where $\mathbf{T_c}$ = time until event **c** occurs, T = time until any competing risk event occurs.
B. Formula in terms of CIC: $\mathbf{CPC_c} = CIC_{\mathbf{c}}/(1 - CIC_{\mathbf{c}'})$ where $CIC_{\mathbf{c}'}$ = CIC from risks other than **c**.
C. Graphs of CPCs can be derived from graphs of CICs.
D. Tests to compare CPCs: Pepe and Mori (1993) – 2 curves; Lunn (1998) – g curves.

VIII. Method 2—The Lunn-McNeil (LM) Approach (pages 455–461)

A. Allows only one Cox PH model to be fit rather than fitting separate modes for each event type (Method 1).

B. **LM** uses an augmented data layout.

1. for ith subject at time t_i, layout has C rows of data, where C = # event-types.
2. Dummy variables D_1, D_2,..., D_C are created to distinguish the C event-types.
3. The Status variable, $e_{\mathbf{c}}$, $\mathbf{c} = 1,\ldots, C$, equals 1 for event-type **c** and 0 otherwise.
4. Predictors are denoted by $X_1,\ldots, X_p$.
5. Example of data layout for Byar dataset.

C. General form of **LM** model (interaction SC model).

$$\underset{g=1,2,\ldots,C}{h^*_{\mathbf{g}}(t,\mathbf{X})} = h^*_{0g}(t)\exp\Big[\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p + \delta_{21}D_2X_1+\delta_{22}D_2X_2+\cdots+\delta_{2p}D_2X_p + \delta_{31}D_3X_1+\delta_{32}D_3X_2+\cdots+\delta_{3p}D_3X_p + \cdots + \delta_{C1}D_CX_1+\delta_{C2}D_CX_2+\cdots+\delta_{Cp}D_CX_p\Big]$$

1. **LM** model for event-type g = 1:
   a. $h^*_1(t,\mathbf{X}) = h^*_{01}(t) \times \exp\Big[\beta_1X_1+\beta_2X_2+\cdots+\beta_pX_p\Big]$
   b. $D_2 = D_3 = \ldots = D_C = 0$
   c. $HR_{\mathbf{g=1}}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\beta_1]$
2. **LM** model for event-type g (> 1):
   a. $h^*_g(t,\mathbf{X}) = h^*_{0g}(t) \times \exp\big[(\beta_1+\delta_{g1})X_1+(\beta_2+\delta_{g2})X_2 + \cdots+\big(\beta_p+\delta_{gp}\big)X_p\Big]$
   b. $HR_{\mathbf{g}}(X_1 = 1 \text{ vs. } X_1 = 0 = \exp[(\beta_1+\delta_{g1})]$

D. **LM** model for Byar data.

1. $h^*_g(t,X) = h^*_{0g}(t) \times \exp[\beta_1\,Rx + \beta_2Age+\cdots+\beta_8SG + \delta_{21}D_2Rx+\delta_{22}D_2Age+\cdots+\delta_{28}D_2SG + \delta_{31}D_3Rx+\delta_{32}D_3Age+\cdots+\delta_{38}D_3SG]$
   $g = 1,2,3$
2. D_2 = CVD and D_3 = OTH are (0,1) dummy variables for 3 event-types.

3. $HR_{\mathbf{Ca}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1]$
   $HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_{21}]$
   $HR_{\mathbf{OTH}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_{31}]$
4. Mathematical formulas for Wald tests and confidence intervals require standard errors for sums $\left(\hat{\beta}_1 + \hat{\delta}_{g1}\right)$ for g = 2 and 3; requires special syntax for computer.
5. Output provided and compared to output from Method 1.

IX. Method 2a: Alternative Lunn–McNeil ($\mathbf{LM_{alt}}$) Approach (pages 461–464)

A. An alternative LM formulation that gives identical output to that obtained from Method 1 (separate models approach).
B. Same data layout as for LM model, but only product terms in $\mathbf{LM_{alt}}$ model.
C. General form of $\mathbf{LM_{alt}}$ (interaction SC) model:

$$
\begin{aligned}
&g = 1, \ldots, C \\
&h'_g(t,\mathbf{X}) = h'_{0g}(t) \\
&\quad \times \exp\left[\delta'_{11}D_1X_1 + \delta'_{12}D_1X_2 + \cdots + \delta'_{1p}D_1X_p\right. \\
&\quad + \delta'_{21}D_2X_1 + \delta'_{22}D_2X_2 + \cdots + \delta'_{2p}D_2X_p \\
&\quad + \cdots \\
&\quad \left. + \delta'_{C1}D_CX_1 + \delta'_{C2}D_CX_2 + \cdots + \delta'_{Cp}D_2X_p\right]
\end{aligned}
$$

D. Hazard ratio formula involves only coefficients of product terms for all g:

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp\left[\delta'_{g1}\right], \quad g = 1, 2, 3$$

   a. Statistical inference information directly provided by $\mathbf{LM_{alt}}$ output.

E. Example of output for $\mathbf{LM_{alt}}$ model using Byar data.

X. Method 1 (Separate Models) versus Method 2 (LM Approach) (pages 465–468)

A. **LM** and $\mathbf{LM_{alt}}$ models allow flexibility to perform statistical inferences about features of competing risks model not conveniently available using separate models (Method 1) approach.
B. **LM** and $\mathbf{LM_{alt}}$ models can assess whether a no-interaction SC model is more appropriate than the initial interaction SC **LM** model.

C. Example of comparison of no-interaction with interaction SC model using Byar data.
D. **LM** and $\mathbf{LM_{alt}}$ models can assess whether an unstratified LM model (called $\mathbf{LM_U}$) is more appropriate than a stratified **LM** model.
E. Example of $\mathbf{LM_U}$ model involving Byar data.

XI. Summary (pages 468–473)

A. Competing risks survival data can be analyzed using Cox PH models and standard computer packages.
B. There are two alternative methods that use a Cox PH model formulation.
   1. Fit separate models for each cause-specific event type, treating the remaining event types as censored.
   2. Use the **Lunn–McNeil (LM)** approach to fit a single model that incorporates the analysis for each cause-specific event.
C. Each of the above approaches requires that competing risks be independent (i.e., independent censoring).
D. Without the independence assumption, methods for competing risks analysis are unavailable.
E. The **Cumulative Incidence Curve (CIC)** or the **Conditional Probability Curve (CPC)** are alternatives to the KM curve, when use of a KM curve has questionable interpretation.

Practice Exercises

Answer questions 1 to 15 as true or false (circle T or F).

T F 1. A competing risk is an event-type (i.e., failure status) that can occur simultaneously with another event of interest on the same subject.
T F 2. An example of competing risks survival data is a study in which patients receiving radiotherapy for head and neck cancer may either die from their cancer or from some other cause of death.
T F 3. if all competing risks in a given study are different causes of death, then it is possible to have both competing risks and recurrent events in the same study.

T F 4. Suppose patients with advanced-stage cancer may die after surgery before their hospital stay is long enough to get a hospital infection. then such deaths from surgery reduce the hospital's burden of infection control.

T F 5. The typical approach for analyzing competing risks using a Cox PH model involves fitting separate models for each competing risk ignoring the other competing risks.

T F 6. Suppose that a cause-specific risk of interest is development of lung metastasis, and a competing risk is local recurrence of a lung tumor. Then a patient who develops a local recurrence is treated as a failure in a competing risk analysis.

T F 7. When there are no competing risks, then any study subject in the risk set at a given time has the same risk for failing as any other subject in the risk set with the same values for covariate predictors at time t.

T F 8. If, when analyzing competing risks survival data, it is assumed that censoring is independent, then a subject in the risk set at time t is as likely to fail from any competing risk as to be lost to follow-up.

T F 9. When a sensitivity analysis indicates that a worst-case scenario gives meaningfully different results from an analysis that assumes independence of competing risks, then there is evidence that the independence assumption is violated.

T F 10. The typical competing risk analysis assumes that competing risks are independent even if this assumption is not true.

T F 11. The Cumulative Incidence Curve (CIC) provides risk estimates for the occurrence of a cause-specific event in the presence of competing risks.

T F 12. CIC = 1 − KM, where KM denotes the Kaplan–Meier curve.

T F 13. A CIC for a cause-specific event that ignores the control of covariates does not require the assumption of independent competing risks.

T F 14. A Cumulative Probability Curve (CPC) gives the probability of experiencing an event c by time t, given that an individual has experienced any of the other competing risks by time t.

T F 15. If $CIC_c = .4$, then $CPC = .4/.6 = .667$.

T F 16. The Lunn–McNeil (LM) approach fits a single stratified Cox model using an augmented dataset to obtain the same results as obtained by fitting separate Cox models for each cause-specific competing risk.

T F 17. An advantage of the Lunn–McNeil (LM) approach over the approach that fits separate Cox models is that the LM approach allows for testing whether a no-interaction SC model might be preferable to an interaction SC model.

T F 18. **Given the LM model** stratified on two cause-specific events, Cancer and CVD:

$$h^*_g(t, \mathbf{X}) = h^*_{0g}(t) \exp[\beta_1 Rx + \beta_2 Age + \delta_1(D \times Rx) + \delta_2(D \times Age)],$$

$$g = 1,\ 2 \text{ where}$$

$$D = 0 \text{ if Ca and} = 1 \text{ if CVD}$$

then

$$HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_1]$$

T F 19. **Given the $\mathbf{LM_{alt}}$** model for two cause-specific events, Cancer and CVD:

$$h'_g(t, \mathbf{X}) = h'_{0g}(t) \times \exp[\delta'_{11} D_1 Rx + \delta'_{12} D_1 Age + \delta'_{21} D_2 Rx + \delta'_{22} D_2 Age],$$

$$g = 1, 2 \text{ where}$$

$D_1 = 1$ if Ca or 0 if CVD, and
$D_2 = 0$ if Ca or 1 if CVD,

then

$$HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\delta'_{21}]$$

T F 20. The $\mathbf{LM_U}$ model that would result if the **LM** model of Question 18 were changed to an unstratified Cox PH model can be written as follows.

$$h^\bullet(t, \mathbf{X}) = h^\bullet_0(t) \exp\left[\beta^\bullet_1 Rx + \beta^\bullet_2 Age + \delta^\bullet_{21}(D \times Rx) + \delta^\bullet_{22}(D \times Age)\right]$$

Consider a hypothetical study of the effect of a bone marrow transplant for leukemia on leukemia-free survival, where transplant failures can be of one of two types: relapse of leukemia and nonrelapse death (without prior relapse of leukemia). Suppose that in hospital A, 100 patients undergo such a transplant and that within the first 4 years post-transplant, 60 die without relapse by year 2 and 20 relapse during year 4. Suppose that in hospital B, 100 patients undergo such a transplant but post-transplant, there are 20 non-relapse deaths by year 1, 15 relapses during year 2, 40 non-relapse deaths between years 3 and 4, and 5 relapses during year 4.

21. What are the competing risks in this study?
22. What is the proportion of initial patients in hospitals A and B, respectively, that have leukemia relapse by 4 years?

The following tables provide the Kaplan–Meier curves for relapse of leukemia for each study.

Hospital A

t_j	n_j	m_j	q_j	$S(t_j)$
0	100	0	60	1
2	40	0	0	1
4	40	20	20	.5

Hospital B

t_j	n_j	m_j	q_j	$S(t_j)$
0	100	0	20	1
1	80	0	0	1
2	80	15	0	0.8125
3	65	0	40	0.8125
4	25	5	20	0.65

23. How have both tables treated the competing risk for nonrelapse death in the calculation of the KM probabilities?
24. Why are the KM probabilities different at 4 years for each hospital?

25. Compute the CIC curves for each hospital using the following tables.

Hospital A

t_f	n_f	m_f	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$CIC(t_f)$
0	100	0	0	–	–	–
2	40	0	0	1	0	0
4	40	20	–	–	–	–

Hospital B

t_f	n_f	m_f	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$CIC(t_f)$
0	100	0	0	–	–	–
1	80	0	0	1	0	0
2	80	15	–	–	–	–
3	65	0	–	–	–	–
4	25	5	–	–	–	–

26. Why are the CIC probabilities the same at 4 years?

Consider a hypothetical study to assess the effect of a new hospital infection control strategy for patients who undergo heart transplant surgery in a given hospital. The exposure variable of interest is a binary variable Group (**G**): $\mathbf{G} = 0$ for those patients receiving heart transplants from 1992 through 1995 when the previous hospital control strategy was used; $\mathbf{G} = 1$ for those patients receiving heart transplants from 1996 through 1999 when the new hospital infection control strategy was adopted. The primary event of interest is getting a hospital infection after surgery. A competing risk is death during recovery from surgery without getting a hospital infection. Control variables being considered are tissue mismatch score (TMS) at transplant and AGE at transplant. The outcome variable of interest is time (DAYS after surgery) until a patient developed a hospital infection.

27. State a cause-specific no-interaction Cox PH model for assessing the effect of group status (**G**) on time until a hospital infection event.
28. When fitting the model given in Question 27, which patients should be considered censored?
29. Describe or provide a table that would show how the data on the ith patient should be augmented for input into a Lunn–McNeil (LM) model for this analysis.

30. State a LM model that can be used with an augmented dataset that will provide identical results to those obtained from using the model of Question 27.
31. For the LM model of Question 30, what is the formula for the hazard ratio for the group effect **G**, controlling for **TMS** and **AGE**.
32. Describe how you would test whether a no-interaction SC LM model would be more appropriate than an interaction SC LM model.
33. State a LM_{alt} model that can be used with an augmented dataset that will provide identical results to those obtained from using the model of Question 27.
34. For the LM_{alt} model of Question 33, what is the formula for the hazard ratio for the group effect **G**, controlling for **TMS** and **AGE?**

Test

The dataset shown below describes a hypothetical study of recurrent bladder cancer. The entire dataset contained 53 patients, each with local bladder cancer tumors who are followed for up to 30 months after transurethral surgical excision. Three competing risks being considered are local recurrence of bladder cancer tumor (**event** = 1), bladder metastasis (**event** = 2), or other metastasis (**event** = 3). The variable **time** denotes survival time up to the occurrence of one of the three events or censorship from loss to follow-up, withdrawal, or end of study. The exposure variable of interest is drug treatment status (**tx**, 0 = placebo, 1 = treatment A), The covariates listed here are initial number of tumors (**num**) and initial size of tumors (**size**) in centimeters.

id	event	time	tx	num	size
1	1	8	1	1	1
2	0	1	0	1	3
3	0	4	1	2	1
4	0	7	0	1	1
5	0	10	1	5	1
6	2	6	0	4	1
7	0	10	1	4	1
8	0	14	0	1	1
9	0	18	1	1	1
10	3	5	0	1	3
11	0	18	1	1	3
12	1	12	0	1	1
13	2	16	1	1	1
14	0	18	0	1	1
15	0	23	1	3	3

(Continued on next page)

(*Continued*)

id	event	time	tx	num	size
16	3	10	0	1	3
17	1	15	1	1	3
18	0	23	0	1	3
19	2	3	1	1	1
20	3	16	0	1	1
21	1	23	1	1	1
22	1	3	0	3	1
23	2	9	1	3	1
24	2	21	0	3	1
25	0	23	1	3	1
26	3	7	0	2	3
27	3	10	1	2	3
28	1	16	0	2	3
29	1	24	1	2	3
30	1	3	0	1	1
31	2	15	1	1	1
32	2	25	0	1	1
33	0	26	1	1	2
34	1	1	0	8	1
35	0	26	1	8	1
36	1	2	0	1	4
37	1	26	1	1	4
38	1	25	0	1	2
39	0	28	1	1	2
40	0	29	0	1	4
41	0	29	1	1	2
42	0	29	0	4	1
43	3	28	1	1	6
44	1	30	0	1	6
45	2	2	1	1	5
46	1	17	0	1	5
47	1	22	1	1	5
48	0	30	0	1	5
49	3	3	1	2	1
50	2	6	0	2	1
51	3	8	1	2	1
52	3	12	0	2	1
53	0	30	1	2	1

1. Suppose you wish to use these data to determine the effect of **tx** on survival time for the cause-specific event of a local recurrence of bladder cancer. State a no-interaction Cox PH model for assessing this relationship that adjusts for the covariates **num** and **size**.

2. When fitting the model given in Question 1, which subjects are considered censored?
3. How would you modify your answers to Questions 1 and 2 if you were interested in the effect of **tx** on survival time for the cause-specific event of finding metastatic bladder cancer?
4. For the model considered in Question 1, briefly describe how to carry out a sensitivity analysis to determine how badly the results from fitting this model might be biased if the assumption of independent competing risks is violated.
5. The following two tables provide information necessary for calculating CIC curves for local recurrence of bladder cancer (**event** = 1) separately for each treatment group. The CIC formula used for both tables is given by the expression

$$\mathbf{CIC_1}(t_f) = \sum_{f'=1}^{f} \hat{I}_1(t'_f) = \sum_{f'=1}^{f} \hat{S}(t_{f'} - 1)\hat{h}_1(t_{f'})$$

where $\hat{h}_1(t_f) = m_{1f}/n_f$, m_{1f} is the number of local recurrent failures at time t_f and $S(t_{f-1})$ is the overall (event-free) survival probability for failure from either of the two competing risks at time t_{f-1}.

tx = 1 (Treatment A)

t_f	n_f	d_{1f}	$\hat{h}_1(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_1(t_f)$	$CIC_1(t_f)$
0	27	0	0	—	—	—
2	27	0	0	1	0	0
3	26	0	0	.9630	0	0
4	24	0	0	.8889	0	0
8	**23**	**1**	**.0435**	**.8889**	**.0387**	**.0387**
9	21	0	0	.8116	0	.0387
10	20	0	0	.7729	0	.0387
15	17	1	.0588	.7343	.0432	.0819
16	15	0	0	.6479	0	.0819
18	14	0	0	.6047	0	.0819
22	12	1	.0833	.6047	.0504	.1323
23	11	1	.0910	.5543	.0504	.1827
24	8	1	.1250	.5039	.0630	.2457
26	7	1	.1429	.4409	.0630	.3087
28	4	0	0	.3779	0	.3087
29	2	0	0	.2835	0	.3087
30	1	0	0	.2835	0	.3087

tx = 1 (Placebo)

t_f	n_f	d_{1f}	$\hat{h}_1(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_1(t_f)$	$CIC_1(t_f)$
0	26	0	0	—	—	—
1	26	1	.0400	1	.0400	.0400
2	24	1	.0417	.9615	.0400	.0800
3	23	2	.0870	.9215	.0801	.1601
5	21	0	0	.8413	0	.1601
6	20	0	0	.8013	0	.1601
7	18	0	0	.7212	0	.1601
10	16	0	0	.6811	0	.1601
12	15	1	.0667	.6385	.0426	.2027
14	13	0	0	.6835	0	.2027
16	12	1	.0833	.5534	.0461	.2488
17	10	1	.1000	.4612	.0461	.2949
18	9	0	0	.4150	0	.2949
21	8	0	0	.4150	0	.2949
23	7	0	0	.3632	0	.2949
25	**6**	**1**	**.1667**	**.3632**	**.0605**	**.3554**
29	4	0	0	.2421	0	.3554
30	2	1	0	.2421	0	.3554

a. Verify the CIC_1 calculation provided at failure time $t_f = 8$ for persons in the treatment group ($\mathbf{tx} = 1$); that is, use the original data to compute $\hat{\mathbf{h}}_1(\mathbf{t}_f)$, $\hat{S}(\mathbf{t}_{f-1})$, $\hat{\mathbf{I}}_1(\mathbf{t}_f)$, and $\mathbf{CIC}_1(\mathbf{t}_f)$, assuming that the calculations made up to this failure time are correct.

b. Verify the CIC_1 calculation provided at failure time $t_f = 25$ for persons in the placebo group ($\mathbf{tx} = 0$).

c. Interpret the CIC_1 values obtained for both the treatment and placebo groups at $t_f = 30$.

d. How can you calculate the CPC_1 values for both treatment and placebo groups at $t_f = 30$?

6. The following output was obtained using separate models for each of the 3 event-types.

Event = 1

Var	DF	Coef	Std.Err.	p > |z|	Haz.ratio
tx	1	−0.6258	0.5445	0.2504	0.535
num	1	0.0243	0.1900	0.8983	1.025
size	1	0.0184	0.1668	0.9120	1.125

Event = 2

Var	DF	Coef	Std.Err.	p > |z|	Haz.ratio
tx	1	−0.0127	0.6761	0.9851	0.987
num	1	−0.1095	0.2281	0.6312	0.896
size	1	−0.6475	0.3898	0.0966	0.523

Event = 3

Var	DF	Coef	Std.Err.	p > |z|	Haz.ratio
tx	1	−0.3796	0.6770	0.5750	0.684
num	1	−0.1052	0.3135	0.7372	0.900
size	1	−0.0238	0.2177	0.9128	0.976

a. What is the effect of treatment on survival from having a local recurrence of bladder cancer, and is it significant?

b. What is the effect of treatment on survival from developing metastatic bladder cancer, and is it significant?

c. What is the effect of treatment on survival from other metastatic cancer, and is it significant?

7. Below is the output from fitting a LM model to the bladder cancer data.

Var	DF	Coef	Std.Err.	p > |z|	Haz.ratio
txd2	1	0.6132	0.8681	0.4800	1.846
txd3	1	0.2463	0.8688	0.7768	1.279
numd2	1	−0.1337	0.2968	0.6523	0.875
numd2	1	−0.1295	0.3666	0.7240	0.879
sized2	1	−0.6660	0.4239	0.1162	0.514
sized3	1	−0.0423	0.2742	0.8775	0.959
tx	1	−0.6258	0.5445	0.2504	0.535
num	1	0.0243	0.1900	0.8983	1.025
size	1	0.0184	0.1668	0.9120	1.125

   a. State the hazard model formula for the LM model used for the above output.
   b. Determine the estimated hazard ratios for the effect of each of the 3 cause-specific events based on the above output.
   c. Verify that the estimated hazard ratios computed in Part b are identical to the hazard ratios computed in Question 6.

8. Below is the output from fitting a $\mathbf{LM_{alt}}$ model to the bladder cancer data.

Var	DF	Coef	Std.Err.	p > |z|	Haz.ratio
txd1	1	−0.6258	0.5445	0.2504	0.535
txd2	1	−0.0127	0.6761	0.9851	0.987
txd3	1	−0.3796	0.6770	0.5750	0.684
numd1	1	0.0243	0.1900	0.8983	1.025
numd2	1	−0.1095	0.2281	0.6312	0.896
numd3	1	−0.1052	0.3135	0.7372	0.900
sized1	1	0.0184	0.1668	0.9120	1.125
sized2	1	−0.6475	0.3898	0.0966	0.523
sized3	1	−0.0238	0.2177	0.9128	0.976

   a. State the hazard model formula for the $\mathbf{LM_{alt}}$ model used for the above output.
   b. Determine the estimated hazard ratios for the effect of each of the 3 cause-specific events based on the above output.
   c. Verify that the estimated hazard ratios computed in Part b are identical to the hazard ratios computed in Questions 6 and 7.

9. State the formula for a no-interaction SC LM model for these data.
10. Describe how you would test whether a no-interaction SC LM model would be more appropriate than an interaction SC LM model.

Answers to Practice Exercises

1. F: Only one competing risk event can occur at a given time.
2. T
3. F: You can die only once.
4. T
5. F: Competing risks must be treated as censored observations, rather than ignored.
6. F: A patient who develops a local recurrence will be treated as censored.
7. F: The statement would be true providing censoring is independent.
8. T
9. F: A sensitivity analysis can never provide explicit evidence about whether the independence assumption is satisfied; it can only suggest how biased the results might be if the assumption is not satisfied.
10. T
11. T
12. F: The formula is correct only if there is one risk. See Section V in the text for the general formula.
13. T
14. F: The correct statement should be: CPC gives the probability of experiencing an event c by time t, given that an individual has **not** experienced any of the other competing risks by time t.

15. F: the correct formula for CPC is:
$\mathbf{CPC}_c = CIC_{\mathbf{c}}/(1 - CIC_{\mathbf{c}'})$ where $CIC_{\mathbf{c}} = .4$ and $CIC_{\mathbf{c}'} =$ CIC from risks other than **c**, where the latter is not necessarily equal to .4.
16. T
17. T
18. T
19. T.
20. F: The correct $\mathbf{LM}_U$ model is

$$h^{\bullet}(t,\mathbf{X}) = h_0^{\bullet}(t)\exp\left[\gamma_1 D + \beta_1^{\bullet}Rx + \beta_2^{\bullet}Age + \delta_{21}^{\bullet}(D \times Rx) + \delta_{22}^{\bullet}(D \times Age)\right]$$

21. The competing risks are (1) relapse of leukemia and (2) nonrelapse death.
22. 20/100 = 0.2.
23. Both tables have treated the competing risks as if they were censored observations.
24. The KM probabilities are different for the two hospitals because the competing risks contribute a different pattern of censorship in the two hospitals.
25. The CIC curves for each hospital are calculated as follows.

Hospital A

t_f	n_f	m_f	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$CIC(t_f)$
0	100	0	0	—	—	—
2	40	0	0	1	0	0
4	40	20	0.5	0.4	0.20	0.20

Hospital B

t_f	n_f	m_f	$\hat{h}_{ca}(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_{ca}(t_f)$	$CIC(t_f)$
0	100	0	0	—	—	—
1	80	0	0	1	0	0
2	80	15	0.1875	0.8	0.15	0.15
3	65	0	0	0.65	0	0.15
4	25	5	0.20	0.25	0.05	0.20

26. The CIC probabilities are the same at 4 years because they give marginal probabilities that are not influenced by the pattern of censorship of the competing risks that are treated as censored. In hospital B, for example, the marginal probability of 0.15 at year 2 is given by the proportion of the initial risk set of 100 subjects that had a relapse of leukemia by year 2, regardless of the number of nonrelapse deaths prior to year 2. Similarly for hospital B, the marginal probability of .20 at year 4 adds to the marginal probability at year 2 the proportion of the initial risk set of 100 subjects that had a relapse between year 2 and year 4, regardless of the number of nonrelapse deaths that occurred between year 2 and 4.
27. $h_{HI}(t,\mathbf{X}) = h_0(t)\exp[\beta_{1HI}G + \beta_{2HI}TMS + \beta_{3HI}AGE]$ where HI denotes a hospital infection event
28. Patients who die after surgery without developing a hospital infection are censored. Also censored are any patients who are either lost to follow-up or withdraw from the study, although such patients are unlikely.
29. Augmented Data for LM Approach

Subj	Stime	Status	D_1	D_2	G	TMS	AGE
i	t_i	e_{1i}	1	0	G_i	TMS_i	AGE_i
i	t_i	e_{2i}	0	0	G_i	TMS_i	AGE_i

where e_{1i} = 1 if the ith subject develops a hospital infection, 0 otherwise
e_{2i} = 1 if ith subject dies after surgery, 0 otherwise
D_1 = indicator for hospital infection event
D_2 = indicator for death after surgery event

30. $$h^*_{\substack{g\\g=1,2}}(t,\mathbf{X}) = h^*_{0g}(t)\exp[\beta_1 G + \beta_2 TMS + \beta_3 AGE + \delta_{21}D_2G + \delta_{22}D_2TMS + \delta_{23}D''_2AGE]$$
31. $HR_{HI}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_1]$

32. Carry out a likelihood ratio test to compare the following two models. Full (SC Interaction LM) model:

$$\underset{g=1,\,2}{h^*_g}(t, \mathbf{X}) = h^*_{0g}(t) \exp[\beta_1 G + \beta_2 TMS + \beta_3 AGE + \delta_{21} D_2 G + \delta_{22} D_2 TMS + \delta_{23} D_2 AGE]$$

Reduced (no-interaction SC LM) model:

$$\underset{g=1,\,2}{h^*_g}(t, \mathbf{X}) = h^*_{0g}(t) \exp[\beta_1 G + \beta_2 TMS + \beta_3 AGE]$$

LR test statistic $= -2 \ln L_R - (-2 \ln L_F)$ is distributed χ^2_3 under H_0: no-interaction model

33. $$\underset{g=1,\,2}{h'_g}(t, \mathbf{X}) = h'_{0g}(t) \exp[\delta'_{11} D_1 G + \delta'_{12} D_1 TMS + \delta'_{13} D_1 AGE + \delta'_{21} D_2 G + \delta'_{22} D_2 TMS + \delta'_{23} D_2 AGE]$$

34. $HR_{HI}(RX = 1 \text{ vs. } Rx = 0) = \exp[\delta'_{11}]$

10 Design Issues for Randomized Trials

D.G. Kleinbaum and M. Klein, *Survival Analysis: A Self-Learning Text, Third Edition,*
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9_10,
© Springer Science+Business Media, LLC 2012

Introduction

In this chapter, we consider issues involved when designing studies with time-to-event outcomes. This includes determining the expected number of events to occur over the study period, the total number of subjects to be recruited into the study, the power of the study, the time period over which to accrue the study subjects, the time period over which enrolled subjects will be followed, and how to adjust sample size requirements to allow for subjects who might be lost to follow-up and/or who might switch therapies from the one they were originally allocated during the study period.

Our primary focus will consider prospective randomized trials of two groups (control and treatment) with equal allocation into each group. We will also consider how to modify sample size and/or power calculations when there is unequal allocation into the two groups.

Abbreviated Outline

The outline below gives the user a preview of the material to be covered by the presentation. A detailed outline for review purposes follows the presentation.

I. **Introduction: Background to Statistical Decisions** (pages 500–501)

II. **Design Considerations: Time-to-Event Outcomes** (pages 502–503)

III. **Determining the Required Number of Events (N_{EV})** (pages 504–505)

IV. **Determining the Required Total Number of Participants (N)** (pages 505–511)

V. **Other Design Issues** (pages 511–514)

VI. **Summary** (pages 514–517)

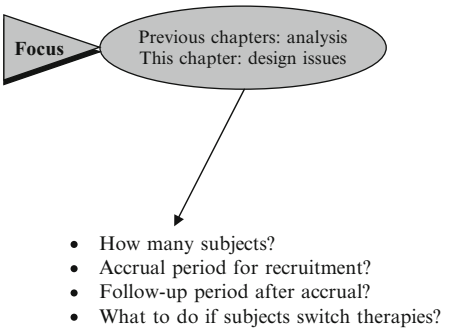
Objectives

Upon completing this chapter, the learner should be able to:

1. Define and illustrate the following terms: *significance level, type I error, type II error, power, effect size, accrual period, follow-up period, loss to follow-up, crossover subject, drop-in, drop-out*.
2. Given α, β, and Δ for a planned time-to-event study in a clinical to compare two treatments, determine the number of events required for the study.
3. Given α, β, Δ, and either the hazard rate, survival probability, or median survival time expected for the control group in a clinical trial to compare two treatments, determine the study size required for an allocation ratio of 1:1 between treatment and control groups.
4. For the same situation as in objective 3, determine the study size required for an allocation ratio of R:1 for a specified R, where $R = N_1/N_0$, N_i is the study size in the i-th group, and N is the total study size.
5. For the same situation as in objective 4, describe and/or illustrate how to adjust the study size to account for possible loss to follow-up.
6. For the same situation as in objective 4, describe and/or illustrate how to adjust the study size to account for subjects who crossover from treatment to control group or from control to treatment group.

Presentation

I. Introduction: Background to Statistical Decisions



The preceding chapters have been focused on developing a methodological and statistical framework for analysis of time to event or survival problems. In this chapter, we concentrate on the issues which are needed when designing studies with time to event outcomes.

These issues typically include questions such as: how many subjects should be recruited into the study; over what period should the study accrue the subjects; for what period after accrual should the subjects be followed-up for an event; what is the consequence of subjects switching therapies from the one they were originally allocated during the study period? Addressing these issues is key to designing a study which has sound scientific/statistical principles and adds to the credibility of the final results among the scientific community.

Review: "power" and "level of significance"

Before we develop these issues, we need to review the key statistical concepts of statistical power and level of significance.

Most survival studies:

- Compare two or more groups
- H_0: survival experience is the same in groups being compared
- H_A: survival experience is different in groups being compared

Most survival studies are designed to compare the time to event of two or more groups. We typically begin with the assumption that the survival experience is *the same* in these groups. This assumption is called the **null hypothesis (H_0)**.

We then seek (through the study) to determine whether there is sufficient evidence to demonstrate that the survival experience in the groups is different. This is termed the **alternative hypothesis (H_A)**.

P-value:

- Gives evidence against H_0
- P small $\Rightarrow$ statistical significance

The amount of evidence found against the null hypothesis is determined by the **P-value** (e.g. from a logrank or proportional hazards tests). If the P-value is small enough, we conclude that there is statistically sufficient evidence to conclude that the groups are different.

How small?
Traditionally:

$$P < 0.05 \Rightarrow \text{Evidence vs. } H_0 \quad (\text{i.e., Reject } H_0)$$

How small is small enough? Traditionally a P-value less than 0.05 is regarded as providing sufficient evidence against the null hypothesis of no difference between the two groups.

α = level of significance:

$$\Pr(\text{Reject } H_0 | H_0) = \Pr(\text{Type I Error}) = \alpha$$

This threshold value, referred to as the **level of significance**, is denoted by α. For a specified value of α, we are saying that there is a probability of α that we may decide to reject the null hypothesis when in fact the null hypothesis is correct. In statistical terms, this type of error is referred to as a **Type I error**.

$$\Pr(\text{Do not reject } H_0 | H_A) = \Pr(\text{Type II Error}) = \beta$$

The other error in statistical decision making is to declare that there is insufficient evidence to say the groups are different (associated with P-values > 0.05) when in fact they are actually different, This is referred to statistically as a **Type II error**.

$$\text{Power} = \Pr(\text{Reject } H_0 | H_A) = 1 - \Pr(\text{Type I error}) = 1 - \beta$$

The **power** of the study is 1- probability of a Type II error, also denoted 1-β, where β is the probability of a Type II error.

$$\text{Sample size} \nearrow \text{ (information)} \Rightarrow \text{Power} \nearrow \text{ (declare } H_A\text{)}$$

The power will typically increase as the sample size increases, i.e., the more information we gather, the more likely we will declare the two groups being compared as different when they truly are different.

Strength of evidence, i.e., **effect size** (Δ):

- Different ways to quantify Δ:
  $\Delta = \theta_1 - \theta_0$ or θ_1/θ_0 where
1. θ_i are hazard rates
2. θ_i are survival probabilities
3. θ_i are mean survival times
- Δ large $\Rightarrow$ smaller sample size required

The final quantity to consider is the strength of evidence of interest, typically referred to as the **effect size**, and denoted as Δ. This measure (Δ) can be either in the form of a difference or ratio of hazard rates, survival probabilities, or median survival times. If the effect size is large (e.g. a 60% increase in disease free survival at 5 years), then fewer subjects would be required to make a decision that the groups are different.

EXAMPLES: RATIO MEASURES OF EFFECT SIZE

a) Reduction in annual event rate from 10% to 5%: $\Delta = \lambda_0/\lambda_1 = 10/5 = 2$
b) Increase in 3-year survival from 74% to 86%: $\Delta = (S_1/S_0) = 0.86/0.74 = 1.16$ or $\Delta = (\ln S_0/\ln S_1) = \ln(0.74)/\ln(0.86) = 2$
c) Increase in median survival from 7 to 14 months: $\Delta = (m_1/m_0) = 14/7 = 2$

Examples of three types of effect sizes that are all ratio measures are shown at the left. We typically define Δ to be greater than 1 in order to reflect a "positive" effect. For example, in a) at the left, a reduction in event rates indicates a positive effect, so the numerator in the ratio formula for Δ contains λ_0 rather than λ_1.

Sample size (N) related to α, $1 - \beta$, and Δ

At this point, we have identified three crucial inputs required for sample size calculation: the level of significance (α), the power (1-β) and the effect size (Δ).

II. Design Considerations: Time-to-Event Outcomes

Our discussion of design and sample size considerations in survival analysis studies will primarily focus on prospective randomized trials of two groups (control and intervention) with equal allocation into each group. We will also consider how to modify sample size and/or power calculations when there is unequal allocation into the two groups.

Primary focus:

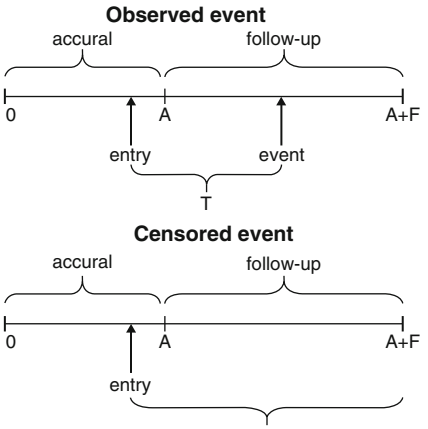
- Prospective randomized trials
- Comparing two groups
- Equal allocation
- Unequal allocation

Assume:

A = accrual period
F = follow-up period

A + F = total study period

Consider a study where we plan to accrue subjects for a period of A (days, months, years) time units, and then observe or follow them up for a further F time units. This defines the total study period to have A + F time units.



The diagrams on the left show how subjects who experience an event or are censored are represented in the total study time A + F. The top diagram illustrates a subject experiencing the event T units after recruitment into the study.

The bottom diagram illustrates a subject not experiencing the event (i.e., censored) after recruitment into the study.

Information on each subject:

- Occurrence of event (yes or no)?
- Time to event or censoring
- Treatment group allocation

For each subject, we require information on whether or not the event of interest had occurred, the time from recruitment into the study to either the time that the event occurs or the time of censoring, and the treatment group to which the subject was allocated.

Design issues:

- Effect size of interest (plausible and clinically relevant)
- Sufficient funds
- Adequate resources
- Sufficient # of patients available to participate

A key design issue for clinical trials already mentioned is the **effect size** of interest for comparing the two treatments. This difference should be both plausible and clinically worthwhile. Other issues include the need for **sufficient funds** to support the study, **adequate resources** to carry out the study (number of institutions participating, etc.), and **sufficient number of patients** with the disease who would be prepared to participate in the proposed study.

Study size (N) versus # **of Events** (N_{EV})

When planning a clinical trial involving time-to-event outcomes, it is important to distinguish the **study size** required from the **number of events** required.

N = total # of study participants
- includes those who get event and those who do not get the event

The **study size** refers to the total number of subjects chosen to participate in the study, including those who get the event and those who do not get the event.

N_{EV} = # of participants who get event during study period

The **number of events** refers to those study participants who actually get the event during the study period.

Process for determining N in time-to-event studies:
Step 1: Determine N_{EV} using α, $1-\beta$, Δ
Step 2: Determine N from N_{EV} by extrapolation using $N = N_{EV}/p_{EV}$
where
p_{EV} = Pr(study participant gets event)

When determining sample size and/or power in such studies, we typically first determine the **expected (or required) number of events** (N_{EV}), after which we then determine the **study size required** (N). This means that α, β and Δ are needed to determine N_{EV}, which then gets "extrapolated" to N by dividing by (a guestimate of) the probability that a study participant gets an event. (Details on "extrapolation" in Section IV.)

In practice:
recruit study participants (N)
rather than
events (N_{EV})- occur after study begins

Note that the recruitment of study participants (rather than only those participants who become cases) is what actually happens in practice, since events do not start occurring until after the study begins.

Section IV (page):
determining N from N_{EV} using assumptions on
study duration
and **distribution of event times**

Nevertheless, the study size required can be determined without waiting for events to start occurring, as we will show in Section IV, by making certain assumptions about study duration (i.e., accrual and follow-up periods) and the (hazard or survival) distribution of event times.

Study duration

- long enough to observe N_{EV}
- specify in advance: accrual and follow-up periods, and N_{EV}

The study **duration** must be determined to be sufficiently long to observe the expected number of events. If the accrual and follow-up periods as well as expected number of events are determined in advance, it is possible to project a sufficiently long study duration.

III. Determining The Required Number of Events (N_{EV})

Assumptions:
1. $N_1 = N_0$ i.e., equal study sizes in each group
2. H_A is two tailed
3. N_1 and N_0 large enough to use normal approximation, i.e., $Z \sim N(0,1)$
4. Proportional hazards satisfied, i.e., λ_1/λ_0 constant over time

For convenience, we describe how to determine the required number of events by making the assumptions shown at the left.

Assumption 4 (PH satisfied) allows ratio measures of effect size measures of any type:

$\Delta = \lambda_1/\lambda_0,\ \Delta = S_1/S_0, \text{or } \Delta = m_1/m_0$

If we assume proportional hazards, then the effect size can be calculated based on any of the three survival attributes: event rates (λ_i), survival probabilities (S_i) or median survival times (m_i).

$$N_{EV} = \left(\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)(\Delta+1)}{(\Delta-1)} \right)^2 \quad \cdots\cdots(1)$$

For a given effect size (Δ), level of significance α, and power ($1-\beta$), the number of events required **for both groups combined** is given (Freedman 1982) by equation (1) on the left.

$$\text{Power} = \Pr(Z < z_{EV}) \cdots\cdots\cdots\cdots\cdots (2)$$

$$\text{where } z_{EV} = \left[\sqrt{N_{EV}} \left(\frac{\Delta - 1}{\Delta + 1} \right) - z_{1-\alpha/2} \right]$$

The power corresponding to a specified number of events can be obtained using equation (2). Note that $\Pr(Z < z_{EV})$ is the cumulative probability below the value z_{EV} of the standard normal distribution.

EXAMPLE

$\Delta\ (= m_1/m_0) = 14/7 = 2$
$\alpha = 0.05,\ z_{1-\alpha/2} = 1.96,\ \beta = 0.20,$
$z_{1-\beta} = 0.84$

$$N_{EV} = \left(\frac{(1.96+0.84)(2+1)}{(2-1)} \right)^2 = 70.56 \approx \boxed{71}$$

Power: $\Pr(Z < z_{EV})$ where

$$z_{EV} = \left[\sqrt{100} \left(\frac{2-1}{2+1} \right) - 1.96 \right] = 1.373$$

so that

$\Pr(Z < 1.373) = \mathbf{0.92}$

As an example, suppose we are interested in the increase in median survival time defined by the effect size $\Delta\ (= m_1/m_0) = 14/7 = 2$. Suppose further we set the level of significance at 5%, so $z_{1-\alpha/2} = 1.96$, and we desire 80% power, so $z_{1-\beta} = 0.84$, where the z values are obtained from the standard normal distribution.

The number of events required using equation (1) is given by the calculation shown at the left; thus, 71 events are required.

Suppose we commence the study and find that the event rate is occurring much faster than anticipated, and we have observed 100 events. Then the power of our study using equation (2) is given by the calculation shown at the left. Thus with 100 events, our study has **92% power** to detect a increase in median survival time from 7 months to 14 months.

Unequal group sizes:

- Requires modifying formula (1)
- One treatment is more expensive than the other
- Gain more experience with a new intervention

Formula (1) can be modified to allow for unequal group sizes.

Unequal group sizes may be considered when one treatment is more expensive than the other or to gain more experience with a new intervention.

$R = N_1/N_0$ where
$N = N_1 + N_0$, N_i=# of events in group i, i=0, 1

$$N_{EV} = \left(\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)(R\Delta+1)}{\sqrt{R}(\Delta - 1)} \right)^2 \quad (3)$$

We let R denote the ratio N_1/N_0, where N_i denote the number of subjects (with and without events) in the i-th treatment group, i=0, 1, so that $N= N_1+N_0$ gives the total sample size for the study.

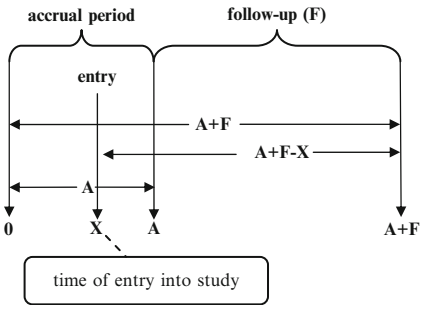
The modified formula for the required number of events (N_{EV}) is modified to equation (3) on the left.

EXAMPLE: R=2

$$N_{EV} = \left(\frac{(1.96 + 0.84)(2(2) + 1)}{\sqrt{2}(2 - 1)} \right)^2 = \mathbf{98}$$

Here, we illustrate calculations using formula (3) for R=2. Note that the value obtained for N_{EV} (i.e., 98) when R = 2 is somewhat higher than the value previously obtained (i.e., 71) when we used equal sample sizes (i.e., R=1).

IV. Determining the Required Total Number of Participants (N)



To calculate the number of participants required in designing a clinical study, some additional pieces of information are required (see left diagram);

1. the **accrual period**, A;
2. the **follow-up period**, F, after the last subject has been entered onto the study;
3. the **median follow-up time** $M_F = A/2 + F$
4. X denotes the **time point at study entry** for any subject entering between time 0 and time A
5. $A + F - X$ denotes the **maximum follow-up since study entry** for a subject entering at time X
6. the **distribution of the time(s) to event *T***:
   specify either
   the survivor functions $S_i(t)$, i=1,2, for each group
   or assume
   constant event rates λ_i, so that $S_i(t) = e^{-\lambda t}$ (exponential distribution).

Determining N (cont'd)

Relating number of events to total sample size:

$$N_{EV} = N \times p_{EV}$$

where

N= total sample size

p_{EV} = Pr(event occurs since study entry)

$$p_{EV} = \Pr(EV|i = 1)\Pr(i = 1) + \Pr(EV|i = 0)\Pr(i = 0) = w_1 p_{EV1} + w_0 p_{EV0}$$

where

$$p_{EVi} = \Pr(EV|i) \text{ and } w_i = \Pr(i),\ i = 0,1$$

If $R = N_1/N_0$ and $N_1+N_0=N$ where N_i is the sample size in group, i=0,1 then

$$w_i = N_i/N,\ i = 0,\ 1$$

Algebra:

$$w_1 = N_1/(N_1 + N_0) = R/(R + 1)$$
$$w_0 = N_0/(N_1 + N_0) = 1/(R + 1)$$

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

Special Case: $R = 1$

$$N = \frac{N_{EV}}{(p_{EV1} + p_{EV0})/2}$$

Alternative formula:
Express (p_{EVI}) in terms of parameters of $S_i(t)$

$$p_{EVi} = 1 - \Pr(\text{no event since study entry}|\ i) = 1 - \Pr(\text{surviving since study entry}|\ i) \approx 1 - S_i(M_F)$$

where

$S_i(t)$ = survival function at time t for group i

$M_F = A/2 + F$ = median follow-up time

A general formula for the total sample size (subjects with and without events) is derived by assuming that the number of events needed (N_{EV}) is the product of the total sample size (N) times the probability (p_{EV}) that a subject will get an event since study entry.

Since we are assuming two treatment groups, we can use basic conditional probability rules to express p_{EV} as the sum of the two components shown at the left.

The term p_{EVi} denotes the probability of an event occurring given a subject is in treatment group i, i=0,1. The term w_i denotes the probability of being in treatment group i.

If the allocation ratio for subjects into the two treatment groups is determined in advance to be $R = N_1/N_0$, where N_i is the total sample size in group i, then w_i, which is effectively a weight, denotes the proportion of the total sample size in group i (i.e., N_i/N).

With algebra, we can divide both the numerator and denominator of w_i by N_0 to express the weights (w_i) in terms of R, as shown on the left.

We can thus write the formula for total sample size (N) in terms of the number of events (N_{EV}) and the allocation ratio R.

As a special case, if the allocation ratio is R=1, i.e., $N_1 = N_0$, then the denominator in the formula for N simplifies to the arithmetic average of the p_{EVi}.

An alternative formula for the total sample size can be obtained if we can express the event probabilities (p_{EVi}) in terms of survival curve parameters.

We first write p_{EVi} as 1 minus the probability of no event occurring since study entry, that is, one minus the probability of surviving since study entry.

The latter survival probability can be approximated as the probability that a subject survives past the median follow-up time (i.e., $M_F = A/2 + F$) for all subjects.

Total Sample Size: Formula 1

$$N = \frac{N_{EV}}{\frac{R}{R+1}\{1 - S_1(M_F)\} + \frac{1}{R+1}\{1 - S_0(M_F)\}}$$

Special case: $S_i(t) = \exp(-\lambda_i t)$, i=0,1

$$N = \frac{N_{EV}}{\frac{R}{R+1}\{1 - e^{-\lambda_1 M_F}\} + \frac{1}{R+1}\{1 - e^{-\lambda_0 M_F}\}}$$

Thus, a general formula for the total sample size can be given in terms of the survival model assumed by the investigators, as shown on the left.

In particular, if the survival function is exponential, the formula can be further specified as shown here.

EXAMPLE

A=2, F=4, so M_F=A/2 + F = 5
R=1
α=0.05, β=0.10
$\lambda_0 = 0.10$, $\Delta = \lambda_0/\lambda_1 = 2$

$$N = \frac{[(1.96 + 1.282)(2+1)/(2-1)]^2}{\frac{1}{1+1}\left\{1 - e^{-2(0.05)(5)}\right\} + \frac{1}{1+1}\left\{1 - e^{-(0.05)(5)}\right\}}$$

$$= \frac{94.595}{0.307} = \mathbf{307.8}$$

(N_{EV} in the numerator; P_{EV} in the denominator)

As an example, suppose the accrual period is 2 years, and the follow-up period is 4 years, so that the median follow-up time (M_F) is 5. Suppose further that R=1, α=.05, β=.10, the control group hazard rate (λ_0) is 0.10 and the effect size ($\Delta = \lambda_0/\lambda_1$) is 2, i.e., we consider a reduction in hazard rates from 0.10/year to 0.05/year.

Substituting this information into the sample size formula, we obtain a total required sample size of 308 subjects, which will yield approximately 95 events, with an approximate overall event probability of 0.31. Note that the latter probability estimate is calculated using the denominator in the sample size formula.

Alternative approach for computing N:

1. Assumes exponential survival
   $S_i(t) = \exp(-\lambda_i t),\ i = 0, 1$
2. Does not use $S_i(M_F)$

We now describe an alternative approach for deriving a total sample size formula that assumes an exponential survival curve, but doesn't use the median follow-up time (M_F) to approximate the calculation of survival probabilities.

Exponential survival function: typically yields reasonable sample size estimates

(note: Cox model has exponential component)

Parametric survival distributions other than exponential are always possible for survival data. Nevertheless, by assuming an exponential function, we typically obtain reasonable sample size estimates. This makes sense because of the wide applicability of the Cox model, whose parametric component has exponential form.

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

Recall that we previously provided a formula for the total sample size (shown again here) that involved the number of events required (N_{EV}), the allocation ratio (R), and the probabilities (p_{EVi}) of an event occurring in each treatment group.

Next: calculating p_{EVi}, i=0,1.

We now describe how to calculate event probabilities (p_{EVi}) in terms of exponential survival parameters.

$$\Pr_i(\text{EV and entry at X}) = \Pr_i(\text{EV} \mid \text{entry at X})\Pr(\text{entry at X})$$

$$\Pr(A \text{ and } B) = \Pr(A|B)\Pr(B)$$

We first write the probability that a subject entering treatment group i at time X gets the event as the product of probabilities. This follows from the basic probability rule that the joint probability of two events A and B equals the conditional probability of A given B times the probability of B.

Assume uniform distribution for each i:

$$X \sim U[0, A] \Rightarrow \Pr(\text{entry at X}) = 1/A$$

so that

$$\Pr_i(\text{EV and entry at X}) = \frac{1}{A}\Pr_i(\text{EV}|\text{entry at X})$$

We can simplify the above formula if we assume that the time X at which any subject enters the study has the uniform distribution over the accrual period.

$$\begin{aligned}&\frac{1}{A}\Pr_i(\text{EV}|\text{entry at X})\\ &\quad = \frac{1}{A}[1 - \Pr_i(\text{No EV}|\text{entry at X})]\\ &\quad = \frac{1}{A}[1 - \Pr_i(\text{Survive past A} + \text{F} - \text{X}|\text{entry at X})]\end{aligned}$$

The probability of an event is equal to 1 minus the probability of no event, which is 1 minus the probability of surviving since entry. Thus, we can rewrite the right hand side of above formula as shown here. Note that A+F−X gives the survival time for a subject entering at time X and surviving to the end of the study (at time A+F).

