

# Statistics for Biology and Health

*Series Editors*

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

David G. Kleinbaum  
Mitchel Klein

# Survival Analysis

A Self-Learning Text

Second Edition

 Springer

David G. Kleinbaum  
Department of Epidemiology  
Rollins School of Public Health at  
Emory University  
1518 Clifton Road NE  
Atlanta GA 30306  
Email: dkleinb@sph.emory.edu