Integration formula for pev_i:

$$p_{EVi} = \frac{1}{A}\left[\int_0^A [1 - S(A + F - X)]\, dx.\right.$$

Recognizing that a subject's entry time X ranges between 0 and A, we can now integrate over this range to obtain p_{EVi}, the probability that any subject entering treatment group i at time X gets the event since study entry. This integral formula is shown at the left.

Let u = A+F−X, so

$X = A \Rightarrow u = F$ and $X = 0 \Rightarrow u = A + F$

Then,

$$p_{EVi} = 1 - \frac{1}{A}\int_F^{A+F} S(u)du$$

Letting u denote A+F−X, and noting that u will range from F to A+F as X ranges between A and 0, we can further rewrite this integral as shown here.

Further simplification:

Assume $S(t) = \exp(-\lambda t)$

$$\begin{aligned}p_{EVi} &= 1 - \frac{1}{A}\int_F^{A+F} \exp(-\lambda_i u)du\\ &= 1 - \frac{1}{\lambda_i A}[-\exp(-\lambda_i u)]_F^{A+F}\\ &= 1 - \frac{1}{\lambda_i A}[\exp(-\lambda_i F) - \exp(-\lambda_i(A + F))]\end{aligned}$$

We can further simplify this formula if we are willing to assume that the survival curve follows an exponential distribution.

We can then carry out the integration and corresponding algebra to obtain an expression for p_{EVi} that involves the constant hazard rate λ_i for treatment group i, the accrual period A, and the follow-up period F as shown on the left.

Total Sample Size: Formula 2

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

where

$$p_{EVi} = 1 - \frac{1}{\lambda_i A}\left[e^{-\lambda_i F} - e^{-\lambda_i (A+F)}\right]$$

λ_i = constant hazard for group i
R = allocation ratio,
A = accrual period,
F = follow-up period

We then obtain the general formula shown at the left for the required number of total participants (N) in terms of the total expected events (N_{EV}), exponential hazard rates (λ_i) in each group, effect size ($\Delta = (\lambda_1/\lambda_0)$, allocation ratio R, accrual period (A), and follow-up period (F).

EXAMPLE

A=2, F=4
R=1
α=0.05, β=0.10
$\lambda_0 = 0.10$, $\Delta = \lambda_0/\lambda_1 = 2$
$N_{EV} = 94.595$

$$p_{EV1} = 1 - \frac{1}{(0.05)(2)}\left[e^{-(0.05)(4)} - e^{-(0.05)(2+4)}\right]$$
$$= 1 - 0.7791 = \mathbf{0.2207}$$

$$p_{EV0} = 1 - \frac{1}{(0.10)(2)}\left[e^{-(0.10)(4)} - e^{-(0.10)(2+4)}\right]$$
$$= 1 - 0.6075 = \mathbf{0.3925}$$

$$N = \frac{94.595}{\frac{1}{2}(0.2207) + \frac{1}{2}(0.3925)}$$
$$= \frac{94.595}{0.3066} = 308.52 \approx \mathbf{309}$$

We again consider the same example we previously used with Formula 1, but now applied to Formula 2. Recall that here we consider a reduction in hazard rates from 0.10/year to 0.05/year, so that the effect size (Δ) is 2. The allocation ratio is assumed to be 1:1, i.e., R = 1.

We previously found for this example that the number of events required (N_{EV}) was 94.595.

Substituting the above information into formula for p_{EVi}, find that $p_{EV1} = 0.2207$ and $p_{EV0} = 0.3925$.

We then substitute these values for p_{EVi} into Formula 2, as shown at the left to obtain a total sample size (N) of 309, which is essentially the same as the 308 obtained from Formula 1.

SECOND EXAMPLE

Patients with metastatic colorectal cancer
Intervention group: molecular targeted therapy
vs.
Control group standard therapy

As a second example, suppose we wish to determine the sample size for a clinical trial looking at a molecular targeted therapy (intervention group) for patients with metastatic colorectal cancer.

SECOND EXAMPLE: (continued)

Study aim: median survival time (m_i) improves from $m_0 = 6$ to $m_1 = 10$

A=15 mos., F=12 mos.
α=0.05, β=0.10
R=2

We are given $m_0 = 6$ and $m_1 = 10$ but we need corresponding λ_0 and λ_1

$$S_i(m_i) = \exp(-\lambda_i m_i) = 0.5,\ i = 0, 1$$
$$\Downarrow$$
$$-\lambda_i m_i = \ln 0.5 = -\ln 2$$
$$\Downarrow$$
$$\lambda_i = \ln 2/m_i$$

$\lambda_0 = \ln 2/6 = \mathbf{0.1155}$

$\lambda_1 = \ln 2/10 = \mathbf{0.0693}$

$\Delta = \lambda_0/\lambda_1 = 0.1155/0.0693 = 1.667.$

$$N_{EV} = \left(\frac{(z_{1-\alpha/2} + z_{1-\beta})(R\Delta+1)}{\sqrt{R}(\Delta - 1)}\right)^2$$
$$= \left(\frac{(1.96 + 1.28)(2(1.667) + 1)}{\sqrt{2}(1.667 - 1)}\right)^2$$
$$= 221.6, \text{ where rounds to}$$
222 required events

The study aim is to increase the median time to disease progression survival from 6 months in the control group to 10 months in the intervention group. Assume a 15-month accrual period and a 12-month follow-up period, 90% power and a 5% significance level. Also subjects are to be allocated in a 1:2 ratio (2 subjects in the intervention for each control), i.e., R=2.

Note that in this example, we are given the desired increase in median survival times (m_i) from control to intervention group, but not the corresponding set of hazard rates (λ_i) for these two groups, which we need to use in our sample size formula.

Nevertheless, we can determine the values for λ_i from their corresponding median survival times m_i by setting the exponential survival probability $S_i(m_i) = \exp(-\lambda_i m_i)$ equal to 0.5 for group i, and then solving for λ_i, as shown at the left. Note that the value of any survival function at its median survival time is, by definition, always equal to 0.5.

Thus, substituting the values of m_0=6 and m_1=10 in the above formula for λ_i, we obtain the values $\lambda_0 = 0.116$ and $\lambda_1 = 0.069$, as shown at the left.

Consequently, the effect size ($\Delta = \lambda_0/\lambda_1$) needed for our sample size formula is 0.1155/0.0693, which equals 1.667.

To determine the sample size required for this study, we first compute the required number of events (N_{EV}), as shown on the left.

Thus, we require 222 total events.

Next: Computing N_1 and N_0 using Formula 2

We now use our Formula 2 (shown again here) to compute the sample sizes required in each treatment group.

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

where

$$p_{EVi} = 1 - \frac{1}{\lambda_i A}\left[e^{-\lambda_i F} - e^{-\lambda_i (A+F)}\right]$$

SECOND EXAMPLE: (continued)

p_{EV1}

$$= 1 - \frac{1}{(0.0693)(15)}\left[e^{-(0.0693)(12)} - e^{-(0.0693)(15+12)}\right]$$

$$= 1 - 0.2707 = \boxed{0.7293}$$

p_{EV0}

$$= 1 - \frac{1}{(0.1155)(15)}\left[e^{-(0.1155)(12)} - e^{-(0.1155)(15+12)}\right]$$

$$= 1 - 0.1188 = 0.8812$$

$$N = \frac{221.6}{\frac{2}{2+1}(0.7293) + \frac{1}{2+1}(0.8812)}$$

$$= 284.12 \approx \boxed{285}$$

$N_1 = [R/(R+1)]N$ and $N_0 = N_1/2$

R=2 and $\mathbf{N = 285 \Rightarrow N_1 = 190}$ and $\mathbf{N_0 = 95}$

The computation of p_{EVi}, i=0,1 is shown on the left.

We obtain $p_{EV1} = 0.7293$

and

$p_{EV0} = 0.8812$.

We now substitute N_{EV}=221.6, $p_{EV1} = 0.7293$, $p_{EV0} = 0.8812$, and R=2 into the formula for N as shown at the left.

We obtain a total required sample size (N) of 285.

Now, given N and R, we can solve for N_1 and N_0 using the formula shown at the left.

Since R =2, so that N_1=2N_0, it follows that N_1=190 and N_0= 95 subjects are required in groups 1 and 0 respectively.

V. Other Design Issues

- Choosing accrual and follow-up times;
- Adjusting for loss to follow-up;
- Adjusting for cross-over (drop-in/drop-out)

Other issues to consider in the design of time-to-event studies are shown at the left.

Accrual and follow-up times:
A = accrual period
F = follow-up period

- Need balance between the choices of accrual and follow-up periods
- Cost of Accrual >> Cost of follow-up

We have seen the accrual period (A) and length of follow-up period (F) contribute in the expressions for the probability of an event as well as to total sample size. In any design, there needs to be a balance between the choices of accrual and follow-up periods. The cost of accruing subjects into a study is, in general, far greater than the cost of follow-up.

Impact of accrual and follow-up times on the study size

Colorectal example: $N_{EV} = 161$,
$\alpha = 0.05$,
$1 - \beta = 0.90$

Study time		N
Accrual	follow-up	
12	12	**206**
12	15	194
12	18	**186**
15	12	200
15	18	184
18	12	**196**
18	15	188
18	18	182

The table on the left illustrates, using the colorectal cancer example, how the study size required (N) varies depending on the duration periods of accrual and follow-up. **This table indicates that a longer follow-up time has a larger impact on sample size than a longer accrual time.**

For example, for an accrual period of 12 months, the sample size needed is reduced from 206 subjects to 186 subjects as the follow-up period is increased from 12 to 18 months. However, for a follow-up period of 12 months, the sample size needed is reduced from 206 subjects to only 196 subjects as the accrual period is increased from 12 to 18 months.

Loss to follow-up

Subjects leave study before experiencing event
Reasons:

- personal circumstances
- experiencing unacceptable toxicity
- loss of interest
- receiving newer therapy

Follow-up losses ↗ $\Rightarrow$ N_{EV} ↘
$\Rightarrow$ Power ↘

In large scale clinical trials, it is usually difficult to retain complete follow-up on all subjects. Inevitably, a proportion of subjects will leave the study before experiencing an event. Some reasons include personal circumstances (moving to another district/city/state/country), experiencing unacceptable toxicity from the therapy, loss of interest, or receiving a newer therapy.

As more subjects leave the study, the number of events will be reduced and the power of the study will be compromised. This can be partially adjusted for by increasing the sample size in an attempt to maintain the number of events.

Adjusting for loss to follow-up: Need
p_{lof} = proportion of subjects expected to be lost to follow-up
N = study size with full follow-up

We can adjust for loss to follow-up in sample size calculations if we have an estimate from previous studies of the proportion of subjects expected to be lost to follow-up ($\mathbf{p_{lof}}$). We also need the study size required with full follow-up ($\mathbf{N}$).

Then

$$\boxed{N_{LOFadj} = N/(1 - p_{lof})}$$

Then the adjusted sample size (N_{LOFadj}) that accounts for the loss to follow-up is obtained by dividing N by $(1 - p_{lof})$.

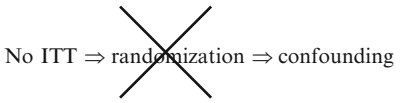
EXAMPLE

$N = 270$, $p_{lof} = 0.25$
$\Downarrow$
$N_{LOFadj} = 270/(1 - 0.25) = \mathbf{360}$

For example, in a study of 270 subjects, it is observed during the study that 25% of subjects are lost to follow-up. To ensure that the study power is maintained, the sample size would need to be increased to 360 subjects.

Cross-over (drop-in/drop-out) adjustments

One approach: **Intention-to-treat (ITT) principle-**
subjects analyzed according to the originally assigned treatment group



Another analysis approach: **Adjust for crossovers**

Crossover: a subject who changes from T to C or from C to T during the trial
where
C = control group,
T = treatment group
T to C ⇒ **drop-out**
C to T ⇒ **drop-in**
N = original total sample size (w/o adjusting for cross-overs)
d_c = proportion of (drop-outs) from T to C
d_t = proportion of (drop-ins) from C to T

Sample-size formula adjusted for ITT:

$$N_{ITTadj} = \frac{N}{(1 - d_c - d_t)^2}$$

Sample Size Inflation factor

	Drop-in rate					
Drop-out rate	**0%**	**1%**	**5%**	**10%**	**15%**	**20%**
0%	1	1.02	1.11	1.23	1.38	1.56
1%	1.02	1.04	1.13	1.26	1.42	1.60
5%	1.11	1.13	1.23	1.38	1.56	1.78
10%	1.23	1.26	1.38	1.56	1.78	2.04
15%	1.38	1.42	1.56	1.78	2.04	2.37
20%	1.56	1.60	1.78	2.04	2.37	2.78

In randomized controlled trials, the analysis is typically based on the **Intention-to-Treat (ITT) principle**. This principle states that subjects will be analyzed to the treatment group to which they were originally assigned at the time of randomization.

Not adhering to the ITT principle would break the randomization to some extent, which could potentially lead to a confounded adjusted estimate.

Another approach to the analysis is to adjust for **cross-over** from one treatment group to another.

If a subject changes groups during the trial, from either treatment (T) to control (C) or from control to treatment, we say that this subject is a "cross-over". A change from T to C is called a "drop-out" and a change from C to T is a "drop-in."

Let **N** denote the originally calculated sample size prior to considering the possibility of cross-overs. Also, let $\mathbf{d_c}$ and $\mathbf{d_t}$ be the potential proportion of subjects who drop into the control and intervention groups; these are guestimates.

Then it can be shown that for an intention-to-treat analysis, to maintain the study power, the total study size should be adjusted to give a **sample size inflation factor**. This adjustment formula (N_{ITTadj}) is shown at the left.

The table at the left gives the sample size inflation required for different levels of *drop-ins/drop-outs* originally assigned at the time of randomization. Thus, subjects which receive the *opposite* treatment to which they were randomized will (in general) *dilute* the effect seen by the intervention due to this cross-over.

EXAMPLE

Exercise regimen → ? → Disease progression in colorectal cancer patients

Patients who stop exercise: **drop-outs**
Patients who start exercise: **drop-ins**

$N = 600, d_c = 0.05, d_t = 0.15$

$\Downarrow$

$N_{ITTadj} = 600/(1 - 0.05 - 0.15)^2$
$= 600 \times 1.56 = \mathbf{936}$

As an example, we consider a study investigating the impact of exercise on time to disease progression in colorectal cancer patients, randomized to an exercise regimen.

Those patients who stop exercising are **dropping out of** the exercise arm (into the control arm). Those who commence an exercise program are **dropping into** the exercise arm (out of the control arm).

Suppose the study size originally required 600 subjects, with 5% expected to drop *out of* the intervention arm, and 15% expected to drop *into* the intervention arm. Then, using ITT, the sample size (see the table on the left) needs to be increased to 936 subjects.

VI. Summary

A summary of this presentation is now provided. We have addressed several issues, which are listed on the left.

Issues:

- How many events to expect?
- How many subjects?
- Impact of length of accrual and follow-up periods?
- Adjusting for loss of follow-up?
- What to do if subjects switch therapies?

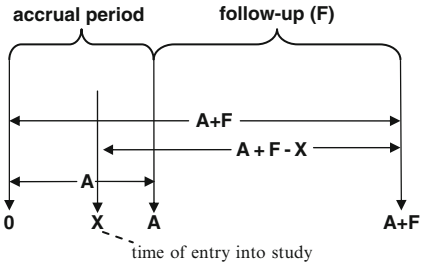
Key statistical concepts:

H_0, H_A, P-value,
Type I error (α), Type II error (β),
Power = $1-\beta$, effect size (Δ).

We began by reviewing key statistical concepts, including describing the terms null hypothesis (H_0), alternative hypothesis (H_A), P-value, Type I error (α), Type II error (β), Power = $1-\beta$, and effect size (Δ).

Δ can be $\theta_1 - \theta_0$ or θ_1/θ_0 where θ_i are hazard rates (λ_i), survival probabilities (S_i), or mean survival times (m_i)

The effect size (Δ) can be either in the form of difference or ratio of hazard rates, survival probabilities, or median survival times.



We consider randomized trials in which subjects are accrued for a period of A (days, months, years) time units, and then followed for a further period of F time units. For each subject, we obtain information on whether or not the event of interest had occurred, the time from study entry (X) to the occurrence of the event or censoring, and the treatment group allocation.

Process for Determining N

Step 1: Determine N_{EV} using α, $1-\beta$, Δ
Step 2: Determine $N = N_{EV}/p_{EV}$
where
p_{EV} = Pr(study participant gets event)

When determining the sample size and/or power, we typically first determine the expected (i.e., required) number of events (N_{EV}), after which we then determine the study size required (N).

The parameters α, β, and Δ are needed to determine N_{EV}, which then gets "extrapolated" to N by dividing by (a guestimate of) the probability that a study participant gets an event (p_{EV}).

Formula for N_{EV}:

$$N_{EV} = \left(\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)(\Delta+1)}{(\Delta-1)} \right)^2 \cdots\cdots (1)$$

Assumes: $N_1 = N_0$, both large, PH assumption

Given α, β, and Δ, the required number of events (N_{EV}) for a trial involving *equal study sizes* in two groups is given by formula (1) shown at the left. This formula assumes that the sample sizes (N_1 and N_0) in each group are large enough to use a normal approximation. The formula also assumes that the PH assumption is satisfied, i.e., Δ is constant over time.

$$\text{Power} = \Pr(Z < z_{EV}) \cdots\cdots\cdots\cdots (2)$$

where

$$z_{EV} = \left[\sqrt{N_{EV}} \left(\frac{\Delta - 1}{\Delta + 1} \right) - z_{1-\alpha/2} \right]$$

The power corresponding to a specified number of events (N_{EV}) can be obtained using formula (2), where $P(Z < z_{EV})$ is the cumulative probability below the value z_{EV} of the standard normal distribution.

Unequal allocations: $R = N_1/N_0$ where $N_1 \neq N_0$

$$N_{EV} = \left(\frac{\left(z_{1-\alpha/2} + z_{1-\beta}\right)(R\Delta+1)}{\sqrt{R}(\Delta-1)} \right)^2 \cdots (3)$$

The formula for N_{EV} can be modified for unequal sample sizes as shown at the left.

Formulae for N: Two versions

Both versions assume
$R = N_1/N_0$ where $N = N_1 + N_0$
and extrapolate from N_{EV} to N using $\mathbf{N = N_{EV}/p_{EV}}$
where p_{EV} = Pr(event occurs)

There are two approaches to determining the **total sample size N** and the corresponding sample sizes N_1 and N_0 in groups 1 and 0, respectively, for $R = N_1/N_0$ where $N_1 + N_0 = N$. Both formulae are determined by extrapolating from NEV to N by dividing by (a guestimate of) the probability that a study participant gets an event (p_{EV}).

Formula using $M_F = A/2 + F$ (i.e., median FU time)

$$N = \frac{N_{EV}}{\frac{R}{R+1}\{1 - S_1(M_F)\} + \frac{1}{R+1}\{1 - S_0(M_F)\}}$$

Special case: $S_i(t) = \exp(-\lambda_i t)$, i=0,1

$$N = \frac{N_{EV}}{\frac{R}{R+1}\{1 - e^{-\lambda_1 M_F}\} + \frac{1}{R+1}\{1 - e^{-\lambda_0 M_F}\}}$$

The formula shown at the left computes p_{EV} making use of the median follow-up time (M_F) for all study subjects, where $M_F = A/2 + F$, with A denoting the accrual period and F denoting the follow-up period for the study.

Formula assuming exponential survival $S_i(t) = \exp(-\lambda_i t)$, i=0,1, and guestimates λ_i, S_i or m_i.

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

where

$$p_{EVi} = 1 - \frac{1}{\lambda_i A}\left[e^{-\lambda_i F} - e^{-\lambda_i (A+F)}\right]$$

λ_i = constant hazard for group i
R = allocation ratio,
A = accrual period,
F = follow-up period

The second formula shown here computes p_{EV} making use of guestimates of the two parameters being compared to define the effect size of interest. These two parameters can be hazards (λ_i, i=0,1), survival probabilities (S_i, i = 0,1) or median survival times (m_i, i=0,1) for each group.

Even if the parameters being guestimated are S_i, or m_i, rather than λ_i, one can transform the guestimate into a corresponding value for λ_i if an exponential survival distribution is assumed. In particular, if the guestimates are median survival times (m_i), it follows that $\lambda_i = \ln 2/m_i$.

$N_1 = [R/(R+1)]N$ and $N_0 = N_1/2$

To obtain N_1 and N_0, from N and R, the formulae shown at the left can be used.

Other issues:

- Choosing accrual and follow-up times;
- Adjusting for loss to follow-up;
- Adjusting for cross-over (drop-in/drop-out)

Three other issues were also considered as shown at the left.

Accrual versus follow-up periods: A = accrual period, F = follow-up period

- Need balance between lengths of accrual and follow-up periods
- Cost of Accrual >> Cost of follow-up

There needs to be a balance between the lengths of accrual and follow-up periods. The cost of accruing subjects into a study is, in general, far greater than the cost of follow-up. In particular, a longer follow-up time has a larger impact on sample size than a longer accrual time (illustrated by example).

Adjusting for loss to follow-up: Need
p_{lof} = proportion of subjects expected to be lost to follow-up
N = study size with full follow-up

$$N_{LOFadj} = N/(1 - p_{lof})$$

We can adjust for possible loss to follow-up in sample size calculations if we have an estimate from previous studies of the proportion of subjects expected to be lost to follow-up (p_{lof}).

The adjusted sample size (N_{LOFadj}) that accounts for the loss to follow-up is obtained by dividing N by ($1-p_{lof}$).

Cross-over (drop-in or drop-out) problems:

Typical approach to analysis uses the **ITT principle**:
subjects analyzed according to the originally assigned treatment group

The typical approach to analysis when study subjects drop-in or drop-out of study groups originally assigned is the **Intention-to-Treat** principle (ITT): analyze subjects according to their original treatment group allocation.

Another analysis approach:
Adjust for crossovers
N = original total sample size (w/o adjusting for cross-overs)
d_c = proportion of (drop-outs) from T to C
d_t = proportion of (drop-ins) from C to T

Another approach to this problem involves adjusting the sample size for potential cross-overs. An adjustment formula can be derived by defining d_c and d_t as the potential proportion of subjects who drop into the control and intervention groups, respectively.

Sample-size formula adjusted for ITT:

$$N_{ITTadj} = \frac{N}{(1 - d_c - d_t)^2}$$

Then it can be shown that for an intention-to-treat analysis, to maintain the study power, the total study size should be adjusted to give a **sample size inflation factor**. This adjustment formula (N_{ITTadj}) is shown at the left.

Detailed Outline

I. Introduction: Background to Statistical Decisions (pages 500–501)

A. Issues addressed in this chapter
   1. How many subjects?
   2. Accrual period for recruitment?
   3. Follow-up period after accrual?
   4. What to do if subjects switch therapies?

B. Review of statistical concepts
   1. List of concepts: H_0, H_A, P-value, Type I error (α), Type II error (β), Power $= 1-\beta$, effect size (Δ).
   2. Δ can be $\theta_1 - \theta_0$ or θ_1/θ_0 where θ_i are hazard rates (λ_i), survival probabilities (S_i), or mean survival times (m_i)

II. Design Considerations: Time-to-Event Outcomes (pages 502–503)

A. Randomized trials scheduling: Subjects are
   1. accrued for a period of A (days, months, years) time units
   2. followed for a further period of F time units.
   3. observed as to whether or not the event or censoring occurred over follow-up from study entry (X)
   4. randomly allocated into treatment groups

B. Study Size (N) versus Number of Events (N_{EV})
   1. N= total # of study participants
   2. N_{EV} = # of participants who get event during the study period
   3. Process for determining N in time-to-event studies:

      Step i: Determine N_{EV} using α, $1-\beta$, Δ

      Step ii: Determine N from N_{EV} using extrapolation formula

      $N = N_{EV}/p_{EV}$ where

      $p_{EV} = \Pr(\text{study participant gets event})$
   4. In practice: determine N first, since N_{EV} occurs after study begins
   5. Determine N from N_{EV} using assumptions on
      i. study duration
      ii. distribution of event times.

III. Determining the Required Number of Events (N_{EV}) (pages 504–505)

A. Formula for N_{EV} (Assumes: $N_1 = N_0$, both large, PH assumption):

$$N_{EV} = \left(\frac{(z_{1-\alpha/2} + z_{1-\beta})(\Delta + 1)}{(\Delta - 1)}\right)^2 \cdots\cdots\cdots\cdots (1)$$

B. Formula for Power, given N_{EV}, α, and Δ:

$$\text{Power} = \Pr(Z < z_{EV}) \quad \cdots\cdots (2)$$

where $z_{EV} = \left[\sqrt{N_{EV}}\left(\frac{\Delta-1}{\Delta+1}\right) - z_{1-\alpha/2}\right]$ and $Z \sim N(0,1)$

C. Formula for N_{EV} (Allows for unequal sample sizes N_1 and N_0 where $N = N_1 + N_0$ and $R = N_1/N_0$):

$$N_{EV} = \left(\frac{(z_{1-\alpha/2} + z_{1-\beta})(R\Delta + 1)}{\sqrt{R}(\Delta - 1)}\right)^2 \quad \cdots\cdots (3)$$

D. Examples for $\alpha = 0.05$, $1-\beta = 0.80$, $\Delta\ (= m_1/m_2) = 2$: $R=1 \Rightarrow N_{EV}=71$; $R=2 \Rightarrow N_{EV}=98$

IV. Determining the Required Total Number of Participants (N) (pages 505–511)

A. Two versions, each assume:

1. $R=N_1/N_0$ where $N = N_1+N_0$
2. N is extrapolated from N_{EV} to N using $N = N_{EV}/p_{EV}$ where $p_{EV} = \Pr(\text{event occurs})$.

B. Version 1: Uses $M_F = A/2 + F$ (i.e., median FU time)

1. Any survival distribution:

$$N = \frac{N_{EV}}{\frac{R}{R+1}\{1 - S_1(M_F)\} + \frac{1}{R+1}\{1 - S_0(M_F)\}}$$

2. Special case: assumes exponential survival $S_i(t) = \exp(-\lambda_i t)$, i=0,1:

$$N = \frac{N_{EV}}{\frac{R}{R+1}\left\{1 - e^{-\lambda_1 M_F}\right\} + \frac{1}{R+1}\left\{1 - e^{-\lambda_0 M_F}\right\}}$$

C. Version 2: Uses guestimates for hazard functions λ_i, i=0,1 and assumes exponential survival $S_i(t) = \exp(-\lambda_i t)$, where λ_i = constant hazard for group i=0,1.

$$N = \frac{N_{EV}}{\frac{R}{R+1}p_{EV1} + \frac{1}{R+1}p_{EV0}}$$

where

$$p_{EVi} = 1 - \frac{1}{\lambda_i A}\left[e^{-\lambda_i F} - e^{-\lambda_i (A+F)}\right]$$

D. Formula for obtaining N_1 and N_0 from N and R: $N_1 = [R/(R+1)]N$ and $N_0 = N_1/2$.

E. Example using A=2, F=4. α=0.05, β=0.10, $\Delta = \lambda_0/\lambda_1 = 2$, R=1.
   1. Version 1: $M_F = 5$, $N \approx 308$
   2. Version 2: $\lambda_0 = 0.10$, $\lambda_1 = 0.05$, $p_{EV1} = 0.2207$, $p_{EV0} = 0.3925$, $N \approx 309$

F. Second Example: A=15, F=12, α=0.05, β=0.10, $m_0 = 6$, $m_1 = 10$, R=2.
   1. Using version 2, must change m_i to λ_i using the relationship $\lambda_i = \ln 2/m_i$.
   2. $m_0 = 6 \Rightarrow \lambda_0 = 0.1155$; $m_1 = 10 \Rightarrow \lambda_1 = 0.0693$.
   3. Computed $N_{EV} \approx 222$.
   4. Using version 2, computed p_{EV1}=0.7293, p_{ev0}=0.8812, and $N \approx 285$.
   5. Obtaining N_1 and N_0 from N and R: N_1=190, N_0=95.

V. Other Design Issues (pages 511–514)

A. Three issues:
   - Choosing accrual and follow-up times
   - Adjusting for loss to follow-up
   - Adjusting for cross-over (drop-in and/or drop-out)

B. Accrual versus follow-up times
   1. Need balance
   2. Cost of accrual >> Cost of follow-up
   3. (Colorectal cancer) Example: longer follow-up time has > impact on sample size than longer accrual time.

C. Adjusting for loss to follow-up.
   1. Formula: $N_{LOFadj} = N/(1-p_{lof})$ where p_{lof} = proportion of subjects expected to be lost to follow-up.
   2. Example: N=270, $p_{lof} = 0.25$ $N_{LOFadj} = 360$.

D. Adjusting for crossovers (drop-in/drop-out).
   1. Typical analysis approach: Intention- to-Treat (ITT); analyze what you randomize
   2. Can adjust sample size to allow for cross-overs.
      a. Formula: $N_{ITTadj} = N/(1 - d_c - d_t)^2$,

   where
   N = original total sample size
   d_c = proportion of (drop-outs) from T to C
   d_t = proportion of (drop-ins) from C to T

   b. Table of sample size inflation factors for different combinations of d_c and d_t.
      i. Subjects receiving opposite treatment from randomized allocation "dilute" effect of intervention unless adjusted for.
   c. Example (Exercise regimen effect on colorectal cancer?): N=600, $d_c = 0.05$, $d_t = 0.15 \Rightarrow N_{ITTadj} = 936$.

VI. Summary (pages 514–517)

Practice Exercises

Answer questions 1–10 as true or false (circle T or F)

T F 1. In a randomized trial comparing 2 groups, the null hypothesis (H_0) assumes that the survival experience in the groups is different.

T F 2. The effect size is typically determined in the form of a difference of hazard rates.

T F 3. Suppose that a time-to-event study is designed to determine if there is reduction in annual hazard rates from 5% to 2.5%. Then the effect size of interest defined in terms of a ratio measure is 2.

T F 4. The P-value is the probability that the null hypothesis is true given the observed study data.

T F 5. Consider a two-group randomized clinical trial, in which the randomization process is not systematically flawed. Then the use of a P-value to compare the distributions of a known risk factor (e.g., gender) in each group is uninformative.

T F 6. When determining sample size for a time-to-event study, one typically first determines the expected (i.e., required) number of events (N_{EV}) prior to determining the study size required (N).

T F 7. When carrying out a time-to-event study, one typically recruits the study participants prior to observing the number of events that actually occur.

T F 8. A formula for the total sample size (N) for a time-to-event study can be derived by multiplying the required number of events (N_{EV}) by the probability (p_{EV}) that a study subject will get an event since study entry, i.e., $N = N_{EV} \times p_{EV}$.

T F 9. Suppose the allocation ratio ($R = N_1/N_0$) for subjects into two treatment groups is 1/2, and it is determined that the total study size (N) required is 300. Then $N_1 = 100$ and $N_0 = 200$.

T F 10. Suppose that it is determined that the total study size required (N) is 300 without considering the possibility of loss to follow-up during the study. If, nevertheless, the investigators expect that 20% of study subjects would be lost to follow-up, Then to adjust for possible loss to follow-up, the sample size should be increased to 1500 subjects.

Consider a randomized trial in which the accrual period (A) is 2 years and the follow-up period (F) is 3 years. Suppose further that the allocation ratio (R) is 2, α=.05, β=.20, and the study aim is to increase the median survival time (m_i) in the control group from 1.5 years to 2.2 years in the intervention group.

11. Assuming an exponential survival rate in both groups, transform the given median survival times (m_i) to corresponding hazard rates (λ_i).
12. Based on your answer to question 11, what is Δ?
13. Determine the number of events (N_{EV}) required for the study.
14. Determine the probability of getting an event (p_{EVi}) for each group, i=0,1.
15. Determine the total sample size (N) required for this study.
16. Determine the sample sizes in each group (i.e., N_0 and N_1).
17. Based on your answer to question 14, how would you adjust your total sample size to consider loss-to-follow-up of 25%?
18. Based on your answer to question 17, determine the sample sizes required in each group (i.e., N_0 and N_1).
19. Based on your answer to question 17, how would you adjust your total sample size to consider cross-over percentages of $d_c = .05$ and $d_t = 0.10$?
20. Based on your answer to question 17, determine the sample sizes required in each group (i.e., N_0 and N_1).

Test

Consider a randomized trial in which the accrual period (A) is 2 years and the follow-up period (F) is 2 years. Suppose further that the allocation ratio (R) is 2, α=.05, β=.10, and the study aim is to reduce hazard rate (λ_i) in the control group from 10% to 5% in the intervention group.

1. Use Formula 1 (see detailed outline) to determine the required number of events (N_{EV}), total study size (N), and sample sizes required in each group (i.e., N_0 and N_1).
2. Use Formula 2 (see detailed outline) to determine the required number of events (N_{EV}), total study size (N), and sample sizes required in each group (i.e., N_0 and N_1).
3. How do the total sample sizes (N) obtained in questions 1 and 2 compare?
4. Based on your answer to question 2, how would you adjust your total sample size to consider loss-to-follow-up of 25%?
5. Based on your answer to question 4, determine the sample sizes required in each group (i.e., N_0 and N_1).
6. Based on your answer to question 4, how would you adjust your total sample size to consider cross-over proportions of $d_c = .05$ and $d_t = 0.10$?
7. Based on your answer to question 6, determine the sample sizes required in each group (i.e., N_0 and N_1).
8. Using the total study size (N) calculated in question 6, these study subjects will need to be recruited at an accrual rate of r=N/A per year, where A=2. If this accrual rate is not feasible, i.e., you couldn't find r subjects per year, how can you adjust your sample size to make your study feasible?

Answers to Practice Exercises

2. F: H_0 assumes that the survival experience in the two groups is the same.
2. F: The effect size can be either in the form of a difference or ratio of hazard rates, survival probabilities, or median survival times.

1. T
2. F: The P-value is the probability of obtaining a result as or more extreme than the result obtained in one's study given that the that the null hypothesis is true.
3. T
4. T
7. T
8. F: $N = N_{EV}/p_{EV}$.
9. T

10. F: The sample size should be increased to 375 subjects.
11. Using the formula $\lambda_i = \ln 2/m_i$, i=0,1, we obtain $m_0 = 1.5 \Rightarrow \lambda_0 = 0.4621$; $m_1 = 2.2 \Rightarrow \lambda_1 = 0.3151$.
12. $\Delta = \lambda_0//\lambda_1 = 0.4621/0.3151 = 1.467$.
13. $N_{EV} = \left(\dfrac{(1.96 + 0.84)(2(1.467) + 1)}{\sqrt{2}(1.467 - 1)}\right)^2 = 129.91$, which rounds to 130 required events.
14. Using Formula 2, we obtain

$$p_{EV1} = 1 - \frac{1}{(0.3151)(2)}\left[e^{-(0.3151)(3)} - e^{-(0.3151)(2+3)}\right]$$
$$= 1 - 0.2883 = 0.7117$$
$$p_{EV0} = 1 - \frac{1}{(0.4621)(2)}\left[e^{-(0.4621)(3)} - e^{-(0.4621)(2+3)}\right]$$
$$= 1 - 0.1632 = 0.8368$$

15. $$N = \frac{130}{\frac{2}{2+1}(0.7117) + \frac{1}{2+1}(0.8368)}$$
$$= 172.55 \approx 173$$

16. $N_1 = [2/(2+1)]173 = 115.33 \approx 115$ and $N_0 = 115.33/2 = 57.67 \approx 58$
17. $N_{LOFadj} = 173/(1 - 0.25) = 230.67 \approx 231$
18. $N_1 = [2/(2+1)]231 = 154$ and $N_0 = 154/2 = 77$
19. $N_{ITTadj} = 231/(1 - 0.05 - 0.10)^2 = 319.72 \approx 320$
20. $N_1 = [2/(2+1)]320 = 213.33 \approx 213$ and $N_0 = 213.33/2 = 106.665 \approx 107$

Computer Appendix: Survival Analysis on the Computer

D.G. Kleinbaum and M.Klein, *Survival Analysis: A Self-Learning Text, Third Edition*,
Statistics for Biology and Health, DOI 10.1007/978-1-4419-6646-9,
© Springer Science+Business Media, LLC 2012

In this appendix, we provide examples of computer programs for carrying out the survival analyses described in this text. This appendix does not give an exhaustive survey of all computer packages currently available, but rather is intended to describe the similarities and differences among four of the most widely used packages. The software packages that we describe are Stata (version 10.0), SAS (version 9.2), SPSS (PASW 18), and R. A complete description of these packages is beyond the scope of this appendix. Readers are referred to the built-in help functions for each program for further information.

Datasets

Most of the computer syntax and output presented in this appendix are obtained from running step-by-step survival analyses on the "addicts" dataset. The other dataset that is utilized in this appendix is the "bladder cancer" dataset for analyses of recurrent events. The "addicts" and "bladder cancer" data are described below and can be downloaded from our website at **http://www.sph.emory.edu/dkleinb/surv3.htm**. On this website, we also provide many of the other datasets that have been used in the examples and exercises throughout this text. The data on our website are provided in five forms (1) as Stata datasets (with a **.dta** extension), (2) as SAS datasets (with a **.sas7bdat** extension), (3) as SPSS datasets (with a **.sav** extension), (4) as R datasets (with an **.rda** extension), and (5) as text datasets (with a **.dat** extension).

Addicts Dataset (addicts.dat)

In a 1991 Australian study by Caplehorn et al., two methadone treatment clinics for heroin addicts were compared to assess patient time remaining under methadone treatment. A patient's survival time was determined as the time (in days) until the person dropped out of the clinic or was censored. The two clinics differed according to its live-in policies for patients. The variables are defined as follows:

ID – Patient ID
SURVT – The time (in days) until the patient dropped out of the clinic or was censored
STATUS – Indicates whether the patient dropped out of the clinic (coded 1) or was censored (coded 0)
CLINIC – Indicates which methadone treatment clinic the patient attended (coded 1 or 2)
PRISON – Indicates whether the patient had a prison record (coded 1) or not (coded 0)
DOSE – A continuous variable for the patient's maximum methadone dose (mg/day)

Bladder Cancer Dataset (bladder.dat)

The bladder cancer dataset contains recurrent event outcome information for eighty-six cancer patients followed for the recurrence of bladder cancer tumor after transurethral surgical excision (Byar and Green 1980). The exposure of interest is the effect of the drug treatment of thiotepa. Control variables are the initial number and initial size of tumors. The data layout is suitable for a counting processes approach. The variables are defined as follows:

ID – Patient ID (may have multiple observations for the same subject)
EVENT – Indicates whether the patient had a tumor (coded 1) or not (coded 0)
INTERVAL – A counting number representing the order of the time interval for a given subject (coded 1 for the subject's first time interval, coded 2 for a subject's second time interval, etc.)
START – The starting time (in months) for each interval
STOP – The time of event (in months) or censorship for each interval
TX – Treatment status (coded 1 for treatment with thiotepa and 0 for the placebo)
NUM – The initial number of tumors
SIZE – The initial size (in centimeters) of the tumor

Software

What follows is a detailed explanation of the code and output necessary to perform the type of survival analyses described in this text. The rest of this appendix is divided into four broad sections, one for each of the following software packages:

A. Stata
B. SAS
C. SPSS
D. R Software

Each of these sections is self-contained, allowing the reader to focus on the particular statistical package of his or her interest.

A. Stata

Analyses using Stata are obtained by typing the appropriate statistical commands in the Stata Command window or in the Stata Do-file Editor window. The key commands used to perform the survival analyses are listed below. These commands are case sensitive and lower-case letters should be used.

stset – Declares data in memory to be survival data. Used to define the "time-to-event" variable, the "status" variable, and other relevant survival variables. Other Stata Commands beginning with **st** utilize these defined variables.

sts list – Produces Kaplan-Meier (KM) or Cox-adjusted survival estimates in the output window. The default is KM survival estimates.

sts graph – Produces plots of Kaplan-Meier (KM) survival estimates. This command can also be used to produce Cox-adjusted survival plots.

sts generate – Creates a variable in the working dataset that contains Kaplan-Meier or Cox adjusted survival estimates.

sts test – Used to perform statistical tests for the equality of survival functions across strata.

stphplot – Produces plots of log-log survival against the log of time for the assessment of the proportional hazards (PH) assumption. The user can request KM log-log survival plots or Cox adjusted log-log survival plots.

stcoxkm – Produces KM survival plots and Cox adjusted survival plots on the same graph.

stcox – Used to run a Cox proportional hazard model, a stratified Cox model, or an extended Cox model (i.e., containing time varying covariates).

stphtest – Performs statistical tests on the PH assumption based on Schoenfeld residuals. Use of this command requires that a Cox model be previously run with the command **stcox** and the **schoenfeld()** option.

streg – Used to run parametric survival models.

Four windows will appear when Stata is opened. These windows are labeled Stata Command, Stata Results, Review, and Variables. The user can click on File → Open to select a working dataset for analysis. Once a dataset is selected, the names of its variables appear in the Variables window. Commands are entered in the Stata Command window. The output generated by commands appears in the Results window after the return key is pressed. The Review window preserves a history of all the commands executed during the Stata session. The commands in the Review window can be saved, copied, or edited as the user desires. Command can also be run from the Review window by double-clicking on the command. Commands can also be saved in a file by clicking on the log button on the Stata tool bar.

Alternatively, commands can be typed, or pasted into the Do-file Editor. The Do-file Editor window is activated by clicking on Window → Do-file Editor or by simply clicking on the Do-file Editor button on the Stata tool bar. Commands are executed from the Do-file Editor by clicking on Tools → Do. The advantage of running commands from the Do-file Editor is that commands need not be entered

and executed one at a time as they do from the Stata Command window. The Do-file Editor serves a similar function as the program editor in SAS. In fact, by typing **#delim** in the Do-file Editor window, the semicolon becomes the delimiter for completing Stata statements (as in SAS) rather than the default carriage return.

The survival analyses demonstrated in Stata are as follows:

1. Estimating survival functions (unadjusted) and comparing them across strata
2. Assessing the PH assumption using graphical approaches
3. Running a Cox PH model
4. Running a stratified Cox model
5. Assessing the PH assumption with a statistical test
6. Obtaining Cox adjusted survival curves
7. Running an extended Cox model
8. Running parametric models
9. Running frailty models
10. Modeling recurrent events

The first step is to activate the addicts dataset by clicking on File → Open and selecting the Stata dataset, **addicts.dta**. Once this is accomplished, you will see the command **use "addicts.dta", clear** in the Review window and Results window. This indicates that the addicts dataset is activated in Stata's memory.

To perform survival analyses, you must indicate which variable is the "time-to-event" variable and which variable is the "status" variable. Rather than program this in every survival analysis command, Stata provides a way to program it once with the **stset** command. All survival commands beginning with **st** utilize the survival variables defined by **stset** as long as the dataset remains in active memory. The code to define the survival variables for the addicts data is as follows:

```
stset survt, failure(status==1) id(id)
```

Following the word **stset** comes the name of the "time-to-event" variable. Options for Stata Commands follow a comma. The first option used is to define the variable and value that indicates an event (or failure) rather than a censorship. Without this option, Stata assumes that all observations had an event (i.e., no censorships). Notice that two equal signs are used to express equality. A single equal sign is used to designate assignment. The next option defines the id variable as the variable, ID. This is unnecessary with the addicts dataset since each observation

represents a different patient (cluster). However if there were multiple observations and multiple events for a single subject (cluster), Stata can provide robust variance estimates appropriate for clustered data.

The **stset** command will add four new variables to the dataset. Stata interprets these variables as follows:

_t – The "time-to-event" variable
_d – The "status variable" (coded 1 for an event and 0 for a censorship)
_t0 – The beginning "time variable." All observations start at time 0 by default
_st – Indicates which variables are used in the analysis. All observations are used (coded 1) by default

To see the first 10 observations printed in the output window, enter the command:

```
list in 1/10
```

The command **stdes** provides descriptive information (output below) of survival time.

```
stdes
```

```
      failure _d:  status == 1
analysis time _t:  survt
              id:  id

                            |---------per subject----------|
Category              total       mean   min   median    max
-------------------------------------------------------------------------------
no. of subjects         238
no. of records          238          1     1        1      1

(first) entry time                   0     0        0      0
(final) exit time              402.5714     2    367.5   1076

subjects with gap         0
time on gap if gap        0          .     .        .      .
time at risk          95812   402.5714     2    367.5   1076

failures                150   .6302521     0        1      1
-------------------------------------------------------------------------------
```

The commands **strate** and **stir** can be used to obtain incident rate comparisons for different categories of specified variables. The **strate** command lists the incident rates by CLINIC while the **stir** command gives rate ratios and rate

differences. Type the following commands one at a time (output omitted):

```
strate clinic
stir clinic
```

For the survival analyses that follow, it is assumed that the command **stset** has been run for the addicts dataset, as demonstrated on the previous page.

1. ESTIMATING SURVIVAL FUNCTIONS (UNADJUSTED) AND COMPARING THEM ACROSS STRATA

To obtain Kaplan-Meier survival estimates use the command **sts list**. The code and output follow:

```
sts list
```

```
      failure _d:  status == 1
analysis time _t:  survt
                id:  id
```

Time	Beg. Total	Fail	Net Lost	Survivor Function	Std. Error	[95% Conf.	Int.]
2	238	0	2	1.0000	.	.	.
7	236	1	0	0.9958	0.0042	0.9703	0.9994
13	235	1	0	0.9915	0.0060	0.9665	0.9979
17	234	1	0	0.9873	0.0073	0.9611	0.9959
19	233	1	0	0.9831	0.0084	0.9555	0.9936
26	232	1	0	0.9788	0.0094	0.9499	0.9911
28	231	0	2	0.9788	0.0094	0.9499	0.9911
29	229	1	0	0.9745	0.0103	0.9442	0.9885
30	228	1	0	0.9703	0.0111	0.9386	0.9857
33	227	1	0	0.9660	0.0118	0.9331	0.9828
.				.			
.				.			
.				.			
905	8	0	1	0.1362	0.0364	0.0748	0.2159
932	7	0	2	0.1362	0.0364	0.0748	0.2159
944	5	0	1	0.1362	0.0364	0.0748	0.2159
969	4	0	1	0.1362	0.0364	0.0748	0.2159
1021	3	0	1	0.1362	0.0364	0.0748	0.2159
1052	2	0	1	0.1362	0.0364	0.0748	0.2159
1076	1	0	1	0.1362	0.0364	0.0748	0.2159

If we wish to stratify by CLINIC and compare the survival estimates side-to-side for specified time points, we use the **by()** and **compare()** option. The code and output follow:

sts list, by(clinic) compare at (0 20 to 1080)

```
        failure _d:  status == 1
  analysis time _t:  survt
                id:  id

                        Survivor Function
clinic                      1          2
-------------------------------------------
time          0        1.0000     1.0000
             20        0.9815     0.9865
             40        0.9502     0.9595
             60        0.9189     0.9459
             80        0.9000     0.9320
            100        0.8746     0.9320
            120        0.8681     0.9179
            140        0.8422     0.9038
            160        0.8093     0.8753
            180        0.7690     0.8466
            200        0.7420     0.8323
            220        0.6942     0.8179
                .
                .
                .
            840        0.0725     0.5745
            860        0.0543     0.5745
            880        0.0543     0.5171
            900        0.0181     0.5171
            920             .     0.5171
            940             .     0.5171
            960             .     0.5171
            980             .     0.5171
           1000             .     0.5171
           1020             .     0.5171
           1040             .     0.5171
           1060             .     0.5171
           1080             .          .
```

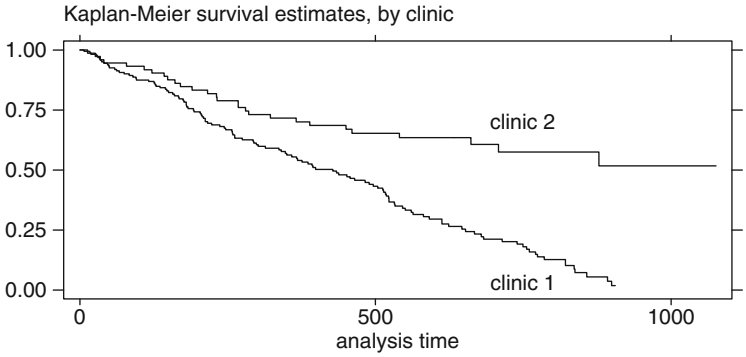
Notice that the survival rate for CLINIC=2 is higher than CLINIC=1. Other survival times could have been requested using the **compare()** option.

To graph the Kaplan-Meier survival function (against time), use the code:

sts graph

The code and output that provide a graph of the Kaplan-Meier survival function stratified by CLINIC follow:

```
sts graph, by(clinic)
```



The **failure** option graphs the failure function (the cumulative risk) rather than the survival (zero to one rather than one to zero). The code follows (output omitted):

```
sts graph, by(clinic) failure
```

The code to run the log rank test on the variable CLINIC (and output) follows:

```
sts test clinic
```

```
      failure _d:  status == 1
analysis time _t:  survt
             id:  id

Log-rank test for equality of survivor functions

           Events    Events
clinic   observed  expected
---------------------------------
1             122     90.91
2              28     59.09
---------------------------------
Total         150    150.00

        chi2(1) =     27.89
        Pr>chi2 =    0.0000
```

The Wilcoxon, Tarone-Ware, Peto, and Flemington-Harrington tests can also be requested. These tests are variations of the log rank test that weight each observation differently. The Wilcoxon test weights the j^{th} failure time by n_i (the number still at risk). The Tarone-Ware test weights the j^{th} failure time by $\sqrt{n_j}$. The Peto test weights the j^{th} failure time by the survival estimate, $\tilde{s}(t_j)$ calculated over all groups combined. This survival estimate, $\tilde{s}(t_j)$, is similar but not exactly equal to the Kaplan-Meier survival estimate. The Flemington-Harington test uses the Kaplan-Meier survival estimate, $\hat{s}(t)$, over all groups to calculate its weights for the j^{th} failure time, $\hat{s}(t_{j-1})^p[1-\hat{s}(t_{j-1})]^q$, so it takes two arguments (p and q). The code follows (output omitted):

```
sts test clinic, wilcoxon
sts test clinic, tware
sts test clinic, peto
sts test clinic, fh(1,3)
```

Notice that the default test for the **sts test** command is the log rank test. The choice of which weighting of the test statistic to use (e.g., log rank or Wilcoxon) depends on which test is believed to provide the greatest statistical power, which in turn depends on how it is believed the null hypothesis is violated. However, one should make an a priori decision on which statistical test to use rather than fish for a desired p-value.

A stratified log rank test for CLINIC (stratified by PRISON) can be run with the strata option. With the stratified approach, the observed minus expected number of events are summed over all failure times for each group within each stratum and then summed over all strata. The code follows (output omitted):

```
sts test clinic, strata(prison)
```

The **sts generate** command can be used to create a new variable in the working dataset containing the KM survival estimates. The following code defines a new variable called SKM (the variable name is the user's choice) that contains KM survival estimates stratified by CLINIC:

```
sts generate skm=s, by(clinic)
```

The **ltable** command produces life tables. Life tables are an alternative approach to Kaplan-Meier that are particularly useful if you do not have individual-level data. The code and output that follows provide life table survival estimates, stratified by CLINIC, at the time points (in days) specified by the **interval()** option:

ltable survt status, by(clinic) interval(60 150 200 280 365 730 1095)

Interval	Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
clinic = 1							
0	.	163	13	4	0.9193	0.0215	0.8650	0.9523
60	150	146	14	6	0.8293	0.0300	0.7609	0.8796
150	200	126	13	3	0.7427	0.0352	0.6661	0.8043
200	280	110	17	2	0.6268	0.0393	0.5446	0.6984
280	365	91	10	6	0.5556	0.0408	0.4720	0.6313
365	730	75	41	15	0.2181	0.0367	0.1509	0.2934
730	1095	19	14	5	0.0330	0.0200	0.0080	0.0902
clinic = 2							
0	.	75	4	2	0.9459	0.0263	0.8624	0.9794
60	150	69	5	3	0.8759	0.0388	0.7749	0.9334
150	200	61	3	0	0.8328	0.0441	0.7242	0.9015
200	280	58	5	1	0.7604	0.0508	0.6429	0.8438
280	365	52	3	2	0.7157	0.0540	0.5943	0.8065
365	730	47	7	23	0.5745	0.0645	0.4385	0.6890
730	1095	17	1	16	0.5107	0.0831	0.3395	0.6584

2. ASSESSING THE PH ASSUMPTION USING GRAPHICAL APPROACHES

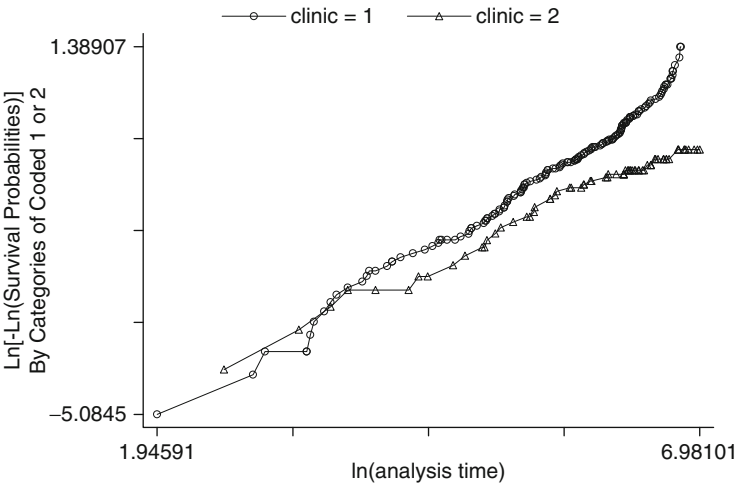
Three graphical approaches for the assessment of the PH assumption for the variable CLINIC are demonstrated:

1) Log-log Kaplan-Meier survival estimates (stratified by CLINIC) plotted against time (or against the log of time)
2) Log-log Cox adjusted survival estimates (stratified by CLINIC) plotted against time
3) Kaplan-Meier survival estimates and Cox adjusted survival estimates plotted on the same graph.

All three approaches are somewhat subjective yet hopefully informative. The first two approaches are based on whether the log log survival curves are parallel for different levels of CLINIC. The third approach is to determine if the Cox adjusted survival curve (not stratified) is close to the KM curve. In other words, are predicted values from the PH model (from Cox) close to the "observed" values using KM?

The first two approaches use the **stphplot** command while the third approach uses the **stcoxkm** command. The code and output for the log-log Kaplan-Meier survival plots follow:

stphplot, by(clinic) nonegative

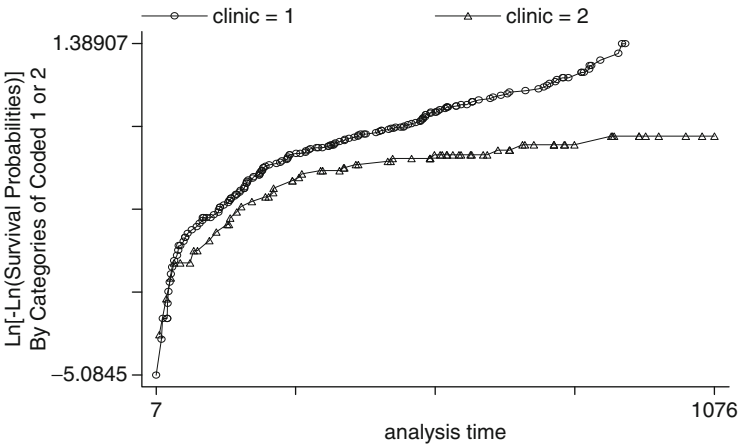


The left side of the graph seems jumpy for CLINIC=1 but it only represents a few events. It also looks like there is some separation between the plots at the later times (right side). The **nonegative** option in the code requests log(-log) curves rather than the default -log(-log) curves. The choice is arbitrary. Without the option, the curves would go downward rather than upward (left-to-right).

Stata (as well as SAS) plot log(survival time) rather than survival time on the horizontal axis by default. As far as checking the parallel assumption, it does not matter if log (survival time) or survival time is on the horizontal axis. However, if the log log survival curves look like straight lines with log(survival time) on the horizontal axis, then there is evidence that the "time-to-event" variable follows a Weibull distribution. If the slope of the line equals one, then there is evidence that the survival time variable (SURVT) follows an exponential distribution – a special case of the Weibull distribution. For these situations, a parametric survival model can be used.

It may be visually more informative to graph the log log survival curves against survival time (rather than log survival time). The **nolntime** option can be used to put survival time on the horizontal axis. The code and output follows:

stphplot, by(clinic) nonegative nolntime

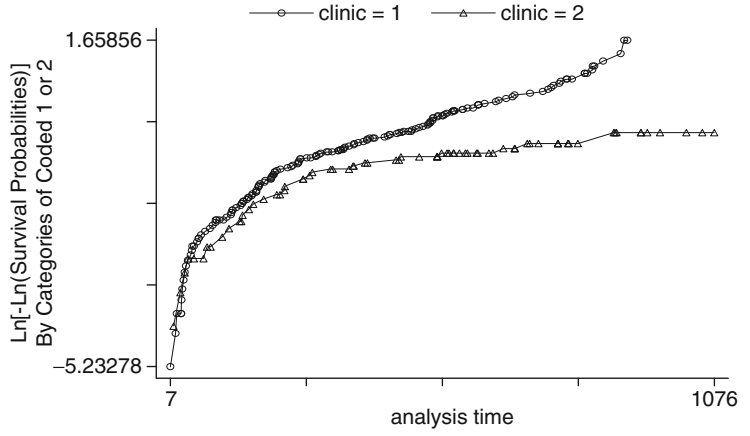


The graph suggests that the curves begin to diverge over time.

The **stphplot** command can also be used to obtain log-log Cox adjusted survival estimates. The code follows:

stphplot, strata(clinic) adjust(prison dose) nonegative nolntime

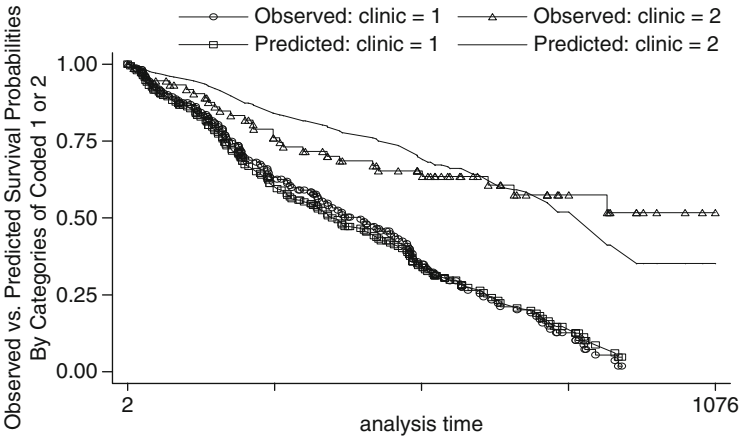
The log-log curves are adjusted for PRISON and DOSE using a stratified COX model on the variable CLINIC. The mean values of PRISON and DOSE are used for the adjustment. The output follows:



The Cox adjusted curves look very similar to the KM curves.

The **stcoxkm** command is used to compare Kaplan-Meier survival estimates and Cox adjusted survival estimates plotted on the same graph. The code and output follow:

```
stcoxkm, by(clinic)
```



The KM and adjusted survival curves are very close together for CLINIC=1 and less so for CLINIC=2. These graphical approaches suggest that there is some violation with the PH assumption. The predicted values are Cox adjusted for CLINIC, and therefore assume the PH assumption. Notice that the predicted survival curves are not parallel by CLINIC even though we are adjusting for CLINIC. It is the log-log survival curves, rather than the survival curves, that are forced to be parallel by Cox adjustment.

The same graphical analyses can be performed with PRISON and DOSE. However, DOSE would have to be categorized since it is a continuous variable.

3. RUNNING A COX PH MODEL

For a Cox PH model, the key assumption is that the hazard is proportional across different patterns of covariates. The first model that is demonstrated contains all three covariates: PRISON, DOSE, and CLINIC. In this model, we are assuming the same baseline hazard for all possible patterns of these covariates. In other words, we are accepting the PH assumption for each covariate (perhaps incorrectly). The code and output follow:

stcox prison clinic dose, nohr

```
        failure _d:  status == 1
  analysis time _t:  survt
                id:  id

Iteration 0:   log likelihood =  -705.6619
Iteration 1:   log likelihood = -674.54907
Iteration 2:   log likelihood =   -673.407
Iteration 3:   log likelihood = -673.40242
Iteration 4:   log likelihood = -673.40242
Refining estimates:
Iteration 0:   log likelihood = -673.40242

Cox regression -- Breslow method for ties

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   64.52
Log likelihood   =  -673.40242          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
     _t
     _d      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .3265108   .1672211     1.95   0.051    -.0012366    .6542581
clinic   -1.00887    .2148709    -4.70   0.000    -1.430009   -.5877304
  dose   -.0353962   .0063795    -5.55   0.000    -.0478997   -.0228926
------------------------------------------------------------------------------
```

The output indicates that it took five iterations for the log likelihood to converge at −673.40242. The iteration history typically appears at the top of Stata model output; however, the iteration history will subsequently be omitted. The final table lists the regression coefficients, their standard errors, a Wald test statistic (z) for each covariate, with corresponding p-value, and 95% confidence interval.

The **nohr** option in the **stcox** command requests the regression coefficients rather than the default exponentiated coefficients (hazard ratios). If you want the exponentiated coefficients, omit the **nohr** option. The code and output follow:

stcox prison clinic dose

```
Cox regression -- Breslow method for ties

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   64.52
Log likelihood   =  -673.40242          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t
          _d  Haz. Ratio  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    1.386123    .231789   1.95   0.051     .9987642    1.923715
clinic    .3646309   .0783486  -4.70   0.000     .2393068    .5555868
  dose     .965223   .0061576  -5.55   0.000     .9532294    .9773675
------------------------------------------------------------------------------
```

This table contains the hazard ratios, its standard errors, and corresponding confidence intervals. Notice that you do not need to supply the "time-to event" variable or the status variable when using the **stcox** command. The **stcox** command uses the information supplied from the **stset** command. A Cox model can also be run using the **cox** command, which does not rely on the **stset** command having previously been run. The code follows:

cox survt prison clinic dose, dead(status)

Notice that with the **cox** command, we have to list the variable SURVT. The **dead()** option is used to indicate that the variable STATUS distinguishes events from censorship. The variable used with the **dead()** option needs to be coded nonzero for events and zero for censorships. The output from the **cox** command follows:

```
Cox regression -- Breslow method for ties
Entry time 0                            Number of obs  =     238
                                        LR chi2(3)     =   64.52
                                        Prob > chi2    =  0.0000
Log likelihood = -673.40242             Pseudo R2      =  0.0457

------------------------------------------------------------------------------
       survt
      status      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .3265108    .1672211   1.95   0.051    -.0012366    .6542581
clinic   -1.00887    .2148709  -4.70   0.000    -1.430009   -.5877304
  dose   -.0353962    .0063795  -5.55   0.000    -.0478997   -.0228926
------------------------------------------------------------------------------
```

The output is identical to that obtained from the **stcox** command except that the regression coefficients are

given by default. The **hr** option for the **cox** command supplies the exponentiated coefficients.

Notice with the output that the default method of handling ties (i.e., when multiple events happen at the same time) is the Breslow method. If you wish to use more exact methods, you can use the **exactp** option (for the exact partial likelihood) or the **exactm** option (for the exact marginal likelihood) in the **stcox** or **cox** command. The exact methods are computationally more intensive and typically have a slight impact on the parameter estimates. However, if there are a lot of events that occur at the same time, then exact methods are preferred. The code and output follow:

stcox prison clinic dose, nohr exactm

```
Cox regression -- exact marginal likelihood

No. of subjects  =        238           Number of obs  =     238
No. of failures  =        150
Time at risk     =      95812
                                        LR chi2(3)     =   64.56
Log likelihood   =  -666.3274           Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t
          _d      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .326581   .1672306     1.95   0.051    -.0011849    .6543469
clinic  -1.009906   .2148906    -4.70   0.000    -1.431084   -.5887285
  dose  -.0353694   .0063789    -5.54   0.000    -.0478718   -.0228669
```

Alternatively, you could use Efron method of handling ties. This is the method that the R statistical package uses as its default. The code follows (output omitted):

stcox prison clinic dose, nohr efron

Suppose you are interested in running a Cox model with two interaction terms with PRISON. The **generate** command can be used to define new variables. The variables CLIN_PR and CLIN_DO are product terms that are defined from CLINIC × PRISON and CLINIC × DOSE. The code follows:

generate clin_pr=clinic*prison
generate clin_do=clinic*dose

Type **describe** or **list** to see that the new variables are in the working dataset.

The following code runs the Cox model with the two interaction terms:

stcox prison clinic dose clin_pr clin_do, nohr

```
Cox regression -- Breslow method for ties

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(5)     =   68.12
Log likelihood   =  -671.59969          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
     _t
     _d      Coef.  Std. Err.      z   P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
 prison   1.191998   .5413685   2.20  0.028   .1309348   2.253061
 clinic   .1746985    .893116   0.20  0.845  -1.575777   1.925174
   dose  -.0193175     .01935  -1.00  0.318  -.0572428   .0186079
clin_pr  -.7379931   .4314868  -1.71  0.087  -1.583692   .1077055
clin_do  -.0138608   .0143275  -0.97  0.333  -.0419422   .0142206
------------------------------------------------------------------------------
```

The **lrtest** command can be used to perform likelihood ratio tests. For example, to perform a likelihood ratio test on the two interaction terms, CLIN_PR and CLIN_DO, in the preceding model, we can save the –2 log likelihood statistic of the full model in the computer's memory by typing the following command:

lrtest, saving(0)

Now, the reduced model (without the interaction terms) can be run (output omitted) by typing:

stcox prison clinic dose

After the reduced model is run, the following command provides the results of the likelihood ratio test comparing the full model (with the interaction terms) to the reduced model:

lrtest

The resulting output follows:

```
Cox: likelihood-ratio test          chi2(2)     =     3.61
                                    Prob > chi2 =   0.1648
```

The p-value of 0.1648 is not significant at the alpha = 0.05 level.

4. RUNNING A STRATIFIED COX MODEL

If the proportional hazard assumption is not met for the variable CLINIC, but is met for the variables PRISON and DOSE, then a stratified Cox analysis can be performed. The **stcox** command can be used to run a stratified Cox model. The following code (with output) runs a Cox model stratified on CLINIC:

stcox prison dose, strata(clinic)

```
Stratified Cox regr. -- Breslow method for ties

No. of subjects  =       238            Number of obs  =     238
No. of failures  =       150
Time at risk     =     95812
                                        LR chi2(2)     =   33.94
Log likelihood   =  -597.714            Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t
          _d  Haz. Ratio  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison   1.475192   .2491827    2.30   0.021     1.059418    2.054138
  dose    .9654655   .0062418   -5.44   0.000      .953309     .977777
-------------------------------------------------------------------------------
                                                  Stratified by clinic
```

The **strata()** option allows up to five stratified variables.

A stratified Cox model can be run including the two interaction terms. Recall that the **generate** command created these variables in the previous section. This model allows for the effect of PRISON and DOSE to differ for different values of CLINIC. The code and output follow:

stcox prison dose clin_pr clin_do, strata(clinic) nohr

```
Stratified Cox regr. -- Breslow method for ties

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(4)     =   35.81
Log likelihood   =  -596.77891          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
     _t
     _d      Coef.  Std. Err.      z    p>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
 prison   1.087282   .5386163    2.02   0.044     .0316135    2.142951
   dose  -.0348039   .0197969   -1.76   0.079    -.0736051    .0039973
clin_pr   -.584771   .4281291   -1.37   0.172    -1.423889    .2543465
clin_do  -.0010622    .014569   -0.07   0.942    -.0296169    .0274925
------------------------------------------------------------------------------
                                                Stratified by clinic
```

Suppose we wish to estimate the hazard ratio for PRISON=1 vs. PRISON=0 for CLINIC=2. This hazard ratio can be estimated by exponentiating the coefficient for prison plus 2 times the coefficient for the clinic-prison interaction term. This expression is obtained by substituting the appropriate values into the hazard in both the numerator (for PRISON=1) and denominator (for PRISON=0) (see below):

$$HR = \frac{h_0(t)\exp[1\beta_1 + \beta_2 DOSE + (2)(1)\beta_3 + \beta_4 CLIN_DO]}{h_0(t)\exp[0\beta_1 + \beta_2 DOSE + (2)(0)\beta_3 + \beta_4 CLIN_DO]}$$
$$= \exp(\beta_1 + 2\beta_3).$$

The **lincom** command can be used to exponentiate linear combinations of parameters. Run this command directly after running the model to estimate the HR for PRISON where CLINIC=2. The code and output follow:

lincom prison+2*clin_pr, hr

```
 ( 1)  prison + 2.0 clin_pr = 0.0
------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   .9210324   .3539571    -0.21   0.831     .4336648    1.956121
------------------------------------------------------------------------------
```

Models can also be run on a subsetted portion of the data using the **if** statement. The following code (with output) runs a Cox model on the data where CLINIC=2:

stcox prison dose if clinic==2

```
Cox regression -- Breslow method for ties

No. of subjects  =          75          Number of obs  =       75
No. of failures  =          28
Time at risk     =       36254
                                         LR chi2(2)     =     9.70
Log likelihood   =  -104.37135          Prob > chi2    =   0.0078

------------------------------------------------------------------------------
     _t
     _d  Haz. Ratio  Std. Err.      z   p>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .9210324   .3539571  -0.21  0.831   .4336648    1.956121
  dose    .9637452   .0118962  -2.99  0.003   .9407088    .9873457
------------------------------------------------------------------------------
```

The hazard ratio estimates for PRISON=1 vs. PRISON=0 (for CLINIC=2) are exactly the same using the stratified Cox approach with product terms and the subsetted data approach (0.9210324).

5. ASSESSING THE PH ASSUMPTION WITH A STATISTICAL TEST

The **stphtest** command can be used to perform a statistical test. A statistical test gives objective criteria for assessing the PH assumption compared to using the graphical approach. This does not mean that this statistical test is better than the graphical approach. It is just more objective. In fact, the graphical approach is generally more informative for descriptively characterizing the form of a PH violation.

The command **stphtest** outputs a PH global test for all the covariates simultaneously and can also be used to obtain a test for each covariate separately with the detail option. To run these tests, you must obtain Schoenfeld residuals for the global test and scaled Schoenfeld residuals for separate tests with each covariate. The idea behind the PH test is that if the PH assumption is satisfied, then the residuals should not be correlated with survival time (or ranked survival time). On the other hand, if the residuals tend to be positive for subjects who become events at a relatively early time and negative for subjects who become events at a relatively late time (or vice versa), then there is evidence that the hazard ratio is not constant over time (i.e., PH assumption is violated).

Before the **stphtest** can be implemented, the **stcox** command needs to be run to obtain the Schoenfeld residuals (with the **schoenfeld()** option) and the scaled Schoenfeld residuals (with the **scaledsch()** option). The names of newly defined variables are in the parentheses: **schoen*** creates SCHOEN1, SCHOEN2, and SCHOEN3 while **scaled*** creates SCALED1, SCALED2, and SCALED3. These variables contain the residuals for PRISON, DOSE, and CLINIC, respectively (the order that the variables were entered in the model). The user is free to type any variable name in the parentheses. The Schoenfeld residuals are used for the global test while the scaled Schoenfeld residuals are used for the testing of the PH assumption for individual variables:

stcox prison dose clinic, schoenfeld(schoen*) scaledsch(scaled*)

Once the residuals are defined, the **stphtest** command can be run. The code and output follow:

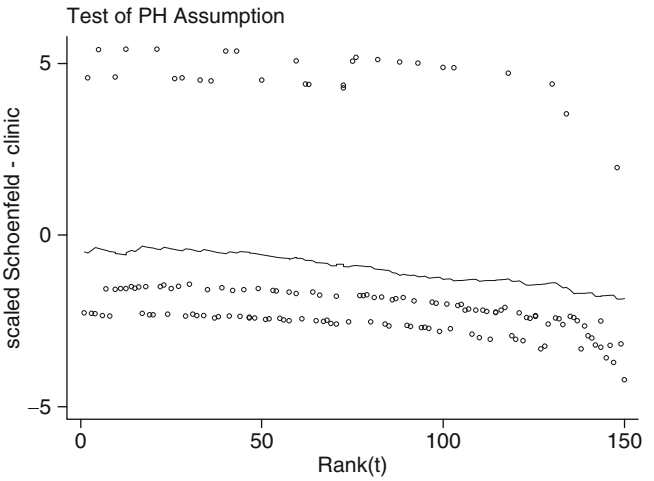
stphtest, rank detail

```
Test of proportional hazards assumption
Time: Rank(t)
------------------------------------------------------------
                    rho   chi2  df  Prob>chi2
------------------------------------------------------------
prison        -0.04645   0.32   1     0.5689
dose           0.08975   1.08   1     0.2996
clinic        -0.24927  10.44   1     0.0012
------------------------------------------------------------
global test             12.36   3     0.0062
------------------------------------------------------------
```

The tests suggest that the PH assumption is violated for CLINIC with the p-value at 0.0012. The tests do not suggest violation of the PH assumption for PRISON or DOSE.

The **plot()** option of the **stphtest** command can be used to produce a plot of the scaled Schoenfeld residuals for CLINIC against survival time ranking. If the PH assumption is met, the fitted curve should look horizontal since the scaled Schoenfeld residuals would be independent of survival time. The code and graph follow:

stphtest, rank plot(clinic)



The fitted curve slopes slightly downward (not horizontal).

6. OBTAINING COX ADJUSTED SURVIVAL CURVES

Adjusted survival curves can be obtained with the **sts graph** command. Adjusted survival curves depend on the pattern of covariates. For example, the adjusted survival estimates for a subject with PRISON=1, CLINIC=1, and DOSE=40 are generally different than for a subject with PRISON=0, CLINIC=2, and DOSE=70. The **sts graph** command produces adjusted baseline survival curves. The following code produces an adjusted survival plot with PRISON=0, CLINIC=0, and DOSE=0 (output omitted):

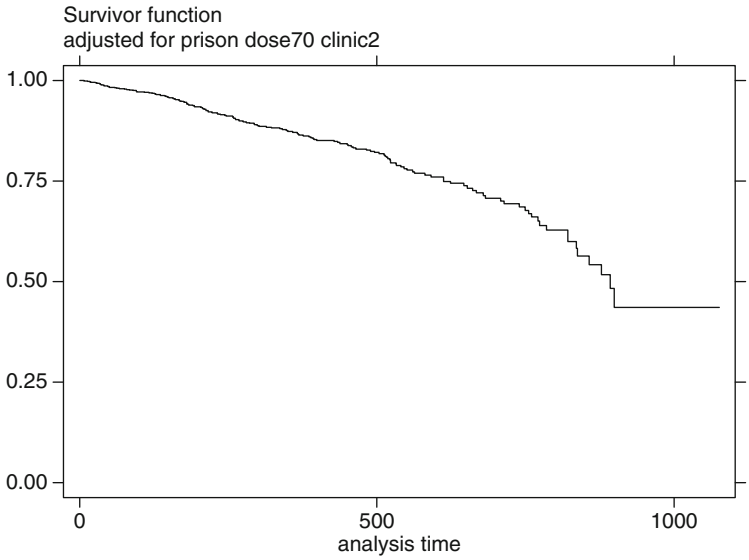
sts graph, adjustfor(prison dose clinic)

It is probably of more interest to create adjusted plots for reasonable patterns of covariates (CLINIC=0 is not even a valid value). Suppose we are interested in graphing the adjusted survival curve for PRISON=0, CLINIC=2, and DOSE=70. We can create new variables with the **generate** command that can be used with the **sts graph** command:

generate clinic2=clinic-2
generate dose70=dose-70

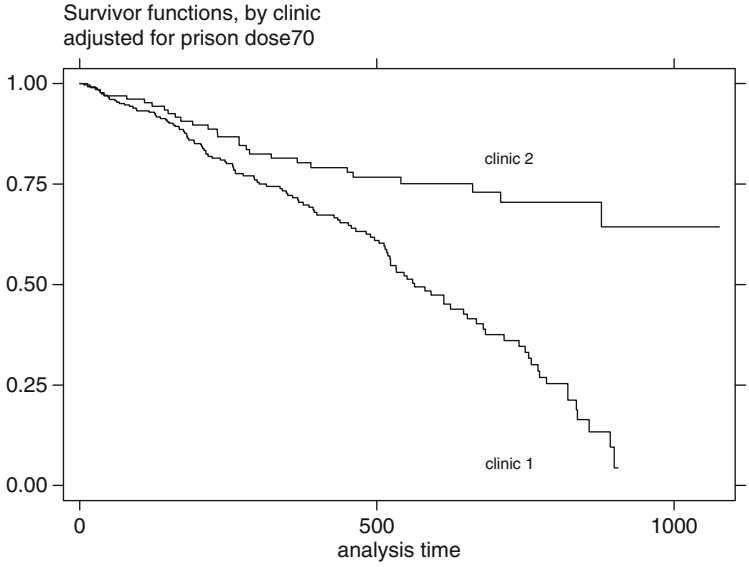
These variables (PRISON, CLINIC2, and DOSE70) produce the desired pattern of covariate when each is set to zero. The following code produces the desired results:

sts graph, adjustfor(prison dose70 clinic2)



Adjusted stratified Cox survival curves can be obtained with the **strata()** option. The following code creates two survival curves stratified by clinic (CLINIC=1, PRISON=0, and DOSE=70) and (CLINIC=2, PRISON=0, and DOSE=70):

sts graph, strata(clinic) adjustfor(prison dose70)



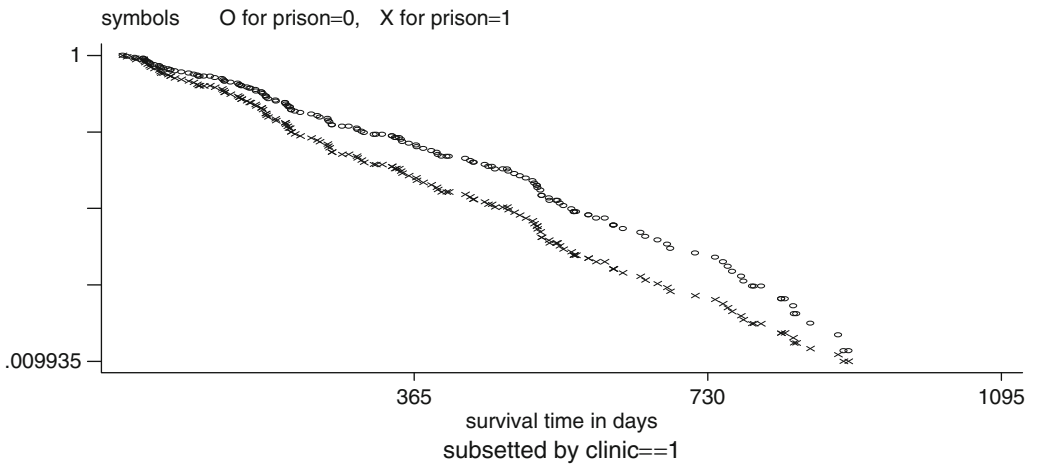
The adjusted curves suggest that there is a strong effect from CLINIC on survival.

Suppose the interest is in comparing adjusted survival plots of PRISON=1 to PRISON=0 stratified by CLINIC. In this setting, the **sts graph** command cannot be used directly since we cannot simultaneously define both levels of prison (PRISON=1 and PRISON=0) as the baseline level (recall **sts graph** plots only the baseline survival function). However, survival estimates can be obtained using the **sts generate** command twice, once where PRISON=0 is defined as baseline and once where PRISON=1 is defined as baseline. The following code creates variables containing the desired adjusted survival estimates:

```
generate prison1=prison-1
sts generate scox0=s, strata(clinic) adjustfor(prison dose70)
sts generate scox1=s, strata(clinic) adjustfor(prison1 dose70)
```

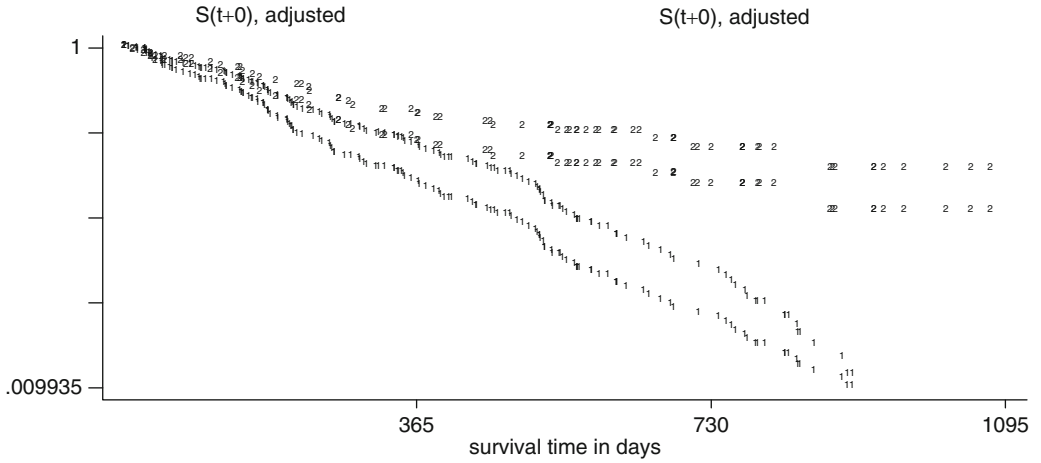
The variables SCOX1 and SCOX0 contain the survival estimates for PRISON=1 and PRISON=0, respectively, adjusting for dose and stratifying by clinic. The **graph** command is used to plot these estimates. If you are using a higher version of Stata than Stata 7.0 (e.g., Stata 8.0), then you should replace the **graph** command with the **graph7** command. The code and output follow:

```
Graph7 scox0 scox1 survt, twoway symbol([clinic] [clinic]) xlabel(365,730,1095)
```



We can also graph PRISON=1 and PRISON=0 subsetting the data where CLINIC=1. The option **twoway** requests a two-way scatter plot. The options **symbol**, **xlabel**, and **title** request the symbols, axis labels, and title, respectively:

graph7 scox0 scox1 survt if clinic==1, twoway symbol(ox) xlabel(365,730,1095) t1(" symbols O for prison=0, X for prison=1") title("subsetted by clinic==1")



7. RUNNING AN EXTENDED COX MODEL

If the PH assumption is not satisfied, a possible strategy is to run a stratified Cox model. Another strategy is to run a Cox model with time-varying covariates (an extended Cox model). The challenge of running an extended Cox model is to choose the appropriate function of survival time to include in the model.

Suppose we want to include a time dependent covariate DOSE times the log of time. This product term could be appropriate if the hazard ratio comparing any two levels of DOSE monotonically increases (or decreases) over time. The **tvc** option( ) of the **stcox** command can be used to declare DOSE a time varying covariate that will be multiplied by a function of time. The specification of that function of time is stated in the **texp** option with the variable **_t** representing time. The code and output for a model containing the time varying covariate, DOSE x ln(_t), follow:

stcox prison clinic dose, tvc(dose) texp(ln(_t)) nohr

```
Cox regression -- Breslow method for ties

No. of subjects  =          238          Number of obs  =     238
No. of failures  =          150
Time at risk     =        95812
                                          LR chi2(4)     =   66.29
Log likelihood   =  -672.51694           Prob > chi2    =  0.0000
------------------------------------------------------------------------------
        _t
        _d      Coef.  Std. Err.      z   p>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
rh
   prison   .3404817   .1674672   2.03  0.042    .012252   .6687113
   clinic -1.018682    .215385  -4.73  0.000  -1.440829  -.5965352
     dose  -.0824307   .0359866  -2.29  0.022  -.1529631  -.0118982
-------------------------------------------------------------------------------
t
     dose   .0085751   .0064554   1.33  0.184  -.0040772   .0212274
-------------------------------------------------------------------------------
note: second equation contains variables that continuously vary
      with respect to time; variables interact with current
      values of ln(_t).
```

The parameter estimate for the time-dependent covariate, DOSE x ln(_t), is 0.0085751; however, it is not statistically significant with a Wald test p-value of 0.184.

A heaviside function can also be used. The following code runs a model with a time-dependent variable equal to CLINIC if time is greater than or equal to 365 days and 0 otherwise.

stcox prison dose clinic, tvc(clinic) texp(_t>=365) nohr

Stata recognizes the expression (_t>=365) as taking the value 1 if survival time is ≥365 days and 0 otherwise. The output follows:

```
Cox regression -- Breslow method for ties

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(4)     =   74.17
Log likelihood   =  -668.57443          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
        _t
        _d     Coef.  Std. Err.      z   p>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
rh
   prison    .377704   .1684024   2.24  0.025   .0476414    .7077666
     dose  -.0355116   .0064354  -5.52  0.000  -.0481247   -.0228985
   clinic  -.4595628   .2552911  -1.80  0.072   -.959924    .0407985
-------------------------------------------------------------------------------
t
   clinic  -1.368665   .4613948  -2.97  0.003  -2.272982    -.464348
-------------------------------------------------------------------------------
note: second equation contains variables that continuously vary
      with respect to time; variables interact with current
      values of _t>=365.
```

Unfortunately, the **texp** option can only be used once in the **stcox** command. This makes it more difficult to run the equivalent model with two heaviside functions. However, it can be accomplished using the **stsplit** command, which adds extra observations to the working dataset. The following code creates a variable called V1 and adds new observations to the dataset:

stsplit v1, at(365)

After the above **stsplit** command is executed, any subject followed more than 365 days is represented by two observations rather than one. For example, the first subject (ID=1) had an event on the 428th day; the first observation for that subject shows no event between 0 and 365 days while the second observation shows an event on the 428th day. The newly defined variable **v1** has the value 365 for observations with survival time exceeding or equal to 365 and 0 otherwise. The following code lists the first ten observations for the requested variables (output follows):

list id _t0 _t _d clinic v1 in 1/10

	id	_t0	_t	_d	clinic	v1
1.	1	0	365	0	1	0
2.	1	365	428	1	1	365
3.	2	0	275	1	1	0
4.	3	0	262	1	1	0
5.	4	0	183	1	1	0
6.	5	0	259	1	1	0
7.	6	0	365	0	1	0
8.	6	365	714	1	1	365
9.	7	0	365	0	1	0
10.	7	365	438	1	1	365

With the data in this form, two heaviside functions can actually be defined in the data using the following code:

```
generate hv2=clinic*(v1/365)
generate hv1=clinic*(1-(v1/365))
```

The following code and output list a sample of the observations (**in 159/167**) with the observation number suppressed (the **noobs** option):

list id _t0 _t clinic v1 hv1 hv2 in 159/167, noobs

id	_t0	_t	clinic	v1	hv1	hv2
100	0	365	1	0	1	0
100	365	749	1	365	0	1
101	0	150	1	0	1	0
102	0	365	1	0	1	0
102	365	465	1	365	0	1
103	0	365	2	0	2	0
103	365	708	2	365	0	2
104	0	365	2	0	2	0
104	365	713	2	365	0	2

With the two heaviside functions defined in the split data, a time dependent model using these functions can be run with the following code (the output follows):

stcox prison dose hv1 hv2, nohr

stcox prison clinic dose hv1 hv2, nohr

```
No. of subjects  =          238          Number of obs  =     360
No. of failures  =          150
Time at risk     =        95812
                                          LR chi2(4)     =   74.17
Log likelihood   =  -668.57443           Prob > chi2    =  0.0000
------------------------------------------------------------------------------
     _t
     _d      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .377704   .1684024    2.24  0.025    .0476414    .7077666
  dose  -.0355116   .0064354   -5.52  0.000   -.0481247   -.0228985
   hv1  -.4595628   .2552911   -1.80  0.072    -.959924    .0407985
   hv2  -1.828228    .385946   -4.74  0.000   -2.584668   -1.071788
```

The **stsplit** command is complicated but it offers a powerful approach for manipulating the data to accommodate time varying analyses.

If you wish to return the data to its previous form, drop the variables that were created from the split and then use the **stjoin** command:

drop v1 hv1 hv2
stjoin

It is possible to split the data at every single failure time, but this uses a large amount of memory. However, if there is only one time varying covariate in the model, the simplest way to run an extended Cox model is by using the **tvc** and **texp** options with the **stcox** command.

One should not confuse an individual's survival time variable (the outcome variable) with the variable used to define the time dependent variable (**_t** in Stata). The individual's survival time variable is a time independent variable. The time of the individual's event (or censorship) does not change. A time-dependent variable, on the other hand, is defined so that it can change its values over time.

8. RUNNING PARAMETRIC MODELS

The Cox PH model is the most widely used model in survival analysis. A key reason why it is so popular is that the distribution of the survival time variable need not be specified. However, if it is believed that survival time follows a particular distribution, then that information can be utilized in a parametric modeling of survival data.

Many parametric models are accelerated failure time (AFT) models. Whereas the key assumption of a PH model is that hazard ratios are constant over time, the key assumption for an AFT model is that survival time accelerates (or decelerates) by a constant factor when comparing different levels of covariates.

The most common distribution for parametric modeling of survival data is the Weibull distribution. The Weibull distribution has the desirable property that if the AFT assumption holds, then the PH assumption also holds. The exponential distribution is a special case of the Weibull distribution. The key property for the exponential distribution is that the hazard is constant over time (not just the hazard ratio). The Weibull and exponential model can be run as a PH model (the default) or an AFT model.

A graphical method for checking the validity of a Weibull assumption is to examine Kaplan-Meier log-log survival curves against log survival time. This is accomplished with the **sts graph** command (see Section 2 of this appendix). If the plots are straight lines, then there is evidence that the distribution of survival times follows a Weibull distribution. If the slope of the line equals one, then the evidence suggests that survival time follows an exponential distribution.

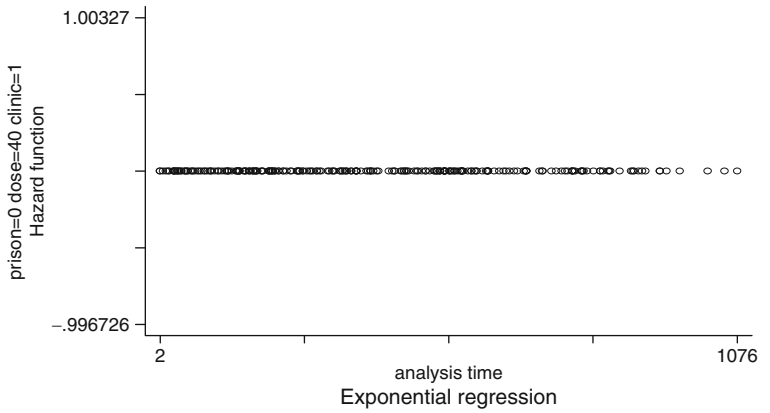
The **streg** command is used to run parametric models. Even though the log log survival curves obtained using the addicts dataset are not straight lines, the data will be used for illustration. First, a parametric model using the exponential distribution will be demonstrated. The code and output follow:

streg prison dose clinic, dist(exponential) nohr

```
Exponential regression -- log relative-hazard form

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   49.91
Log likelihood   =  -270.47929          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .2526491   .1648862     1.53   0.125    -.070522    .5758201
  dose   -.0289167   .0061445    -4.71   0.000   -.0409596   -.0168738
clinic   -.8805819    .210626    -4.18   0.000   -1.293401   -.4677625
 _cons   -3.684341   .4307163    -8.55   0.000   -4.528529   -2.840152
------------------------------------------------------------------------------
```

The distribution is specified with the **dist()** option. The **stcurv** command can be used following the **streg** command to obtain fitted survival, hazard, or cumulative hazard curves. The following code obtains the estimated hazard function for PRISON=0, DOSE=40, and CLINIC=1:

```
stcurv, hazard at (prison=0 dose=40 clinic=1)
```



The graph illustrates the fact that the hazard is constant over time if survival time follows an exponential distribution.

Next, a Weibull distribution is run using the **streg** command:

```
streg prison dose clinic, dist(weibull) nohr
```

```
Weibull regression -- log relative-hazard form

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   60.89
Log likelihood   =  -260.98467          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
          _t      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison    .3144143   .1659462     1.89   0.058    -.0108342    .6396628
  dose   -.0334675    .006255    -5.35   0.000    -.0457272   -.0212079
clinic   -.9715245   .2122826    -4.58   0.000    -1.387591   -.5554582
 _cons   -5.624436   .6588041    -8.54   0.000    -6.915668   -4.333203
-------------------------------------------------------------------------------
 /ln_p    .3149526   .0675583     4.66   0.000     .1825408    .4473644
-------------------------------------------------------------------------------
     p    1.370194    .092568                      1.200263    1.564184
   1/p    .7298235   .0493056                      .6393109    .8331507
------------------------------------------------------------------------------
```

Notice that the Weibull output has a parameter p that the exponential distribution does not have. The hazard function for a Weibull distribution is λpt^{p-1}. If p = 1, then the Weibull distribution is also an exponential distribution (h (t) = λ). Hazard ratio parameters are given by default for the Weibull distribution. If you want the parameterization for an AFT model, then use the **time** option.

The code and output for a Weibull AFT model follow:

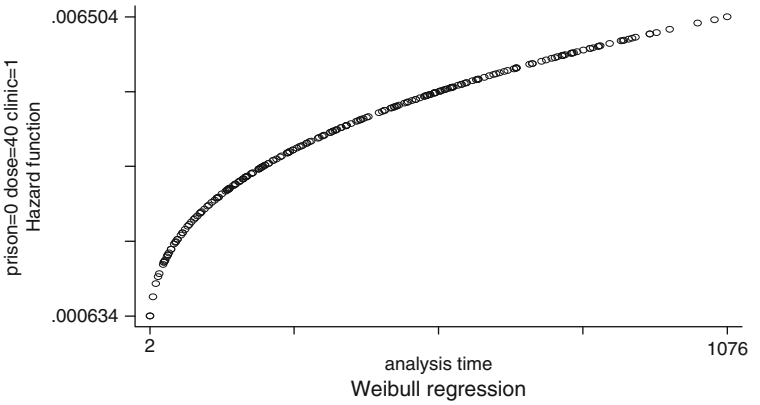
streg prison dose clinic, dist(weibull) time

```
Weibull regression -- accelerated failure-time form

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   60.89
Log likelihood   =  -260.98467          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
prison  -.2294669   .1207889    -1.90   0.057    -.4662088    .0072749
  dose   .0244254   .0045898     5.32   0.000     .0154295    .0334213
clinic   .7090414   .1572246     4.51   0.000     .4008867    1.017196
 _cons   4.104845   .3280583    12.51   0.000     3.461863    4.747828
-------------------------------------------------------------------------------
 /ln_p   .3149526   .0675583     4.66   0.000     .1825408    .4473644
-------------------------------------------------------------------------------
       p   1.370194    .092568                   1.200263    1.564184
     1/p   .7298235   .0493056                   .6393109    .8331507
------------------------------------------------------------------------------
```

The relationship between the hazard ratio parameter β_j and the AFT parameter α_j is $\beta_j = -\alpha_j p$. For example, using the coefficient estimates for PRISON in the Weibull PH and AFT models yields the relationship 0.3144 = (−0.2295)(1.37).

The **stcurv** can again be used following the **streg** command to obtain fitted survival, hazard, or cumulative hazard curves. The following code obtains the estimated hazard function for PRISON=0, DOSE=40, and CLINIC=1:

stcurv, hazard at (prison=0 dose=40 clinic=1)



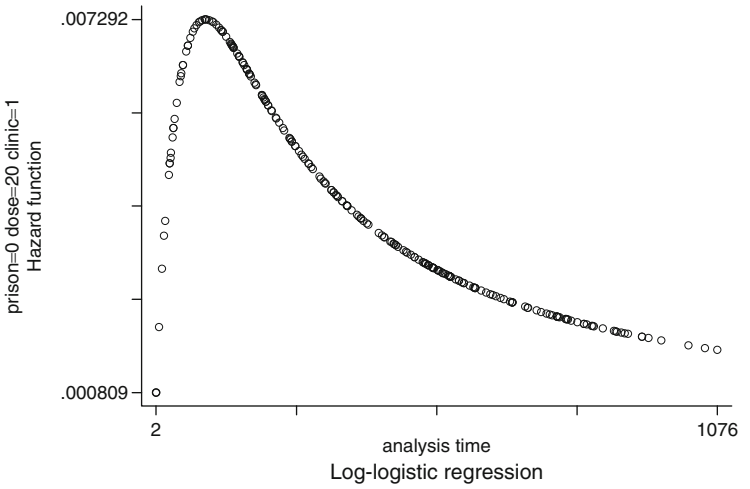
The plot of the hazard is monotonically increasing. With a Weibull distribution, the hazard is constrained such that it cannot increase and then decrease. This is not the case with the log logistic distribution as demonstrated in the next example. The log logistic model is not a PH model, so the default model for the **streg** command is an AFT model. The code and output follow:

streg prison dose clinic, dist(loglogistic)

```
Log-logistic regression -- accelerated failure-time form

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   52.18
Log likelihood   =  -270.42329          Prob > chi2    =  0.0000

------------------------------------------------------------------------------
          _t      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
      prison  -.2912719   .1439646    -2.02   0.043    -.5734373   -.0091065
        dose   .0316133   .0055192     5.73   0.000     .0207959    .0424307
      clinic   .5805977   .1715695     3.38   0.001     .2443276    .9168677
       _cons   3.563268   .3894467     9.15   0.000     2.799967     4.32657
-------------------------------------------------------------------------------
    /ln_gam  -.5331424   .0686297    -7.77   0.000    -.6676541   -.3986306
-------------------------------------------------------------------------------
       gamma   .5867583    .040269                      .5129104    .6712386
------------------------------------------------------------------------------
```

Note that Stata calls the shape parameter gamma for a log-logistic model. The code to produce the graph of the hazard function for PRISON=0, DOSE=40, and CLINIC=1 follows:

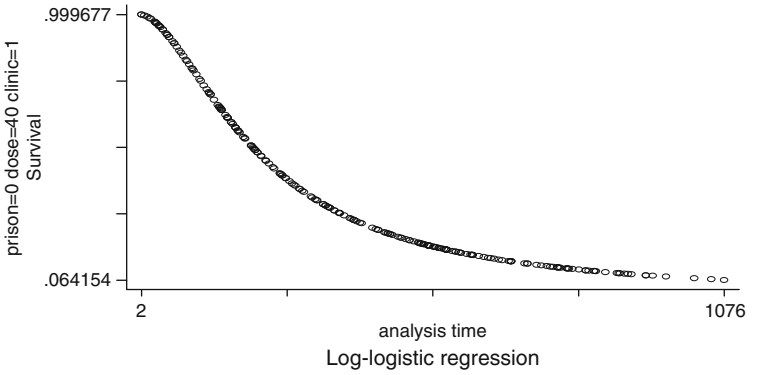
stcurv, hazard at (prison=0 dose=40 clinic=1)



The hazard function (in contrast to the Weibull hazard function) first increases and then decreases.

The corresponding survival curve for the log logistic distribution can also be obtained with the **stcurve** command:

stcurv, survival at (prison=0 dose=40 clinic=1)



If the AFT assumption holds for a log logistic model, then the proportional odds assumption holds for the survival function (although the PH assumption would not hold). The proportional odds assumption can be evaluated by plotting of the log odds of survival (using KM estimates) against the log of survival time. If the plots are straight lines for each pattern of covariates, then the log-logistic distribution is reasonable. If the straight lines are also parallel, then the proportional odds and AFT assumptions also hold. The following code will plot the estimated log odds of survival against the log of time by CLINIC (output omitted):

```
sts generate skm=s, by(clinic)
generate logodds=ln(skm/(1-skm))
generate logt=ln(survt)
graph7 logodds logt, twoway symbol([clinic] [clinic])
```

Another context for thinking about the proportional odds assumption is that the odds ratio estimated by a logistic regression does not depend on the length of the follow-up. For example, if a follow-up study was extended from 3 to 5 years, then the underlying odds ratio comparing two patterns of covariates would not change. If the proportional odds assumption is not true, then the odds ratio is specific to the length of follow-up.

Both the log-logistic and Weibull models contain an extra shape parameter that is typically assumed constant. This assumption is necessary for the PH or AFT assumption to hold for these models. Stata provides a way of modeling the shape parameter as a function of predictor variables by use of the **ancillary** option in the **streg** command (see Chapter 7 under the heading "Other Parametric Models"). The following code runs a log-logistic model in which the shape parameter gamma is modeled as a function of CLINIC while λ is modeled as a function of PRISON and DOSE:

```
streg prison dose, dist(loglogistic) ancillary(clinic)
```

The output follows:

```
Log-logistic regression -- accelerated failure-time form

No. of subjects  =         238                 Number of obs   =       238
No. of failures  =         150
Time at risk     =       95812
                                               LR chi2(2)      =     38.87
Log likelihood   =  -272.65273                 Prob > chi2     =    0.0000

------------------------------------------------------------------------------
          _t      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
_t
      prison  -.3275695   .1405119    -2.33  0.020  -.6029677  -.0521713
        dose   .0328517   .0054275     6.05  0.000    .022214   .0434893
       _cons   4.183173   .3311064    12.63  0.000   3.534216    4.83213
-------------------------------------------------------------------------------
ln_gam
      clinic   .4558089   .1734819     2.63  0.009   .1157906   .7958273
       _cons  -1.094496   .2212143    -4.95  0.000  -1.528068  -.6609238
------------------------------------------------------------------------------
```

Notice there is a parameter estimate for CLINIC as well as an intercept (_cons) under the heading ln_gam (the log of gamma). With this model, the estimate for gamma depends on whether CLINIC=1 or CLINIC=2. There is no easy interpretation for the predictor variables in this type of model, which is why it is not commonly used. However, for any specified value of PRISON, DOSE, and CLINIC, the hazard and survival functions can be estimated by substituting the parameter estimates into the expressions for the log-logistic hazard and survival functions.

Other distributions supported by **streg** are the generalized gamma, the lognormal, and the Gompertz distributions.

9. RUNNING FRAILTY MODELS

Frailty models contain an extra random component designed to account for individual-level differences in the hazard otherwise unaccounted for by the model. The frailty, α, is a multiplicative effect on the hazard assumed to follow some distribution. The hazard function conditional on the frailty can be expressed as $h(t|\alpha) = \alpha[h(t)]$.

Stata offers two choices for the distribution of the frailty: the gamma and the inverse-Gaussian, both of mean 1 and variance theta. The variance (theta) is a parameter estimated by the model. If theta = 0, then there is no frailty.

For the first example, a Weibull PH model is run with PRISON, DOSE, and CLINIC as predictors. A gamma distribution is assumed for the frailty component. The models in this section were run using Stata 8.0. The code follows:

```
streg dose prison clinic, dist(weibull) frailty(gamma) nohr
```

The **frailty()** option requests that a frailty model be run. The output follows:

```
Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(4)     =   60.89
Log likelihood   =  -260.98467          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
          _t      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
        dose  -.0334635   .0062553  -5.35  0.000  -.0457237  -.0212034
      prison   .3143786    .165953   1.89  0.058  -.0108833   .6396405
      clinic  -.9714998   .2122909  -4.58  0.000  -1.387582  -.5554173
       _cons  -5.624342   .6588994  -8.54  0.000  -6.915761  -4.332923
-------------------------------------------------------------------------------
       /ln_p   .3149036   .0675772   4.66  0.000   .1824548   .4473525
  /ln_the  -15.37947   722.4246  -0.02  0.983  -1431.306   1400.547
-------------------------------------------------------------------------------
         p   1.370127   .0925893                 1.20016   1.564166
       1/p   .7298592   .0493218                .6393185   .8332223
     theta   2.09e-07   .0001512                       0          .
------------------------------------------------------------------------------
Likelihood ratio test of theta = 0: chibar2(01) = 0.00
  Prob>=chibar2 = 1.000
```

Notice that there is one additional parameter (theta) compared to the model run in the previous section. The estimate for theta is 2.09 times 10^{-7} or 0.000000209 which is essentially zero. A likelihood ratio test for the inclusion of theta is provided at the bottom of the output and yields a chi-square value of 0.00 and a p-value of 1.000. The frailty has no effect on the model and need not be included.

The next model will be the same as the previous except that CLINIC will not be included. One might expect a frailty component to play a larger role if an important covariate, such as CLINIC, is not included in the model. The code and output follow:

streg dose prison, dist(weibull) frailty(gamma) nohr

```
Weibull regression -- log relative-hazard form
                      Gamma frailty

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   36.00
Log likelihood   =  -273.42782          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
          _t      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
       dose  -.0358231    .010734  -3.34  0.001  -.0568614  -.0147849
     prison   .2234556   .2141028   1.04  0.297  -.1961783   .6430894
      _cons  -6.457393   .6558594  -9.85  0.000  -7.742854  -5.171932
-------------------------------------------------------------------------------
      /ln_p   .2922832   .1217597   2.40  0.016   .0536385   .5309278
  /ln_the  -2.849726   5.880123  -0.48  0.628  -14.37456   8.675104
-------------------------------------------------------------------------------
           p   1.339482    .163095                 1.055103   1.700509
         1/p   .7465571   .0909006                 .5880591   .9477747
       theta   .0578602    .340225                 5.72e-07    5855.31
------------------------------------------------------------------------------
Likelihood ratio test of theta = 0: chibar2(01) = 0.03
  Prob>=chibar2 = 0.432
```

The variance (theta) of the frailty is estimated at 0.0578602. Although this estimate is not exactly zero as in the previous example, the p-value for the likelihood ratio test for theta is nonsignificant at 0.432. So the addition of frailty did not account for CLINIC being omitted from the model.

Next, the same model is run except that the inverse-Gaussian distribution is used for the frailty rather than the gamma distribution. The code and output follow:

streg dose prison, dist(weibull) frailty(invgaussian) nohr

```
Weibull regression -- log relative-hazard form
                      Inverse-Gaussian frailty

No. of subjects  =         238          Number of obs  =     238
No. of failures  =         150
Time at risk     =       95812
                                        LR chi2(3)     =   35.99
Log likelihood   =  -273.43201          Prob > chi2    =  0.0000
------------------------------------------------------------------------------
          _t      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
        dose  -.0353908   .0096247  -3.68  0.000  -.0542549  -.0165268
      prison   .2166456   .1988761   1.09  0.276  -.1731445   .6064356
       _cons  -6.448779   .6494397  -9.93  0.000  -7.721658  -5.175901
-------------------------------------------------------------------------------
       /ln_p   .2875567   .1122988   2.56  0.010   .0674551   .5076583
     /ln_the  -3.137696   7.347349  -0.43  0.669  -17.53824   11.26284
-------------------------------------------------------------------------------
           p   1.333166   .1497129                1.069782   1.661396
         1/p   .7500941   .0842347                .6019034   .9347697
       theta   .0433827   .3187475                2.42e-08   77873.78
-------------------------------------------------------------------------------
Likelihood ratio test of theta = 0: chibar2(01) = 0.02
  Prob>=chibar2 = 0.443
```

The p-value for the likelihood ratio test for theta is 0.443 (at the bottom of the output). The results in this example are very similar whether assuming the inverse-Gaussian or the gamma distribution for the frailty component.

An example of shared frailty applied to recurrent event data is shown in the next section.

10. MODELING RECURRENT EVENTS

The modeling of recurrent events is illustrated with the bladder cancer dataset (**bladder.dta**) described at the start of this appendix. Recurrent events are represented in the data with multiple observations for subjects having multiple events. The data layout for the bladder cancer dataset is suitable for a counting process approach with time intervals defined for each observation (see Chapter 8). The following code prints the 12th–20th observation, which contains information for four subjects. The code and output follow:

list in 12/20

	id	event	interval	start	stop	tx	num	size
12.	10	1	1	0	12	0	1	1
13.	10	1	2	12	16	0	1	1
14.	10	0	3	16	18	0	1	1
15.	11	0	1	0	23	0	3	3
16.	12	1	1	0	10	0	1	3
17.	12	1	2	10	15	0	1	3
18.	12	0	3	15	23	0	1	3
19.	13	1	1	0	3	0	1	1
20.	13	1	2	3	16	0	1	1

There are three observations for ID=10, one observation for ID=11, three observations for ID=12, and two observations for ID=13. The variables START and STOP represent the time interval for the risk period specific to that observation. The variable EVENT indicates whether an event (coded 1) occurred. The first three observations indicate that the subject with ID=10 had an event at 12 months, another event at 16 months, and was censored at 18 months.

Before using Stata's survival commands, the **stset** command must be used to define the key survival variables. The code follows:

stset stop, failure(event==1) id(id) time0(start) exit(time.)

We have previously used the **stset** command on the "addicts" dataset, but more options from **stset** are included here. The **id()** option defines the subject variable (i.e., the cluster variable), the **time0()** option defines the variable that begins the time interval, and the **exit(time .)** option tells Stata that there is no imposed limit on the length of follow-up time for a given subject (e.g., subjects are not out of the risk set after their first event). With the **stset** command, Stata creates the variables **_t0**, **_t**, and **_d**, which Stata automatically recognizes as survival variables representing the time interval and event status. Actually, the **time0()** option could have been omitted from this **stset** command and by default Stata would have created the starting time variable, **_t0**, in the correct counting process format as long as the **id()** option was used (otherwise **_t0** would default to zero). The following code (and output) lists the 12^{th}–20^{th} observation with the newly created variables:

list id _t0 _t _d tx in 12/20

	id	_t0	_t	_d	tx
12.	10	0	12	1	0
13.	10	12	16	1	0
14.	10	16	18	0	0
15.	11	0	23	0	0
16.	12	0	10	1	0
17.	12	10	15	1	0
18.	12	15	23	0	0
19.	13	0	3	1	0
20.	13	3	16	1	0

A Cox model with recurrent events using the counting process approach can now be run with the **stcox** command. The predictors are treatment status (TX), initial number of tumors (NUM), and the initial size of tumors (SIZE). The **robust** option requests robust standard errors for the coefficient estimates. Omit the **nohr** option if you want the exponentiated coefficients. The code and output follow:

stcox tx num size, nohr robust

```
Cox regression -- Breslow method for ties

No. of subjects  =          85        Number of obs   =     190
No. of failures  =         112
Time at risk     =        2711
                                      Wald chi2(3)    =   11.25
Log likelihood   =  -460.07958        Prob > chi2     =  0.0105

                (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
  _t              Robust
  _d      Coef.  Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
  tx  -.4070966   .2432658    -1.67   0.094    -.8838889    .0696956
 num   .1606478   .0572305     2.81   0.005     .0484781    .2728174
size  -.0400877   .0726459    -0.55   0.581     -.182471    .1022957
------------------------------------------------------------------------------
```

The interpretation of these parameter estimates is discussed in Chapter 8

A stratified Cox model can also be run using the data in this format with the variable INTERVAL as the stratified variable. The stratified variable indicates whether subjects were at risk for their 1st, 2nd, 3rd, or 4th event. This approach is called a Stratified CP approach in Chap. 8 and is used if the investigator wants to distinguish the order in which recurrent events occur. The code and output follow:

stcox tx num size, nohr robust strata(interval)

```
stratified Cox regr. -- Breslow method for ties

No. of subjects  =          85        Number of obs  =     190
No. of failures  =         112
Time at risk     =        2711
                                      Wald chi2(3)   =    7.11
Log likelihood   =  -319.85912        Prob > chi2    =  0.0685

                (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
  _t             Robust
  _d      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
  tx  -.3342955   .1982339   -1.69  0.092  -.7228268    .0542359
 num   .1156526   .0502089    2.30  0.021    .017245    .2140603
size  -.0080508   .0604807   -0.13  0.894  -.1265908    .1104892
------------------------------------------------------------------------------
                                                Stratified by interval
```

Interaction terms between the treatment variable (TX) and the stratified variable could be created to examine whether the effect of treatment differed for the 1st, 2nd, 3rd, or 4th event. (Note that in this dataset, subjects have a maximum of 4 events).

Another stratified approach (called Gap Time) is a slight variation of the Stratified CP approach. The difference is in the way the time intervals for the recurrent events are defined. There is no difference in the time intervals when subjects are at risk for their first event. However, with the Gap Time approach, the starting time at risk gets reset to zero for each subsequent event. The following code creates data suitable for running a Gap Time recurrent event model.

```
generate stop2 =_t - _t0
stset stop2, failure(event==1) exit(time .)
```

The **generate** command defines a new variable called STOP2 representing the length of the time interval for each observation. The **stset** command is used with STOP2 as the outcome variable (**_t**). By default, Stata sets the variable **_t0** to zero. The following code (and output) lists the 12th through 20th observations for selected variables.

```
list id _t0 _t _d tx in 12/20
```

```
        id      _t0      _t      _d      tx
12.     10        0      12       1       0
13.     10        0       4       1       0
14.     10        0       2       0       0
15.     11        0      23       0       0
16.     12        0      10       1       0
17.     12        0       5       1       0
18.     12        0       8       0       0
19.     13        0       3       1       0
20.     13        0      13       1       0
```

Notice that the **id()** option was not used with the **stset** command for the Gap Time approach. This means that Stata does not know that multiple observations correspond to the same subject. However, the **cluster()** option can be used directly in the **stcox** command to request that the analysis be clustered by ID (i.e., by subject). The following code runs a stratified Cox model using the Gap Time approach with the **cluster()** and **robust** options. The code and output follow:

```
stcox tx num size, nohr robust strata(interval) cluster(id)
```

```
Stratified Cox regr. -- Breslow method for ties

No. of subjects  =         190        Number of obs  =     190
No. of failures  =         112
Time at risk     =        2711
                                      Wald chi2(3)   =   11.99
Log likelihood   =  -363.16022        Prob > chi2    =  0.0074
              (standard errors adjusted for clustering on id)
------------------------------------------------------------------------------
  _t               Robust
  _d      Coef.  Std. Err.      z    P>|z|  [95% Conf. Interval]
-------------------------------------------------------------------------------
  tx  -.2695213   .2093108   -1.29  0.198  -.6797628    .1407203
 num   .1535334   .0491803    3.12  0.002   .0571418    .2499249
size   .0068402   .0625862    0.11  0.913  -.1158265     .129507
-------------------------------------------------------------------------------
                                          Stratified by interval
```

The results using the Gap Time approach vary slightly from that obtained using the Stratified CP approach.

Next, we demonstrate how a shared frailty model can be applied to recurrent event data. Frailty is included in recurrent event analyses to account for variability due to unobserved subject-specific factors that may lead to within-subject correlation.

Before running the model, we rerun the **stset** command shown earlier in this section to get the data back to the form suitable for a counting process approach. The code follows:

```
stset stop, failure(event==1) id(id) time0(start) exit(time .)
```

Next a parametric Weibull model is run with a gamma-distributed shared frailty component using the **streg** command. We use the same three predictors for comparability with the other models presented in this section. The code follows:

```
streg tx num size, dist(weibull) frailty(gamma) shared(id) nohr
```

The **dist()** option requests the distribution for the parametric model. The **frailty()** option requests the distribution for the frailty and the **shared()** option defines the cluster variable, ID. For this model, observations from the same subject share the same frailty. The output follows:

```
Weibull regression -- log relative-hazard form
                      Gamma shared frailty

No. of subjects  =          85           Number of obs   =     190
No. of failures  =         112
Time at risk     =        2711
                                         LR chi2(3)      =    8.04
Log likelihood   =  -184.73658           Prob > chi2     =  0.0453

------------------------------------------------------------------------------
          _t      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------------------------------------------------------------------------
          tx   -.4583219   .2677275    -1.71   0.087    -.9830582    .0664143
         num    .1847305   .0724134     2.55   0.011     .0428028    .3266581
        size   -.0314314   .0911134    -0.34   0.730    -.2100104    .1471476
       _cons   -2.952397   .4174276    -7.07   0.000     -3.77054   -2.134254
-------------------------------------------------------------------------------
      /ln_p   -.1193215   .0898301    -1.33   0.184    -.2953852    .0567421
    /ln_the   -.7252604   .5163027    -1.40   0.160    -1.737195    .2866742
-------------------------------------------------------------------------------
           p    .8875224   .0797262                      .7442449    1.058383
         1/p    1.126732   .1012144                      .9448377    1.343644
       theta    .4841985    .249993                      .1760134     1.33199
------------------------------------------------------------------------------
Likelihood ratio test of theta=0: chibar2(01) = 7.34
  Prob>=chibar2 = 0.003
```

The model output is discussed in Chapter 8.

The counting process data layout with multiple observations per subject need not only apply to recurrent event data, but can also be used for a more conventional survival analyses in which each subject is limited to one event. A subject with four observations may be censored for the first three observations before getting the event in the time interval represented by the fourth observation. This data layout is particularly suitable for representing time-varying exposures, which may change values over different intervals of time (see the **stsplit** command in Section 7 of this appendix).

B. SAS

Analyses are carried out in SAS by using the appropriate SAS procedure on a SAS dataset. The key SAS procedures for performing survival analyses are:

PROC LIFETEST – This procedure is used to obtain Kaplan-Meier survival estimates and plots. It can also be used to output life table estimates and plots. It will generate output for the log rank and Wilcoxon test statistics if stratifying by a covariate. A new SAS dataset containing survival estimates can be requested.

PROC PHREG – This procedure is used to run the Cox proportional hazards model, a stratified Cox model, and an extended Cox model with time-varying covariates. It can also be used to create a SAS dataset containing adjusted survival estimates. These adjusted survival estimates can then be plotted using PROC GPLOT.

PROC LIFEREG – This procedure is used to run parametric accelerated failure time AFT models.

Analyses on the "addicts" dataset will be used to illustrate these procedures. The "addicts" dataset was obtained from a 1991 Australian study by Caplehorn et al. and contains information on 238 heroin addicts. The study compared two methadone treatment clinics to assess patient time remaining under methadone treatment. The two clinics differed according to its live-in policies for patients. A patient's survival time was determined as the time (in days) until the person dropped out of the clinic or was censored. The variables are defined at the start of this appendix.

All of the SAS programming code will be written in capital letters for readability. However, SAS is *not* case sensitive. If a program is written with lower-case letters, SAS reads them as upper case. The number of spaces between words (if more than one) has no effect on the program. Each SAS programming statement ends with a semicolon.

The addicts dataset is stored as a permanent SAS dataset called **addicts.sas7bdat**. A LIBNAME statement is needed to indicate the path to the location of the SAS dataset. In our examples, we assume the file is located on the C drive. The LIBNAME statement includes a reference name as well as the path. We call the reference name REF. The code is as follows:

```
LIBNAME REF 'C:\';
```

The user is free to define his/her own reference name. The path to the location of the file is given between the quotation marks. The general form of the code is:

```
LIBNAME Your reference name 'Your path to file location';
```

PROC CONTENTS, PROC PRINT, PROC UNIVARIATE, PROC FREQ, and PROC MEANS can be used to list or describe the data. SAS code can be run in one batch or highlighted and submitted one procedure at a time. Code can be submitted by clicking on the submit button on the toolbar in the Editor window. The code for using these procedures follows (output omitted):

```
PROC CONTENTS DATA=REF.ADDICTS;RUN;
PROC PRINT DATA=REF.ADDICTS;RUN;
PROC UNIVARIATE DATA=REF.ADDICTS;VAR SURVT;RUN;
PROC FREQ DATA=REF.ADDICTS;TABLES CLINIC PRISON;RUN;
PROC MEANS DATA=REF.ADDICTS;VAR SURVT;CLAS CLINIC;RUN;
```

Notice that each SAS statement ends with a semicolon. If each procedure is submitted one at a time, then each procedure must end with a RUN statement. Otherwise one RUN statement at the end of the last procedure is sufficient. With the LIBNAME statement, SAS recognizes a two-level file name: the reference name and the file name without an extension. For our example, the SAS file name is REF.ADDISTS. Alternatively, a temporary SAS dataset could be created and used for these procedures.

Text that you do not wish SAS to process can be written as a comment:

```
/* A comment begins with a forward slash followed by a
   star and ends with a star followed by a forward slash. */
* A comment can also be created by beginning with a star
  and ending with a semicolon;
```

The survival analyses demonstrated in SAS are as follows:

1. Demonstrating PROC LIFETEST to obtain Kaplan-Meier and life table survival estimates (and plots).
2. Running a Cox PH model with PROC PHREG.
3. Running a stratified Cox model.
4. Assessing the PH assumption with a statistical test.
5. Obtaining Cox adjusted survival curves.
6. Running an extended Cox model (i.e., containing time varying covariates).
7. Running parametric models with PROC LIFEREG.
8. Modeling recurrent events

1. DEMONSTRATING PROC LIFETEST TO OBTAIN KM AND LIFE TABLE SURVIVAL ESTIMATES (AND PLOTS)

PROC LIFETEST produces Kaplan-Meier survival estimates with the METHOD=KM option. The PLOTS=(S) option plots the estimated survival function. The TIME statement defines the time-to-event variable (SURVT) and the value for censorship (STATUS=0). The code follows (output omitted):

```
PROC LIFETEST DATA=REF.ADDICTS METHOD=KM PLOTS=(S);
TIME SURVT*STATUS(0);
RUN;
```

Use a STRATA statement in PROC LIFETEST to compare survival estimates for different groups (e.g., strata clinic). The PLOTS=(S, LLS) option produces log-log curves as well as survival curves. If the PH assumption is met, the log-log survival curves will be parallel. The STRATA statement also provides the log rank test and Wilcoxon test statistics. The code follows:

```
PROC LIFETEST DATA=REF.ADDICTS METHOD=KM PLOTS=(S,LLS);
TIME SURVT*STATUS(0);
STRATA CLINIC;
RUN;
```

PROC LIFETEST yields the following edited output:

The LIFETEST Procedure (stratified)

Stratum 1: CLINIC = 1

Product-Limit Survival Estimates

SURVT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00	1.0000	0	0	0	163
2.00*	.	.	.	0	162
7.00	0.9938	0.00617	0.00615	1	161
17.00	0.9877	0.0123	0.00868	2	160
.	.	.	.	.	.
.	.	.	.	.	.
836.00	0.0869	0.9131	0.0295	118	6
837.00	0.0725	0.9275	0.0279	119	5

Stratum 2: CLINIC = 2

Product-Limit Survival Estimates

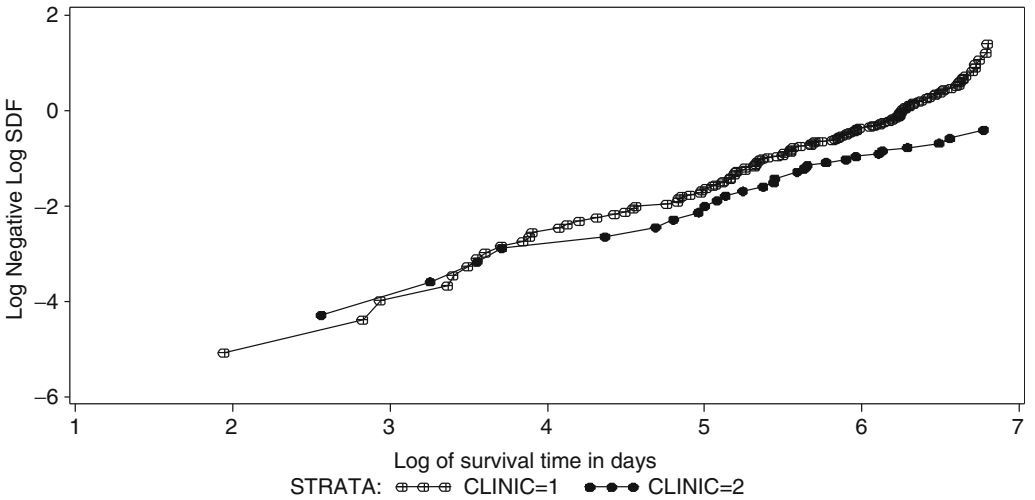
SURVT	Survival	Failure	Survival Standard Error	Number Failed	Number Left
0.00	1.0000	0	0	0	75
2.00*	.	.	.	0	74
13.00	0.9865	0.0135	0.0134	1	73
26.00	0.9730	0.0270	0.0189	2	72
.	.	.	.	.	.
.	.	.	.	.	.

Test of Equality over Strata

Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	27.8927	1	<.0001
Wilcoxon	11.6268	1	0.0007
-2Log(LR)	26.0236	1	<.0001

Both the log rank and Wilcoxon test yield highly significant chi-square test statistics. The Wilcoxon test is a variation of the log rank test weighting the observed minus expected score of the j^{th} failure time by n_j (the number still at risk at the j^{th} failure time).

The requested log-log plots from PROC LIFETEST follow:



SAS (as well as Stata and R) plots log(survival time) rather than survival time on the horizontal axis by default for log-log curves. As far as checking the parallel assumption, it does not matter if log(survival time) or survival time is on the horizontal axis. However, if the log-log survival curves look like straight lines with log(survival time) on the horizontal axis, then there is evidence that the "time-to-event" variable follows a Weibull distribution. If the slope of the line equals one, then there is evidence that the survival time variable follows an exponential distribution – a special case of the Weibull distribution. For these situations, a parametric survival model can be used.

You can gain more control over how variables are plotted, by creating a dataset that contains the survival estimates. Use the OUTSURV= option in the PROC LIFETEST statement to create a SAS data containing the KM survival estimates. The option OUTSURV=DOG creates a dataset called dog (make up your own name) containing the survival estimates in a variable called SURVIVAL. The code follows:

```
PROC LIFETEST DATA=REF.ADDICTS METHOD=KM OUTSURV=DOG;
TIME SURVT*STATUS(0);
STRATA CLINIC;
RUN;
```

Data dog contains the survival estimates but not the log(-(log)) of the survival estimates. Data cat is created in the following code from data dog (using the statement SET DOG) and defines a new log-log variable called LLS.

```
DATA CAT;
SET DOG;
LLS=LOG(-LOG(SURVIVAL));
RUN;
```

In SAS, the LOG function returns the natural log, not the log base 10.

PROC PRINT prints the data in the output window.

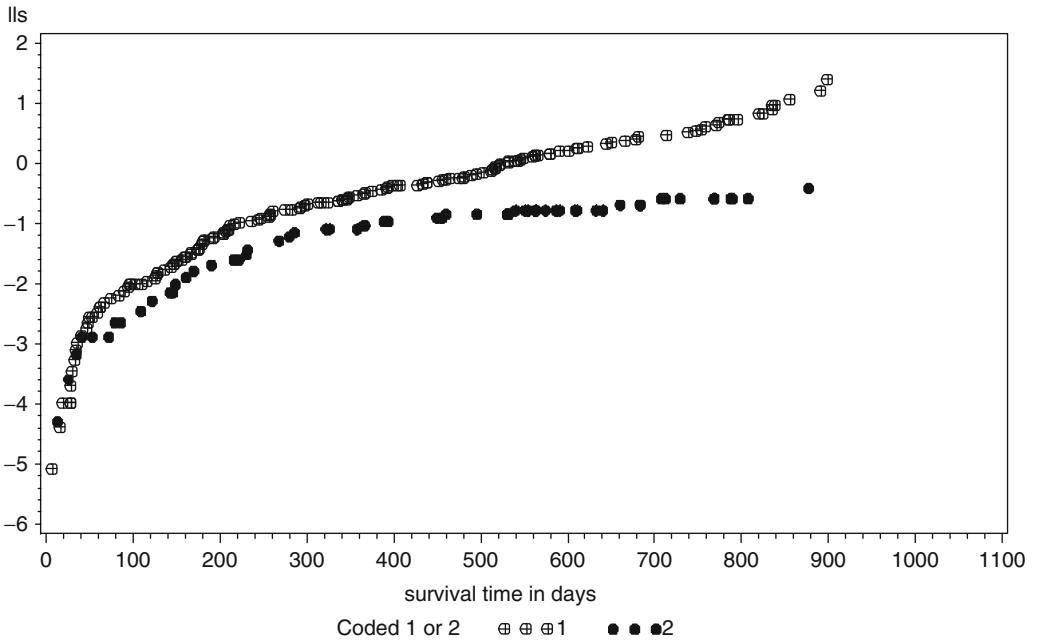
```
PROC PRINT DATA=CAT; RUN;
```

The first 10 observations from PROC PRINT are listed below:

Obs	CLINIC	SURVT	_CENSOR_	SURVIVAL	LLS
1	1	0	0	1.00000	.
2	1	2	1	1.00000	.
3	1	7	0	0.99383	-5.08450
4	1	17	0	0.98765	-4.38824
5	1	19	0	0.98148	-3.97965
6	1	28	1	0.98148	-3.97965
7	1	28	1	0.98148	-3.97965
8	1	29	0	0.97523	-3.68561
9	1	30	0	0.96898	-3.45736
10	1	33	0	0.96273	-3.27056

The PLOT LLS*SURVT=CLINIC statement puts the variable LLS (the log-log survival variables) on the vertical axis and SURVT on the horizontal axis, stratified by CLINIC. The SYMBOL option can be used to choose plotting colors for each level of clinic. The code and output for plotting the log log curves by CLINIC follow:

```
SYMBOL COLOR=BLUE;
SYMBOL2 COLOR=RED;

PROC GPLOT DATA=CAT;
PLOT LLS*SURVT=CLINIC;
RUN;
```



The plot has survival time (in days) rather than the default log(survival time). The log-log survival plots look parallel for CLINIC the first 365 days but then seem to diverge. This information can be utilized when developing an approach for modeling CLINIC with a time dependent variable in an extended Cox model.

You can also obtain survival estimates using life tables. This method is useful if you do not have individual level survival information but rather have group survival information for specified time intervals. The user determines the time intervals using the INTERVALS= option. The code follows (output omitted):

```
PROC LIFETEST DATA=REF.ADDICTS
  METHOD=LT INTERVALS= 50 100 150
  200 TO 1000 BY 100 PLOTS=(S);
TIME SURVT*STATUS(0);
RUN;
```

2. RUNNING A COX PROPORTIONAL HAZARD MODEL WITH PROC PHREG

PROC PHREG is used to request a Cox proportional hazards model. The code follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE CLINIC;
RUN;
```

The code SURVT*STATUS(0), in the MODEL statement specifies the time-to-event variable (SURVT) and the value for censorship (STATUS=0). Three predictors are included in the model: PRISON, DOSE, and CLINIC. The option RL in the MODEL statement of PROC PHREG provides 95% confidence intervals for the hazard ratio estimates. The PH assumption is assumed to follow for each of these predictors (perhaps incorrectly). The output produced by PROC PHREG follows:

The PHREG Procedure
Model Information

Data Set	REF.ADDICTS	
Dependent Variable	SURVT	survival time in days
Censoring Variable	STATUS	status (0=censored, 1=endpoint)
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
238	150	88	36.97

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1411.324	1346.805
AIC	1411.324	1352.805
SBC	1411.324	1361.837

Analysis of Maximum Likelihood Estimates

Parameter	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PRISON	1	0.32650	0.16722	3.8123	0.0509	1.386	0.999	1.924
DOSE	1	-0.03540	0.00638	30.7844	<.0001	0.965	0.953	0.977
CLINIC	1	-1.00876	0.21486	22.0419	<.0001	0.365	0.239	0.556

The table above lists the parameter estimates for the regression coefficients, their standard errors, a Wald chi-square test statistic for each predictor, and corresponding p-value. The column labeled HAZARD RATIO gives the estimated hazard ratio per one-unit change in each predictor by exponentiating the estimated regression coefficients. The final two columns give the 95% confidence limits for this hazard ratio.

You can use the TIES=EXACT option in the model statement rather than run the default TIES=BRESLOW option that was used in the previous model. The TIES=EXACT option is a computationally intensive method to handle events that occur at the same time. If many events

occur simultaneously in the data, then the TIES=EXACT option is preferred. Otherwise, the difference between this option and the default is slight. The TIES=EFRON option is another tie-handling approach that SAS offers. The TIES=EFRON is the default method used in R.

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE CLINIC/TIES=EXACT RL;
RUN;
```

The output follows:

The PHREG Procedure

Model Information

Data Set	REF.ADDICTS	
Dependent Variable	SURVT	survival time in days
Censoring Variable	STATUS	status (0=censored, 1=endpoint)
Censoring Value(s)	0	
Ties Handling	EXACT	

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence	Limits
PRISON	1	0.32657	0.16723	3.8135	0.0508	1.386	0.999	1.924
DOSE	1	-0.03537	0.00638	30.7432	<.0001	0.965	0.953	0.977
CLINIC	1	-1.00980	0.21488	22.0832	<.0001	0.364	0.239	0.555

The parameter estimates and their standard errors vary only slightly from the previous model without the TIES=EXACT option. Notice that the type of ties-handling approach is listed in the table called MODEL INFORMATION in the output.

Suppose we wish to assess interaction between PRISON and CLINIC and between PRISON and DOSE. We can define two interaction terms in a new temporary SAS dataset (called addicts2) and then run a model containing those terms. Product terms for CLINIC times PRISON (called CLIN_PR) and CLINIC time DOSE (called CLIN_DO) are defined in the following data step:

```
DATA ADDICTS2;
SET REF.ADDICTS;
CLIN_PR=CLINIC*PRISON;
CLIN_DO=CLINIC*DOSE;
RUN;
```

The interaction terms (called CLIN_PR and CLIN_DO) are then added to the model. The CONTRAST statement can be used to test the two interaction terms simultaneously with a generalized Wald test. After the word CONTRAST is a user-supplied label in quotes (i.e., the user's option what to put in quotes). Then the tested covariates (the product terms) are listed followed by a 1 and separated by a comma (see code below):

```
PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)= PRISON DOSE CLINIC CLIN_PR CLIN_DO;
CONTRAST "TEST INTERACTION" CLIN_PR 1, CLIN_DO 1;
RUN;
```

The PROC PHREG output follows:

The PHREG Procedure

Model Information

Data Set	WORK.ADDICTS2	
Dependent Variable	SURVT	survival time in days
Censoring Variable	STATUS	status (0=censored, 1=endpoint)
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	1411.324	1343.199
AIC	1411.324	1353.199
SBC	1411.324	1368.253

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	1.19200	0.54137	4.8480	0.0277	3.294
DOSE	1	-0.01932	0.01935	0.9967	0.3181	0.981
CLINIC	1	0.17469	0.89312	0.0383	0.8449	1.191
CLIN_PR	1	-0.73799	0.43149	2.9253	0.0872	0.478
CLIN_DO	1	-0.01386	0.01433	0.9359	0.3333	0.986

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
TEST INTERACTION	2	3.5803	0.1669

The estimates of the hazard ratios (left column) may be deceptive when product terms are in the model. For example, by exponentiating the estimated coefficient for PRISON at exp (1.19200) = 3.284, we obtain the estimated hazard ratio for PRISON=1 versus PRISON=0, where DOSE=0 and CLINIC=0. This is a meaningless hazard ratio since CLINIC is coded 1 or 2 and DOSE is always greater than zero (all patients are on methadone). In the next section (on stratified Cox models), we illustrate how a CONTRAST statement can be used to obtain more meaningful hazard ratio estimates for models with interaction terms. The CONTRAST statement can be used to obtain a linear combination of parameter estimates in addition to the generalized Wald test shown above.

The Wald chi-square p-values for the two product terms are 0.0872 for CLIN_PR and 0.3333 for CLIN_DO. The generalized Wald chi-square p-values for testing both product terms simultaneously is 0.1669. Alternatively, a likelihood ratio test can simultaneously test both product terms by subtracting the –2 log-likelihood statistic for the full model (with the two product terms) from the reduced model (without the product terms). The –2 log likelihood statistic can be found on the output in the table called MODEL FIT STATISTICS and under the column called WITH COVARIATES. The –2 log likelihood statistic is 1,343.199 for the full model and 1,346.805 for the reduced model. The test is a two degree of freedom test since 2 product terms are simultaneously tested.

The PROBCHI function in SAS can be used to obtain p-values for chi-square tests. The code follows:

```
DATA TEST;
REDUCED = 1346.805;
FULL = 1343.199;
DF = 2;
P_VALUE = 1 - PROBCHI(REDUCED-FULL,DF);
RUN;

PROC PRINT DATA=TEST;RUN;
```

Note that you must write 1 minus the PROBCHI function to obtain the area under the right side of the chi-square probability density function. The output from the PROC PRINT follows:

Obs	REDUCED	FULL	DF	P_VALUE
1	1346.81	1343.20	2	0.16480

The p-value for the likelihood ratio test for both product terms is 0.16480, a similar result to the p-value that was obtained from the generalized Wald test (0.1669). Both of these tests are two degree of freedom tests since the two interaction terms are simultaneously tested.

3. RUNNING A STRATIFIED COX MODEL

Suppose we believe that the variable CLINIC violates the proportional hazards assumption but the variables PRISON and DOSE follow the PH assumption within each level of CLINIC. A stratified Cox model on the variable CLINIC can be run with PROC PHREG using the STRATA CLINIC statement. The code follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE/RL;
STRATA CLINIC;
RUN;
```

The output of the parameter estimates follows:

The PHREG Procedure
Analysis of Maximum Likelihood Estimates

		Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	95% Hazard Ratio Confidence Limits	
PRISON	1	0.38877	0.16892	5.2974	0.0214	1.475	1.059	2.054
DOSE	1	-0.03514	0.00647	29.5471	<.0001	0.965	0.953	0.978

Notice there is no parameter estimate for CLINIC since CLINIC is the stratified variable. The hazard ratio for PRISON=1 vs. PRISON=0 is estimated at 1.475. This hazard ratio is assumed not to depend on CLINIC since an interaction term between PRISON and CLINIC was not included in the model.

Suppose we wish to assess interaction between PRISON and CLINIC as well as DOSE and CLINIC in a Cox model stratified by CLINIC. We can define interaction terms in a new SAS dataset (called addicts2) and then run a model containing these terms.

```
DATA ADDICTS2;
SET REF.ADDICTS;
CLIN_PR=CLINIC*PRISON;
CLIN_DO=CLINIC*DOSE;
RUN;
```

Note with the interaction model that the hazard ratio for PRISON=1 versus PRISON=0 for CLINIC=1 controlling for

DOSE is $\exp(\beta_1 + \beta_3)$, and the hazard ratio for PRISON=1 versus PRISON=0 for CLINIC=2 controlling for DOSE is $\exp(\beta_1 + 2\beta_3)$. This latter calculation is obtained by substituting the appropriate values into the hazard in both the numerator (for PRISON=1) and denominator (for PRISON=0) (see below):

$$HR = \frac{h_0(t)\exp[1\beta_1 + \beta_2 DOSE + (2)(1)\beta_3 + \beta_4 CLIN_DO]}{h_0(t)\exp[0\beta_1 + \beta_2 DOSE + (2)(0)\beta_3 + \beta_4 CLIN_DO]} = \exp(\beta_1 + 2\beta_3).$$

A CONTRAST statement with the ESTIMTES= option can be used with PROC PHREG when we wish to obtain estimates of a linear combination of parameter estimates. We can also use the CONTRAST statement to test the two interaction terms simultaneously with a generalized Wald test as we illustrated in the previous section.

The code below runs a stratified Cox model (STRATA CLINIC) including two interaction terms in the model. Three CONTRAST statements are used: the first to estimate the hazard ratio for PRISON among those with CLINIC=1, $\exp(\beta_1 + \beta_3)$; the second to estimate the hazard ratio for PRISON among those with CLINIC=2, $\exp(1 + 2\beta_3)$; and the third to test the two interaction terms with a two degree of freedom generalized Wald test. The ESTIMATE=EXP option in the first two CONTRAST statements requests that the parameter estimates be exponentiated. The code in the second CONTRAST statement PRISON **1** CLIN_PR **2**/ESTIMATE=EXP; requests the estimate for $\exp(\beta_1 + 2\beta_3)$. The β_1 corresponds to PRISON and the beta3 (β)corresponds to the third variable in the model, CLIN_PR. The code follows:

```
PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)= PRISON DOSE  CLIN_PR CLIN_DO;
STRATA CLINIC;
CONTRAST 'HR FOR PRISON AMONG CLINIC=1' PRISON 1 CLIN_PR 1/ESTIMATE=EXP;
CONTRAST 'HR FOR PRISON AMONG CLINIC=2' PRISON 1 CLIN_PR 2/ESTIMATE=EXP;
CONTRAST "TEST INTERACTION" CLIN_PR 1, CLIN_DO 1;
RUN;
```

Notice that when we stratify by CLINIC, we do not put the variable CLINIC in the model statement. However, the interaction terms CLIN_PR and CLIN_DO are put in the model statement while CLINIC is put in the strata statement. The output follows:

The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	1.08716	0.53861	4.0741	0.0435	2.966
DOSE	1	-0.03482	0.01980	3.0930	0.0786	0.966
CLIN_PR	1	-0.58467	0.42813	1.8650	0.1721	0.557
CLIN_DO	1	-0.00105	0.01457	0.0052	0.9427	0.999

Contrast Rows Estimation and Testing Results

Contrast	Type	Estimate	Standard Error	Confidence Limits	
HR FOR PRISON AMONG CLINIC=1	EXP	1.6528	0.3119	1.1419	2.3925
HR FOR PRISON AMONG CLINIC=2	EXP	0.9211	0.3540	0.4337	1.9563

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
HR FOR PRISON AMONG CLINIC=1	1	7.0918	0.0077
HR FOR PRISON AMONG CLINIC=2	1	0.0457	0.8307
TEST INTERACTION	2	1.8650	0.3936

The hazard ratio (PRISON=1 vs PRISON=0) is estimated at 1.6528 among CLINIC=1 and 0.9211 among CLINIC=2. The generalized Wald test for testing both interaction terms simultaneously (a 2 df test: 1 $\beta_3 = 0$, 1 $\beta_4 = 0$) yields a p-value of 0.3936.

An alternative approach allowing for interaction with CLINIC and the other covariates is obtained by running two models: one subsetting on the observations where CLINIC=1 and the other subsetting on the observations where CLINIC=2. The code and output follow:

```
PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)=PRISON DOSE;
WHERE CLINIC=1;
TITLE COX MODEL RUN ONLY ON DATA WHERE CLINIC=1;
RUN;
```

A WHERE statement in a SAS procedure subsets the number of observations for analyses. A TITLE statement can also be added to the procedure. The output containing the parameter estimates subsetting on the observations where CLINIC=1 follows:

COX MODEL RUN ONLY ON DATA WHERE CLINIC=1
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	0.50249	0.18869	7.0918	0.0077	1.653
DOSE	1	-0.03586	0.00774	21.4761	<.0001	0.965

Similarly, the code and output containing the parameter estimates subsetting on the observations where CLINIC=2:

```
PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)=PRISON DOSE;
WHERE CLINIC=2;
TITLE COX MODEL RUN ONLY ON DATA WHERE CLINIC=2;
RUN;
```

COX MODEL RUN ONLY ON DATA WHERE CLINIC=2
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
PRISON	1	-0.08226	0.38430	0.0458	0.8305	0.921	0=none, 1=prison record
DOSE	1	-0.03693	0.01234	8.9500	0.0028	0.964	methadone dose (mg/day)

The estimated hazard ratio for PRISON=1 versus PRISON=0 is 0.921 among CLINIC=2 controlling for DOSE. This result is consistent with the stratified Cox model previously run in which all the product terms with CLINIC were included in the model.

4. ASSESSING THE PH ASSUMPTION WITH A STATISTICAL TEST

The following SAS program makes use of the addicts dataset to demonstrate how a statistical test of the PH assumption is performed for a given covariate (Harrel and Lee 1986). This is accomplished by finding the correlation between the Schoenfeld residuals for a particular covariate and the ranking of individual failure times. If the PH assumption is met, then the correlation should be near zero. The p-value for testing this correlation can be obtained from PROC CORR (or PROC REG). The Schoenfeld residuals for a given model can be saved in a SAS dataset using PROC PHREG. The ranking of events by failure time can be saved in a SAS dataset using PROC RANKED. The null hypothesis is that the PH assumption is not violated.

First, we run a model containing CLINIC, PRISON, and DOSE. The output statement creates a SAS dataset, the OUT= option defines an output dataset, and the RESSCH= statement is followed by user-defined variable names, so that the output dataset contains the Schoenfeld residuals. The order of the names corresponds to the order of the independent variables in the model statement. The actual variable names are arbitrary. The name we chose for the dataset is RESID and the names we chose for the variables containing the Schoenfeld residuals for CLINIC, PRISON, and DOSE are RCLINIC, RPRISON, and RDOSE. The code follows:

The code follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=CLINIC PRISON DOSE;
OUTPUT OUT=RESID RESSCH=RCLINIC RPRISON RDOSE;
RUN;

PROC PRINT DATA=RESID;RUN;
```

The first 10 observations of the PROC PRINT are printed below. The three columns on the right are the variables containing the Schoenfeld residuals.

Obs	SURVT	STATUS	CLINIC	PRISON	DOSE	RCLINIC	RPRISON	RDOSE
1	428	1	1	0	50	-0.18715	-0.40641	-8.2100
2	275	1	1	1	55	-0.15841	0.55485	-2.6277
3	262	1	1	0	55	-0.16453	-0.45197	-2.5635
4	183	1	1	0	30	-0.14577	-0.48727	-26.0823
5	259	1	1	1	65	-0.16306	0.54313	7.3701
6	714	1	1	0	55	-0.25853	-0.50074	-8.5347
7	438	1	1	1	65	-0.19292	0.58106	6.6072
8	796	0	1	1	60	.	.	.
9	892	1	1	0	50	-0.34478	-0.22372	-15.9088
10	393	1	1	1	65	-0.17712	0.57376	6.6886

Next, create a SAS dataset that deletes censored observations (i.e., only contains observations that fail).

```
DATA EVENTS;
SET RESID;
IF STATUS=1;
RUN;
```

Use PROC RANK to create a dataset containing a variable that ranks the order of failure times. The user supplies the name of the output dataset using the OUT= option. The variable to be ranked is SURVT. The RANKS statement precedes a user-defined variable name for the rankings of failure times. The user-defined names are arbitrary. The name we chose for this variable is TIMERANK. The code follows:

```
PROC RANK DATA=EVENTS OUT=RANKED TIES=MEAN;
VAR SURVT;
RANKS TIMERANK;
RUN;

PROC PRINT DATA=RANKED;RUN;
```

PROC CORR is used to get the correlations between the ranked failure time variable (called TIMERANK in this example) and the variables containing the Schoenfeld residuals of CLINIC, PRISON, and DOSE (called RCLINIC, RPRISON, and RDOSE, respectively, in this example). The NOSIMPLE option suppresses the printing of summary statistics. If the PH assumption is met for a particular covariate, then the correlation should be near zero. The p-value obtained from PROC CORR which tests whether this correlation is zero is the same p-value we use for testing the PH assumption. The code follows:

```
PROC CORR DATA=RANKED NOSIMPLE;
VAR RCLINIC RPRISON RDOSE;
WITH TIMERANK;
RUN;
```

The PROC CORR output follows:

```
                    The CORR Procedure

      Pearson Correlation Coefficients, N = 150
              Prob > |r| under H0: Rho=0

                              RCLINIC    RPRISON     RDOSE

TIMERANK                    -0.26153   -0.07970   0.07733
Rank for Variable SURVT       0.0012     0.3323    0.3469
```

The sample correlations with their corresponding p-values printed underneath are shown above. The p-values for CLINIC, PRISON, and DOSE are 0.0012, 0.3323, and 0.3469, respectively, suggesting that the PH assumption is violated for CLINIC, but reasonable for PRISON and DOSE.

The same p-values can be obtained by running linear regressions with each predictor (one at a time) using PROC REG and examining the p-values for the regression coefficients. The code below will produce output containing the p-value for CLINIC:

```
PROC REG DATA=RANKED;
MODEL TIMERANK=RCLINIC;
RUN;
```

The output produced by PROC REG follows:

The REG Procedure

Parameter Estimates

Variable	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	75.49955	3.43535	21.98	<.0001
RCLINIC	-28.38848	8.61194	-3.30	0.0012

The p-value for CLINIC (0.0012) is printed in the column on the right and matches the p-value that was obtained using PROC CORR.

5. OBTAINING COX ADJUSTED SURVIVAL CURVES

We use the BASELINE statement in PROC PHREG to create an output dataset containing Cox adjusted survival estimates for a specified pattern of covariates. The particular pattern of covariates of interest must first be created in a SAS dataset that is subsequently used as the input dataset for the COVARIATES= option in the BASELINE statement of PROC PHREG. Each pattern of covariates yields a different survival curve (assuming nonzero effects). Adjusted log(-log) survival plots can also be obtained for assessing the PH assumption. This will be illustrated with three examples:

Ex1 – Run a PH model using PRISON, DOSE, and CLINIC and obtain adjusted survival curves where PRISON=0, DOSE=70, and CLINIC=2.

Ex2 – Run a stratified Cox model (by CLINIC). Obtain two adjusted survival curves using the mean value of PRISON and DOSE for CLINIC=1 and CLINIC=2. Use the log log curves to assess the PH assumption for CLINIC adjusted for PRISON and DOSE.

Ex3 – Run a stratified Cox model (by CLINIC) and obtain adjusted survival curves for PRISON=0, DOSE=70 and for PRISON=1, DOSE=70. This yields four survival curves in all, two for CLINIC=1 and two for CLINIC=2.

Basically, there are three steps:

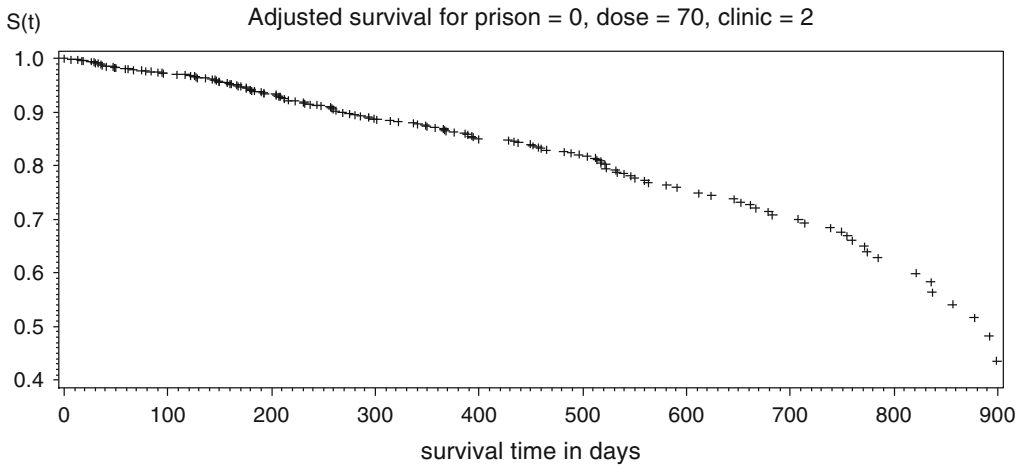
1) Create the input dataset containing the pattern (values) of covariates used for the adjusted survival curves.
2) Run a Cox model with PROC PHREG using the BASELINE statement to input the dataset from step (1) and output a dataset containing the adjusted survival estimates.
3) Plot the adjusted survival estimates from the output dataset created in step (2).

For Ex1, we create an input dataset (called IN1) with one observation where PRISON=0, DOSE=70, and CLINIC=2. We then run a model and create an output dataset (called OUT1) containing a variable with the adjusted survival estimates (called S1). Finally, the adjusted survival curve is plotted using PROC GPLOT. The code follows:

```
DATA IN1;
INPUT PRISON DOSE CLINIC;
CARDS;
0 70 2
;

PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC;
BASELINE COVARIATES=IN1 OUT=OUT1 SURVIVAL=S1/NOMEAN;
RUN;

PROC GPLOT DATA=OUT1;
PLOT S1*SURVT;
TITLE Adjusted survival for prison=0, dose=70, clinic=2;
RUN;
```

The BASELINE statement in PROC PHREG specifies the input dataset, the output dataset, and the name of the variable containing the adjusted survival estimates. The NOMEAN option suppresses the survival estimates using the mean values of PRISON, DOSE, and CLINIC. The next example (Ex2) will not use the NOMEAN option.

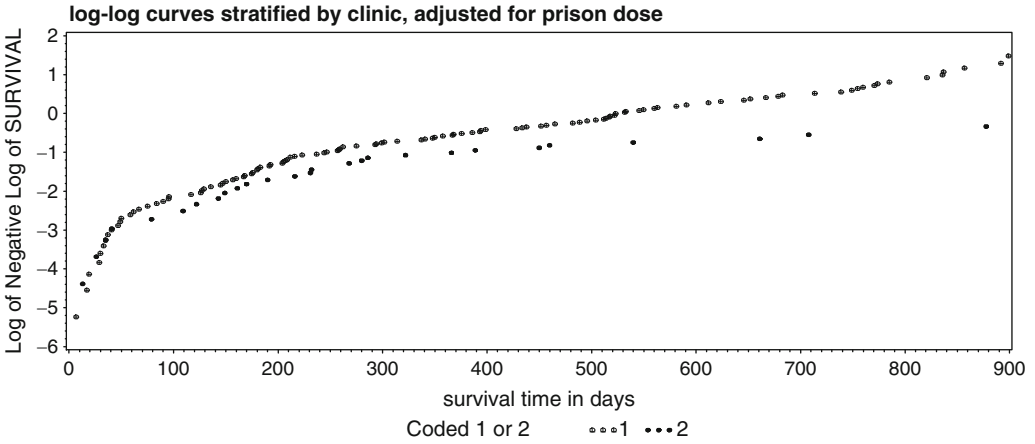
The output for PROC GPLOT follows:



For Ex2, we wish to create and output dataset (called OUT2) that contains the adjusted survival estimates from a Cox model stratified by CLINIC using the mean values of PRISON and DOSE. An input dataset need not be specified since by default the mean values of PRISON and DOSE will be used if the NOMEAN option is not used in the BASELINE statement. The code follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0) = PRISON DOSE CLINIC;
BASELINE OUT=OUT2 SURVIVAL=S2 LOGLOGS=LS2;
RUN;

PROC GPLOT DATA=OUT2;
PLOT S2*SURVT=CLINIC;
TITLE adjusted survival stratified by clinic;
RUN;

PROC GPLOT DATA=OUT2;
PLOT LS2*SURVT=CLINIC;
TITLE log-log curves stratified by clinic, adjusted for
  prison, dose;
RUN;
```

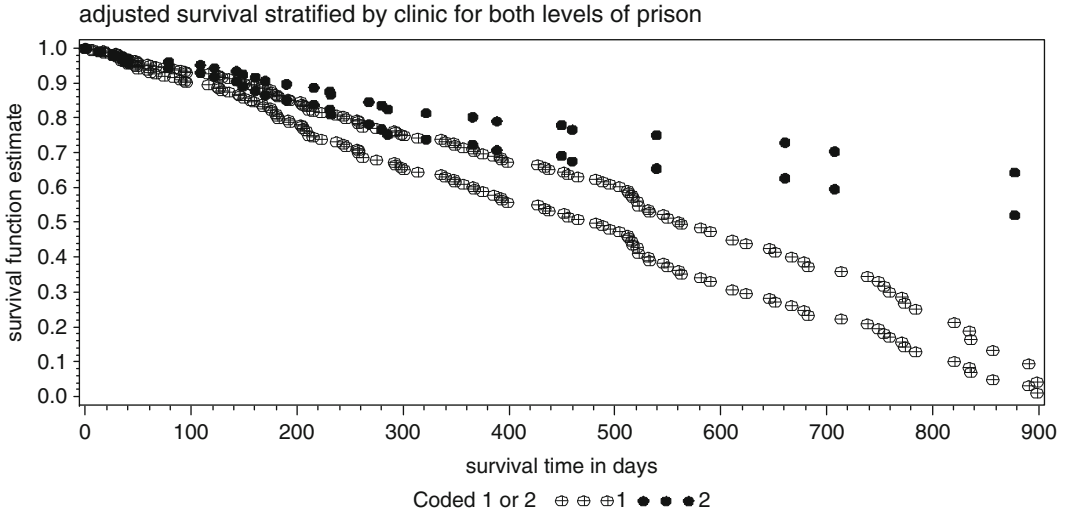
The code, PLOT LS2*SURVT=CLINIC, in the 2nd PROC GPLOT will plot LS2 on the vertical axis, SURVT on the horizontal axis, stratified by CLINIC on the same graph. The variable LS2 was created in the BASELINE statement of PROC PHREG and contains the adjusted log-log survival estimates. The PROC GPLOT output for the log-log survival curves stratified by CLINIC adjusted for PRISON and DOSE follows:



The adjusted log-log plots look similar to the unadjusted log-log Kaplan-Meier plots shown earlier, in that the plots look reasonably parallel before 365 days but then diverge, suggesting that the PH assumption is violated after 1 year.

For Ex3, a stratified Cox (by CLINIC) is run and adjusted curves are obtained for PRISON=1 and PRISON=0 holding DOSE=70. An input dataset (called IN3) is created with two observations for both levels of PRISON with DOSE=70. An output dataset (called OUT3) is created with the BASELINE statement that contains a variable (called S3) of survival estimates for all four curves (two for each stratum of CLINIC). The code follows:

```
DATA IN3;
INPUT PRISON DOSE;
CARDS;
1 70
0 70
;

PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE;
STRATA CLINIC;
BASELINE COVARIATES=IN3 OUT=OUT4 SURVIVAL=S3/NOMEAN;
RUN;

PROC GPLOT DATA=OUT3;
PLOT S3*SURVT=CLINIC;
TITLE adjusted survival stratified by clinic for both levels
  of prison;
RUN;
```



For the graph above, the PH assumption is not assumed for CLINIC since that is the stratified variable. However, the PH assumption is assumed for PRISON within each stratum of CLINIC (i.e., CLINIC=1 and CLINIC=2).

6. RUNNING AN EXTENDED COX MODEL

Models containing time-dependent variables are run using PROC PHREG. Time dependent variables are created with programming statements within the PROC PHREG procedure. Sometimes, users incorrectly define time-dependent variables in the data step. This leads to wrong estimates because the time variable used in the data step (SURVT) is actually time-independent and therefore different than the time variable (also called SURVT) used to define time-dependent variables in the PROC PHREG statement. See the discussion on the extended Cox likelihood in Chapter 6 for further clarification of this issue.

We have evaluated the PH assumption for the variable CLINIC by plotting KM log-log curves and Cox-adjusted log log curves stratified by CLINIC and checking whether the curves were parallel. We could do similar analyses with the variables PRISON and DOSE although with DOSE we would need to categorize the continuous variable before comparing plots for different strata of DOSE.

If it is expected that the hazard ratio for the effect of DOSE increases (or decreases) monotonically with time, we could add a continuous time-varying product term with DOSE and some function of time. The model defined below contains a time-varying variable (LOGTDOSE) defined as the product of DOSE and the natural log of time (SURVT). In some sense, a violation of the PH assumption for a particular variable means that there is an interaction between that variable and time. Note that the variable LOGTDOSE is defined within the PHREG procedure and not in the data step. The code follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON CLINIC DOSE LOGTDOSE;
LOGTDOSE=DOSE*LOG(SURVT);
RUN;
```

The output produced by PROC PHREG follows:

The PHREG Procedure
Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	0.34047	0.16747	4.1333	0.0420	1.406
CLINIC	1	-1.01857	0.21538	22.3655	<.0001	0.361
DOSE	1	-0.08243	0.03599	5.2468	0.0220	0.921
LOGTDOSE	1	0.00858	0.00646	1.7646	0.1841	1.009

The Wald test for the time-dependent variable LOGTDOSE yields a p-value of 0.1841. A nonsignificant p-value does not necessarily mean that the PH assumption is reasonable for DOSE. Perhaps, a different defined time-dependent variable would have been significant (e.g., DOSE × (TIME – 100)). Also, the sample-size of the study is a key determinant of the power to reject the null, which in this case means rejection of the PH assumption.

Next, we consider time-dependent variables for CLINIC. The next two models use heaviside functions that allow a different hazard ratio to be estimated for CLINIC before and after 365 days. The first model uses two heaviside functions in the model (HV1 and HV2) but not CLINIC. The second model uses one heaviside function (HV) but also includes CLINIC in the model. These two models yield the same hazard ratio estimates for CLINIC but are coded differently. The code and output for the model with two heaviside functions follows:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE HV1 HV2;
IF SURVT < 365 THEN HV1 = CLINIC; ELSE HV1 = 0;
IF SURVT >= 365 THEN HV2 = CLINIC; ELSE HV2 = 0;
CONTRAST 'TEST EQUALITY OF HEAVISIDES' HV1 1 HV2 -1;
RUN;
```

Analysis of Maximum Likelihood Estimates

Parameter	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	0.37770	0.16840	5.0304	0.0249	1.459
DOSE	-0.03551	0.00644	30.4503	<.0001	0.965
HV1	-0.45956	0.25529	3.2405	0.0718	0.632
HV2	-1.82823	0.38595	22.4392	<.0001	0.161

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
TEST EQUALITY OF HEAVISIDES	1	8.7993	0.0030

The parameter estimates for HV1 and HV2 can be used directly to obtain the estimated hazard ratio for CLINIC=2 vs CLINIC=1 before and after 365 days. The estimated hazard ratio for CLINIC at 100 days is exp(–0.45956) = 0.632 and the estimated hazard ratio for CLINIC at 400 days is exp(–1.82823) = 0.161. The CONTRAST statement provides a Wald test on the equality of two heaviside coefficients ($\beta_3 = \beta_4$ or $\beta_3 - \beta_4 = 0$). If the two heaviside coefficients were equal, then the hazard ratios for CLINIC would not depend on time. So the test could be viewed as a test of one form of PH violation. The p-value for the test is highly significant at 0.0030, suggesting that the PH assumption is violated for CLINIC.

The code and output for an equivalent model with one heaviside function are shown below:

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=CLINIC PRISON DOSE HV;
IF SURVT >= 365 THEN HV = CLINIC; ELSE HV = 0;
CONTRAST 'HR FOR CLINIC <365 days' CLINIC 1/ESTIMATE=EXP;
CONTRAST 'HR FOR CLINIC >=365 days' CLINIC 1 HV 1/ESTIMATE=EXP;
RUN;
```

Analysis of Maximum Likelihood Estimates

Parameter	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
CLINIC	-0.45956	0.25529	3.2405	0.0718	0.632
PRISON	0.37770	0.16840	5.0304	0.0249	1.459
DOSE	-0.03551	0.00644	30.4503	<.0001	0.965
HV	-1.36866	0.46139	8.7993	0.0030	0.254

Contrast Rows Estimation and Testing Results

Contrast	Type	Estimate	Standard Error	Confidence Limits	
HR FOR CLINIC <365 days	EXP	0.6316	0.1612	0.3829	1.0416
HR FOR CLINIC >=365 days	EXP	0.1607	0.0620	0.0754	0.3424

Notice that the variable CLINIC is included in this model and the coefficient for the time-dependent heaviside function, HV, does not contribute to the estimated hazard ratio until day 365. The estimated hazard ratio for CLINIC at 100 days is exp(–0.45956) = 0.6316 while the estimated hazard ratio for CLINIC at 400 days is exp((–0.45956) + (–1.36866)) = 0.1607 as calculated using the ESTIMATE=EXP option in the CONTRAST statement. These results are consistent with the estimates obtained from the model with two heaviside functions. A Wald test for the variable HV shows a statistically significant *p*-value of 0.003 suggesting a violation of the PH assumption for CLINIC. This is the same test as that obtained with the CONTRAST statement using the model with two heaviside functions.

Suppose it is believed that the hazard ratio for CLINIC=2 versus CLINIC=1 is constant over the first year but then monotonically increases (or decreases) after the first year. The following code defines a model allowing for a time-varying covariate called CLINTIME (defined in the code) which contributes to the hazard ratio for CLINIC after 365 days (output omitted):

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=CLINIC PRISON DOSE CLINTIME;
IF SURVT < 365 THEN CLINTIME=0;
ELSE IF SURVT >= 365 THEN CLINTIME = CLINIC*(SURVT-365);
RUN;
```

SAS is flexible in the way it can accommodate the modeling of time-varying covariates from different data formats. To illustrate this point, consider an example that was discussed in Chapter 6. The data below (Data D1) contain one observation for Jane who had an event at 49 months (MONTHS=49 and STATUS=1). Her dose of medication at the beginning of follow-up was 60 mg (DOSE1=60 and TIME1=0). At the 12th month of follow-up, her dose was changed to 120 mg (DOSE2=120 and TIME2=12). At the 30th month of follow-up, her dose was changed to 150 mg (DOSE3=120 and TIME3=30).

(Data D1) DOSE changes at three time points for Jane

(Data D1) DOSE changes at 3 time points for Jane

ID	MONTHS	STATUS	DOSE1	TIME1	DOSE2	TIME2	DOSE3	TIME3
Jane	49	1	60	0	120	12	150	30

If dosage is measured at multiple time points, then we would want to treat dose as a time varying covariate. We are assuming that Jane's observation is one out of many individuals Å. The following code would run an extended Cox model for data formatted as above:

```
PROC PHREG DATA=D1;
MODEL MONTHS*STATUS(0)=T_DOSE;
IF MONTHS<=TIME2 THEN T_DOSE=DOSE1;
ELSE IF MONTHS<=TIME3 THEN T_DOSE=DOSE2;
ELSE T_DOSE=DOSE3;
RUN;
```

The time dependent variable T_DOSE is defined below the MODEL statement and is defined in terms of DOSE1, DOSE2, and DOSE3 at specified points in time.

Alternatively, the data can be transposed in a counting process format such that Jane would have three observations to accommodate her three values of dosage over her risk period.

The following code transposes the data (D1) into a counting process format (D2):

```
DATA D2;
SET D1;
START=TIME1;STOP=TIME2;EVENT=0;DOSE=DOSE1;OUTPUT;
START=TIME2;STOP=TIME3;EVENT=0;DOSE=DOSE2;OUTPUT;
START=TIME3;STOP=MONTHS;EVENT=STATUS;DOSE=DOSE3;OUTPUT;
DROP MONTHS DOSE1 DOSE2 DOSE3 TIME1 TIME2 TIME3 STATUS;
RUN;
```

Now the data (D2) is transposed to contain three observations for Jane, allowing DOSE to be represented as a time-dependent variable. For the first time interval (START=0, STOP=12), Jane's dose was 60 mg. For the second time interval (12–30 months), Jane's dose was 120 mg. For the third time interval (30–49 months), Jane's dose was 150 mg. The data indicate that Jane had an event at 49 months (STOP=49 and STATUS=1). Jane's three observations are printed below:

```
PROC PRINT DATA=D2;RUN;
```

ID	START	STOP	EVENT	DOSE
JANE	0	12	0	60
JANE	12	30	0	120
JANE	30	49	1	150

The code to run the model with the data in counting process format is shown below:

```
PROC PHREG DATA=D2;
MODEL (START,STOP)*EVENT(0)=DOSE;
RUN;
```

Using PROC PHREG on data in the counting process format is discussed in more detail when we discuss the modeling recurrent events in SAS (Section 8).

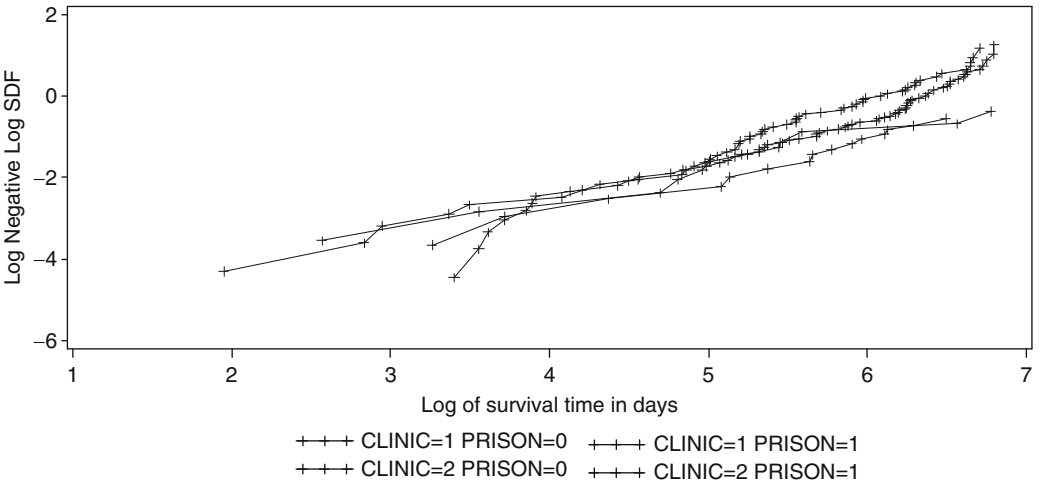
7. RUNNING PARAMETRIC MODELS WITH PROC LIFEREG

PROC LIFEREG runs parametric AFT models rather than PH models. Whereas the key assumption of a PH model is that hazard ratios are constant over time, the key assumption for an AFT model is that survival time accelerates (or decelerates) by a constant factor when comparing different levels of covariates.

The most common distribution for parametric modeling of survival data is the Weibull distribution. The hazard function for a Weibull distribution is $\lambda p t^{p-1}$. If p = 1, then the Weibull distribution is also an exponential distribution. The Weibull distribution has a desirable property, in that if the AFT assumption holds then the PH assumption also holds. The exponential distribution is a special case of the Weibull distribution. The key property for the exponential distribution is that the hazard is constant over time ($h(t) = \lambda$). In SAS, the Weibull and exponential model are run only as AFT models.

The Weibull distribution has the property that the log-log of the survival function is linear with the log of time. PROC LIFETEST can be used to plot Kaplan-Meier log-log curves against the log of time. If the curves are approximately straight lines (and parallel), then the assumption is reasonable. Furthermore, if the straight lines have a slope of 1, then the exponential distribution is appropriate. The code below produces log-log curves stratified by CLINIC and PRISON that can be used to check the validity of the Weibull assumption for those variables:

```
PROC LIFETEST DATA=REF.ADDICTS METHOD=KM PLOTS=(LLS);
TIME SURVT*STATUS(0);
STRATA CLINIC PRISON;
RUN;
```



The log-log curves do not look straight but for illustration we shall proceed as if the Weibull assumption were appropriate. First, an exponential model will be run with PROC LIFEREG. In this model, the Weibull shape parameter (p) is forced to equal 1, which forces the hazard to be constant.

```
PROC LIFEREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC/DIST=EXPONENTIAL;
RUN;
```

The DIST=EXPONENTIAL option in the MODEL statement requests the Weibull distribution. The output of parameter estimates obtained from PROC LIFEREG follows:

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	3.6843	0.4307	2.8402	4.5285	73.17	<.0001
PRISON	1	-0.2526	0.1649	-0.5758	0.0705	2.35	0.1255
DOSE	1	0.0289	0.0061	0.0169	0.0410	22.15	<.0001
CLINIC	1	0.8806	0.2106	0.4678	1.2934	17.48	<.0001
Scale	0	1.0000	0.0000	1.0000	1.0000		
Weibull Shape	0	1.0000	0.0000	1.0000	1.0000		

The exponential model assumes a constant hazard. This is indicated in the output by the value of the Weibull shape parameter (1.0000). The output can be used to calculate the estimated hazard for any subject given a pattern of covariates. For example, a subject with PRISON=0, DOSE=50, and CLINIC=2 has an estimated hazard of $\exp\{-(3.6843 + 50(0.0289)) + 2(0.8806)\} = 0.001$. Note that SAS gives the parameter estimates for the AFT form of the exponential model. Multiply the estimated coefficients by negative one to get estimates consistent with the PH parameterization of the model (see Chapter 7).

Next, a Weibull AFT model is run with PROC LIFEREG.

```
PROC LIFEREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC/DIST=WEIBULL;
RUN;
```

The DIST=WEIBULL option in the MODEL statement requests the Weibull distribution. The output for the parameter estimates follows:

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	4.1048	0.3281	3.4619	4.7478	156.56	<.0001
PRISON	1	-0.2295	0.1208	-0.4662	0.0073	3.61	0.0575
DOSE	1	0.0244	0.0046	0.0154	0.0334	28.32	<.0001
CLINIC	1	0.7090	0.1572	0.4009	1.0172	20.34	<.0001
Scale	1	0.7298	0.0493	0.6393	0.8332		
Weibull Shape	1	1.3702	0.0926	1.2003	1.5642		

The Weibull shape parameter is estimated at 1.3702. SAS calls the reciprocal of the Weibull shape parameter, the Scale parameter, estimated at 0.7298. The acceleration factor comparing CLINIC=2 to CLINIC=1 is estimated at exp(0.7090) = 2.03. So, the estimated median survival time (time off heroin) is double for patients enrolled in CLINIC=2 compared to CLINIC=1.

To obtain the hazard ratio parameters from the Weibull AFT model, multiply the Weibull shape parameter by the negative of the AFT parameter (see Chapter 7). For example, the HR estimate for CLINIC=2 vs CLINIC=1 controlling for the other covariates is exp(1.3702(-0.7090)) = 0.38.

Next, a log-logistic AFT model is run with PROC LIFEREG.

```
PROC LIFEREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC/DIST=LLOGISTIC;
RUN;
```

The output of the log-logistic parameter estimates follows:

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr >ChiSq
Intercept	1	3.5633	0.3894	2.8000	4.3266	83.71	<.0001
PRISON	1	-0.2913	0.1440	-0.5734	-0.0091	4.09	0.0431
DOSE	1	0.0316	0.0055	0.0208	0.0424	32.81	<.0001
CLINIC	1	0.5806	0.1716	0.2443	0.9169	11.45	0.0007
Scale	1	0.5868	0.0403	0.5129	0.6712		

From this output, the acceleration factor comparing CLINIC=2 to CLINIC=1 is estimated at exp(0.5806) = 1.79. If the AFT assumption holds for a log-logistic model, then the proportional odds assumption holds for the survival function (although the PH assumption will not hold). The proportional odds assumption can be evaluated by plotting the log odds of survival (using KM estimates) against the log of survival time. If the plots are straight lines for each pattern of covariates, then the log-logistic distribution is reasonable. If the straight lines are also parallel, then the proportional odds and AFT assumptions also hold.

A SAS dataset containing the KM survival estimates can be created using PROC LIFETEST (see Section 1 of this appendix). Once this variable is created, a dataset containing variables for the estimated log odds of survival and the log of survival time can also be created. PROC GPLOT can then be used to plot the log odds of survival against survival time.

Another context for thinking about the proportional odds assumption is that the odds ratio estimated by a logistic regression does not depend on the length of the follow-up. For example, if a follow-up study was extended from 3 to 5 years, then the underlying odds ratio comparing two patterns of covariates would not change. If the proportional odds assumption is not true, then the odds ratio is specific to the length of follow-up.

An AFT model is a multiplicative model with respect to survival time or equivalently an additive model with respect to the log of time. In the previous example, the median survival time was estimated as 1.79 **times** longer for CLINIC=2 compared to CLINIC=1. In that example, survival time was assumed to follow a log-logistic distribution or equivalently the log of survival time was assumed to follow a logistic distribution.

SAS allows additive failure time models to be run (see chapter 7 under the heading "Other Parametric Models"). The NOLOG option in the MODEL statement of PROC LIFEREG suppresses the default log link function which means that time, rather than log(time), is modeled as a linear function of the regression parameters. The following code requests an additive failure time model in which time follows a logistic (not log-logistic) distribution:

```
PROC LIFEREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC/DIST=LLOGISTIC NOLOG;
RUN;
```

Even though the option DIST=LLOGISTIC appears to request that survival time follows a log-logistic distribution. The NOLOG option actually means that survival time is assumed to follow a logistic distribution. (Note that the NOLOG option in Stata means – something completely different using the **streg** command – that the iteration log file not be shown in the output.) The output from the additive failure time model follows:

Analysis of Parameter Estimates

Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	-358.482	114.0161	-581.949	-135.014	9.89	0.0017
PRISON	1	-89.7816	42.9645	-173.990	-5.5727	4.37	0.0366
DOSE	1	10.3893	1.6244	7.2055	13.5731	40.91	<.0001
CLINIC	1	214.2525	53.1204	110.1385	318.3665	16.27	<.0001
Scale	1	172.4039	11.3817	151.4792	196.2191		

The parameter estimate for CLINIC is 214.2525. The interpretation for this estimate is that the median survival time (or time to any fixed value of $S(t)$) is estimated at 214 days more for CLINIC=2 compared to CLINIC=1. In other words, you add 214 days to the estimated median survival time for CLINIC=1 to get the estimated median survival time for CLINIC=2. This contrasts with the previous AFT model in which you multiply estimated median survival time for CLINIC=1 by 1.79 to get the estimated median survival time for CLINIC=2. The additive model can be viewed as a shifting of survival time while the AFT model can be viewed as a scaling of survival time.

If survival time follows a logistic distribution and the additive failure time assumption holds, then the proportional odds assumption also holds. The logistic assumption can be evaluated by plotting the log odds of survival (using KM estimates) against time (rather than against the log of time as analogously used for the evaluation of the log-logistic assumption). If the plots are straight lines for each pattern of covariates, then the logistic distribution is reasonable. If the straight lines are also parallel, then the proportional odds and additive failure time assumptions hold.

Other distributions supported by PROC LIFEREG are the generalized gamma (DIST=GAMMA) and lognormal (DIST=LNORMAL) distributions. If the NOLOG option is specified with the DIST=LNORMAL option in the model statement, then survival time is assumed to follow a normal distribution.

8. MODELING RECURRENT EVENTS

The modeling of recurrent events is illustrated with the bladder cancer dataset (**bladder.sas7bdat**) described at the start of this appendix. Recurrent events are represented in the data with multiple observations for subjects having multiple events. The data layout for the bladder cancer dataset is suitable for a counting process approach with time intervals defined for each observation (see Chapter 8). The following code prints the 12^{th}–20^{th} observation, which contains information for four subjects. The code follows:

```
PROC PRINT DATA=REF.BLADDER (FIRSTOBS= 12 OBS=20);
RUN;
```

The output follows:

OBS	ID	EVENT	INTERVAL	START	STOP	TX	NUM	SIZE
12	10	1	1	0	12	0	1	1
13	10	1	2	12	16	0	1	1
14	10	0	3	16	18	0	1	1
15	11	0	1	0	23	0	3	3
16	12	1	1	0	10	0	1	3
17	12	1	2	10	15	0	1	3
18	12	0	3	15	23	0	1	3
19	13	1	1	0	3	0	1	1
20	13	1	2	3	16	0	1	1

There are three observations for ID=10, one observation for ID=11, three observations for ID=12, and two observations for ID=13. The variables START and STOP represent the time interval for the risk period specific to that observation. The variable EVENT indicates whether an event (coded 1) occurred. The first three observations indicate that the subject with ID=10 had an event at 12 months, another event at 16 months, and was censored at 18 months.

PROC PHREG can be used for survival data using a counting process data layout. The following code runs a model with three predictors – treatment status (TX), initial number of tumors (NUM), and the initial size of tumors (SIZE) – included in the model:

```
PROC PHREG DATA=BLADDER COVS(AGGREGATE);
MODEL (START,STOP)*EVENT(0)=TX NUM SIZE;
ID ID;
RUN;
```

The code (START,STOP)*EVENT(0) in the MODEL statement indicates that the time intervals for each observation are defined by the variables START and STOP and that EVENT=0 denotes a censored observation. The ID statement defines ID as the variable representing each subject. The COVS(AGGREGATE) option in the PROC PHREG statement requests robust standard errors for the parameter estimates. The output generated by PROC PHREG follows:

The PHREG Procedure

Model Information

Data Set	WORK.BLADDER
Dependent Variable	START
Dependent Variable	STOP
Censoring Variable	EVENT
Censoring Value(s)	0
Ties Handling	BRESLOW

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio
TX	-0.40710	0.24183	1.209	2.8338	0.0923	0.666
NUM	0.16065	0.05689	1.185	7.9735	0.0047	1.174
SIZE	-0.04009	0.07222	1.028	0.3081	0.5788	0.961

Coefficient estimates are provided with robust standard errors. The column under the heading StdErrRatio provides the ratio of the robust to the non-robust standard errors. For example, the standard error for the coefficient for TX (0.24183) is 1.209 greater than the standard error would be if we had not requested robust standard errors (i.e., omit the COVS(AGGREGATE) option). The robust standard errors are estimated slightly different compared to the corresponding model in Stata or R.

A stratified Cox model can also be run using the data in this format with the variable INTERVAL as the stratified variable. The stratified variable indicates whether the subject was at risk for their 1st, 2nd, 3rd, or 4th event. This approach is called a Stratified CP approach in Chapter 8 and is used if the investigator wants to distinguish the order in which recurrent events occur. The code for a stratified Cox follows:

```
PROC PHREG DATA=BLADDER COVS(AGGREGATE) ;
MODEL (START,STOP)*EVENT(0)=TX NUM SIZE;
ID ID;
RUN;
```

The only additional code from the previous model is the STRATA statement, indicating that the variable INTERVAL is the stratified variable. The output containing the parameter estimates follows:

Analysis of Maximum Likelihood Estimates

Variable	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio
TX	-0.33430	0.19706	0.912	2.8777	0.0898	0.716
NUM	0.11565	0.04991	0.930	5.3690	0.0205	1.123
SIZE	-0.00805	0.06012	0.827	0.0179	0.8935	0.992

Interaction terms between the treatment variable (TX) and the stratified variable could be created to examine whether the effect of treatment differed for the 1st, 2nd, 3rd, or 4th event.

Another stratified approach (called Gap Time) is a slight variation of the stratified counting process approach. The difference is in the way the time intervals for the recurrent events are defined. There is no difference in the time intervals when subjects are at risk for their first event. However, with the Gap Time approach, the starting time at risk gets reset to zero for each subsequent event. The following code creates data suitable for using the gap-time approach:

```
DATA BLADDER2;
SET REF.BLADDER;
START2=0;
STOP2=STOP-START;
RUN;
```

The new dataset (bladder2) copies the data from re.bladder and creates two new variables for the time interval: START2, which is always set to zero and STOP2, which is the length of the time interval (i.e., STOP–START). The following code uses these newly created variables to run a Gap Time model with PROC PHREG:

```
PROC PHREG DATA=BLADDER2 COVS(AGGREGATE);
MODEL (START2,STOP2)*EVENT(0)=TX NUM SIZE;
ID ID;
STRATA INTERVAL;
RUN;
```

The output follows:

Variable	Parameter Estimate	Standard Error	StdErr Ratio	Chi-Square	Pr > ChiSq	Hazard Ratio
TX	-0.26952	0.20808	1.002	1.6778	0.1952	0.764
NUM	0.15353	0.04889	0.938	9.8620	0.0017	1.166
SIZE	0.00684	0.06222	0.889	0.0121	0.9125	1.007

The results using the Gap Time approach vary slightly from that obtained using the Stratified CP approach.

The counting process data layout with multiple observations per subject need not only apply to recurrent event data, but can also be used for a more conventional survival analyses in which each subject is limited to one event. A subject with four observations may be censored for the first three observations before getting the event in the time interval represented by the fourth observation. This data layout is particularly suitable for representing time-varying exposures (i.e., exposures which change values over different intervals of time).

C. SPSS

Analyses are carried out in SPSS by using the appropriate SPSS procedure on an SPSS dataset. Most users select procedures by pointing and clicking the mouse through a series of menus and dialog boxes. The code, or command syntax, generated by these steps can be viewed and edited.

Analyses on the "addicts" dataset will be used to illustrate these procedures. The addicts dataset was obtained from a 1991 Australian study by Caplehorn et al. and contains information on 238 heroin addicts. The study compared two methadone treatment clinics to assess patient time remaining under methadone treatment. The two clinics differed according to its live-in policies for patients. A patient's survival time was determined as the time (in days) until the person dropped out of the clinic or was censored. The variables are defined at the start of this appendix.

After getting into SPSS, open the dataset **addicts.sav**. The data should appear on your screen. This is now your working dataset. To obtain a basic descriptive analysis of the outcome variable (SURVT), click on Analyze → Descriptive Statistics → Descriptive from the drop-down menus to reach the dialog box to specify the analytic variables. Select the SURVT from the list of variables and enter it into the variable box. Click on OK to view the output. Alternatively, you can click on Paste (rather than OK) to obtain the corresponding SPSS syntax. The syntax can then be submitted (by clicking the button under Run), edited, or saved for another session. The syntax created is as follows (output omitted):

```
DESCRIPTIVES
  VARIABLES=survt
  /STATISTICS=MEAN STDDEV MIN MAX.
```

There are some analyses that SPSS only performs by submitting syntax rather than using the point and click approach (e.g., running an extended Cox model with two time-varying covariates). Each time the point and click approach is presented, the corresponding syntax will also be presented.

To obtain more detailed descriptive statistics on survival time stratified by CLINIC, click on Analyze → Descriptive Statistics → Explore from the drop-down menus. Select SURVT from the list of variables and enter it into the Dependent List and then select CLINIC and enter it into the Factor List. Click on OK to see the output. The syntax created from clicking on Paste (rather than OK) is as follows (output omitted):

```
EXAMINE
  VARIABLES=survt BY clinic
  /PLOT BOXPLOT STEMLEAF
  /COMPARE GROUP
  /STATISTICS DESCRIPTIVES
  /CINTERVAL 95
  /MISSING LISTWISE
  /NOTOTAL.
```

Survival analyses can be performed in SPSS by selecting Analyze → Survival. There are then four choices for selection: Life Tables, Kaplan-Meier, Cox Regression, and Cox w/ Time-Dep Cov. The key SPSS procedures for survival analysis are the KM and COXREG procedures.

The survival analyses demonstrated in SPSS are as follows:

1. Estimating survival functions (unadjusted) and comparing them across strata
2. Assessing the PH assumption using Kaplan-Meier log-log survival curves
3. Running a Cox PH model
4. Running a stratified Cox model and obtaining Cox adjusted log-log curves
5. Assessing the PH assumption with a statistical test
6. Running an extended Cox model

SPSS (version PASW 18) does not provide commands to run parametric survival models, frailty models, or models using a counting process data layout for recurrent events.

1. ESTIMATING SURVIVAL FUNCTIONS (UNADJUSTED) AND COMPARING THEM ACROSS STRATA

To obtain Kaplan-Meier survival estimates, select Analyze → Survival → Kaplan-Meier. Select the SURVT from the variable list and enter it into the Time box, then select the variable STATUS and enter it into the Status box. You will then see a question mark in parentheses after the status variable, indicating that the value of the event needs to be entered. Click the Define Event button and insert the value 1 in the box since the variable STATUS is coded 1 for events and 0 for censorships. Click on Continue and then OK to view the output. The syntax, obtained from clicking on Paste (rather than OK), is as follows (output omitted):

```
KM
  survt /STATUS=status(1)
  /PRINT TABLE MEAN.
```

The stream of output of these KM estimates is quite long. If you wish to edit the output, try right clicking inside the output and then select Edit Content. You then have a choice to select In Viewer or In Separate Window. Click on one of these depending on if you want to open a separate window for your edited output.

To obtain KM survival estimates and plots by CLINIC as well as log rank (and other) test statistics, select Analyze → Survival → Kaplan-Meier and then select SURVT as the time-to-event variable and STAUS as the status variable as described above. Enter CLINIC into the Factor box and click the Compare Factor button. You have a choice of three test statistics for testing the equality of survival functions across CLINIC. Select all three (log rank, Breslow, and Tarone–Ware) for comparison and click Continue. Select the Options button to request plots. There are four choices (unfortunately, log-log survival plots are not included). Select Survival to obtain KM plots by clinic. Click Continue and then OK to view the output.

The syntax follows:

```
KM
  survt BY clinic /STATUS=status(1)
  /PRINT TABLE MEAN
  /PLOT SURVIVAL
  /TEST LOGRANK BRESLOW TARONE
  /COMPARE OVERALL POOLED.
```

The output containing the KM estimates for the first five events or censorship times from CLINIC=1 and CLINIC=2 as well for the log rank, Breslow, and Tarone–Ware tests follow:

Survival Analysis for SURVT Survival time (days)

Survival Table

clinic	Time	Status	Cumulative Proportion Surviving at the Time		N of Cumulative Events	N of Remaining Cases
			Estimate	Std. Error		
1	7.000	endpoint	.994	.006	1	161
2	17.000	endpoint	.988	.009	2	160
3	19.000	endpoint	.981	.011	3	159
4	28.000	censored	.	.	3	158
5	28.000	censored	.	.	3	157
.						
.						

```
Factor CLINIC = 2.00
```

2.00	1	13.000	endpoint	.986	.013	1	73
	2	26.000	endpoint	.973	.019	2	72
	3	35.000	endpoint	.959	.023	3	71
	4	41.000	endpoint	.946	.026	4	70
	5	53.000	censored	.	.	4	69

```
      .
      .
Test Statistics for Equality of Survival Distributions for CLINIC

              Statistic        df       Significance
```

Note that what SPSS calls the Breslow test statistic is equivalent to what Stata (and SAS) call the Wilcoxon test statistic.

Life Table estimates can be obtained by selecting Analyze → Survival → Life Tables. The time-to-event and status variables are defined similarly as described above for KM estimates. However with life tables, SPSS presents a Display Time Intervals box. This allows the user to define the time intervals used for the life table analysis. For example, 0 to 1,000/100 would define 10 time intervals of equal length. Life table plots can similarly be requested as described above for the KM plots.

2. ASSESSING THE PH ASSUMPTION USING KAPLAN-MEIER LOG-LOG SURVIVAL CURVES

SPSS does not provide unadjusted KM log-log curves by directly using the point and click approach with the KM command. SPSS does provide adjusted log log curves from running a stratified Cox model (described later in the stratified Cox section). A log-log curve equivalent to the unadjusted KM log-log curve can be obtained in SPSS by running a stratified Cox without including any covariates in the model. In this section, however, we illustrate how new variables can be defined in the working dataset and then used to plot unadjusted log-log KM plots.

First, a variable will be created containing the KM survival estimates. Then, another new variable will be created containing the log-log of the survival estimates. Finally, the log-log survival estimates will be plotted against survival time to see if the curves for CLINIC=1 and CLINIC=2 are parallel. Each step can be done with the point and click approach or by typing in the code directly.

A variable containing the survival estimates can be created by selecting Analyze → Survival → Kaplan-Meier and then selecting SURVT as the time-to-event variable, STAUS as the status variable, and CLINIC as the factor variable as described above. Then click the Save button. This opens a dialogue box called Kaplan-Meier Save New Variables. Check Survival and click on Continue and then on Paste. The code that is created is as follows:

```
KM
   survt BY clinic /STATUS=status(1)
   /PRINT TABLE MEAN
   /SAVE SURVIVAL.
```

By submitting this code, a new variable containing the KM estimates called **SUR_1** is created. To create a new variable called **lls** containing the log(-log) of **SUR_1**, submit the following code:

```
COMPUTE lls = LN(-LN (SUR_1)).
   EXECUTE.
```

The above code could also be generated by selecting Transform → Compute Variable and defining the new variable in the dialogue box. To plot **lls** against survival time, submit the code:

```
GRAPH
   /SCATTERPLOT(BIVAR)=survt WITH lls BY clinic
   /MISSING=LISTWISE.
```

This final piece of code could also be run by selecting Graphs → Legacy Dialogue → Scatter/Dot → and then clicking on Simple Scatter and then Define in the Scatter/Dot dialogue box. Select LLS for the Y-axis, SURVT for the X-axis, and CLINIC in the Set Marker By box. Clicking on paste creates the code or clicking OK submits the program. A plot of LLS against log(SURVT) could similarly be created. Parallel curves support the PH assumption for CLINIC.

3. RUNNING A COX PH MODEL

A Cox PH model can be run by selecting Analyze → Survival → Cox Regression. Select the SURVT from the variable list and enter it into the Time box, then select the variable STATUS and enter it into the Status box. You will then see a question mark in parentheses after the status variable, indicating that the value of the event needs to be entered. Click the Define Event button and insert the value 1 in the box since the variable STATUS is coded 1 for events and 0 for censorships. Click on Continue and select PRISON, DOSE, and CLINIC from the variable list and enter them into the Covariates box. You can click on Plots or Options to explore some of the options (e.g., 95% CI for exp(β)). Click OK to view the output or click on Paste to see the code. The code follows:

```
COXREG
  survt /STATUS=status(1)
  /METHOD=ENTER prison dose clinic
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

Note that the PH assumption is assumed to hold for all three covariates using this Cox model (the output follows).

Omnibus Tests of Model Coefficients[a,b]

−2Log Likelihood	Overall (score)			Change From Previous Step			Change From Previous Block		
	Chi-square	df	Sig.	Chi-square	df	Sig.	Chi-square	df	Sig.
1346.805	56.273	3	.000	64.519	3	.000	64.519	3	.000

[a] Beginning Block Number 0, initial Log Likelihood function: −2 Log likelihood: 1411.324
[b] Beginning Block Number 1. Method = Enter

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PRISON	.327	.167	3.813	1	.051	1.386
DOSE	−.035	.006	30.785	1	.000	.965
CLINIC	−1.009	.215	22.045	1	.000	.365

4. RUNNING A STRATIFIED COX MODEL AND OBTAINING COX ADJUSTED LOG-LOG CURVES

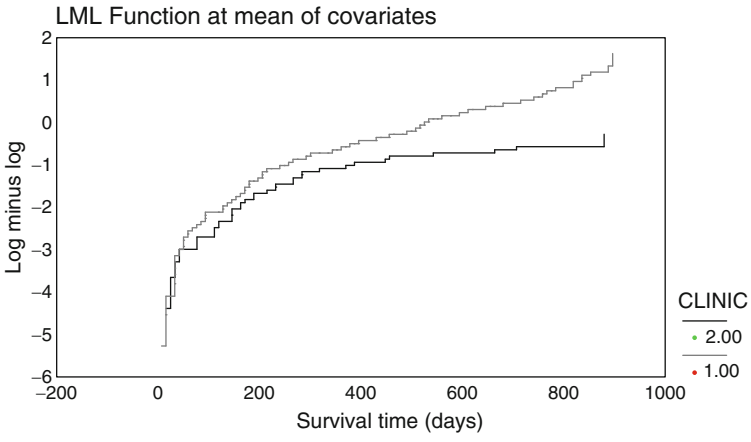
A stratified Cox model is run by selecting Analyze → Survival → Cox Regression. Select the SURVT from the variable list and enter it into the Time box. Select the variable STATUS and enter it into the Status box and then define the value of the event as 1. Put the variables PRISON and DOSE in the Covariates box and the variable CLINIC in the Strata box. The Cox model will be stratified by CLINIC. Click the Plots button and check Log minus log as the plot type and then click on Continue. Click on OK to view the output or click on Paste to see the code. The code follows:

```
COXREG
  survt /STATUS=status(1)
  /STRATA=clinic
  /METHOD=ENTER prison dose
  /PLOT LML
  /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

The output containing the parameter estimates and the adjusted log log plots follows:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PRISON	.389	.169	5.298	1	.021	1.475
DOSE	−.035	.006	29.552	1	.000	.965



Notice that there are parameter estimates for PRISON and DOSE but not CLINIC since CLINIC is the stratified variable. The Cox-adjusted log log plots are fitted using the mean values of PRISON and DOSE and are used to evaluate the PH assumption for CLINIC.

Suppose rather than using the mean value of DOSE for the adjusted log log plots, you wish to obtain adjusted plots in which DOSE=70. Run the same code as before, up to clicking on the Plots button and checking Log minus log as the plot type. Instead, click on DOSE(Mean) in the window called Covariate Values Plotted at. Underneath the heading called Change Value, click on the word Value, type in the value 70, and then click on the button called Change. Now, the variable in the window should be called DOSE(70) rather than DOSE(Mean). Click on Continue and then OK to view the output.

5. ASSESSING THE PH ASSUMPTION WITH A STATISTICAL TEST

SPSS does not easily accommodate a statistical test on the PH assumption using the Schoenfeld residuals. However, it can be programmed using several steps. The steps are as follows:

1. Run a Cox PH model to obtain the Schoenfeld residuals for all the covariates. These residuals are saved as new variables in the working dataset.
2. Delete observations that were censored.
3. Create a variable that contains the ranked order of survival time. For example, the subject who had the fourth event gets a value of 4 for this variable.
3. Run correlations on the survival rankings with the Schoenfeld residuals.
4. The p-value for testing whether the correlation between the ranked survival time and the covariate's Schoenfeld residuals is zero is the same p-value used to test the PH assumption. The null hypothesis is that the PH assumption is not violated.

First, run a Cox PH model with CLINIC, PRISON, and DOSE. Click on the Save button before submitting the model. A dialogue box appears that is called Cox Regression: Save Model Variables. Check Partial Residuals and click on Continue. This creates three new variables in the working dataset called **PR1_1**, **PR2_1**, and **PR3_1**, which are the partial residuals (Schoenfeld residuals) for CLINIC, PRISON, and DOSE, respectively. Click OK to run the model (or Paste to generate the code).

Next, delete all censored observations (i.e., only keep observations in which STATUS=1). To do this, select Data → Select Cases. Then check If condition is satisfied, and then click on If. Type status=1 in the dialogue box and click on Continue. Check Delete unselected cases in the box called Output. Click OK and only observations with events will be kept in the dataset. (Remember to go back to the addicts dataset that contains the censored observations when you continue work through the other sections that use the addicts data.)

Create the variable that contains the ranking of survival times by selecting Transform → Ranked Cases. Select the SURVT into the Variables box. Click on Rank Types, check Ranks, and click on Continue and then click on Ties, check Mean, and click Continue. Click OK and a new variable (called **Rsurvt**) will be created containing the ranked survival time.

Finally, obtain correlations (and their p-values) between the ranked survival and the Schoenfeld residuals. Select Analyze → Correlate → Bivariate. Move the ranked survival time variable as well as the three partial residual variables into the variable box. Check Pearson (for Pearson correlations) and Two-tailed for a two-tail test of significance and click OK to see the output. The code that is generated from these steps follows:

```
COXREG
    survt /STATUS=status(1)
    /METHOD=ENTER clinic prison dose
    /SAVE= PRESID
    /CRITERIA=PIN(.05) POUT(.10) ITERATE(20) .

FILTER OFF.
    USE ALL.
    SELECT IF(status=1).
    EXECUTE

RANK
    VARIABLES=survt (A) /RANK /PRINT=YES
    /TIES=MEAN.

CORRELATIONS
    /VARIABLES=Rsurvt PR1_1 PR2_1 PR3_1
    /PRINT=TWOTAIL NOSIG
    /MISSING=PAIRWISE.
```

The output containing the correlations follows:

Correlations

		RANK of SURVT	Partial residual for CLINIC	Partial residual for PRISON	Partial residual for DOSE
RANK of SURVT	Pearson Correlation	1	-.262**	-.080	.077
	Sig. (2-tailed)	.	.001	.332	.347
	N	150	150	150	150
Partial residual for CLINIC	Pearson Correlation	–.262**	1	.010	.023
	Sig. (2-tailed)	.001	.	.904	.776
	N	150	150	150	150
Partial residual for PRISON	Pearson Correlation	–.080	.010	1	.171*
	Sig. (2-tailed)	.332	.904	.	.037
	N	150	150	150	150
Partial residual for DOSE	Pearson Correlation	.077	.023	.171*	1
	Sig. (2-tailed)	.347	.776	.037	.
	N	150	150	150	150

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

The p-values for the correlations are the p-values for the PH test. In the output, examine the row labeled RANK of SURVT Sig(2-tailed). Notice that the null hypothesis is rejected for CLINIC (p = 0.001) but not for PRISON (p = 0.332) or DOSE (p = 0.347).

6. RUNNING AN EXTENDED COX MODEL

An extended Cox model with exactly one time-dependent covariate can be run using the point and click approach. Suppose we want to include a time-dependent covariate DOSE times the log of survival time. This product term could be appropriate if the hazard ratio comparing any two levels of DOSE monotonically increases (or decreases) over time. Select Analyze → Survival → Cox w/ Time-Dep Cov. This opens a dialogue called Expression for T_COV_. The user defines a time-dependent variable (called T_COV_) in this box. A variable T_ is included in the variable list. This is the variable that represents time-varying survival (as opposed to SURVT which is an individual's fixed time of event). We wish to define T_COV_ to be the log of T_ × DOSE. Enter the expression LN(T_)*dose into the dialogue box and click on the Model button. Now, run a Cox model that includes the covariates: PRISON, CLINIC, DOSE, and T_COV_. The code generated is as follows:

```
TIME PROGRAM.
   COMPUTE T_COV_ = LN(T_) * dose.

COXREG
   survt /STATUS=status(1)
   /METHOD=ENTER prison clinic dose T_COV_
   /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

The output containing the parameter estimates follows:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PRISON	.340	.167	4.134	1	.042	1.406
CLINIC	−1.019	.215	22.369	1	.000	.361
DOSE	−.082	.036	5.247	1	.022	.921
T_COV_	.009	.006	1.765	1	.184	1.009

The variable T_COV_ represents the time-dependent variable included in the model, which in this example is DOSE times the log of survival time.

A heaviside function for CLINIC can similarly be created. We can define a time dependent variable equal to CLINIC if time is greater than or equal to 365 days and 0 otherwise. Select Analyze → Survival → Cox w/ Time-Dep Cov. Define T_COV to be (T_ × 365) × clinic. After clicking on the Model button, run a Cox model that includes PRISON, DOSE, CLINIC, and T_COV_. The code generated is as follows:

```
TIME PROGRAM.
   COMPUTE T_COV_ = (T_ > = 365)* clinic.

COXREG
   survt /STATUS=status(1)
   /METHOD=ENTER prison clinic dose T_COV_
   /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

Note that SPSS recognizes the expression (T_ × 365) as taking the value 1 if survival time is ≥365 days and 0 otherwise.

The output follows:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PRISON	.378	.168	5.030	1	.025	1.459
CLINIC	−.460	.255	3.241	1	.072	.632
DOSE	−.036	.006	30.450	1	.000	.965
T_COV_	−1.369	.461	8.799	1	.003	.254

Notice the variable CLINIC is included in this model and the time-dependent heaviside function, T_COV_, does not contribute to the estimated hazard ratio until day 365. The estimated hazard ratio for CLINIC at 100 days is exp(-0.460) = 0.632 while the estimated hazard ratio for CLINIC at 400 days is exp((-0.460) + (-1.369)) = 0.161.

It may be of interest to define two heaviside functions (with CLINIC) and not include CLINIC in the model. This is essentially the same model as the one described above with one heaviside function. However, the coding of two heaviside functions makes it somewhat computationally more convenient for estimating the two hazard ratios for CLINIC (HR for <365 days and HR for ≥ 365 days). Unfortunately, SPSS allows just one time-dependent variable (i.e., T_COV_) using the point and click approach. However, by examining the code created for the single heaviside function, there is only a slight adjustment needed to create code for two heaviside functions. The following code creates two heaviside functions (called HV1 and HV2) and runs a model containing PRISON, DOSE, HV1, and HV2:

```
TIME PROGRAM.
COMPUTE hv1= (T_ < 365)* clinic.
COMPUTE hv2= (T_ >= 365)* clinic.

COXREG
   survt /STATUS=status(1)
   /METHOD=ENTER prison dose hv1 hv2
   /CRITERIA=PIN(.05) POUT(.10) ITERATE(20).
```

The output follows:

Variables in the Equation

	B	SE	Wald	df	Sig.	Exp(B)
PRISON	.378	.168	5.030	1	.025	1.459
DOSE	−.036	.006	30.450	1	.000	.965
HV1	−.460	.255	3.241	1	.072	.632
HV2	−1.828	.386	22.439	1	.000	.161

The parameter estimates for HV1 and HV2 can be used directly to obtain the estimated hazard ratio for CLINIC=2 vs CLINIC=1 before and after 365 days. The estimated hazard ratio for CLINIC at 100 days is exp(-0.460) = 0.632 and the estimated hazard ratio for CLINIC at 400 days is expexp(-1.828) = 0.161. These results are consistent with the estimates obtained from the previous model with one heaviside function.

D. R Software

R is available for free and can be downloaded from the Comprehensive R Archive Network (CRAN) at its home site at http://www.r-project.org/. Analyses are carried out in R by applying functions on R data (stored as R objects). R functions are stored in packages. Only when a package is loaded, its contents are available. The base packages are installed when you download R. Packages that are not base packages need to be installed separately.

Once you open R, you'll see a prompt: Type 1+1 and press enter. You'll (hopefully) see the answer 2 returned at the line below. Alternatively, you can type commands in a script by clicking on File → New script. A new script window will open up. By typing commands in this window, you can submit batches of code at one time by highlighting the code and clicking on Edit → Run line or by clicking on Edit → selection. Programming in a script window serves a similar function as the program editor in SAS or the Do-file Editor in Stata, in that code can be submitted as a block rather than one line at a time.

To see which packages are installed at your site, type and enter **library( )**. To run many of the functions needed to perform survival analyses, you will need to install the survival package (not generally a base package).

To install the survival package, click on Packages → Install package(s). You will see a heading called CRAN mirror with a listing of many different countries under that heading. Click on one of these (e.g., USA (AZ)) and then scroll down and click on survival and then click OK. The survival package (with its many survival functions) should now be installed. Type **library(survival)** and press enter, and the survival package will be ready to use. As a check, type the word kidney and hit enter. A dataset called kidney (which is part of the survival package) should print on your screen. Once the survival package is installed, you do not have to reinstall it each session. However, you will need to type **library(survival)** each session before you run the survival functions contained in the package.

Before discussing survival analyses in R, it may be useful to give a brief overview on some of the ways data are stored in R. In particular, we describe four classes of data storage: **vectors**, **matrices**, **dataframes**, and **lists**. If you type and enter the code below, R will create a numerical vector with five elements:

```
c(1,7,12,6,3)
```

The **c** function combines its arguments to form a vector. We can store this vector as an object under the name (identifier), **x1**:

```
x1=c(1,7,12,6,3)
```

Type **x1** and press enter, and you will see the vector **x1** printed as output. The code and output are shown below:

```
x1
1 7 12 6 3
```

We can identify elements from the vector **x1** by placing brackets **[ ]** after **x1**. For example, **x1[2]** will identify the 2nd element of **x1**. The code **x1[1:3]** will identify the first three elements of **x1** and the code **x1[x1>6]** will identify the elements in **x1** greater than 6. The code and output for these three examples follow:

```
x1[2]
7
x1[1:3]
1  7  12
x1[x1>6]
7  12
```

The **:** operator (used in the 2nd example) creates a sequence of integers incremented by 1. Next, we create four more vectors called **x2**, **x3**, **x4**, and **x5**:

```
x2=2*(1:5)
x3=2*x1 + x2
x4=x1>z6
x5=c("blue","green","red","green","purple")
```

The code **x2=2*(1:5)** creates the vector 2, 4, 6, 8, and 10 which we named **x2**. The vector **x3** results from arithmetic operations of **x1** and **x2** ($2 \times$ **x1** $+$ **x2**). The vectors **x1**, **x2**, and **x3** are each numeric vectors. If you apply the **mode** function on **x1** (i.e., type **mode(x1)** and press enter), it will return the word "numeric" as output. The vector **x4** is a logical vector. The **mode** function will return the word "logical" if you submit the code **mode(x4)**. The elements of a vector of mode logical are either "TRUE" or "FALSE." Enter the code **x4** (output below):

```
x4
FALSE TRUE TRUE FALSE FALSE
```

The 2nd and 3rd elements of **x4** are TRUE because the 2nd and 3rd elements of **x1** are greater than 6. The vector **x5** is a character vector. R is case sensitive, so naming the vector **x5** is not the same as naming it **X5**.

We can create a numeric **matrix** (called **y**) using the vectors **x1**, **x2**, and **x3** as columns of the matrix by applying the **cbind** function:

```
y=cbind(x1,x2,x3)
```

Enter the code **class(y)** and the word "matrix" will be returned as output. Enter the code **mode(y)** and the word "numeric" will be returned since **y** is a numeric matrix. You cannot mix numeric and character vectors in a matrix.

Type **y** and press enter, and the matrix will be printed (shown below):

```
y
     x1 x2 x3
[1,]  1  2  4
[2,]  7  4 18
[3,] 12  6 30
[4,]  6  8 20
[5,]  3 10 16
```

A dataframe provides a more general class of data storage than a matrix in R because a dataframe can contain a mix of numeric, character, and logical variables. A dataframe in R is similar to a dataset in Stata, SAS, or SPSS, in that it can store different types of variables. The **data.frame** function can be used to combine vectors and matrices as follows:

z=data.frame(x1,x2,x3,x4,x5) or, equivalently, **z=data.frame(y,x4,x5)**

Type **z** and press enter, and the dataframe will be printed (shown below):

```
z
   x1 x2 x3    x4     x5
1   1  2  4 FALSE   blue
2   7  4 18  TRUE  green
3  12  6 30  TRUE    red
4   6  8 20 FALSE  green
5   3 10 16 FALSE purple
```

Brackets [] can be used to access particular rows and/or columns of a dataframe or matrix. Enter the code: **z[2,5]** and the 2nd row, 5th column will be printed from the matrix **z** (the element "green" in this example). If you want to access the first three rows (observations) of the fifth column, type the code **z[1:3,5]** or equivalently **z[c(1,2,3),5]**. If you want to access the entire 5th column, enter **z[,5]**. Alternatively, since the 5th column (or variable) is named **x5**, you can access the entire 5th column by entering the code **z$x5**. The **$** in this example points to the variable named **x5** from the dataframe named **z**.

A list offers a more general type of data storage than the vector, matrix, or dataframe, and can include any of those data objects as part of the list. The following code creates a list called **w** that contains a character vector of length 2 as its first element, the vector **x1** as its second element, the matrix **y** as its third element, and the dataframe **z** as its fourth element:

w=list(c("hello","good-bye"),x1,y,z)

Double brackets **[[ ]]** can be used to access particular elements of a list. If you want to access the dataframe **z** from the list **w**, enter the code **w[[4]]** since **z** is the fourth element of **w**. If you want to access the first row third column of the fourth element of **w** from the list, enter the following code:

w[[4]] [1,3]

The 1st row 3rd column of the 4th element of w has the value 4.

Survival Functions in R

Once the survival package has been installed you will have access to the survival functions needed to perform the survival analyses in this appendix. Enter the code **library (survival)** each session to access these functions. Some of the key survival functions are listed below:

Surv – Used to define the "time-to-event" and "status" outcome variables. This function creates a survival object that can be used as the outcome variable for other survival functions in R

survfit – Produces KM or Cox-adjusted survival estimates or survival estimates from a previously fitted parametric model

survdiff – Used to perform statistical tests for the equality of survival functions across strata

coxph – Used to run a Cox PH model, a stratified Cox model, or an extended Cox model

cox.zph – Performs statistical tests on the PH assumption based on Schoenfeld residuals

survSplit – Creates a new dataset in the counting process format, with a start time, stop time, and event status for each record. Splits single observations into multiple observations given survival data and specified cut times

survreg – Used to run parametric survival models

Generic functions in R such as the **summary** function and the **plot** function are often used in conjunction with these survival functions in order to produce survival estimates and plots.

R documentation (online help) for these functions can be obtained by typing and submitting a question mark and then the name of the function as one word. For example, to access R documentation on the **coxph** function, submit the code **?coxph**.

The survival analyses demonstrated in R are as follows:

1. Estimating survival functions (unadjusted) and comparing them across strata.
2. Assessing the PH assumption using graphical approaches.
3. Running a Cox PH model.
4. Running a stratified Cox model.
5. Assessing the PH assumption with a statistical test.
6. Obtaining Cox-adjusted survival curves.

7. Running an extended Cox model.
8. Running parametric models.
9. Running frailty models.
10. Modeling recurrent events.

We use the addicts dataset for illustration. The **load** function is used to access an R dataframe that has been saved as a file. Suppose the addicts dataset has been saved on your C drive as C:\craddicts.rda. The following code will load the addicts data:

```
load("C:\craddicts.rda")
```

To print the addicts dataset, enter the code:

```
addicts
```

To print the first five observations, enter the code:

```
addicts[1:5, ]
```

All 6 variables (columns) are printed because there was no entry after the comma. Equivalently, we could have entered the code **addicts[1:5,1:6]**. The output follows:

```
  id clinic status survt prison dose
1  1      1      1   428      0   50
2  2      1      1   275      1   55
3  3      1      1   262      0   55
4  4      1      1   183      0   30
5  5      1      1   259      1   65
```

The time-to-event variable in the addicts dataset is named SURVT and the variable indicating whether a subject had an event or was censored is named STATUS. The function **Surv** creates a survival object in R linking these two outcome variables (code shown below):

```
Surv(addicts$survt,addicts$status==1)
```

The first argument is the time-to-event variable which is accessed from the addicts dataframe with the **$** notation (**addicts$survt**). The second argument (**addicts$status==1**) indicates an event occurs (as opposed to a censorship) when the status variable equals 1. Notice that two equal signs are used to express equality. A single equal sign is used to designate assignment in R. A portion of the output from the **Surv** function is shown below:

```
 [1]  428  275  262  183  259  714  438  796+ 892  393  161+ 836
[13] 523  612  212  399  771  514  512  624   209  341  299  826+
[25] 262  566+ 368  302  602+ 652  293  564+ 394  755  591  787+
```

The output above shows the survival times for the first 36 subjects in the addicts data (out of 238). A plus (+) sign after their time indicates censorship rather than event.

This survival object created by the **Surv** function is often used in R as the response variable for survival analyses. Next we demonstrate survival analyses in R by specific topics.

1. ESTIMATING SURVIVAL FUNCTIONS (UNADJUSTED) AND COMPARING THEM ACROSS STRATA

Kaplan-Meier survival estimates are obtained in R with the use of three functions. The **Surv** function (described above) is used within the **survfit** function, which is then used within the **summary** function. The code follows:

```
summary(survfit(Surv(addicts$survt,addicts$status==1)~1))
```

To better understand how this code works we'll break down each function. The code: **Y=Surv(addicts$survt, addicts$status==1)** creates a survival object called **Y** that is used as the response variable in the analysis. Now consider the code **Y~1** This syntax is called a formula. Formulas are used as arguments in many functions in R, particularly those that specify statistical models. **Y~1** requests an intercept only model. In other words, we are not conditioning on any other variable. Later in this section we stratify on the variable CLINIC and use the formula **Y~ addicts$clinic**. A formula needs to be supplied as the argument of the **survfit** function (shown below):

```
kmfit1=survfit(Y~1)
```

An object, which we named **kmfit1**, was created with the **survfit** function. Enter the code **kmfit1** and press enter (output shown below):

```
kmfit1
```

records	n.max	n.start	events	median	0.95LCL	0.95UCL
238	238	238	150	504	399	560

The output contains descriptive information on the number of records, the number at risk at time 0, the number of events, and the median estimated survival time with a 95% confidence interval. The **summary** function can then be used to get Kaplan-Meier survival estimates for all event times. The code **summary(kmfit1)** is equivalent to the code **summary(survfit(Surv(addicts$survt, addicts$status==1)~1))** shown above. The output follows:

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
7	236	1	0.996	0.00423	0.9875	1.000
13	235	1	0.992	0.00597	0.9799	1.000
17	234	1	0.987	0.00729	0.9731	1.000
19	233	1	0.983	0.00840	0.9667	1.000
26	232	1	0.979	0.00937	0.9606	0.997
29	229	1	0.975	0.01026	0.9546	0.995
30	228	1	0.970	0.01107	0.9488	0.992
.						
.						
.						
821	20	2	0.225	0.03675	0.1635	0.310
836	17	1	0.212	0.03690	0.1506	0.298
837	16	1	0.199	0.03689	0.1380	0.286
857	14	1	0.184	0.03688	0.1246	0.273
878	13	1	0.170	0.03667	0.1116	0.260
892	10	1	0.153	0.03675	0.0958	0.245
899	9	1	0.136	0.03639	0.0807	0.230

The **summary** function can also produce survival estimates for specified survival times (e.g., at day 365) with the **times=** option. Code and output follow:

summary(kmfit1,times=365)

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
365	122	87	0.606	0.0331	0.545	0.675

If we wish to stratify by the variable CLINIC and compare the Kaplan-Meier survival estimates at specified times, we can first create an object (called **kmfit2**, where the name is arbitrary) from the **survfit** function:

kmfit2=survfit(Y~addicts$clinic)

To get survival estimates at specified times (every 100 days) for each level of CLINIC, enter the code:

summary(kmfit2,times=c(0,100,200,300,400,500,600,700,800,900,1000))

The output follows:

```
                addicts$clinic=1
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0    163       0   1.0000  0.0000       1.00000        1.000
 100    137      20   0.8746  0.0262       0.82467        0.928
 200    110      20   0.7420  0.0353       0.67601        0.814
 300     87      20   0.6046  0.0399       0.53120        0.688
 400     68      14   0.5025  0.0415       0.42741        0.591
 500     53       9   0.4319  0.0418       0.35719        0.522
 600     30      16   0.2951  0.0403       0.22570        0.386
 700     20       8   0.2113  0.0383       0.14818        0.301
 800     10       8   0.1268  0.0326       0.07660        0.210
 900      1       7   0.0181  0.0172       0.00283        0.116

                addicts$clinic=2
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   0     75       0    1.000  0.0000         1.000        1.000
 100     66       5    0.932  0.0294         0.876        0.991
 200     58       7    0.832  0.0442         0.750        0.924
 300     50       7    0.730  0.0530         0.633        0.842
 400     43       3    0.685  0.0558         0.584        0.804
 500     39       2    0.653  0.0577         0.549        0.776
 600     27       1    0.634  0.0590         0.528        0.761
 700     19       1    0.606  0.0625         0.495        0.742
 800     11       1    0.575  0.0669         0.457        0.722
 900      7       1    0.517  0.0812         0.380        0.703
1000      3       0    0.517  0.0812         0.380        0.703
```

Survival estimates are supplied for each 100th day. For CLINIC=1, survival times stopped at 900 rather than 1000 as requested because no subject was at risk on day 1000. The second argument of the summary function requesting a vector of survival times could have been equivalently written: **summary(kmfit2,times=100*(0:10))**. The output would be identical if this alternative syntax had been used.

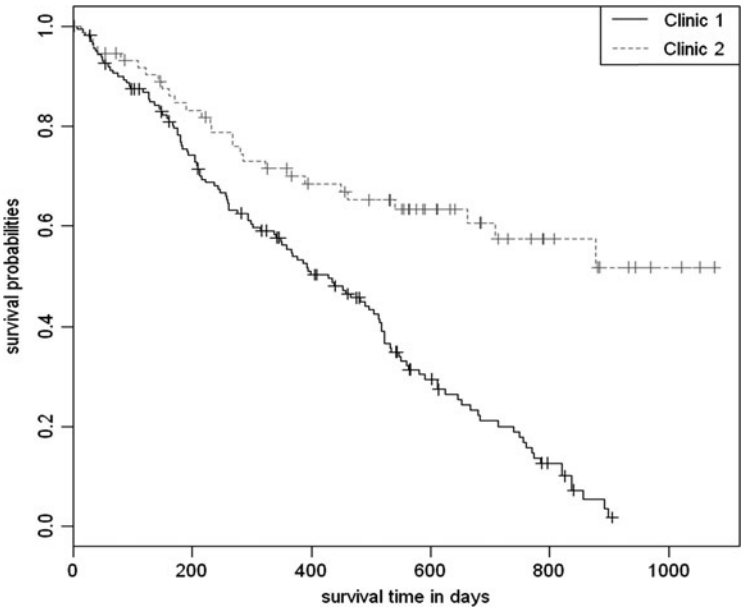
KM survival plots can be obtained using the **plot** function:

```
plot(kmfit2)
```

There are many plotting options that can be applied with the **plot** function. The code below requests different line types (**lty=**) and different colors (**col=**) for CLINIC=1 and CLINIC=2 as well as labels for the X and Y axes (**xlab=** and **ylab=**). If the code **col( )** is submitted, then R returns a list of over 600 colors that can be selected with the **col=** option. The **legend** function is used to add a legend. The first argument, "**topright**," places the legend at the top right part of the graph. The code and output follow:

plot(kmfit2, lty = c("solid", "dashed"), col=c("black","grey"), xlab="survival time in days",ylab="survival probabilities")

legend("topright", c("Clinic 1","Clinic 2"), lty=c("solid","dashed"), col=c("black","grey"))



The plot indicates that subjects from CLINIC=2 have a higher rate of survival than subjects from CLINIC=1.

The **survdiff** function can be used to implement a log rank test on the variable CLINIC (the code follows):

survdiff(Surv(survt,status)~clinic, data=addicts)

The second argument of the **survdiff** function, **data= addicts**, indicates that the variables come from the addicts dataset. Alternatively, you could use the code:

survdiff(Surv(addicts$survt,addicts$status)~addicts$clinic)

As a third alternative, the **attach** function can be used to indicate that all subsequent variable names apply to the addicts dataset (R will search the addicts dataset for variables). The **detach** function can be used to remove a dataset from the search path.

attach(addicts)
survdiff(Surv(survt,status)~clinic)

The output follows:

```
            N Observed Expected (O-E)^2/E (O-E)^2/V
clinic=1 163      122     90.9      10.6      27.9
clinic=2  75       28     59.1      16.4      27.9

 Chisq= 27.9  on 1 degrees of freedom, p= 1.28e-07
```

The log rank statistic is highly significant with a *p*-value of 0.000000128 (i.e., 1.28e-07).

Variations of the log rank test can be obtained by using the **rho=** option as an argument in the **survdiff** function. The contribution of the j^{th} failure time to the test statistic is weighted by $s(t_j)^{rho}$, where $s(t_j)$ represents the KM survival estimates at time t_j. If rho=0, then each failure time is equally weighted since $s(t_j)^0 = 1$ and the resulting test is the log rank test. If rho=1, then the weights for each failure time are the KM survival estimate at that failure time since $s(t_j)^1 = s(t_j)$. This test is equivalent to the Peto & Peto modification of the Gehan–Wilcoxon test. The code and output with rho=1 follows:

survdiff(Surv(survt,status)~clinic,data=addicts,rho=1)

```
            N Observed Expected (O-E)^2/E (O-E)^2/V
clinic=1 163     77.3     61.4      4.08      15.8
clinic=2  75     19.9     35.7      7.03      15.8

 Chisq= 15.8  on 1 degrees of freedom, p= 7.18e-05
```

The results of the test in which rho=1 yield a chi-square value of 15.8 with a p-value of 0.0000718. This is a somewhat different result than the log rank test but still shows a highly significant effect of CLINIC on survival.

A stratified log rank test for CLINIC (stratified by PRISON) can be run with the + **strata(prison)** term included in the model formula. With the stratified approach, the observed minus expected number of events are summed over all failure times for each group within each stratum and then summed over all strata. The code and output follow:

survdiff(Surv(survt,status) ~ clinic + strata(prison),data=addicts)

```
            N Observed Expected (O-E)^2/E (O-E)^2/V
clinic=1 163      122     91.7      10.0      26.9
clinic=2  75       28     58.3      15.8      26.9

 Chisq= 26.9  on 1 degrees of freedom, p= 2.1e-07
```

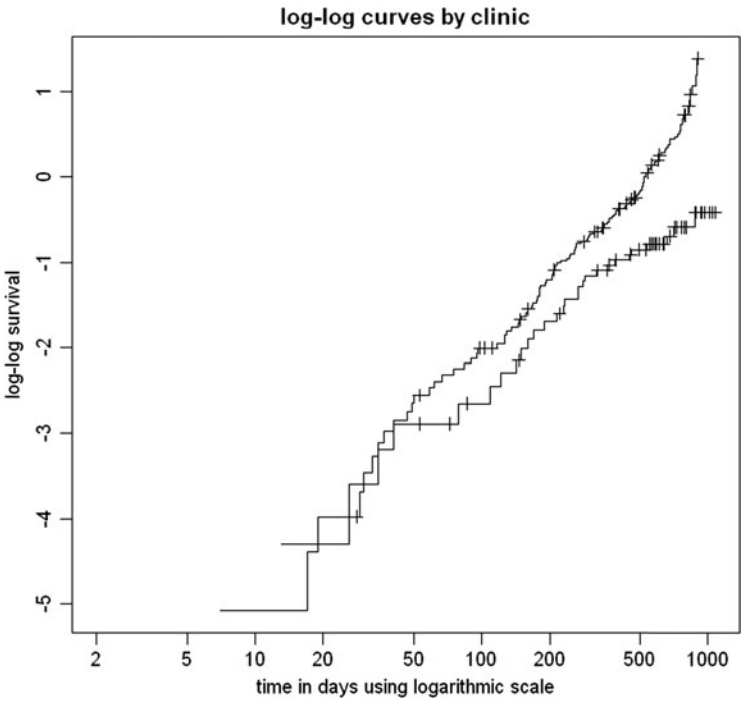
The formula includes the term **+ strata(prison)** in the **survdiff** function. The result of this test is very similar to that obtained from the log rank test without stratifying on PRISON.

2. ASSESSING THE PH ASSUMPTION USING GRAPHICAL APPROACHES

The proportional hazards assumption for CLINIC can be assessed by plotting log-log Kaplan Meier survival estimates against time (or against the log of time) and evaluating whether the curves are reasonably parallel. Recall that a survival object, called **kmfit2**, was created in the previous section with the **survfit**. The code **plot(survfit2)** was used to plot the survival estimates against time. The **fun="cloglog"** option in the **plot** function requests that log-log survival plot be plotted against time (on the log scale). The code follows:

```
plot(kmfit2,fun="cloglog",xlab="time in days using logarithmic scale",ylab="log-log survival", main="log-log curves by clinic")
```

The **xlab=** and **ylab=** request labels for the *x*- and *y*-axes and the **main=** option requests a title. **fun="cloglog"** requests the complimentary log log function. The output follows:



The plot suggests that the proportional hazards assumption is violated as the log-log survival curves are not parallel. The **fun=** option (**fun** denotes function) plots time on a logarithmic scale. It is not so straightforward if you want the log-log survival estimates plotted against time with time not on a logarithmic scale. However, it may be useful to program this task in order to illustrate how analytic output can be saved, manipulated, and plotted. To do that, we first save the survival estimates as an object (which we will call **kmfit3**) using the **summary** function:

```
kmfit3=summary(kmfit2)
```

If we submit the code **names(kmfit3)**, then the column names of the object **kmfit3** are printed. We are interested in the columns that indicate each subject's survival time, KM survival estimate, and level of clinic (1 or 2). With the **names** function, we can see that these columns are called **time**, **surv**, and **strata**. We can examine any of these three columns by submitting **kmfit3$strata**, **kmfit3$time**, or **kmfit3$surv** as code. A dataframe (called **kmfit4**) consisting of these three columns as variables can be created with the **data.frame** function:

```
kmfit4=data.frame(kmfit3$strata,kmfit3$time,kmfit3$surv)
names(kmfit4)=c("clinic","time","survival")
```

The **names** function is used (above) on **kmfit4** to overwrite the default variable names. Next, we'll print the first 5 observations of **kmfit4**:

```
kmfit4[1:5, ]
```

	clinic	time	survival
1	addicts$clinic=1	7	0.9938272
2	addicts$clinic=1	17	0.9876543
3	addicts$clinic=1	19	0.9814815
4	addicts$clinic=1	29	0.9752300
5	addicts$clinic=1	30	0.9689785

We are interested in separating out CLINIC=1 and CLINIC=2. Below, we create two dataframes (**clinic1** and **clinic2**) from **kmfit4**:

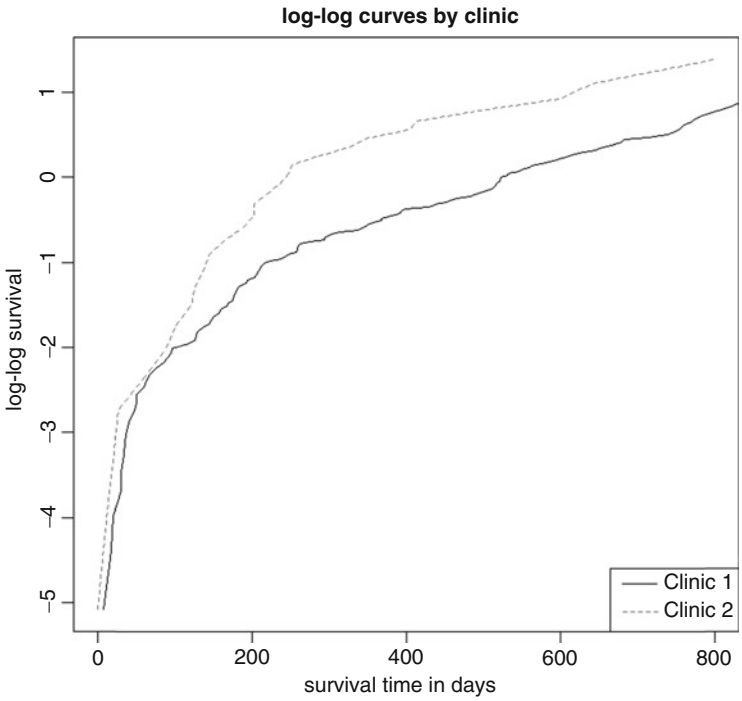
```
clinic1=kmfit4[kmfit4$clinic=="addicts$clinic=1", ]
clinic2=kmfit4[kmfit4$clinic=="addicts$clinic=2", ]
```

The dataframes **clinic1** and **clinic2** contain the survival times and survival estimates for those in CLINIC=1 and CLINIC=2, respectively. We can now use the **plot** function

to plot the log-log survival curves against time (with time not plotted on the log scale). The code follows:

```
plot(clinic1$time,log(-log(clinic1$survival)),xlab="survival time in days",ylab="log-log survival",xlim=c(0,800),col="black",type='l',lty="solid",main="log-log curves by clinic")
par(new=T)
plot(clinic2$time,log(-log(clinic2$survival)),axes=F,xlab="survival time in days",ylab="log-log survival",col="grey50",type='l',lty= "dashed")
legend("bottomright", c("Clinic 1", "Clinic 2"), lty = c("solid", "dashed"),col=c("black","grey50"))
par(new=F)
```

In the first plot, time (**clinic1$time**) is plotted on the x axis, and the log(-log) of survival (**clinc1$survival**) is plotted on the y axis using the dataframe **clinic1**. The code **par (new=T)** requests that the first plot not get erased when the second plot is requested (i.e., the two plots will be overlayed). The **par** function is used to set or query graphical parameters. The second **plot** function is similar to the first except that the data that is plotted are from the dataframe **clinic2**. A legend is added with the **legend** function and finally **par(new=F)** sets the graphical parameter **new** back to its default value of false (so that these plots will be erased when the next plot is requested). The output follows:



The plot suggests that the proportional hazards assumption is violated for CLINIC.

3. RUNNING A COX PH MODEL

The **coxph** function is used to run a Cox proportional hazards model. First, the response variable is created with the **Surv** function and then a Cox PH model containing the variables CLINIC, PRISON, and DOSE is run with the **coxph** function. The code and the **coxph** output follow:

```
Y=Surv(addicts$survt,addicts$status==1)
coxph(Y~ prison + dose + clinic,data=addicts)
```

```
          coef exp(coef) se(coef)     z       p
prison  0.3266     1.386  0.16722  1.95 5.1e-02
dose   -0.0354     0.965  0.00638 -5.54 2.9e-08
clinic -1.0099     0.364  0.21489 -4.70 2.6e-06

Likelihood ratio test=64.6  on 3 df, p=6.23e-14  n= 238
```

The output contains the regression coefficients, the exponentiated coefficients (estimated hazard ratios), as well as the standard errors, z-tests, and corresponding p-values for the coefficients. Additional output including 95% confidence intervals can be obtained by applying the **summary** function to the **coxph** function (code and output shown below):

```
summary(coxph(Y~ prison + dose + clinic,data=addicts))
```

```
           coef exp(coef)  se(coef)      z Pr(>|z|)
prison  0.326555  1.386184  0.167225  1.953   0.0508 .
dose   -0.035369  0.965249  0.006379 -5.545 2.94e-08 ***
clinic -1.009896  0.364257  0.214889 -4.700 2.61e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

       exp(coef) exp(-coef) lower .95 upper .95
prison    1.3862     0.7214    0.9988    1.9238
dose      0.9652     1.0360    0.9533    0.9774
clinic    0.3643     2.7453    0.2391    0.5550

Rsquare= 0.238   (max possible= 0.997 )
Likelihood ratio test= 64.56  on 3 df,   p=6.228e-14
Wald test            = 54.12  on 3 df,   p=1.056e-11
Score (logrank) test = 56.32  on 3 df,   p=3.598e-12
```

The second table of the output gives the estimated hazard ratio, under the column exp(coef), for CLINIC=2 vs CLINIC=1 at 0.3643 with 95% CI (0.2391, 0.5550). Under the column exp(-coef), we see that the estimated hazard ratio for CLINIC=1 vs CLINIC=2 is 2.7453 (the reciprocal of 0.3643).

If any of the events in the data occur at the same time, there are several options for handling ties in the Cox likelihood. R offers three approaches with the **coxph** function 1) the Efron method (the default), 2) the Breslow method, and 3) the exact method. Generally, these methods have little impact on the estimates but model results obtained from different software packages may differ depending on the default tie handling method. R uses the Efron method as the default while Stata, SAS and SPSS use the Breslow method as the default. The **method=** option in the **coxph** function is used to specify the method for handling ties (code shown below, output omitted):

coxph(Y~ prison + dose + clinic,data=addicts, method="efron")
coxph(Y~ prison + dose + clinic,data=addicts, method="breslow")
coxph(Y~ prison + dose + clinic,data=addicts, method="exact")

Next we include two interaction (product) terms with PRISON and test the significance of the interaction terms simultaneously with a likelihood ratio test. The following code creates two objects (called **mod1** and **mod2**) that contain information obtained from the **coxph** function for the no interaction model (**mod1** – the reduced model) and the model with the two interaction terms (**mod2** – the full model)

mod1=coxph(Y ~ prison + dose + clinic,data=addicts)
mod2=coxph(Y ~ prison + dose + clinic + clinic*prison + clinic*dose, data=addicts)

Enter the code **mod2** to see the output with the interaction terms (code and output shown below):

mod2

```
                coef exp(coef) se(coef)      z     p
prison         1.1920     3.294   0.5414  2.202 0.028
dose          -0.0193     0.981   0.0194 -0.998 0.320
clinic         0.1747     1.191   0.8931  0.196 0.840
prison:clinic -0.7380     0.478   0.4315 -1.710 0.087
dose:clinic   -0.0139     0.986   0.0143 -0.967 0.330

Likelihood ratio test=68.1  on 5 df, p=2.52e-13  n= 238
```

The rest of this section gets a little complicated but we include it to demonstrate how analytic output in R can be accessed and then manipulated.

The objects **mod1** and **mod2** contain information that we may wish to utilize. Type the code **names(mod2)** to see the names of the elements in **mod2** (code and output shown below):

names(mod2)

```
 [1] "coefficients"      "var"               "loglik"
 [4] "score"             "iter"              "linear.predictors"
 [7] "residuals"         "means"             "method"
[10] "n"                 "terms"             "assign"
[13] "wald.test"         "y"                 "formula"
[16] "call"
```

The 3rd element of **mod2** is named "loglik." We can access the data stored under this name by entering the code **mod2$loglik** or equivalently **mod2[[3]]** since loglik is the 3rd element in the list (code and output follow):

mod2$loglik
-705.6619 -671.5997

The second element of **mod2$loglik** is −671.5997, which is the log likelihood of the two interaction terms in the model. The first element of -704.6619 is the log likelihood of a model that contains none of the predictors (not of interest right now).

Next we wish to perform a likelihood ratio test on the two interaction terms. To calculate the test statistic, we need to subtract the log-likelihood of the full model (with the interaction terms) from the reduced model (without the interaction terms) and multiply that difference by negative 2. We can obtain this by entering the following code:

(-2)*(mod1$loglik[2]-mod2$loglik[2])

We get the output: 3.605457, which is the likelihood ratio test statistic. Under the null, this test statistic follows a chi-square distribution with two degrees of freedom. We can use the **pchisq** function to obtain a p-value for this test. The code **1** × **pchisq(3.605457,2)** returns the p-value for a two degree of freedom chi-square test. In summary, the following code will produce a p-value for the likelihood ratio test (output follows):

LRT=(-2)*(mod1$loglik[2]-mod2$loglik[2])
Pvalue = 1 - pchisq(LRT, 2)
Pvalue
0.1648485

The *p*-value of 0.168485 is not significant at the 0.05 level of significance.

One of the powerful features of R is the ability for users to define their own functions. We illustrate this feature by defining our own function that performs a likelihood ratio test from two Cox models (a full and reduced model). The following code creates a function which we call **lrt.surv**. This function requests the user to enter three arguments (1) the name of the full model, (2) the name of the reduced model, and (3) the degrees of freedom for the test. The function will return the p-value for the likelihood ratio test.

An R function called **function** is used to define a new function. The three arguments for this function we call **mod.full**, **mod.reduced**, and **df**. The code that R will use to calculate the function output is contained within brackets **{ }** after the arguments are listed. The argument in the R function called **return** informs R of the output that we wish to return from this function (in this example, the p-value for the likelihood ratio test). The code follows:

```
lrt.surv=function(mod.full,mod.reduced,df) {
lrts=(-2)*(mod.full$loglik[2]- mod.reduced$loglik[2])
pvalue=1-pchisq(lrts,df)
return(pvalue)
}
```

Once this code is submitted any user can obtain a p-value from a likelihood ratio test from two Cox models by invoking the function **lrt.surv**. We invoke this new function by performing the same likelihood ratio test that we previously ran for the objects **mod1** and **mod2**. The code and output follow:

```
lrt.surv(mod1, mod2, 2)
0.1648485
```

The p-value is the same as that which we obtained earlier. The function **lrt.surv** is more general and now available to simply obtain p-values for *other* likelihood ratio tests that compare two (full and reduced) Cox models.

4. RUNNING A STRATIFIED COX MODEL

If the proportional hazards assumption is violated for the variable CLINIC but met for PRISON and DOSE, a stratified Cox model can be performed with CLINIC the stratified variable. The **coxph** function includes a **strata()** option in the model formula. First we define the response variable **Y** with the **Surv** function and then the **coxph** function is used to run a stratified Cox model (code and output shown below):

Y=Surv(addicts$survt,addicts$status==1)
coxph(Y∼ prison + dose + strata(clinic),data=addicts)

```
          coef exp(coef) se(coef)     z       p
prison  0.3896     1.476  0.16893  2.31 2.1e-02
dose   -0.0351     0.965  0.00646 -5.43 5.6e-08

Likelihood ratio test=33.9 on 2 df, p=4.32e-08  n= 238
```

Interaction terms for CLINIC can be included directly in the model formula by including product terms using the **:** operator (**clinic:prison** and **clinic:dose**) (code and output follow):

coxph(Y∼ prison + dose + clinic:prison + clinic:dose + strata(clinic),data=addicts)

```
                  coef exp(coef) se(coef)       z     p
prison         1.08584     2.962   0.5386  2.0159 0.044
dose          -0.03464     0.966   0.0198 -1.7495 0.080
prison:clinic -0.58299     0.558   0.4281 -1.3617 0.170
dose:clinic   -0.00116     0.999   0.0146 -0.0799 0.940

Likelihood ratio test=35.8  on 4 df, p=3.22e-07  n= 238
```

Suppose we wish to estimate the hazard ratio for PRISON=1 vs. PRISON=0 for CLINIC=2. This hazard ratio can be estimated by exponentiating the coefficient for prison plus 2 times the coefficient for the CLINIC* PRISON interaction term. This expression is obtained by substituting the appropriate values into the hazard in both the numerator (for PRISON=1) and denominator (for PRISON=0) (see below):

$$HR = \frac{h_0(t)\exp[1\beta_1 + \beta_2 DOSE + (2)(1)\beta_3 + \beta_4 CLINIC \times DOSE]}{h_0(t)\exp[10 + \beta_2 DOSE + (2)(0)\beta_3 + \beta_4 CLINIC \times DOSE]} = \exp(\beta_1 + 2\beta_2).$$

The resulting hazard ratio, $\exp(\beta_1 + 2\beta_2)$, is an exponentiated linear combination of parameters. Unfortunately, R does not have a **lincom** command that Stata provides or an **estimate** statement that SAS provides in order to calculate a linear combination of parameter estimates. However an approach that can be used in any statistical software package for such a situation is to recode the variable(s) of interest such that the desired estimate is no longer a linear combination of parameter estimates.

In this example, we are interested in a hazard ratio PRISON=1 versus PRISON=0 for CLINIC=2. We can define a new variable CLINIC $\times$ 2 so when CLINIC=2, CLINIC $\times$ 2=0.

Addicts$clinic2=addicts$clinic-2
summary(coxph(Y~ prison+dose+clinic2:prison+
clinic2:dose+strata(clinic2),data=addicts))

The first line of code defines a new variable CLINIC2. CLINIC2 is used in the stratified Cox model rather than CLINIC. We are interested in the hazard ratio for PRISON=1 vs PRISON=0 for CLINIC2=0. When CLINIC2=0, the product terms cancel and the hazard ratio reduces to $\exp(\beta_1)$.

The second line of code applies the **summary** function to the **coxph** function. The summary function applied in this way produces additional output including 95% confidence intervals for the hazard ratios. The output follows:

```
 n= 238

                    coef exp(coef)  se(coef)      z Pr(>|z|)
prison          -0.080143  0.922985  0.384305 -0.209  0.83481
dose            -0.036964  0.963711  0.012346 -2.994  0.00275 **
prison:clinic2 -0.582989  0.558227  0.428135 -1.362  0.17329
dose:clinic2   -0.001164  0.998837  0.014570 -0.080  0.93632
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

               exp(coef) exp(-coef) lower .95 upper .95
prison            0.9230      1.083    0.4346    1.9603
dose              0.9637      1.038    0.9407    0.9873
prison:clinic2    0.5582      1.791    0.2412    1.2919
dose:clinic2      0.9988      1.001    0.9707    1.0278

Rsquare= 0.14   (max possible= 0.994 )
Likelihood ratio test= 35.77  on 4 df,   p=3.222e-07
Wald test            = 34.09  on 4 df,   p=7.138e-07
Score (logrank) test = 34.97  on 4 df,   p=4.706e-07
```

The estimate for $\exp(\beta_1)$ can be found in the second table, exp(coef) for prison = 0.9203. The lower and upper confidence limits are 0.4346 and 1.9603, respectively. If we did not recode the variable CLINIC the problem would have been more complicated in that we would have had to use variance–covariance matrix (which can be obtained with the **vcov** function) to calculate a 95% confidence interval for this hazard ratio.

5. ASSESSING THE PH ASSUMPTION WITH A STATISTICAL TEST

The **cox.zph** function is designed to perform a statistical test on the proportional hazards assumption. This statistical test is a test of correlation between the Schoenfeld residuals and survival time (or ranked survival time). A correlation of zero supports the proportional hazards assumption (the null hypothesis). First, we define the response variable **Y** with the **Surv** function and then the **coxph** function is used to run a Cox proportional hazards model with the variables PRISON, DOSE, and CLINIC:

```
Y=Surv(addicts$survt,addicts$status==1)
mod1=coxph(Y~prison + dose + clinic, data=addicts)
```

The object called **mod1** is created from the **coxph** function. This object is the first argument for the **cox.zph** function. The code to run the test of the proportional hazards assumption follows:

```
cox.zph(mod1,transform=rank)
```

The second argument requests that ranked survival times be tested against the Schoenfeld residuals rather than the actual survival times (the default). The output follows:

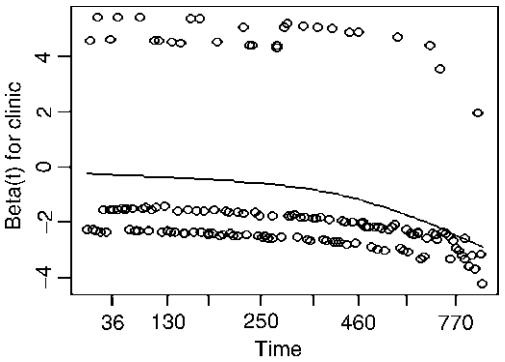
```
           rho  chisq       p
prison -0.0462  0.322 0.57068
dose    0.0905  1.096 0.29521
clinic -0.2498 10.495 0.00120
GLOBAL      NA 12.425 0.00606
```

The output shows that the correation between the Schoenfeld residuals for the variable CLINIC (3rd row) and ranked survival time is -0.2498 with a p-value of 0.00120. The significant p-value offers evidence that the proportional hazards assumption is not satisfied for the variable CLINIC. The p-values for PRISON and DOSE are not significant suggesting that there is not enough evidence to

reject the proportional hazards assumption for PRISON and DOSE. The global test (4^{th} row) tests the proportional hazards assumption for the entire model (i.e., for all three predictor variables simultaneously) and is significant with p = 0.00606. The global test offers evidence that the proportional hazards assumption is violated for this model.

We can plot the Schoenfeld residuals against each individual's failure time with the **plot** function and an object created from the **cox.zph** function as the first argument. The argument **var=clinic**, specifies that the residuals should pertain the variable CLINIC. The argument **se=F**, suppresses the printing of confidence limits for the fitted curve. The code and output follow:

plot(cox.zph(mod1,transform=rank),se=F,var='clinic')



If the PH assumption is met then the fitted curve should look horizontal because the Schoenfeld residuals would be independent of survival time. However, the fitted curve slopes downward.

6. OBTAINING COX-ADJUSTED SURVIVAL CURVES

Cox adjusted survival estimates and plots can be obtained by applying the **summary** or **plot** function to an object created from the function **survfit**. The first step is to run the Cox model with the **coxph** function:

Y=Surv(addicts$survt,addicts$status==1)
mod1=coxph(Y ~ prison + dose + clinic, data=addicts)

Adjusted survival curves generally depend on the pattern of covariates. Suppose we are interested in plotting the survival curve for the pattern PRISON=0, DOSE=70, and CLINIC=2. First, we need to create a dataset (or dataframe)

with the **data.frame** function with one observation. Code and output follows:

pattern1=data.frame(prison=0,dose=70,clinic=2)

pattern1

```
prison dose clinic
0        70      2
```

This one observation dataframe is called **pattern1**. To obtain Cox adjusted survival estimates apply the **survfit** function within the **summary** function as shown below:

summary(survfit(mod1,newdata=pattern1))

The first argument of the **survfit** function is the object called **mod1** created with **coxph** function. The second argument supplies the dataframe containing the pattern of covariates of interest (called **pattern1**).

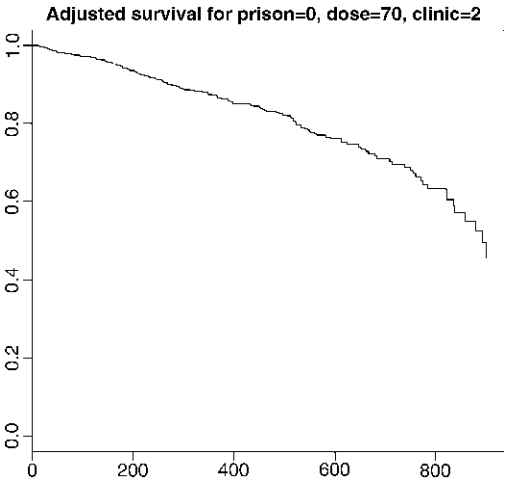
The output follows:

```
time n.risk n.event survival std.err lower 95% CI upper 95% CI
   7    236       1    0.999 0.00105        0.997        1.000
  13    235       1    0.998 0.00154        0.995        1.000
  17    234       1    0.997 0.00193        0.993        1.000
  19    233       1    0.996 0.00229        0.991        1.000
  26    232       1    0.995 0.00263        0.990        1.000
  29    229       1    0.994 0.00296        0.988        1.000
  30    228       1    0.993 0.00328        0.986        0.999
  33    227       1    0.992 0.00359        0.985        0.999
  35    226       2    0.989 0.00419        0.981        0.998
      .
      .
      .
 857     14       1    0.549 0.07953        0.414        0.730
 878     13       1    0.526 0.08204        0.387        0.714
 892     10       1    0.496 0.08580        0.354        0.697
 899      9       1    0.456 0.09090        0.309        0.674
```

To obtain a Cox adjusted survival curve for the same pattern of covariates, apply the plot function in the same manner that the summary function was applied above. The code follows:

plot(survfit(mod1,newdata=pattern1),conf.int=F,main="Adjusted survival for prison=0, dose=70, clinic=2")

The **conf.int=F** option suppresses the plotting of the confidence limits. The option **conf.int=T** (the default) would plot the 95% confidence limits. The **main=** option requests a title for the plot. The output follows:



Stratified Cox adjusted survival curves can be obtained by first running a stratified Cox model (stratified by CLINIC):

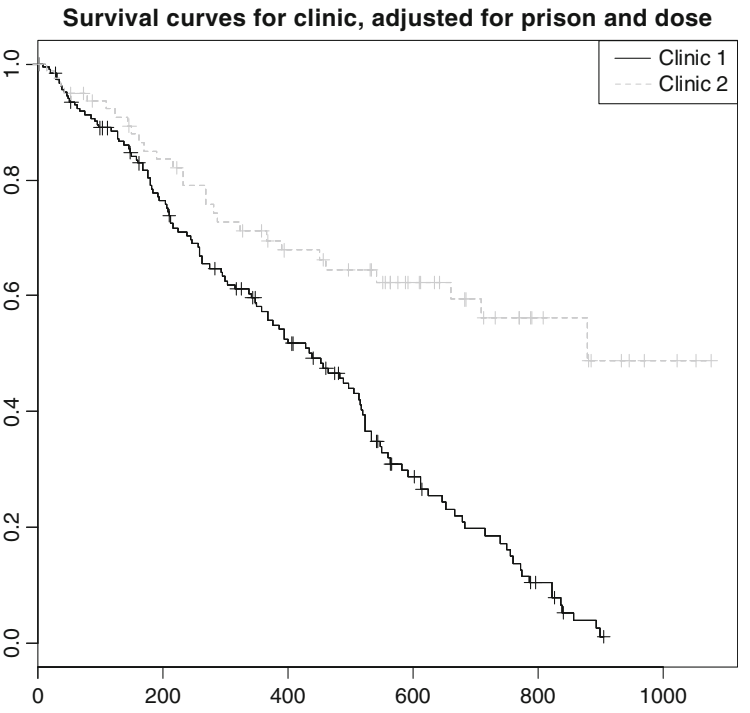
mod3=coxph(Y∼ prison + dose + strata(clinic),data=addicts)

To obtain stratified Cox adjusted curves controlling for PRISON and DOSE, we create a one observation dataframe with the mean values of 0.46 for PRISON and 60.4 for DOSE:

pattern2=data.frame(prison=.46,dose=60.40)

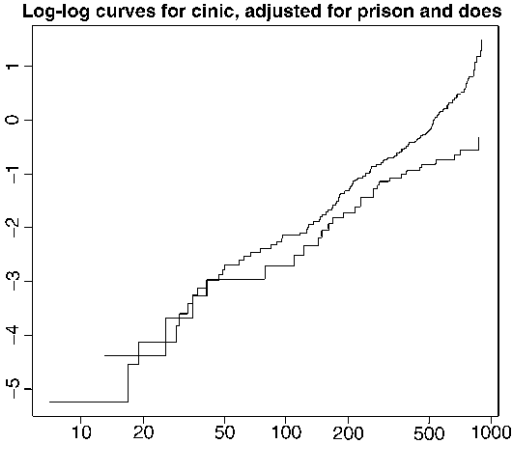
Now apply the **plot** function to the **survfit** function as shown in the last example. The code and output follow:

plot(survfit(mod3,newdata=pattern2), conf.int=F, lty = c("solid", "dashed"), col=c("black","grey"), main="Survival curves for clinic, adjusted for prison and dose")
legend("topright", c("Clinic 1","Clinic 2"), lty=c("solid","dashed"), col=c("black","grey"))



The **fun=** option in the plot function can be used to plot log-log survival curves. The code and output follow:

```
plot(survfit(mod3,newdata=pattern2),fun="cloglog", main=
"Log-log curves for cinic, adjusted for prison and dose")
```



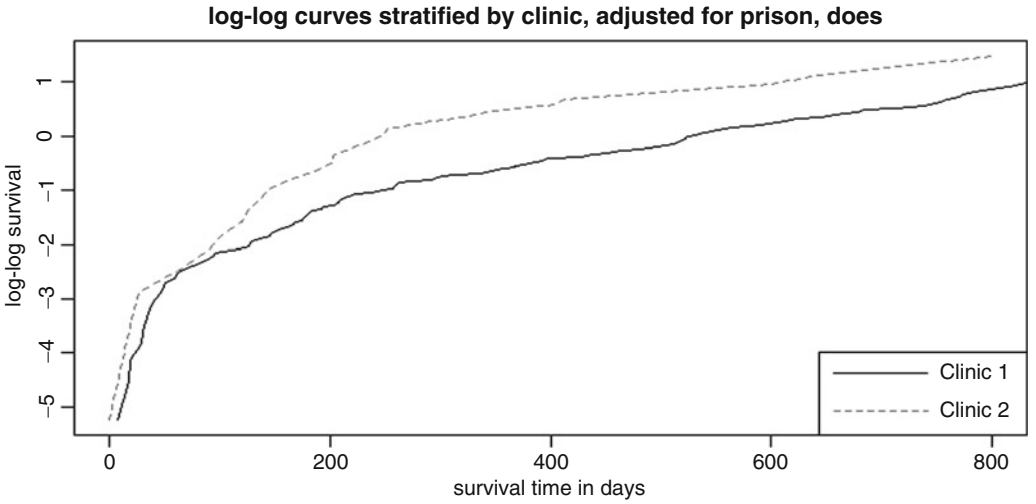
The **fun=** option plots time on a logarithmic scale. It is not so straightforward if you want the log-log plot against time with time not on a logarithmic scale. This was shown in Sect. 2 for KM log log curves. First, the adjusted survival

estimates can be saved in an object we'll call **sum.mod3** (shown below):

```
sum.mod3=summary(survfit(mod3,newdata=pattern2))
```

Now, if it desired to plot the log-log plot against time with time not on a logarithmic scale, similar code can be used as was shown in Section 2, except replace the object we had called **kmfit3** in Section 2 with the object created above, called **sum.mod3**. The code and plot follows:

```
sum.mod4=data.frame(sum.mod3$strata,sum.mod3
$time,sum.mod3$surv)
colnames(sum.mod4)=c("clinic","time","survival")
clinic1=sum.mod4[sum.mod4$clinic=="clinic=1", ]
clinic2=sum.mod4[sum.mod4$clinic=="clinic=2", ]

plot(clinic1$time,log(-log(clinic1$survival)),xlab="survival
time in days",ylab="log-log survival",xlim=c(0,800),col=
"black",type='l',lty="solid", main="log-log curves stratified by
clinic, adjusted for prison, dose")

par(new=T)

plot(clinic2$time,log(-log(clinic2$survival)),axes=F,xlab=
"survival time in days",ylab="log-log survival",col="grey50",
type='l',lty="dashed")

legend("bottomright", c("Clinic 1", "Clinic 2"), lty = c("solid",
"dashed"),col=c("black","grey50"))

par(new=F)
```



7. RUNNING AN EXTENDED COX MODEL

In contrast to Stata, SAS, and SPSS, in order to run an extended Cox model in R, the analytic dataset must be in the counting process (start, stop) format. Unfortunately, the addicts dataset is not in that format, so it needs to be altered in order to include a time-varying covariate. This can be accomplished with the **survSplit** function. The **survSplit** function can create a dataset that provides multiple observations for the same subject allowing a subject's covariate to change values from observation to observation. The user supplies the time cutpoint(s).

The most general choice for time cutpoints that can accommodate the modeling of any time-varying covariate is a vector of time cutpoints that includes all event times in the data. The variable SURVT in the addicts dataset contains each individual's time-to-event or time-to-censorship. The following code creates a new analytic dataset (called **addicts.cp**) which puts the addicts data in the counting process format using the **survSplit** function:

```
addicts.cp=survSplit(addicts,cut=addicts$survt[addicts$status==1],
end="survt", event="status",start="start",id="id")
```

The first argument of the **survSplit** function specifies the dataframe (addicts) to be manipulated into the counting process format. The **cut= addicts$survt[addicts$status==1]** option specified that the time cutpoints are indicated by the SURVT variable subsetted where the STATUS variable equals 1 (i.e., keeping the event times but omitting censorship times). The **event="status"** option specifies STATUS as the variable indicating whether the individual had an event or was censored. The **start="start"** option creates a new variable called START. This newly defined variable for the starting times for each observation is necessary for the data to be in counting process (start, stop) format. The **end="survt"** option defines SURVT as the stop variable (i.e., the time-to-event variable). The option **id="id"** indicates that ID is the variable that identifies each individual. The **survSplit** function creates multiple observations for individuals at risk at multiple time points. The dataset **addicts.cp** created above contains 18,708 observations from the 238 observations in the addicts dataset (use the **nrow** function and the code **nrow(addicts.cp)**) to return the number of observations.

Suppose the PH assumption was violated for the variable DOSE and we were interested in defining a time-varying covariate as the product of DOSE and the natural log of

time (SURVT). This variable can easily be defined if the dataset is in counting process form with time cutpoints at each event time as shown below:

```
addicts.cp$logtdose=addicts.cp$dose*log(addicts.cp$survt)
```

We now have a new variable in the dataset (called LOGTDOSE=ln(DOSE)*T) that varies over time. We print the dataset for one individual (id=106) who had an event at time=35 days. Rather than print all the variables, we request a subset of them with the **c** function:

```
addicts.cp[addicts.cp$id==106,c('id','start','survt','status',
'dose','logtdose')
```

id	start	survt	status	dose	logtdose
106	0	7	0	40	77.8364
106	7	13	0	40	102.5980
106	13	17	0	40	113.3285
106	17	19	0	40	117.7776
106	19	26	0	40	130.3239
106	26	29	0	40	134.6918
106	29	30	0	40	136.0479
106	30	33	0	40	139.8603
106	33	35	1	40	142.2139

The variable LOGTDOSE is time dependent as its values increase with time as expected. The variable SURVT lists all the event times in the addicts dataset up to day 35 when this individual had an event. Notice STATUS=1 when the event occurred and STATUS=0 prior to the event. Next we run an extended Cox model including the predictors PRISON, DOSE, and CLINIC and the time-dependent variable LOGTDOSE:

```
coxph(Surv(addicts.cp$start,addicts.cp$survt,addicts.cp$status) ~
prison + dose + clinic + logtdose + cluster(id),data=addicts.cp)
```

The **Surv** function now takes three arguments: the start variable (called START), the stop variable (called SURVT), and the status variable (called STATUS). The term **cluster (ID)** in the model formula indicates that there are multiple observations (clusters) from the same subject and requests that robust standard errors be produced for the coefficient estimates. These robust standard errors are designed to account for the non-independence of observations from the same subject. The model output follows:

```
              coef exp(coef) se(coef) robust se     z       p
prison     0.34063     1.406  0.16747   0.15972  2.13 3.3e-02
dose      -0.08262     0.921  0.03598   0.02960 -2.79 5.3e-03
clinic    -1.01988     0.361  0.21542   0.23637 -4.31 1.6e-05
logtdose  0.00862     1.009  0.00645   0.00525  1.64 1.0e-01

Likelihood ratio test=66.3  on 4 df, p=1.34e-13  n= 18708
```

The Wald test z statistic of 1.64 (p = 1.0e-01 or p=0.10) is not significant for LOGTDOSE, providing no evidence that the proportional hazards assumption is violated for DOSE.

Next we run an extended Cox model with heaviside functions for CLINIC defined about the time cutpoint of 365 days. We could use the dataset that we just created, **addicts.cp**, but since there is now only one cutpoint, we illustrate how to create a dataset in counting process format with only one cutpoint. The new dataset (called **addicts.cp365**) will have 360 observations compared to 18,708 in the dataset we previously had created called **addicts.cp**. The code follows:

```
addicts.cp365=survSplit(addicts,cut=365,end="survt",
event="status",start="start",id="id")
```

The **cut=365** option in the **survSplit** function requests that day 365 be the only cutpoint. Next we create the two time-dependent variables (HV1 and HV2). HV1 is defined to equal the value of CLINIC if survival time is less than 365 days and 0 otherwise. HV2 is defined to equal 0 if survival time is less than 365 days and equal the value of CLINIC otherwise (code follows):

```
addicts.cp365$hv1=addicts.cp365$clinic*(addicts.cp365$start<365)
addicts.cp365$hv2=addicts.cp365$clinic*(addicts.cp365$start>=365)
```

The conditional statements in the code **(addicts.cp365 $start<365)** and **(addicts.cp365$start>=365)**, take the values of 1 if true and 0 if false and are then multiplied by the variable CLINIC to define HV1 and HV2.

Next we'll sort the dataset by the variables ID and START. This is not a necessary step but it is easier to view and understand the data when multiple observations from the same subject are consecutive. The **order** function sorts the dataset:

```
addicts.cp365=addicts.cp365[order(addicts.cp365$id,addicts.cp365$start), ]
```

Next we print the first 10 observations for selected variables:

addicts.cp365[1:10,c('id'start','survt','status','clinic','hv1','hv2')]

```
    id start survt status clinic hv1 hv2
   1     0   365      0      1   1   0
   1   365   428      1      1   0   1
  10     0   365      0      1   1   0
  10   365   393      1      1   0   1
 100     0   146      0      2   2   0
 101     0   365      0      2   2   0
 101   365   450      1      2   0   2
 102     0   365      0      2   2   0
 102   365   555      0      2   0   2
 103     0   365      0      2   2   0
```

Notice the sorted order of the ID variable is 1, 10, and 100 rather than 1, 2, and 3. The ID variable is a character rather than numeric variable and is sorted in "alphabetical" rather than numerical order. The first subject (ID=1) had an event at 428 days, so was censored (STATUS=0) during the first time interval (0, 365) but had an event (STATUS=1) during the second interval (365, 428). This subject has the value CLINIC=1, thus has the time-dependent values HV1=1 and HV2=0 over the first interval and HV1=0 and HV2=1 over the second interval.

Before running an extended Cox model with these heaviside functions we define an object (called **Y365**) for the response variable using the **Surv** function. This object is then used in the **coxph** model formula. It is not necessary to explicitly define this object and we did not do so for the previous extended Cox model that we ran containing LOGTDOSE, but the code is more readable with the notation for the response variable simplified. The code follows:

Y365=Surv(addicts.cp365$start,addicts.cp365$survt, addicts.cp365$status)

Next we run the model with two heaviside functions (code and output follow):

coxph(Y365 ~ prison + dose + hv1 + hv2 + cluster(id), data=addicts.cp365)

```
          coef exp(coef) se(coef) robust se      z       p
prison  0.3780     1.459  0.16841   0.16765  2.25 2.4e-02
dose   -0.0355     0.965  0.00643   0.00652 -5.44 5.3e-08
hv1    -0.4594     0.632  0.25529   0.25998 -1.77 7.7e-02
hv2    -1.8305     0.160  0.38595   0.39838 -4.59 4.3e-06

Likelihood ratio test=74.2  on 4 df, p=2.89e-15  n= 360
```

The estimated hazard ratio (CLINIC=2 vs. CLINIC=1) is 0.632 for days <365 and 0.160 for days ≧365 (found in the second numeric column under exp(coef)). If we wish to match the SAS, Stata, and SPSS output, we could run the model without robust standard errors and use the **method="breslow"** to handle simultaneous events (ties) in the Cox likelihood. The code follows (output omitted):

coxph(Y365 ~ prison + dose + hv1 + hv2,data=addicts.cp365,method="breslow")

To run an equivalent model with one heaviside function, we need to include the CLINIC variable in the model (code and output shown below):

coxph(Y365 ~ prison + dose + clinic + hv2 + cluster(id),data=addicts.cp365)

```
          coef exp(coef) se(coef) robust se     z       p
prison  0.3780     1.459  0.16841   0.16765  2.25 2.4e-02
dose   -0.0355     0.965  0.00643   0.00652 -5.44 5.3e-08
clinic -0.4594     0.632  0.25529   0.25998 -1.77 7.7e-02
hv2    -1.3711     0.254  0.46140   0.47054 -2.91 3.6e-03

Likelihood ratio test=74.2  on 4 df, p=2.89e-15  n= 360
```

The coefficient estimates are different with this model compared to the model with two heaviside functions but the estimated hazard ratios are the same. The estimated hazard ratio (CLINIC=2 vs. CLINIC=1) is 0.632 for days <365 (exponentiate the coefficient for CLINIC). In order to estimate the hazard ratio for days ≥ 365, we need to sum the coefficient estimates for CLINIC and HV2 and then exponentiate (exp(-0.4594 + -1.3711)) = 0.160). The significant p-value for the estimated coefficient for HV2 of (p = 3.6e-10 or p = 0.0036) suggests that the hazard ratios for CLINIC for the two different time periods are not equal. In other words, the significant p-value provides evidence that the proportional hazard assumption is violated for CLINIC.

8. RUNNING PARAMETRIC MODELS

The **survreg** function in R runs parametric accelerated failure time (AFT) models. Whereas the key assumption of a proportional hazards (PH) model is that hazard ratios are constant over time, the key assumption for an AFT model is that survival time accelerates (or decelerates) by a constant factor when comparing different levels of covariates.

The most common distribution for parametric modeling of survival data is the Weibull distribution. The hazard function for a Weibull distribution is $\lambda p t^{p-1}$. If p = 1, then the Weibull distribution is also an exponential distribution. The Weibull distribution has the desirable property in that if the AFT assumption holds then the PH assumption also holds. The exponential distribution is a special case of the Weibull distribution. The key property for the exponential distribution is that the hazard is constant over time (($h(t) = \lambda$). In R, the Weibull and exponential model are run only as AFT models.

The Weibull distribution has the property that the log-log of the survival function is linear with the log of time. Recall in Section 2 (assessing the PH assumption graphical approach) that the **fun="cloglog"** option in the **plot** function requested Kaplan-Meier log-log survival plot be plotted against time (on the log scale) for the variable CLINIC. The curves from this plot can be used to evaluate the Weibull assumption. If the survival curves are approximately straight lines (and parallel), then the Weibull assumption is reasonable for CLINIC. Furthermore, if the straight lines have a slope of 1, then the exponential distribution is appropriate. We repeat and condense the code that was given in Section 2 (see outputted plot in Section 2):

```
plot(survfit(Y~addicts$clinic), fun="cloglog",xlab="time in days using log-
arithmic scale",ylab="log-log survival", main="log-log curves by clinic")
```

The log–log curves in Section 2 do not look straight but for illustration, we shall proceed as if the Weibull assumption were appropriate. First an exponential model is run with the **survreg** function. In this model, the Weibull shape parameter (p) is forced to equal 1, which forces the hazard to be constant. We'll save the results in an object called **modpar1**:

```
modpar1=survreg(Surv(addicts$survt,addicts$status) ~ prison + dose +
clinic,data=addicts,dist="exponential")
```

Next we apply the summary function to the object we just created (code and output shown below):

summary(modpar1)

```
               Value Std. Error     z        p
(Intercept)  3.6843    0.43072  8.55 1.19e-17
prison      -0.2526    0.16489 -1.53 1.25e-01
dose         0.0289    0.00614  4.71 2.52e-06
clinic       0.8806    0.21063  4.18 2.91e-05

Scale fixed at 1

Exponential distribution
Loglik(model)= -1094   Loglik(intercept only)= -1118.9
        Chisq= 49.91 on 3 degrees of freedom, p= 8.3e-11
Number of Newton-Raphson Iterations: 5
n= 238
```

The key assumption of an exponential model is that the hazard is constant over time. This is indicated in the output by the statement "Scale fixed at 1" listed under the tables of parameter estimates. The output can be used to estimate the hazard ratio for any subject given a pattern of covariates. Note that R outputs the parameter estimates for the AFT form of the exponential model. Multiply the estimated coefficients by one to get estimates consistent with the PH parameterization of the model (see Chapter. 7). For example, the estimated hazard ratio comparing PRISON=1 vs PRISON=0 is exp(0.2526) = 1.29. The corresponding acceleration factor for an exponential model is just the reciprocal of the hazard ratio, exp(-0.2526) = 0.78. Having a prison record accelerates the time to event by a factor of 0.78.

Next a Weibull AFT model is run with the **survreg** function. The results are saved in an object called **modpar2**:

modpar2=survreg(Surv(addicts$survt,addicts$status) ~ prison + dose + clinic,data=addicts,dist="weibull")

Next we apply the summary function to the object **modpar2** (code and output follow):

summary(modpar2)

```
               Value Std. Error     z        p
(Intercept)   4.1048    0.32806 12.51 6.37e-36
prison       -0.2295    0.12079 -1.90 5.75e-02
dose          0.0244    0.00459  5.32 1.03e-07
clinic        0.7090    0.15722  4.51 6.49e-06
Log(scale)   -0.3150    0.06756 -4.66 3.13e-06

Scale= 0.73

Weibull distribution
Loglik(model)= -1084.5   Loglik(intercept only)= -1114.9
        Chisq= 60.89 on 3 degrees of freedom, p= 3.8e-13
n= 238
```

The Weibull shape parameter is the reciprocal of what R calls the Scale parameter (estimated at 0.73). An estimate for the Weibull shape parameter can be obtained by taking the reciprocal, 1/0.73 = 1.37. The acceleration factor comparing CLINIC=2 to CLINIC=1 is estimated at exp(0.7090) = 2.03. So, the estimated median survival time (time off heroin) is double for patients enrolled in CLINIC=2 compared to CLINIC=1.

We can use the model results and the **predict** function to estimate the median (or any other quantile) time to event for any specified pattern of covariates. For example, we can obtain the 25^{th}, 50^{th}, and 75^{th} percentile of survival time estimated from the Weibull model results that we saved in the object **modpar2** for an individual who has the covariate pattern PRISON=1, DOSE=50, and CLINIC=1. The code follows:

```
pattern1=data.frame(prison=1,dose=50,clinic=1)
pct=c(.25,.50,.75)
days=predict(modpar2,newdata=pattern1,type="quantile",p=pct)
cbind(pct,days)
```

The first statement in the code creates a dataframe of one observation specifying the pattern of covariates of interest. This dataframe (called **pattern1**) could have contained more than one observation if we were interested in comparing different patterns of covariates. The next statement creates a vector (called **pct**) which contains the percentiles of interest (25^{th}, 50^{th}, and 75^{th}). The third statement creates an object (called **days**) that contains output from the **predict** function. The first argument of the **predict** function is the object we called **modpar2** that contains the

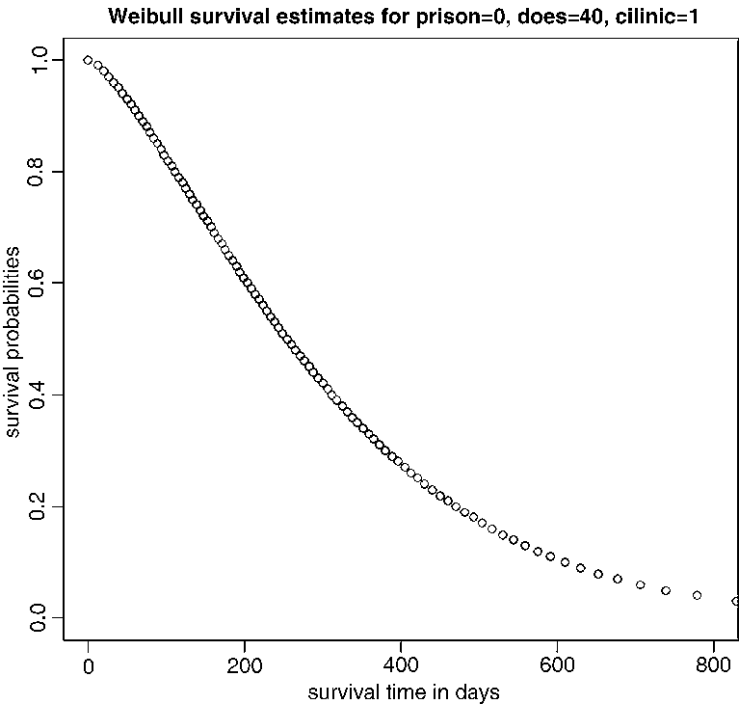
Weibull model results. The second argument, **newdata= pattern1**, inputs the pattern of covariates of interest. The third argument, **type="quantile"**, requests that quantiles be output. The fourth argument, **p=pct**, inputs the vector of quantiles that we created in the line of code above it. The last statement of code uses the **cbind** function to combine the vectors **pct** and **days** side-by-side in columns. The output follows:

```
pct      days
0.25   133.8074
0.50   254.2196
0.75   421.6070
```

The estimated median survival time is 254.2196 days. We can use similar code to plot the survival curve for an individual who has the covariate pattern PRISON=1, DOSE=50, and CLINIC=1 using the Weibull model results. The code follows:

```
pct2=0:100/100
days2=predict(modpar2,newdata=pattern1,
type="quantile",p=pct2)
survival=1-pct2
plot(days2,survival,xlab="survival time in days",ylab= "survival probabilities",main="Weibull survival estimates for prison=0, dose=40,clinic=1",xlim=c(0,800))
```

The first statement creates a vector called **pct2** that contains a sequence of percentiles between 0 and 1 incremented by 0.01,(0, 0.01, 0.02,...,0.99, 1). The second statement creates an object, called **days2**, containing output from the **predict** function. The third argument creates a vector called **survival** which reverses the order of **pct2**. Finally, the **plot** function plots the vectors **days2** on the horizontal axis and **survival** on the vertical axis. Axis labels and a title are added using **plot** function options. The output follows:



Next a log-logistic AFT model is run with the **survreg** function. The results are saved in an object called **modpar3**:

modpar3=survreg(Surv(addicts$survt,addicts$status)∼ prison + dose + clinic,data=addicts,dist="loglogistic")

Next, we apply the summary function to the object **modpar3** (code and output shown below):

summary(modpar3)

```
               Value Std. Error     z        p
(Intercept)   3.5633    0.38945  9.15 5.72e-20
prison       -0.2913    0.14396 -2.02 4.31e-02
dose          0.0316    0.00552  5.73 1.02e-08
clinic        0.5806    0.17157  3.38 7.14e-04
Log(scale)   -0.5331    0.06863 -7.77 7.95e-15

Scale= 0.587

Log logistic distribution
Loglik(model)= -1093.9   Loglik(intercept only)= -1120
        Chisq= 52.18 on 3 degrees of freedom, p= 2.7e-11
Number of Newton-Raphson Iterations: 4
n= 238
```

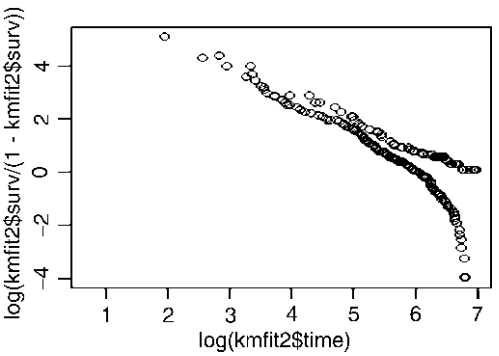
From this output, the acceleration factor comparing CLINIC=2 to CLINIC=1 is estimated as exp(0.5806) = 1.79. If the AFT assumption holds for a log logistic model, then the proportional odds assumption holds for the survival function (although the PH assumption will not hold). The proportional odds assumption can be evaluated by plotting the log odds of survival (using KM estimates) against the log of survival time. If the plots look like straight lines for each pattern of covariates then the log logistic distribution is reasonable. If the straight lines are also parallel then the proportional odds and AFT assumptions also hold.

In Section. 2, we created an object which we called **kmfit2** that contained the Kaplan-Meier survival estimates. We repeat the code to recreate that object:

```
kmfit2=survfit(Surv(addicts$survt,addicts$status)~addicts$clinic)
```

The vector **kmfit2$time** contains the survival times and the vector **kmfit2$surv** contains the KM survival estimates by CLINIC. The **plot** function can be used to plot log odds of survival, $\log[(S/(1 \times S)]$, against the log of survival time. The code and output follow:

```
plot(log(kmfit2$time),log(kmfit2$surv/(1-kmfit2$surv)))
```



The curves do not look like straight lines or parallel so the proportional odds assumption for CLINIC looks to be violated. We had run the log-logistic model earlier for illustration, even though the graph suggests that it is not the appropriate model.

Other distributions supported by the **survreg** function are the normal (dist="gaussian") and the lognormal (dist="lognormal") distributions.

9. RUNNING FRAILTY MODELS

Frailty models contain an extra random component designed to account for individual-level differences in the hazard otherwise unaccounted for by the model. The frailty, α, is a multiplicative effect on the hazard assumed to follow some distribution. The hazard function conditional on the frailty can be expressed as $h(t|\alpha) = \alpha[h(t)]$.

R offers three choices for the distribution of the frailty: the gamma, Gaussian, and t distributions. The variance (theta) of the frailty component is a parameter typically estimated by the model. If theta = 0, then there is no frailty.

First, we rerun a stratified Cox model without frailty (previously shown in Section. 4). The stratified variable is CLINIC while PRISON and DOSE are predictor variables. A stratified Cox model is appropriate if the PH assumption is violated for CLINIC and met for PRISON and DOSE and our interest is in estimating a hazard ratio for PRISON or DOSE. The code and output follow:

Y=Surv(addicts$survt,addicts$status==1)
coxph(Y~ prison + dose + strata(clinic),data=addicts)

```
          coef exp(coef) se(coef)     z       p
prison  0.3896     1.476  0.16893  2.31 2.1e-02
dose   -0.0351     0.965  0.00646 -5.43 5.6e-08

Likelihood ratio test=33.9 on 2 df, p=4.32e-08  n= 238
```

The estimated hazard ratio for PRISON=1 versus PRISON=0 is exp(0.3896) = 1.476. Next we illustrate how to include a frailty component in this model. The code follows:

coxph(Y~ prison + dose + strata(clinic) + frailty(id, distribution= "gamma"), data=addicts)

The term + **frailty(id, distribution="gamma")** is included in the model formula. The first argument of the **frailty** function is the variable **id** and indicates that the unmeasured heterogeneity (the frailty) is at the individual level. The second argument indicates that the distribution of the random component is the gamma distribution. The output follows:

```
                          coef    se(coef) se2     Chisq DF   p
prison                    0.3900 0.16916  0.16893  5.32 1.00 2.1e-02
dose                     -0.0352 0.00647  0.00647 29.51 1.00 5.6e-08
frailty(id, distribution                           0.34 0.32 3.1e-01

Iterations: 5 outer, 41 Newton-Raphson
     Variance of random effect= 0.00227   I-likelihood = -597.5
Degrees of freedom for terms= 1.0 1.0 0.3
Likelihood ratio test=34.6  on 2.32 df, p=5.26e-08  n= 238
```

Under the table of parameter estimates the output indicates that the variance of random effect = 0.00227. The *p*-value for the frailty component of 3.1e-01= 0.31 is provided in the third row and right column of the table and indicates that the frailty component is not significant. We conclude that the variance of the random component is zero for this model (i.e., there is no frailty). The parameter estimates for PRISON and DOSE changed minimally in this model compared to the model previously run without the frailty.

Now, suppose the variable CLINIC was unmeasured. Next we consider a Cox model (without frailty) that does not contain CLINIC. The code and output follow:

coxph(Y~ prison + dose, data=addicts)

```
          coef   exp(coef) se(coef)    z     p
prison  0.1897     1.209    0.164   1.15 2.5e-01
dose   -0.0361     0.965    0.006  -6.01 1.8e-09

Likelihood ratio test=38.2  on 2 df, p=5.04e-09  n= 238
```

The estimated hazard ratio for PRISON=1 versus PRISON=0 is exp(0.1897) = 1.209 as compared to exp (0.3896) = 1.476 that was observed in the model that contained CLINIC as a stratified variable. In previous sections CLINIC was shown to be an important predictor that violates the proportional hazards assumption. If CLINIC was unaccounted for (as in the model above), there may be a source of unobserved heterogeneity that a frailty component might address. The next model omits CLINIC but includes a frailty component and the predictors PRISON and DOSE. The code and output follow:

coxph(Y∼ prison + dose + frailty(id, distribution="gamma"), data=addicts)

```
                          coef    se(coef) se2     Chisq DF   p
prison                    0.4144 0.22160  0.17590   3.5  1.0 6.1e-02
dose                     -0.0517 0.00845  0.00699  37.4  1.0 9.6e-10
frailty(id, distribution                           100.5 69.3 8.6e-03

Iterations: 6 outer, 44 Newton-Raphson
     Variance of random effect= 0.65   I-likelihood = -685.4
Degrees of freedom for terms=  0.6  0.7 69.3
Likelihood ratio test=190  on 70.7 df, p=6.17e-13  n= 238
```

The variance of the frailty component is estimated at 0.65 compared to 0.00227 for the model that we showed previously that contained CLINIC as the stratified variable. The *p*-value for the frailty is highly significant at 8.6e–3 = 0.0086. The hazard ratio for the effect of PRISON is exp (0.4144) = 1.51. The **summary** function can be applied to the **coxph** function to get R to exponentiate the parameter estimates (with 95% CI) when a frailty component is included in a Cox model. The code and output follow:

summary(coxph(Y∼ prison + dose + frailty(id, distribution="gamma"), data=addicts))

```
                         coef     se(coef)  se2     Chisq  DF    p
prison                    0.4144 0.22160  0.17590   3.5  1.0 6.1e-02
dose                     -0.0517 0.00845  0.00699  37.4  1.0 9.6e-10
frailty(id, distribution                           100.5 69.3 8.6e-03

       exp(coef) exp(-coef) lower .95 upper .95
prison      1.51      0.661     0.980     2.337
dose        0.95      1.053     0.934     0.966

Iterations: 6 outer, 44 Newton-Raphson
     Variance of random effect= 0.65   I-likelihood = -685.4
Degrees of freedom for terms=  0.6  0.7 69.3
Rsquare= 0.551   (max possible= 0.997 )
Likelihood ratio test= 190  on 70.7 df,   p=6.17e-13
Wald test            = 38.8  on 70.7 df,   p=1
```

It is interesting that the estimated hazard ratio for PRISON (1.51) obtained in this model (without CLINIC but with the frailty component) is closer to the corresponding hazard ratio obtained from the model that included CLINIC (1.476) compared to the one that did not include CLINIC (1.209). In this example, the frailty component might be accounting to some extent for the fact that CLINIC was omitted from the model.

10. MODELING RECURRENT EVENTS

The modeling of recurrent events is illustrated with the bladder cancer dataset (**bladder.rda**) described at the start of this appendix. Recurrent events are represented in the data with multiple observations for subjects having multiple events. The data layout for the bladder cancer dataset is in the counting process (start, stop) format with time intervals defined for each observation (see Chapter 8). The **load** function is used to access an R dataframe that has been saved as a file. Suppose the bladder dataset has been saved on your C drive as C:\crbladder.rda. The following code will load the bladder data:

```
load("C:\\bladder.rda")
```

The following code prints the 12^{th}–20^{th} observation, which contains information for four subjects:

```
bladder[12:20, ]
```

The output follows:

```
    id  event interval inttime start stop tx num size
12  10     1        1      12     0   12  0   1    1
13  10     1        2       4    12   16  0   1    1
14  10     0        3       2    16   18  0   1    1
15  11     0        1      23     0   23  0   3    3
16  12     1        1      10     0   10  0   1    3
17  12     1        2       5    10   15  0   1    3
18  12     0        3       8    15   23  0   1    3
19  13     1        1       3     0    3  0   1    1
20  13     1        2      13     3   16  0   1    1
```

There are three observations for ID=10, one observation for ID=11, three observations for ID=12, and two observations for ID=13. The variables START and STOP represent the time interval for the risk period specific to that observation. The variable EVENT indicates whether an event (coded 1) occurred. The first three observations indicate that the subject with ID=10 had an event at 12 months, another event at 16 months, and was censored at 18 months.

Recall we analyzed data in the counting process format when we ran extended Cox models (Section 7). In that section we saw how a subject's covariate can change values from time-interval to time-interval. With the bladder dataset, the (start,stop) data format provides a way to indicate that a subject experienced multiple events.

As mentioned in the beginning of our discussion of R, the code **library(survival)** must be submitted at each session before survival functions in R can be accessed.

```
library(survival)
```

The **coxph** function can be used to run Cox models with recurrent events. First, we'll define a response variable using the **Surv** function (called **Y**):

```
Y=Surv(bladder$start,bladder$stop,bladder$event==1)
```

As we have seen in Section 7, the **Surv** function requires three arguments with data in the counting process format: the start variable (called START), the stop variable (called STOP), and the status variable (called EVENT). The code **bladder$event==1** indicates that an event is coded 1. R recognizes the value 1 as the default coding of an event, so it was not necessary to state this explicitly in the **Surv** function as we did. Next, a recurrent-events Cox model is run with the predictors: treatment status (TX), initial number of tumors (NUM), and the initial size of tumors (SIZE):

```
coxph(Y ~ tx + num + size + cluster(id), data=bladder)
```

The term + **cluster(id)** in the model formula requests robust standard errors for the parameter estimates. The model output follows:

```
        coef exp(coef) se(coef) robust se      z      p
tx   -0.4116     0.663   0.1999    0.2488 -1.655 0.0980
num   0.1637     1.178   0.0478    0.0584  2.801 0.0051
size -0.0411     0.960   0.0703    0.0742 -0.554 0.5800

Likelihood ratio test=14.7  on 3 df, p=0.00213  n=190
```

The treatment variable (TX) is coded 1 for treatment with thiotepa and 0 for the placebo. The estimated hazard ratio (TX=1 vs. TX=0) is 0.663 (with a p-value of 0.0980). There are two sets of standard errors presented in the table under the columns labeled: se(coef) and robust se. The p-values and z-test statistics in this table are calculated using the robust standard errors. We could obtain additional model output (including 95% CIs) by applying the **summary** function to the **coxph** function.

A stratified Cox model can also be run using the data in this format with the variable INTERVAL as the stratified variable. The stratified variable indicates whether the subject was at risk for their first, second, third, or fourth event.

This approach is called a **Stratified CP** recurrent event model (see Chap. 8) and is used if the investigator wants to distinguish the order in which recurrent events occur. The bladder data is in the proper format to run this model. The code and output follow:

coxph(Y ~ tx + num + size + strata(interval) + cluster (id),data=bladder)

```
         coef exp(coef) se(coef) robust se      z    p
tx    -0.3335     0.716   0.2162    0.2048 -1.628 0.10
num    0.1196     1.127   0.0533    0.0514  2.328 0.02
size -0.0085     0.992   0.0728    0.0616 -0.138 0.89

Likelihood ratio test=6.51  on 3 df, p=0.0893  n=190
```

The only additional code from the previous model is the term + **strata(interval)** in the model formula which indicates that INTERVAL is the stratified variable. Interaction terms between the treatment variable (TX) and the stratified variable could be created to examine whether the effect of treatment differed for the 1^{st}, 2^{nd}, 3^{rd}, or 4^{th} event.

Another stratified approach (called Gap Time) is a slight variation of the Stratified CP approach. The difference is in the way the time intervals for the recurrent events are defined. There is no difference in the time intervals when subjects are at risk for their first event. However, with the Gap Time approach, the starting time at risk gets reset to zero for each subsequent event. To run a Gap Time model, we need to create two new (start, stop) variables in the bladder dataset, which we'll call START2 and STOP2:

bladder$start2=0
bladder$stop2=bladder$stop – bladder$start

The first of the two newly defined variables (START2) is always zero. The second (STOP2) is defined as the time between each event (STOP–START). To print a subset of these variables, we can use the **data.frame** function. The **attach** function allows variables in the bladder dataset to be listed without the **bladder$** prefix (code and output for printing the 12^{th}–20^{th} observation below).

attach(bladder)
data.frame(id,event,start,stop,start2,stop2)[12:20,]

	id	event	start	stop	start2	stop2
12	10	1	0	12	0	12
13	10	1	12	16	0	4
14	10	0	16	18	0	2
15	11	0	0	23	0	23
16	12	1	0	10	0	10
17	12	1	10	15	0	5
18	12	0	15	23	0	8
19	13	1	0	3	0	3
20	13	1	3	16	0	13

Next we need to reset our response variable using the **Surv** function by changing our time intervals from (START, STOP) to (START2, STOP2):

Y2=Surv(bladder$start2,bladder$stop2,bladder$event)

Next we run a Gap Time model with the bladder data using similar code that was used for the Stratified CP model except we use **Y2** rather than **Y** as our response variable. The code and output follow:

coxph(Y2 ~ tx + num + size + strata(interval) + cluster(id),data=bladder)

	coef	exp(coef)	se(coef)	robust se	z	p
tx	-0.27900	0.757	0.2073	0.2156	-1.294	0.2000
num	0.15805	1.171	0.0519	0.0509	3.103	0.0019
size	0.00742	1.007	0.0700	0.0643	0.115	0.9100

Likelihood ratio test=9.33 on 3 df, p=0.0252 n=190

The results using the Gap Time approach varies slightly from that obtained using the Stratified CP approach.

Test Answers

Chapter 1

True-False Questions:

1. T
2. T
3. T
4. F: Step function.
5. F: Ranges between 0 and 1.
6. T
7. T
8. T
9. T
10. F: Median survival time is longer for group 1 than for group 2.
11. F: Six weeks or greater.
12. F: The risk set at 7 weeks contains 15 persons.
13. F: Hazard ratio.
14. T
15. T
16. $h(t)$ gives the instantaneous potential per unit time for the event to occur given that the individual has survived up to time t;$k(t)$ is greater than or equal to 0; $h(t)$ has no upper bound.
17. Hazard functions
    - give insight about conditional failure rates;
    - help to identify specific model forms (e.g., exponential, Weibull);
    - are used to specify mathematical models for survival analysis.
18. Three goals of survival analysis are:
    - to estimate and interpret survivor and/or hazard functions;
    - to compare survivor and/or hazard functions;
    - to assess the relationship of explanatory variables to survival time.

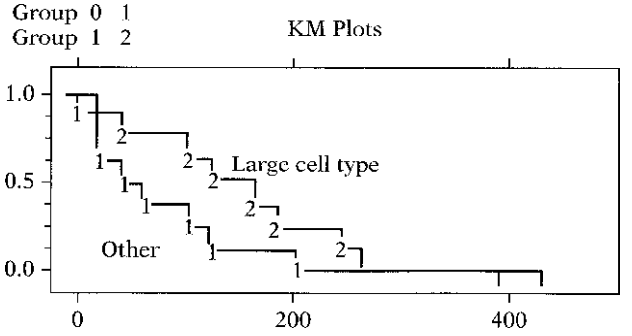
19.

	$t_{(j)}$	m_j	q_j	$R(t_{(j)})$
Group 1:	0	0	0	25 persons survive $\geq$ 0 years
	1.8	1	0	25 persons survive $\geq$ 1.8 years
	2.2	1	0	24 persons survive $\geq$ 2.2 years
	2.5	1	0	23 persons survive $\geq$ 2.5 years
	2.6	1	0	22 persons survive $\geq$ 2.6 years
	3.0	1	0	21 persons suivive $\geq$ 3.0 years
	3.5	1	0	20 persons survive $\geq$ 3.5 years
	3.8	1	0	19 persons survive $\geq$ 3.8 years
	5.3	1	0	18 persons survive $\geq$ 5.3 years
	5.4	1	0	17 persons survive $\geq$ 5.4 years
	5.7	1	0	16 persons survive $\geq$ 5.7 years
	6.6	1	0	15 persons survive $\geq$ 6.6 years
	8.2	1	0	14 persons survive $\geq$ 8.2 years
	8.7	1	0	13 persons survive $\geq$ 8.7 years
	9.2	2	0	12 persons survive $\geq$ 9.2 years
	9.8	1	0	10 persons survive $\geq$ 9.8 years
	10.0	1	0	9 persons survive $\geq$ 10.0 years
	10.2	1	0	8 persons survive $\geq$ 10.2 years
	10.7	1	0	7 persons survive $\geq$ 10.7 years
	11.0	1	0	6 persons survive $\geq$ 11.0 years
	11.1	1	0	5 persons survive $\geq$ 11.1 years
	11.7	1	3	4 persons survive $\geq$ 11.7 years

20. a. Group 1 has a better survival prognosis than group 2 because group 1 has a higher average survival time and a correspondingly lower average hazard rate than group 2.
    b. The average survival time and average hazard rates give overall descriptive statistics. The survivor curves allow one to make comparisons over time.

Chapter 2

1. a. KM plots and the log rank statistic for the cell type 1 variable in the vets.data dataset are shown below.

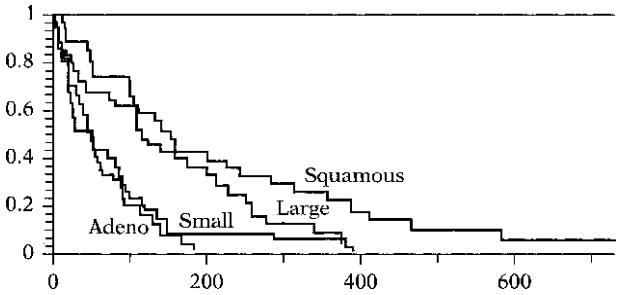


Group	Events observed	Events expected
1	102	93.45
2	26	34.55
Total	128	128.00

Log rank = chi2(1) = 3.02
p-value = Pr > chi2 = 0.0822

The KM curves indicate that persons with large cell type have a consistently better prognosis than persons with other cell types, although the two curves are essentially the same very early on and after 250 days. The log rank test is not significant at the .05 level, which gives somewhat equivocal findings.

b. KM plots and the log rank statistic for the four categories of cell type are shown below.



The KM curves suggest that persons with adeno or small cell types have a poorer survival prognosis than persons with large or squamous cell types. Moreover, there does not appear to be a meaningful difference between adeno or small cell types. Also, persons with squamous cell type seem to have, on the whole, a better prognosis than persons with large cell type.

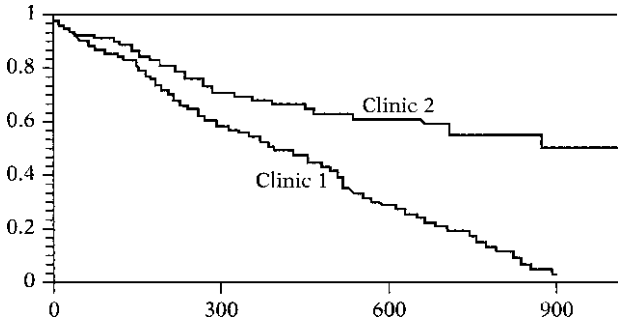
Computer results from Stata giving log rank statistics are now shown.

Group	Events observed	Events expected
1	26	34.55
2	26	15.69
3	45	30.10
4	31	47.65
Total	128	128.00

Log rank = chi2(3) = 25.40
P-value = Pr > chi2 = 0.0000

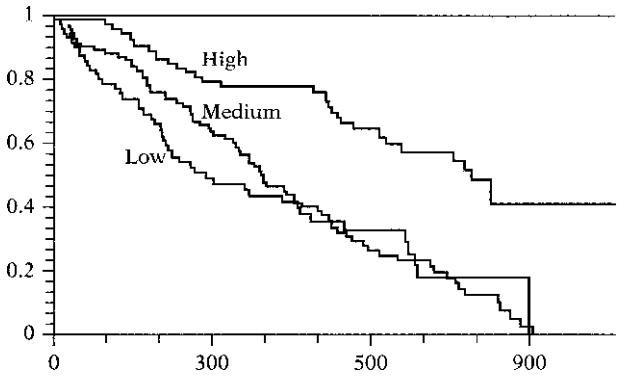
The log-rank test yields highly significant p-values, indicating that there is some overall difference between all four curves; that is, the null hypothesis that the four curves have a common survival curve is rejected,

2. a. KM plots for the two clinics are shown below. These plots indicate that patients in clinic 2 have a consistently better prognosis for remaining under treatment than do patients in clinic 1. Moreover, it appears that the difference between the two clinics is small before one year of follow-up but diverges after one year of follow up.



b. The log rank statistic (27.893) and Wilcoxon statistic (11.63) are both significant well below the .01 level, indicating that the survival curves for the two clinics are significantly different. The log rank statistic is nevertheless much larger than the Wilcoxon statistic, which makes sense because the log rank statistic emphasizes the later survival experience, where the two survival curves are far apart, whereas the Wilcoxon statistic emphasizes earlier survival experience, where the two survival curves are closer together.

c. If methadone dose is categorized into high (70+), medium (55–70) and low (<55), we obtain the KM curves shown below.



The KM curves indicate that persons with high doses have a consistently better survival prognosis (i.e.. maintenance) than persons with medium or low doses. The latter two groups are not very different from each other, although the medium dose group has a somewhat better prognosis up to the first 400 days of follow-up.
The log rank test statistic is shown below for the above categorization scheme.

Group	Events observed	Events expected
0	45	30.93
1	74	54.09
2	31	64.99
Total	150	150.00

Log rank = chi2(2) = 33.02
P-value = Pr>chi2 = 0.0000

The test statistic is highly significant, indicating that these three curves are not equivalent.

Chapter 3

1. a. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 T1 + \beta_2 T_2 + \beta_3 PS + \beta_4 DC + \beta_5 BF + \beta_6(T1 \times PS) + \beta_7(T2 \times PS) + \beta_8(T1 \times DC) + \beta_9(T2 \times DC) + \beta_{10}(T1 \times BF) + \beta_{11}(T2 \times BF)]$

   b. **Intervention A: X*** = (1, 0, PS, DC, BF, PS, 0, DC, 0, BE 0)
   Intervention C: X = (− 1, −1, PS, DC, BF, –PS, –PS, –DC, –DC, –BF, –BF)

$$\mathrm{HR} = \frac{h(t,\mathbf{X}^*)}{h(t,\mathbf{X})} = \exp[2\beta_1 + \beta_2 + 2\beta_6 PS + \beta_7 PS + 2\beta_8 DC + \beta_9 DC + 2\beta_{10} BF + \beta_{11} BF]$$

c. H_0: $\beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$ in the full model.
Likelihood ratio test statistic: $-2 \ln \hat{L}_R - (-2\mathrm{In}\hat{L}_F)$, which is approximately χ^2_6 under H_0, where R denotes the reduced model (containing no product terms) under H_0, and F denotes the full model (given in Part la above)

d. The two models being compared are:
Full model (F): $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 Tl + \beta_2 T2 + \beta_3 PS + \beta_4 DC + \beta_5 BF]$

Reduced model (R): $h(t,\mathbf{X}) = h_0(t)\exp[\beta_3 \text{PS} + \beta_4 \text{DC} + \beta_5 \text{BF}]$
H_0: $\beta_1 = \beta_2 = 0$ in the full model
Likelihood ratio test statistic: $-2\ln\hat{L}_R - (-2\ln\hat{L}_F)$, which is approximately χ^2_2 under H_0.

e.

Intervention A:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_1 + (\overline{\text{PS}})\hat{\beta}_3 + (\overline{\text{DC}})\hat{\beta}_4 + (\overline{\text{BF}})\hat{\beta}_5]}$

Intervention B:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_2 + (\overline{\text{PS}})\hat{\beta}_3 + (\overline{\text{DC}})\hat{\beta}_4 + (\overline{\text{BF}})\hat{\beta}_5]}$

Intervention C:
$\hat{S}(t,\mathbf{X}) = [\hat{S}_0(t)]^{\exp[-\hat{\beta}_1 - \hat{\beta}_2 + (\overline{\text{PS}})\hat{\beta}_3 + (\overline{\text{DC}})\hat{\beta}_4 + (\overline{\text{BF}})\hat{\beta}_5]}$

2. a. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 \text{ CHR} + \beta_2 \text{ AGE} + \beta_3(\text{CHR} \times \text{AGE})]$

b. H_0: $\beta_3 = 0$
LR statistic = 264.90 − 264.70 = 0.21; χ^2 with 1 d.f. under H_0; not *significant*.
Wald statistic gives a chi-square value of .01, also not significant. Conclusions about interaction: the model should not contain an interaction term.

c. When AGE is controlled (using the gold standard model 2), the hazard ratio for the effect of CHR is exp(.8051) = 2.24, whereas when AGE is not controlled, the hazard ratio for the effect of CHR (using Model 1) is exp(.8595) = 2.36. Thus, the hazard ratios are not appreciably different, so AGE is not a confounder.
Regarding precision, the 95% confidence interval for the effect of CHR in the gold standard model (Model 2)is given by exp[.8051 ± 1.96(.3252)] = (1.183, 4.231) whereas the corresponding 95% confidence interval in the model without AGE (Model 1) is given by exp[.8595 ± 1.96(.3116)] = (1.282, 4.350). Both confidence intervals have about the same width, with the latter interval being slightly wider. Thus, controlling for AGE has little effect on the final point and interval estimates of interest.

d. If the hazard functions cross for the two levels of the CHR variable, this would mean that none of the models provided is appropriate, because each model assumes that the proportional hazards assumption is met for each predictor in the model. If hazard functions cross for CHR, however, the proportional hazards assumption cannot be satisfied for this variable.

e. for $CHR = 1 : \hat{S}(t, X) = \left[\hat{S}_0(t)\right]^{\exp\left[0.8051+0.0856(\overline{\text{AGE}})\right]}$
For $CHR = 0 : \hat{S}(t, X) = \left[\hat{S}_0(t)\right]^{\exp\left[0.0856(\overline{\text{AGE}})\right]}$

f. Using Model 1, which is the best model, there is evidence of a moderate effect of CHR on survival time, because the hazard ratio is about 2.4 with a 95% confidence interval between 1.3 and 4.4, and the Wald text for significance of this variable is significant below the .01 level.

a3. Full model (F = Model 1): $h(t,\mathbf{X})=h_0(t)\exp[\beta_1 Rx + \beta_2 \text{Sex} + \beta_3 \log \text{WBC} + \beta_4(Rx \times \text{Sex}) + \beta_5(Rx \times \log \text{WBC})]$
Reduced model (R = model 4):
$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 Rx + \beta_2 \text{Sex} + \beta_3 \text{logWBC}]$
H_0: $\beta_4 - \beta_5 = 0$
LR statistic = 144.218 − 139.030 = 5.19; χ^2 with 2 d.f. under H_0; not significant at 0.05, though signilicant at 0.10. The chunk test indicates some (though mild) evidence of interaction.

b. Using either a Wald test (p-value = .776) or a LR test, the product term $Rx \times$ log WBC is clearly not significant, and thus should be dropped from Model 1. Thus, Model 2 is preferred to Model 1.

c. Using Model 2, the hazard ratio for the effect of Rx is given by HR − $(h(t,\mathbf{X}^*))/(h(t,\mathbf{X})) = \exp[0.405 + 2.013\ \text{Sex}]$

d. Males (Sex = 0) : $\widehat{HR} = \exp[0.405] = 1.499$
Females (Sex = 1) : $\widehat{HR} = \exp[0.405 + 2.013(1)] = 11.223$

e. Model 2 is preferred to Model 3 if one decides that the coefficients for the variables Rx and $Rx \times$ Sex are meaningfully different for the two models. It appears that such corresponding coefficients (0.405 vs. 0.587 and 2.013 vs. 1.906) are different. The estimated hazard ratios tor Model 3 are 1.799 (males) and 12.098 (females), which are different, but not very different from the estimates computed in Part 3d for Model 2. If it is decided that there is a meaningful difference here, then we would conclude that log WBC is a confounder; otherwise log WBC is not a confounder. Note that the log WBC variable is significant in Model 2 ($P = .000$), but this addresses precision and not confounding. When in doubt, as in this case, the safest thing to do (for validity reasons) is to control for log WBC.

f. Model 2 appears to be best, because there is significant interaction of $Rx \times$ Sex ($P = .023$) and because log WBC is a likely confounder (from Part e).

Chapter 4

1. The $P(PH)$ values in the printout provide GOF statistics for each variable adjusted for the other variables in the model These $P(PH)$ values indicate that the clinic variable does not satisfy the PH assumption ($P << .01$), whereas the prison and dose variables satisfy the PH assumption ($P > .10$).
2. The log–log plots shown are parallel. However, the reason why they are parallel is because the clinic variable has been included in the model, and log–log curves for any variable in a PH model must always be parallel. If, instead, the clinic variable had been stratified (i.e., not included in the model), then the log–log plots comparing the two clinics adjusted for the prison and dose variables might not be parallel.
3. The log–log plots obtained when the clinic variable is stratified (i.e., using a stratified Cox PH model) are not parallel. They intersect early on in follow-up and diverge from each other later in follow-up. These plots therefore indicate that the PH assumption is not satisfied for the clinic variable.
4. Both graphs of log–log plots for the prison variable show curves that intersect and then diverge from each other and then intersect again. Thus, the plots on each graph appear to be quite nonparallel, indicating that the PH assumption is not satisfied for the prison variable. Note, however, that on each graph, the plots are quite close to each other, so that one might conclude that, allowing for random variation, the two plots are essentially coincident; with this latter point of view, one would conclude that the PH assumption is satisfied for the prison variable.
5. The conclusion of nonparallel log–log plots in Question 4 gives a different result about the PH assumption for the prison variable than determined from the GOF tests provided in Question 1. That is, the log–log plots suggest that the prison variable does not satisfy the PH assumption, whereas the GOF test suggests that the prison variable satisfies the assumption. Note, however, if the point of view is taken that the two plots are close enough to suggest coincidence, the graphical conclusion would be the same as the GOF conclusion. Although the final decision is somewhat equivocal here, we prefer to conclude that the PH assumption is satisfied for the prison variable because this is strongly indicated from the GOF test and questionably counterindicated by the log–log curves.

6. Because maximum methadone dose is a continuous variable, we must categorize this variable into two or more groups in order to graphically evaluate whether it satisfies the PH assumption. Assume that we have categorized this variable into two groups, say, low versus high. Then, **observed** survival plots can be obtained as KM curves for low and high groups separately To obtain **expected** plots, we can fit a Cox model containing the dose variable and then substitute suitably chosen values for dose into the formula for the estimated survival curve. Typically, the values substituted would be either the mean or median (maximum) dose in each group.
   After obtaining observed and expected plots for low and high dose groups, we would conclude that the PH assumption is satisfied if corresponding observed and expected plots art; not widely discrepant from each other. If a noticeable discrepancy is found for at least one pair of observed versus expected plots, we conclude that the PH assumption is not satisfied.
7. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 \text{ clinic} + \beta_2 \text{ prison} + \beta_3 \text{ dose} + \delta_1 (\text{clinic} \times g(t)) + \delta_2 (\text{prison} \times g(t)) + \delta_3 (\text{dose} \times g(t))]$

   where $g(t)$ is some function of time. The null hypothesis is given by H_0: $\delta_1 = \delta_2 = \delta_3 = 0$. The test statistic is a likelihood ratio statistic of the form $LR = -2\ln L_R - (-2\text{In}L_F)$ where R denotes the reduced (PH) model obtained when all δs are 0, and F denotes the full model given above. Under H_0, the LR statistic is approximately chi-square with 3 d.f.
8. Drawbacks of the extended Cox model approach:

   - Not always clear how to specify $g(t)$; different choices may give different conclusions;
   - Different modeling strategies to choose from, for example, might consider $g(t)$ to be a polynomial in t and do a backward elimination to eliminate nonsignificant higher-order terms; alternatively, might consider $g(t)$ to be linear in t without evaluating higher-order terms.
     Different strategies may yield different conclusions.

9. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1 \text{ clinic} + \beta_2\text{prison} + \beta_3 \text{ dose} + \delta_1(\text{clinic} \times g(t))]$ where $g(t)$ is some function of time. The null hypothesis is given by H_0: $\delta_1 = 0$, and the test statistic is either a Wald statistic or a likelihood ratio statistic. The LR statistic would be of the form $LR = -2 \text{ In } L_R - (-2\text{In } L_F)$, where R denotes the reduced (PH) model obtained when $\delta_1 = 0$, and F denotes the full model given above. Either statistic is approximately chi-square with 1 d.f. under the null hypothesis.
10. $t > 365$ days: $HR = \exp[\beta_1 + \delta_1]$
    $t \leq 365$ days: $HR = \exp[\beta_1]$
    If δ_1 is not equal to zero, then the model does not satisfy the PH assumption for the clinic variable. Thus, a test of H_0: $\delta_1 = 0$ evaluates the PH assumption; a significant result would indicate that the PH assumption is violated. Note that if δ_1 is not equal to zero, then the model assumes that the hazard ratio is not constant over time by giving a different hazard ratio value depending on whether t is greater than 365 days or t is less than or equal to 365 days.

Chapter 5

1. By fitting a stratified Cox (SC) model that stratifies on clinic, we can compare adjusted survival curves for each clinic, adjusted for the prison and dose variables. This will allow us to visually describe the extent of clinic differences on survival over time. However, a drawback to stratifying on clinic is that it will not be possible to obtain an estimate of the hazard ratio for the effect of clinic, because clinic will not be included in the model.
2. The adjusted survival surves indicate that clinic 2 has a better survival prognosis than clinic 1 consistently over time. Moreover, it seems that the difference between the effects of clinic 2 and clinic 1 increases over lime.
3. $h_g(t, \mathbf{X}) = h_{0_g}(t)\exp[\beta_1 \text{ prison} + \beta_2 \text{ dose}], \; g = 1, 2$
   This is a no-interaction model because the regression coefficients for prison and dose are the same for each stratum.
4. Effect of prison, adjusted for clinic and dose: $\widehat{HR} = 1.475$, 95% CI: (1.059, 2.054). It appears that having a prison record gives a 1.475 increased hazard for failure than not having a prison record. The p-value is 0.021, which is significant at the 0.05 level.
5. Version 1: $h_g(t,\mathbf{X}) = h_{0_g}(t)\exp[\beta_{1g} \text{ prison} + \beta_{2g} \text{ dose}], \; g = 1, 2$
   Version 2: $h_g(t,\mathbf{X}) = h_{0_g}(t)\exp[\beta_1 \text{ prison} + \beta_2 \text{ dose} + \beta_3 (\text{clinic} \times \text{prison}) + \beta_4 (\text{clinic} \times \text{dose})], \; g = 1, 2$

6. $g = 1$ (clinic 1):
   $h_1(t,\mathbf{X}) = h_{01}(t)\exp[(0.502)\text{prison} + (-0.036)\text{ dose}]$
   $g = 2$ (clinic 2):
   $h_2(t,\mathbf{X}) = h_{02}(t)\exp[(-0.083)\text{prison} + (-0.037)\text{dose}]$
7. The adjusted survival curves stratified by clinic are virtually identical for the no-interaction and interaction models. Consequently, both graphs (no-interaction versus interaction) indicate the same conclusion that clinic 2 has consistently larger survival (i.e., retention) probabilities than clinic 1 as time increases.
8. H_0: $\beta_3 = \beta_4 = 0$ in the version 2 model (i.e., the no-interaction assumption is satisfied). $LR = -2\text{In } L_R - (-2 \text{ In } L_F)$ where R denotes the reduced (no-interaction) model and F denotes the full (interaction) model. Under the null hypothesis, LR is approximately a chi square with 2 degrees of freedom.
   Computed $LR = 1195.428 - 1193.558 = 1.87$; p-value = 0.395; thus, the null hypothesis is not rejected and we conclude that the no interaction model is preferable to the interaction model.

Chapter 6

1. For the chemo data, the –log-log KM curves intersect at around 600 days; thus the curves are not parallel, and this suggests that the treatment variable does not satisfy the PH assumption.
2. The P (PH) value for the Γx variable is 0, indicating that the PH assumption is not satisfied for the treatment variable based on this goodness-of-fit test.
3. $h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(T\,x)g_1(t) + \beta_2(T\,x)g_2(t) + \beta_3(T\,x)g_3(t)]$
   where
   $$g_1(t) = \begin{cases} 1 & \text{if } 0 \le t < 250 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$
   $$g_2(t) = \begin{cases} 1 & \text{if } 25\,0 \le t < 500 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$
   $$g_3(t) = \begin{cases} 1 & \text{if } t \ge 500 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$
4. Based on the printout the hazard ratio estimates and corresponding p-values and 95% confidence intervals are given as follows for each time interval:

	Haz. Ratio	p> |z|	[95% Conf.	Interval]
$0 \leq t < 250$ days:	0.221	0.001	0.089	0.545
$250 \leq t < 500$ days:	1.629	0.278	0.675	3.934
$t \geq 500$ days:	1.441	0.411	0.604	3.440

The results show a significant effect of treatment below 250 days and a nonsignificant effect of treatment in each of the two intervals after 250 days. Because the coding for treatment was 1 = chemotherapy plus radiation versus 2 = chemotherapy alone, the results indicate that the hazard for chemotherapy plus radiation is 1/0.221 = 4.52 times the hazard for chemotherapy alone. The hazard ratio inverts to a value less than 1 (in favor of chemotherapy plus radiation after 250 days), but this result is nonsignificant. Note that for the significant effect of 1/0.221 = 4.52 below 250 days, the 95% confidence interval ranges between 1/0.545= 1.83 and 1/0.089= 11.24 when inverted, which is a very wide interval.

5. Model with two Heaviside functions:
$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(Tx)g_1(t) + \beta_2(Tx)g_2(t)]$
where

$$g_1(t) = \begin{cases} 1 & \text{if } 0 \leq t < 250 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } t \geq 250 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$

Model with one Heaviside function:
$h(t,\mathbf{X}) = h_0(t)\exp[\beta_1(Tx) + \beta_2(Tx)g_1(t)]$
where $g_1(t)$ is defined above.

6. The results for two time inteivals give hazard ratios that are on the opposite side of the null value (i.e., 1). Below 250 days, the use of chemotherapy plus radiation is, as in the previous analysis, 4.52 times the hazard when chemother apy is used alone. This result is significant and the same confidence interval is obtained as before. Above 250 days, the use of chemotherapy alone has 1.532 times the hazard of chemotherapy plus radiation, but this result is nonsignificant.

Chapter 7

1. F: They are multiplicative models, although additive on the log scale.
2. T
3. T
4. F: If the AFT assumption holds in a log-logistic model, the proportional odds assumption holds.
5. F: An acceleration factor greater than one suggests the exposure is beneficial to survival.
6. T
7. T
8. T
9. F: ln(T) follows an extreme value minimum distribution.
10. F: The subject is right-censored.
11. $$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2 PRISON + \alpha_3 DOSE + \alpha_4 PRISDOSE]}{\exp[\alpha_0 + \alpha_1(1) + \alpha_2 PRISON + \alpha_3 DOSE + \alpha_4 PRISDOSE]}$$
$$= \exp(\alpha_1)$$
$$\hat{\gamma} = \exp(0.698) = 2.01$$
$$95\%\ \text{CI} = \exp[0.698 \pm 1.96(0.158)] = (1.47,\ 2.74)$$

    The point estimate for the acceleration factor (2.01) suggests that the survival time (time off heroin) is double for those enrolled in CLINIC = 2 compared to CLINIC = 1. The 95% confidence interval does not include the null value of 1.0 indicating a statistically significant preventive effect for CLINIC = 2 compared to CLINIC = 1.
12. $$\text{HR} = \frac{\exp[\beta_0 + \beta_1(2) + \beta_2 PRISON + \beta_3 DOSE + \beta_4 PRISDOSE]}{\exp[\beta_0 + \beta_1(1) + \beta_2 PRISON + \beta_3 DOSE + \beta_4 PRISDOSE]}$$
$$= \exp(\beta_1)$$
$$\hat{HR} = \exp(-0.957) = 0.38$$
$$95\%\ \text{CI} = \exp[-0.957 \pm 1.96(0.213)] = (0.25,\ 0.58)$$

    The point estimate of 0.38 suggests the hazard of going back on heroin is reduced by a factor of 0.38 for those enrolled in CLINIC = 2 compared to CLINIC = 1. Or from the other perspective: the estimated hazard is elevated for those in CLINIC = 1 by a factor of exp (+0 957) = 2.60.
13. $\beta_1 = -\alpha_1 p$ for CLINIC, so $\hat{\beta}_1 = -(0.698\text{X}1.370467) = -0.957$, which matches the output for the PH form of the model.

14. The product term PRISDOSE is included in the model as a potential confounder of the effect of CLINIC on survival. It is not an effect modifier because under this model the hazard ratio or acceleration factor for CLINIC does not depend on the value of PRISDOSE. The PRISDOSE term would cancel in the estimation of the hazard ratio or acceleration factor (see Questions 11 and 12). On the other hand, a product term involving CLINIC would be a potential effect modifier.
15. Using the AFT form of the model:

$$\frac{1}{\lambda^{1/p}} = \exp[\ \alpha_0 + \ \alpha_1\,\text{CLINIC} + \ \alpha_2\,\text{PRISON} + \ \alpha_3\,\text{DOSE} + \ \alpha_4\,\text{PRISDOSE}]$$

Median survival time for CLINIC = 2, PRISON = 1, DOSE = 50, PRISDOSE = 100:

$$\text{t} = [-\ln\ \text{S(t)}]^{1/p} \times \frac{1}{\lambda^{1/p}} = [-\ln(0.5)]^{1/p} \times \exp[\beta_0 + 2\beta_1 + \beta_2 + 50\beta_3 + 100\beta_4]$$

$\hat{t}$ (median) = 403.66 days (obtained by substituting parameter estimates from output).

16. Using the same approach as the previous question: Median survival time for CLINIC = 1, PRISON = 1, DOSE = 50, PRISDOSE = 100:
$\text{t} = [-\ln(0.5)]^{1/p} \times \exp[\beta_0 + \text{l}\beta_1 + \beta_2 + 50\beta_3 + 100\delta_4]$
$\hat{t}$ (median) = 200.85 days.
17. The ratio of the median survival times is 403.66/200.85 = 2.01. This is the estimated acceleration factor for CLINIC = 2 vs. CLINIC = 1 calculated in Question 11. Note that if we used any survival probability (i.e., any quantile of survival time), not just S(ι) -= 0.5 (the median), we would have obtained the same ratio.
18. The addition of the frailty component did not change any of the other parameter estimates nor did it change the log likelihood of −260.74854.
19. If the variance of the frailty is zero (theta = 0), then the frailty has no effect on the model. A variance of zero means the frailty (α) is constant at 1. Frailty is defined as a multiplicative random effect on the hazard $\text{h(t}|\alpha) = \alpha\text{h(t)}$. If $\alpha = 1$ then $\text{h(t}|\alpha) = \text{h(t)}$, and there is no frailty.

Chapter 8

1. a. **Survival time (say, in weeks) to the first event (stratum 1)**:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	2	0	0	{B,L}
12	2	1	0	{B,L}
20	1	1	0	{L}

b. For each approach, the observation for the first event is identical.

c. **Survival rime (say, in weeks) from the first to the second event (stratum 2) using the Stratified CP approach**:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	0	0	0	–
16	1	1	0	{B}
23	1	1	0	{L}

d. **Survival time (say, in weeks) from the first to the second event (stratum 2) using the Gap Time approach**:

$t_{(f)}$	n_f	m_f	q_f	$R_{(f)})$
0	2	0	0	{B,L}
3	2	1	0	{B,L}
4	1	1	0	{B}

e. **Survival time (say, in weeks) from the first to the second event using the Marginal approach**:

$t_{(f)}$	n_f	m_f	q_f	$R(t_{(f)})$
0	2	0	0	{B,L}
16	2	1	0	{B.L}
23	1	1	0	{L}

f. Correct choice is iii.
Bonnie is at risk for a second event between times 12 to 16.
Lonnie is at risk for a second event between times 20 to 23.
Neither is in the risk set for the other's second event.

g. Correct choice is ii.
Bonnie is at risk for a second event between times 0 to 4.
Lonnie is at risk for a second event between times 0 to 3.
Bonnie is in the risk set when Lonnie gets her second event.

h. Correct choice is i.
Bonnie is at risk for a second event between times 0 to 16.
Lonnie is at risk for a second event between times 0 to 23.
Lonnie is in the risk set when Bonnie gets her second event.

2. a. Cox PH Model for **CP** approach to Defibrillator Study:

$$h(t,\mathbf{X}) = h_0(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma\ \mathbf{smoking}]$$

where **tx** = 1 if treatment A, 0 if treatment B. **smoking status** = 1 if ever smoked, 0 if never smoked.

b. Using the **CP** approach, there is no significant effect of treatment status adjusted for smoking. The estimated hazard ratio for the effect of treatment is 1.09, the corresponding P-value is 0.42 and a 95% CI for the hazard ratio is (0.88, 1.33).

c. No-interaction SC model for **Marginal** approach:

$$h_g[t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta\ \mathbf{tx} + \gamma\ \mathbf{smoking}],\ g = 1, 2,3$$

Interaction SC model for **Marginal** approach:

$$h_g[t,\mathbf{X}) = h_{0g}(t)\mathbf{exp}[\beta_g\ \mathbf{tx} + \gamma_g\ \mathbf{smoking}],\ g = 1, 2, 3$$

d. $\mathbf{LR} = -2\ln L_R - (-21n\ L_F)$ is approximately χ^2 with 4 df under
$\mathbf{H_0}$:no-interaction SC model is appropriate, where R denotes the reduced (no interaction SC) model and **F** denotes the full (interaction SC) model

e. The use of a no-interaction model does not allow you to obtain stratum-specific HR estimates, even though you are assuming that strata are important.

f. The **CP** approach makes sense for these data because recurrent defibrillator (shock) events on the same subject are the same kind of event no matter when it occurred.

g. You might use the **Marginal** approach if you determined that different recurrent events on the same subject were different because they were of different order.

h. The number in the risk set (n_f) remains unchanged through day 68 because every subject who failed by this time was still at risk for a later event.

i. Subjects 3,6, 10,26, and 31 all fail for the third time at day 98 and are not followed afterwards.

j. Subjects 9, 15, and 28 fail for the second time at 79 days, whereas subject #16 is censored at 79 days.

k. Subjects 4, 14, 15, 24, and 29 were censored between days 111 and 112.

l. Subject #5 gets his first event at 45 days and his second event at 68 days, after which he drops out of the study. This subject is the first of the 36 subjects to drop out of the study, so the number in the risk set changes from 36 to 35 after 68 days.

m. None of the above.

n. The product limit formula is not applicable to the **CP** data; in particular, $P(T > t|T \geq t)$ does not equal "# failing in time interval /# in the risk set at start of interval."

o. Use the information provided in Table T.2 to complete the data layouts for plotting the following survival curves.

i. $S_1(t) = Pr(T_1 > t)$ where T_1 = time to first event from study entry

$t_{(f)}$	n_f	m_f	q_f	$S(t_p) = S(t_{f-1}) \times Pr(T_1 > t \mid T_1 \geq t)$
0	36	0	0	1.00
33	36	2	0	0.94
34	34	3	0	0.86
36	31	3	0	0.78
37	28	2	0	0.72
38	26	4	0	0.61
39	22	5	0	0.47
40	17	1	0	0.44
41	16	1	0	0.42
43	15	1	0	0.39
44	14	1	0	0.36
45	13	2	0	0.31
46	11	2	0	0.25
48	9	1	0	0.22
49	8	1	0	0.19
51	**7**	**2**	**0**	**0.19 × 5/7 = 0.14**
57	**5**	**2**	**0**	**0.14 × 3/5 = 0.08**
58	**3**	**2**	**0**	**0.08 × 1/3 = 0.03**
61	**1**	**1**	**0**	**0.03 × 0/1 = 0.00**

ii. **Gap Time** $S_{2c}(t) = Pr(T_{2c} > t)$ where T_{2c} = time to second event from first event.

$t_{(f)}$	n_f	m_f	q_f	$S_2(t_{(f)}) = S_2(t_{(f-1)}) \times Pr(T_2 > t \mid T_2 \geq t)$
0	36	0	0	1.00
5	36	1	0	0.97
9	35	1	0	0.94
18	34	2	0	0.89
20	32	1	0	0.86
21	31	2	1	0.81
23	28	1	0	0.78
24	27	1	0	0.75

(*Continued on next page*)

(*Continued*)

$t_{(f)}$	n_f	m_f	q_f	$S_2(t_{(f)}) = S_2(t_{(f-1)}) \times \Pr(T_2 > t \mid T_2 \geq t)$
25	26	1	0	0.72
26	25	2	0	0.66
27	23	2	0	0.60
28	21	1	0	0.58
29	20	1	0	0.55
30	19	1	0	0.52
31	18	3	0	0.43
32	15	1	0	0.40
33	14	5	0	0 26
35	9	1	0	0.23
39	8	2	0	0.17
40	**6**	**2**	**0**	**0.17 × 4/6 = 0.12**
41	**4**	**1**	**0**	**0.12 × 3/4 = 0.09**
42	**3**	**1**	**0**	**0.09 × 2/3 = 0.06**
46	**2**	**1**	**0**	**0.06 × 1/2 = 0.03**
47	**1**	**1**	**0**	**0.03 × 0/1 = 0.00**

iii. **Marginal** $S_{2m}(t) = \Pr(T_{2m} > t)$ where T_{2m} = time to second event from study entry.

$t_{(f)}$	n_f	m_f	q_f	$S(t_{(f)}) = S_2(t_{(f-1)}) \times \Pr(T_2 > t \mid T_2 \geq t)$
0	36	0	0	1.00
63	36	2	0	0.94
64	34	3	0	0.86
65	31	2	0	0.81
66	29	3	0	0.72
67	26	4	0	0.61
68	22	2	0	0.56
69	20	1	0	0.53
70	19	1	0	0.50
71	18	1	0	0.47
72	17	2	0	0.42
73	15	1	0	0.39
74	14	1	0	0.36
76	13	1	0	0.33
77	12	1	0	0.31
78	11	2	0	0.25
79	**9**	**3**	**1**	**0.25 × 6/9 = 0.17**
80	**5**	**2**	**0**	**0.17 × 3/5 = 0.10**
81	**3**	**2**	**0**	**0.10 × 1/3 = 0.03**
97	**1**	**1**	**0**	**0.03 × 0/1 = 0.00**

p. The survival curves corresponding to the above data layouts will differ because they are describing different survival functions. In particular, the composition of the risk set differs in all three data layouts and the ordered survival times being plotted are different as well.

Chapter 9

1. Cause-specific no interaction model for local recurrence of bladder cancer (event = 1):

$$h_1(t,\mathbf{X}) = h_{01}(t)\exp[\beta_{11}tx + \beta_{21}num + \beta_{31}size]$$

2. Censored subjects have bladder metastasis (**event** = 2) or other metastasis (**event** = 3).
3. Cause-specific no-interaction model for bladder metastasis (**event** =2):

$$h_2(t,\mathbf{X}) = h_{02}(t)\exp[\beta_{12}tx + \beta_{22} + \beta_{32}size]$$

where censored subjects have local recurrence of bladder cancer (**event** = 1) or other metastasis (**event** = 3).

4. A sensitivity analysis would consider worst-case violations of the independence assumption. For example, subjects censored from failing from events = 2 or 3 might be treated in the analysis as either all being event-free (i.e., change event status to 0 and time to 53) or all experiencing the event of interest (i.e., change event status to 1 and leave time as is).
5. a. Verify the CIC_1 calculation provided at failure time t_f=8 for persons in the treatment group (**tx** = 1):

$$\hat{\mathbf{h}}_1(8) = 1/23 = 0.0435$$
$$\hat{\mathbf{S}}(4) = \hat{\mathbf{S}}(3)\mathbf{Pr}(T > 4|T \geq 4) = 0.9630(1 - 2/26)$$
$$= 0.9630(0.9231) = 0.8889$$
$$\hat{\mathbf{i}}_1(8) = \hat{\mathbf{h}}_1(8)\hat{\mathbf{S}}(4) = 0.0435(.8889) = 0.0387$$
$$\mathbf{CIC}_1(8) = \mathbf{CIC}_1(4) + 0.0387 = 0 + 0.0387 = 0.0387$$

   b. Verify the $\mathbf{CIC}_1$ calculation provided at failure time t_f= 25 for persons in the placebo group (**tx** = 0):

$$\hat{\mathbf{h}}_1(25) = 1/6 = 0.1667$$
$$\hat{\mathbf{S}}(23) = \hat{\mathbf{S}}(21)\mathbf{Pr}(T > 23|T \geq 23) = 0.4150(1 - 1/8)$$
$$= 0.4150(0.875) = 0.3631$$
$$\hat{\mathbf{I}}_1(25) = \hat{\mathbf{h}}_1(25)\hat{\mathbf{S}}(23) = 0.1667(.3631) = 0.0605$$
$$\mathbf{CIC}_1(25) = \mathbf{CIC}_1(23) + 0.0605 = 0.2949 + 0.0605$$
$$= 0.3554$$

   c. interpret the $\mathbf{CIC}_1$ values obtained for both the treat ment and placebo groups at $t_f = 30$.
   For tx = 1, $\mathbf{CIC}_1(t_f = 30) = 0.3087$ and for tx = 0, $\mathbf{CIC}_1(t_f = 30) = 0.3554$.

Thus, for treated subjects (tx = 1), the cumulative risk (i.e., marginal probability) for local bladder cancer recurrence is about 30.1 % at 30 months when allowing for the presence of competing risks for bladder metastasis or other metastasis.

For placebo subjects (tx = 1), the cumulative risk (i.e., marginal probability) for local bladder cancer recurrence is about 35.5% at 30 months when allowing for the presence of competing risks for bladder metastasis or other metastasis.

The placebo group therefore has a 5% increased risk of failure than the treatment group by 30 months of follow-up.

d. Calculating the $\mathbf{CPC_1}$ values for both treatment and placebo groups at $t_f = 30$:

The formula relating CPC to CIC is given by $\mathbf{CPC_c} = CIC_c/(1 - CTC_{c'})$ where CIC_c = CIC for cause-specific risk event = 1 and $CIC_{c'}$ = CIC from risks for events = 2 or 3 combined

For tx = l, $\mathbf{CIC_1}(t_f = 30) = 0.3087$ and for tx = 0, $CIC_1(t_f = 30) = 0.3554$.

The calculation of $CIC_{c'}$ involves recoding the event variable to 1 for subjects with bladder metastasis or other metastasis and 0 otherwise and then computing $CIC_{c'}$. Calculation of $\mathbf{CIC_{c'}}$ involves the following calculations.

tx = 1 (Treatment A)

t_f	n_f	d_{1f}	$\hat{h}_1(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_1(t_f)$	$CIC_{1'}(t_f)$
0	27	0	0	—	—	—
2	27	1	**.0370**	**l**	**.0370**	**.0370**
3	26	2	**.0769**	**.9630**	**.0741**	**.1111**
4	24	0	0	.8889	0	**.1111**
8	**23**	**1**	**.0435**	**.8889**	**.0387**	**.1498**
9	21	1	**.0476**	**.8116**	**.0386**	**.1884**
10	20	1	**.0500**	**.7729**	**.0386**	**.2270**
15	17	1	**.0588**	**.7343**	**.0432**	**.2702**
16	15	1	**.0667**	**.6479**	**.0432**	**.3134**
18	14	0	0	.6047	0	.3134
22	12	0	0	.6047	0	.3134
23	11	0	0	.5543	0	.3134
24	8	0	0	.5039	0	.3134
26	7	0	0	.4409	0	.3134
28	4	1	**.2500**	**.3779**	**.0945**	**.4079**
29	2	0	0	.2835	0	.4079
30	1	0	0	.2835	0	.4079

tx = 0 (Placebo)

t_f	n_f	d_{1f}	$\hat{h}_1(t_f)$	$\hat{S}(t_{f-1})$	$\hat{I}_1(t_f)$	$CIC_{\mathbf{1}'}(t_f)$
0	26	0	0	—	—	—
1	26	0	0	1	0	0
2	24	0	0	.9615	0	0
3	23	0	0	.9215	0	0
5	21	1	**.0476**	**.8413**	**.0400**	**.0400**
6	20	2	**.1000**	**.8013**	**.0801**	**.1201**
7	18	1	**.0556**	**.7212**	**.0401**	**.1602**
10	16	1	**.0625**	**.6811**	**.0426**	**.2028**
12	15	1	**.0667**	**.6385**	**.0426**	**.2454**
14	13	0	0	6835	0	.2454
16	12	1	**.0833**	**.5534**	**.0461**	**.2915**
17	10	0	0	.4612	0	.2915
18	9	0	0	.4150	0	.2915
21	8	1	**.1250**	**.4150**	**.0519**	**.3434**
23	7	0	0	.3632	0	.3434
25	6	1	**.1667**	**.3632**	**.0605**	**.4039**
29	4	0	0	.2421	0	.4039
30	2	0	0	.2421	0	.4039

From these tables, for tx = 1, $CIC_{\mathbf{1}'}((t_f) = 30) = 0.4079$, and for tx = 0, $CIC_{\mathbf{1}'}((t_f) = 30) = 0.4039$.
Thus, for tx =1, $\mathbf{CPC_1}((t_f)=30) = 0.3087/(1 - 0.4079) = \mathbf{0.5213}$, and for tx = 0, $\mathbf{CPC_1}((t_f) = 30) = 0.3554/(1 - 0.4039) = \mathbf{0.5962}$.

6. a. $HR_1(\text{tx} = 1 \text{ vs. tx} = 0) = 0.535 (= 1/1.87)$, p-value = 0.250, N.S.
   b. $HR_2(\text{tx} = 1 \text{ vs. tx} = 0) = 0.987$, p-value = .985, N.S.
   c. $HR_3(\text{tx} = 1 \text{ vs. tx} = 0) = 0.684\ (= 1/1.46)$, p-value = .575, N.S.
7. a. Hazard model formula for the LM model:

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 \text{tx} + \beta_2 \text{num} + \beta_3 \text{size} + \delta_1(\text{txd2}) + \delta_2(\text{numd2}) + \delta_3(\text{sized2}) + \delta_4(\text{txd3}) + \delta_5(\text{numd3}) + \delta_6(\text{sized3})]$$

$g = 1, 2, 3$

where
d2 = 1 if bladder metastasis and 0 otherwise, and
d3 = 1 if or other metastasis and 0 otherwise

b. Hazard ratios for the effect of each of the 3 cause-specific events:

$$\begin{aligned}
HR_1(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.6258) \\
&= 0.535(= 1/1.87) \\
HR_2(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.6258 + .6132) \\
&= 0.987(= 1/1.01) \\
HR_3(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.6258 + .2463) \\
&= 0.684(= 1/1.46)
\end{aligned}$$

c. Corresponding HRs are identical.

8. a. Hazard model formula for the LM_{alt} model:

$$\begin{aligned}
h'_g(t, \mathbf{X}) \quad &= h'_{0g}(t) \exp[\ \delta'_{11}\,\text{txd1} + \delta'_{12}\,\text{numd1} + \delta'_{13}\,\text{sized1} \\
g = 1, 2, 3 \qquad &+ \delta'_{21}\text{txd2} + \delta'_{22}\text{numd2} \\
&+ \delta'_{23}\text{sized2} + \delta'_{31}\text{txd3} \\
&+ \delta'_{32}\text{numd3} + \delta'_{33}\text{sized3}]
\end{aligned}$$

where
d1 = 1 if local bladder cancer recurrence and 0 otherwise
d2 = 1 if bladder metastasis and 0 otherwise, and
d3 = 1 if or other metastasis and 0 otherwise

b. Hazard ratios for the effect of each of the three cause-specific events:
output.

$$\begin{aligned}
HR_1(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.6258) \\
&= 0.535(= 1/1.87) \\
HR_2(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.0127) \\
&= 0.987(= 1/1.01) \\
HR_3(tx = 1 \text{ vs. } tx = 0) &= \exp(-0.3796) \\
&= 0.684(= 1/1.46)
\end{aligned}$$

c. Corresponding hazard ratios arc identical.

9. No interaction SC LM model:

$$\begin{aligned}
h^*_g(t, \mathbf{X}) \quad &= h^*_{0g}(t) \exp[\ \beta_1\,\text{tx} + \beta_2\,\text{num} + \beta_8\,\text{size}] \\
g = 1, 2, 3
\end{aligned}$$

Assumes $HR_1(X) = HR_2(X) = HR_3(X)$ for any X variable e.g., Rx = 0 vs. Rx = 1:
$HR_1(tx) = HR_2(tx) = HR_3(tx) = \exp[\beta_1]$

10. Carry out the following likelihood ratio test:

$$H_0: \ \delta_{gj} = 0 \quad g = 2, 3; \quad j = 1, 2, 3$$

where δ_{gj} is coefficient of $D_g X_j$ in the interaction SC LM model
$LR = 2\log L_R - (-2\text{Log}L_F)$ approx χ^2_6 under H_0
R = no-interaction SC (reduced) model
F = interaction SC (full) model

Chapter 10

1. Example: A=2, F=2, so Mt=A/2 + F =3, R=2
   α=0.05, β=0.10
   $\lambda_0 = 0.10$, $\lambda_1 = 0.05$, $\Delta = \lambda_0/\lambda_1 = 2$

   $N_{EV} = \{(1.96 + 1.282)[2(2) + 1]/[\sqrt{2}(2 - 1)]\}^2$
   $= 131.382 \approx 132$

   Using Formula 1:

   $$N = \frac{131.382}{\frac{2}{2+1}\{1 - e^{-2(0.05)(3)}\} + \frac{1}{2+1}\{1 - e^{-(0.05)3}\}}$$

   (N_{EV} ← 131.382)

   $$= \frac{131.832}{0.2192} = 601.4 \approx 602$$

   (P_{EV} ← 0.2192)

   $N_1 = [2/3]601.4 = 400.93 \approx 401$ and
   $N_0 = 400.9/2 = 200.45 \approx 200$

2. Nev =131.383 = 132 from question 1.

   $$P_{EV1} = 1 - \frac{1}{(0.05)(2)}\left[e^{-(0.05)(2)} - e^{-0.05)(2+2)}\right]$$
   $$= 1 - 0.8611 = 0.1389$$

   $$P_{EV0} = 1 - \frac{1}{(0.10)(2)}\left[e^{-(0.10)(2)} - e^{-(0.10)(2+2)}\right]$$
   $$= 1 - 0.7421 = 0.2579$$

   $$N = \frac{131.382}{\frac{2}{2+1}(0.1389) + \frac{1}{2+1}(0.2579)}$$
   $$= \frac{131.832}{0.1786} = 738.14 \approx 739$$

   $N_1 = [2/3]738.14 = 492.09 \approx 492$ and
   $N_0 = 492.09/2 = 246.04 \approx 246$.

3. The results using Formulae 1 and 2 are somewhat different since Formula 1 yields N=602 whereas Formula 2 yields N=739. Formula 1 uses the median follow-up time M_F in the computation of p_{EVi} whereas Formula 2 computes p_{EVi} by assuming that the time X at which any subject enters the study has the uniform distribution over the accrual period.
4. $N_{LOFadj} = 739/(1 - 0.25) = 985.33 \approx 986$
5. $N_1 = [2/3]985.33 = 656.89 \approx 657$ and $N_0 = 656.89/2 = 328.44 \approx 328$
6. $N_{ITTadj} = 986/(1 - 0\ 05\ -0.10)^2 = 1364.71 \approx 1365$

7. $N_1 = [2/(2+l)]1364.71 = 909.81 \approx 910$ and
   $N_0 = 909.81/2 = 454.91 \approx 455$
8. From question 6, the required accrual rate is $r = N/A = 1365/2 = 682.5 \approx 683$ subjects per year. If this accrual rate is not feasible, but r* was considered feasible, then you can adjust your sample size by reducing the accrual period to $A^* = N/r^*$. For example, if the maximum for r is $r_{max} = 600$/yr, then the required accrual period is modified from A=2 to $A^* = 2.275$ years.

   Now suppose, we keep N_{EV} (=131.382), F=2, R=2, α=0.05, β=0.10, $\lambda_0 = 0.10$, $\lambda_1 = 0.05$, and $\Delta = \lambda_0/\lambda_1 = 2$ all constant, but increase the accrual time to A*=2.275 years. Then we would need to re-compute p_{EV1}, p_{EV0} and N to obtain p_{EV1}= 0.2677, p_{EV0}= 0.1447, and N = 579.541 (prior to adjusting for LOF and Crossovers), which is modified to N* = 1069.51 after adjusting for 25% LOF rate, 5% d_c rate and 10% d_t, rate. For this modified sample size, the modified required accrual rate is $r^* = N^*/A^* = 1069.71/2.275 = 470.11$, which is less than $r_{max} = 600$, so that the study is feasible.

   Note, however, it is also possible to obtain a feasible study if the accrual period remains at A=2, but the follow-up period increases to, say F=4, again keeping N_{EV}(=131.382), R = 2, $\alpha = 0.05$, $\beta = 0.10$, $\lambda_0$0.10, $\lambda_1 = 0.05$, and $\Delta = \lambda_0/\lambda_1 = 2$ all constant. This will require re-computing p_{EV1}, p_{EV0} and N again, followed by adjustments for LOF and Crossovers. In particular, if F is increased (to say F=4), then p_{EV1} and p_{EV0} should correspondingly increase from previously calculated values because the probability for an event occurring should increase if follow-up time is increased.

References

Andersen P.K., Borgan O., Gill R.D., and Keiding N. 1993. *Statistical Models Based on Counting Processes*. Springer Publishers, New York.

AREDS Research Group. 2003. Potential public health impact of age-related eye disease study results. *Arch Opthalmol*, 121: 1621–1624.

Arriagada R., Rutqvist L.E., Kramar A., and Johansson H. 1992. Competing risks determining event-free survival in early breast cancer. *Br. J. Cancer*, 66(5): 951–957.

Berkson J. and Gage R.P. 1952. Survival curve for cancer patients following treatment. *J. Amer. Statist. Assoc.*, 47, 501–515.

Boag J.Q. 1949. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. Roy. Statist. Soc.*, 11, 15–53.

Brookmeyer R. and Crowley J. 1982. A Confidence Interval for the Median Survival Time. *Biometrics* 38 (1): 29–41.

Byar D. 1980. The Veterans Administration study of chemoprophylaxis for recurrent stage I bladder tumors: Comparisons of placebo, pyridoxine, and topical thiotepa. In *Bladder Tumors and Other Topics in Urological Oncology*. Plenum Publishers, New York: 363–370.

Byar D. and Green S. 1980. The Choice of treatment for Cancer Patients based on Covariate Information. *Bulletin du Cancer* 67: 4, 477–490.

Byar D. and Corle D. 1977. Selecting optimal treatment in clinical trials using covariate information. *J Chronic Dis*, 30, 445–459.

Cantor A. 1992. Sample size calculations for the log-rank test: A Gompertz model approach. *J. Clin Epidemiol*, 45, 1131–1136.

Caplehorn J., et al. 1991. Methadone dosage and retention of patients in maintenance treatment. *Med. J. Aust.*, 154, 195–199.

Clayton D. 1994. Some Approaches to the Analysis of Recurrent Event Data. *Statistical Methods in Medical Research*. 3: 244–262.

Cox D.R. and Oakes D. 1984. *Analysis of Survival Data*. Chapman and Hall, London.

Crowley J. and Hu M. 1977. Covariance analysis of heart transplant data. *J. Amer. Stat. Assoc.*, 72, 27–36.

Dixon W.J. 1990. *BMDP Statistical Software Manual*. Berkeley, CA, University of California Press.

Fine J. and Gray R. 1999. A proportional hazards model for the subpopulation of a competing risk. *J. Amer. Stat. Assoc.*, 94, 496–509.

Freedman L.S. 1982. Tables of the number of patients required for clinical trials using the logrank test. *Statistics in Medicine*. 1, 121–129.

Freireich E.O., et al. 1963. The effect of 6-mercaptopmine on the duration of steroid induced remission in acute leukemia. *Blood*, 21, 699–716.

Gebski V. 1997. *Analysis of Censored and Correlated Data (ACCORD)*. Data Analysis and Research Technologies, Eastwood, NSW, Australia.

George S.L. and Desu M.M. 1974. Planning the size and duration of a clinical trial studying the time to some critical event. *J. Chron. Dis.* 27: 15–24.

Goldman A.I. 1984. Survivorship analysis when cure is a possibility: A Monte Carlo study. *Statist. in Med.*, 3: 153–163.

Grambsch P.M. and Therneau T.M. 1994 Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81: 515–526.

Gray R.J. 1988. A class of k-sample tests for comparing the cumulative incidence of a competing risk. *Annals of Stat* 16, 1141–1154.

Gutierrez R.G. 2002. Parametric frailty and shared frailty survival models. *Stata J.* 2: 22–24.

Harrell F. and Lee K. 1986. *Proceedings of the Eleventh Annual SASW User's Group International*: 823–828.

Harris E. and Albert A. 1991. *Survivorship Analysis for Clinical Studies*. Marcel Dekker Publishers, New York.

Hosmer D.W. and Lemeshow S. 2008. *Applied Survival Analysis- 2nd Edition*. John Wiley &Sons, New York.

Kalbfleisch J.D. and Prentice R.L. 2002. *The Statistical Analysis of Failure Time Data- Second Edition*. John Wiley and Sons, New York.

Kaplan E.L. and Meier P. 1958. Nonparametric Estimation from Incomplete Observations. *J. Amer. Statist. Assoc*, 53: 457–481.

Kay R. 1986. Treatment effects in competing-risk analysis of prostate cancer data, *Biometrics* 42, 203–211.

Klein J.P. and Moeschberger M.L. 2003. *Survival Analysis- Techniques for Censored and Truncated Data, 2nd Edition*. Springer Publishers, New York.

Kleinbaum D.G. and Klein M. 2010. *Logistic Regression- A Self Learning Text-Third Edition* (Chapter 14). Springer Publishers, New York.

Kleinbaum D.G., Kupper L.L., and Morgenstern H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. John Wiley and Sons, New York.

Kleinbaum D.G., Kupper L.L., Nizam A., and Muller, K.A. 2008. *Applied Regression Analysis and Other Multivariable Methods, Fourth Edition*. Cengage Learning, Inc, Florence, KY.

Korn E.L., Graubard B.I., and Midthune D. 1997. Time-to-event analysis of longitudinal follow-up for a survey: choice of the time scale. *Am. J. Epid* 145: 72–80.

Krall J.M., Uthoff V.A., and Harley J.B. 1975. A step-up procedure for selecting variables associated with survival data. *Biometrics*, 31, 49–57.

Lachlin J.M. 1981. Introduction to Sample Size Determination and Power Analysis for Clinical Trials, *Controlled Clinical Trials* 2, 93–113.

Lee E.T. 1980. *Statistical Methods for Survival Data Analysis*. Wadsworth Publishers, Belmont, CA.

Lin D.Y. and Wei L.J. 1989. The robust inference for the Cox proportional hazards model. *J. Amer. Statist. Assoc.* 84: 1074–1078.

Lunn M. 1998. Applying k-sample tests to conditional probabilities for competing risks in a clinical trial. *Biometrics* 54, 1662–1672.

Lunn M. and McNeil D. 1995. Applying Cox regression to competing risks. *Biometrics* 51, 524–532.

Makuch R.W. and Parks W.P. 1988. Statistical methods for the analysis of HIV-1 core polypeptide antigen data in clinical studies. *AIDS Research and Human Retroviruses* 4: 305–316.

Pasternack B.S. and Gilbert H.S. 1971. Planning the duration of long-term survival time studies designed for accrual by cohorts. *J. Chronic Dis.* 24: 681–700

Pencina M.J., Larson M.G., and D'Agostino R.B. 2007. Choice of time scale and its effect on significance of predictors in longitudinal studies. *Statist. in Med.* 26: 1343–1359.

Pepe M.S. and Mori M. 1993. Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statist. in Med.* 12, 737–751.

Prentice R.L. and Marek P. 1979. A qualitative discrepancy between censored data rank tests. *Biometrics*, 35(4): 861–867

Prentice R.L., Williams B.J., and Peterson A.V. 1981. On the Regression Analysis of Multivariate Failure Time Data. *Biometrika* 68 (2):373–79.

Rubinstein L., Gail M., and Santner T. 1981. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chron. Dis.* 34: 469–479.

Schoenbach V.J., Kaplan B.H., Fredman L., and Kleinbaum D.G. 1986. Social ties and mortality in Evans County, Georgia. *Amer. J. Epid.*, 123:4, 577–591.

Schoenfeld D. 1982. Partial residuals for the proportional hazards model. *Biometrika*, 69, 51–55.

Stablein D., Carter W., and Novak J. 1981. Analysis of survival data with non-proportional hazard functions. *Controlled Clinical Trials*, 2, 149–159.

Tai B.C., Machin D., White I., and Gebski V. 2001. Competing risks analysis of patients with osteosarcoma: a comparison of four different approaches. *Stat. Med.* 20(5): 661–684.

Thiebaut A.C. and Benichou J. 2004. Choice of time-scale in Cox's model analysis of epidemiologic cohort data: a simulation study. *Stat. Med.* 23: 3803–3820.

Wei L.J., Lin D.Y., and Weissfeld L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions, *J. Amer. Statist. Assoc.*, 84 (408).

Zeger S.L. and Liang C.Y. 1986. Longitudinal Data Analysis for Discrete and Continuous Outcomes. *Biometrics* 42, pp 121–130.

Index

A

Accelerated failure time (AFT)
  assumption, 298–300
  models, 297–298, 314–316, 341, 345–349
Acceleration factor, 299–300
  exponential, 301–303
  with frailty, 337, 340
  log-logistic, 313
  Weibull, 308
Accrual period, 505, 507–512, 514, 516, 518, 520, 522
Addicts dataset, 526
  data analysis, 260–264, 280
  with R programming 620–663
  with SAS programming, 570–607
  with SPSS programming, 607–620
  with STATA programming, 527–570
Additive failure time model, 317
Adjusted survival curves
  log-log plots, 174–175, 189
  observed *vs.* expected plots, 175–180
  stratified Cox procedure, 208
  using Cox PH model, 120–123, 144, 147
AFT. *See* Accelerated failure time
Age as time scale, 131, 134, 142, 144, 147
Age-Related Eye Disease Study (AREDS), 391–395
Age-truncated, 138
  Cox models for, 138–142, 144, 148
Akaike's information criterion (AIC), 318
Average hazard rate, 28

B

Baseline hazard function, 108–109, 111, 145
Biased results, 438
Binary regression, 322
Bladder cancer dataset, 527

Bladder cancer patients
comparison of results for, 385–389
counting process for first, 30 subjects, 371
hypothetical subjects, 368–369
interaction model results for, 386
no-interaction model results for, 386
Byar data, 433–434
cause-specific competing risk analysis, 435–437
Lunn–McNeil models, 455, 461

C

Cause-specific hazard function, 434
Censored data, 5–8, 37–41
interval-censored, 318, 321
left-censored, 7, 318
right-censored, 7, 318
Censoring, 5
informative (dependent), 408–410
non-informative (independent), 405–409
Closed cohort, 134–135, 139
Competing risks, 4, 8, 426, 430
CIC, 444
CPC, 453
examples of data, 474–476
independence assumption, 437
Lunn–McNeil models, 455, 461
separate models for different event types, 434–437
Complementary log-log
binary model, 325
link function, 324
Conditional failure rate, 12
Conditional probability curves (CPC), 453–455
Conditional survival function, 327
Confidence intervals
for hazard ratio when interaction in PH model, 117–119, 143, 146
for KM curves, 78–79, 81, 86
for median survival time, 80, 82, 86
Confounding effect, 30–31
Counting process (CP) approach, 366, 368–379, 385–389, 392, 398–400, 402–404, 408, 410
example, 368–369, 373–375, 377–382
general data layout, 370–371
Counting process format, 20–23, 46–47
for age as time scale analysis, 142, 144
for extended Cox model, 271–273
Cox adjusted survival curves
using SAS, 588–589
using SPSS, 614–615
using Stata, 547–550
Cox likelihood, 127–131
extended for time dependent variables, 223–225
Cox PH cause-specific model, 434
Cox proportional hazards (PH) model
adjusted survival curves using, 120–123
computer example using, 100–108
extension of (*see* Extended Cox model)
formula for, 108–110
maximum likelihood estimation of, 112–114
popularity of, 110–112
review of, 244–246
using SAS, 576–580
using SPSS, 613
using Stata, 538–543
CP approach. *See* Counting process approach
CPC. *See* Conditional probability curves
Crossover observation, 517, 521
Cumulative incidence, viii
Cumulative incidence curves (CIC), 427, 444–455

D

Data layout for computer
augmented (Lunn–McNeil approach) data layout for, 456
counting process data layout for, 370–371
general data layout for, 16–23
marginal approach data layout for, 380–381
Datasets, 526–527
Decreasing Weibull model, 14
Discrete survival analysis, 325
Drop-in/drop-out observation, 517, 521

E

Effect size, 501–502, 507, 509, 514, 518
Empirical estimation, 376

Estimated-ln(-ln) survivor curves, 166
Estimated survivor curves, 29
Evans County Study
  Cox proportional hazards (PH) model application to, 154–156
  Kaplan-Meier survival curves for, 87–89
  multivariable example using, 33–35
  ordered failure times for, 53–54
  survival data from, 149–152
Event, 4
  types, different, separate models for, 434–437
Expected *vs.* observed plots, 175–180
Exponential regression, 13
  accelerated failure-time form, 300–304
  log relative-hazard form, 295–297
Extended Cox likelihood, 269–274
Extended Cox model, 126, 249
  application to Stanford heart transplant data, 265–269
  application to treatment of heroin addiction, 260–264
  hazard ratio formula for, 251–253
  time-dependent variables, 249–251
  using SAS, 593–598
  using SPSS, 617–620
  using Stata, 550–554

F
Failure, 4
  rate, conditional, 12
Flemington–Harrington test, 75
Frailty
  component, 327
  effect, 332
  models, 326–340
    using R, 657–659
    using Stata, 561–564

G
Gamma distribution, 328
Gamma frailty, 333
Gap time model, 379–382, 385–388, 393, 399, 405
Gastric carcinoma data, 285–286
Generalized gamma model, 316
General stratified Cox (SC) model, 208–209
GOF. *See* Goodness-of-fit
Gompertz model, 317
Goodness-of-fit (GOF)
  testing approach, 181–183
  tests, 166
Greenwood's formula, 78–82, 86

H
Hazard function, 9, 10
  cause-specific, 434
  probability density function and, 294–295
Hazard ratio, 36–37, 49
  confidence interval in Cox PH model with interaction, 117–119, 143
  formula for Cox PH model, 114–117, 143, 146
  formula for extended Cox model, 251–253
Heaviside function, 257

I
Increasing Weibull model, 14
Independence assumption, 437–443
Independent censoring, 37–42, 49
  in competing risks, 437–443
Information matrix, 378
Instantaneous potential, 11
Intention-to-treat (ITT) principle, 517, 521
Interactions, 31
  confidence interval in Cox PH model with interaction, 117–119, 143, 146
Interval-censored data, 8, 44, 318–326
Inverse–Gaussian distribution, 328

K
Kaplan-Meier (KM) curves, 56
  example of, 61–65
  general features of, 66–67
  log-log survival curves, 171
KM curves. *See* Kaplan-Meier curves

L
Left-censored data, 7–8, 132–133, 318, 321
Left truncation, 132–133, 136–140, 144, 147

Leukemia remission-time data, 18–20, 30
  Cox proportional hazards (PH) model application, 100–108
  exponential survival, 295–297
  increasing Weibull for, 14
  Kaplan–Meier survival curves for, 61–65, 89–90
    log-log KM curves for, 171–175
  recurrent event data for, 367
  stratified Cox (SC) model application to, 204–216
Likelihood function
  for Cox PH model, 127–131, 145, 244
  for extended Cox model, 269–274, 281
  for parametric models, 318–321, 349–350
  for stratified Cox (SC) model, 223–225, 230
Likelihood ratio (LR) statistic, 103
LM approach. *See* Lunn–McNeil approach
Logit link function, 324
Log-logistic regression, 309–314
  accelerated failure-time form, 352
Log-log
  plots, 167–175
  survival curves, 167–175
Lognormal survival models, 14, 316
Log-rank test, 56
  alternatives to, 73–78
  for several groups, 71–73
  for two groups, 67–71
Loss to follow-up, 512, 516, 520
LR statistic. *See* Likelihood ratio statistic
Lunn–McNeil (LM) approach, 433, 455–461
  alternative, 461–464

M

Macular degeneration data set, 391–395
  marginal probability, 446
  results for, 393
Maximum likelihood (ML) estimation of Cox PH model, 112–114
Median follow-up time, 505, 507, 516, 519
Multiplicative model, 317

N

No-interaction assumption in stratified Cox model, 210–216
Non-informative censoring, 37, 41–42, 49, 437

O

Observed *vs.* expected plots, 175–180
Open cohort, 135, 140, 144

P

Parametric approach using shared frailty, 389–391
Parametric survival models
  defined, 292
  examples
    exponential model, 295–297, 300–304
    log-logistic model, 309–314
    Weibull model, 304–309
  likelihood function, 318–321
  other models, 316–318
  SAS use, 598–603
  Stata use, 554–561
Pepe-Mori test, 471
Peto test, 73
PH assumption
  assessment using
    goodness of fit test with Schonfield residuals, 181–183
    Kaplan-Meier log-log survival curves, 611–612
    observed *vs.* expected plots, 175–180
    SAS, 585–588
    SPSS, 615–617
    Stata, 535–538, 545–547
    time-dependent covariates, 183–187, 253–259
  evaluating, 161–200
  meaning of, 123–127
PH model, Cox. *See* Cox proportional hazards (PH) model
Precision, 106
Probability, 12
  density function, 294–295
PROC LIFEREG (SAS), 598–603
PROC LIFETEST (SAS), 572–576
PROC PHREG, 576–580
Product limit formula, 56
Proportional odds (PO) assumption, 324

R
R software 620–663
  assessing PH assumption
    using graphical approaches, 631–633
    using statistical tests, 640–641
  estimating survival functions, 626–631
  modeling recurrent events, 660–663
  obtaining Cox adjusted survival curves with, 641–645
  running Cox proportional hazards (PH) model, 634–637
  running extended Cox model, 646–650
  running frailty models, 657–659
  running parametric models, 651–656
  running stratified Cox (SC) model, 638–640
Random censoring, 37–41, 49
Recurrent event survival analysis, 363–423
  counting process approach, 368–376
  definition of recurrent events, 4, 364
  examples of recurrent event data, 366–368
  other approaches for analysis, 379–385
  parametric approach using shared frailty, 389–391
  SAS modeling, 603–604
  Stata modeling, 564–570
  survival curves with, 395–398
Right-censored data, 8, 321
Risk set, 26
Robust estimation, 376–378
Robust standard error, 379
Robust variance, 377

S
Sample size inflation factor, 513, 517, 521
SAS, 570–607
  assessing PH assumption, with statistical tests, 585–588
  demonstrating PROC LIFETEST, 572–576
  modeling recurrent events, 603–607
  obtaining Cox adjusted survival curves, 588–592
  running Cox proportional hazards (PH) model, with PROC PHREG, 576–580
  running extended Cox model, 593–598
  running parametric models, with PROC LIFEREG, 598–603
  running stratified Cox (SC) model, 581–584
Schoenfeld residuals, 181–182, 586–587
Score residuals, 378
Semi-parametric model, 109–110, 293
Sensitivity analys.is (with competing risks), 440–443
Shape parameter, 304
Shared frailty, 338–340, 390
  recurrent events analysis using, 389–391
Shared frailty model, 338
SPSS, 607–620
  assessing PH assumption
    with statistical tests, 615–617
    using Kaplan-Meier log-log survival curves, 611–612
  estimating survival functions, 626–628
  running Cox proportional hazards (PH) model, 613
  running extended Cox model, 617–620
  running stratified Cox (SC) model, 614–615
Stanford Heart Transplant Study
  extended Cox model application to, 265–269
  transplants *vs.* nontransplants, 265–269
Stata, 527–570
  assessing PH assumption
    using graphical approaches, 535–538
    using statistical tests, 545–547
  estimating survival functions, 531–535
  modeling recurrent events, 564–570
  obtaining Cox adjusted survival curves with, 547–550
  running Cox proportional hazards (PH) model, 538–543
  running extended Cox model, 550–554
  running frailty models, 561–564
  running parametric models, 554–561
  running stratified Cox (SC) model, 543–545
Step functions, 10
Strata variable, 379

Stratification variables, several, 216–221
Stratified Cox (SC) model, 204–208, 395–398
  for analyzing recurrent event data, 377–385
  conditional approaches, 379–383
  general, 208–209
  graphical view of, 221–222
  marginal approach, 379–383
  using SAS, 581–583
  using SPSS, 614–615
  using Stata, 543–545
Stratified CP model, 379–383, 385–389, 393–395, 405–410, 415
Sub-distribution hazard function, 451–452, 478
Sub-distribution (Fine and Gray) hazard model, 451
Survival curves
  adjusted, 120
    using Cox PH model, 117–123
  Cox adjusted (*see* Cox adjusted survival curves)
  with recurrent events, 395–398
Survival functions
  conditional, 327
  estimation
    R, 626–631
    SAS, 572–576
    SPSS, 609–611
    Stata, 531–535
  probability density function and, 294–295
  unconditional, 327
Survival time, 4
  variable, 15
Survivor function, 9

T
Tarone–Ware test statistic, 74
Time-dependent covariates, assessing PH assumption using, 183–187
Time-dependent variables, 164
  definition and examples of, 246–249
  extended Cox model for, 249–251
Time-independent variables, PH assumption and, 254–259
Time-on-study, 131–142, 148

U
Unconditional survival function, 327
Unshared frailty, 338

V
Veterans Administration Lung Cancer Data
  Kaplan-Meier survival curves for, 72–73
  model with no frailty, 328
  proportional hazards assumption evaluation for, 115–118
  with several stratification variables, 216–219
  stratified Cox (SC) model application, 231–234

W
Wald statistic, 103
Weibull model, 304–309
Weibull regression
  accelerated failure-time form, 354, 357
  log relative-hazard form, 355, 357
Wilcoxon test, 74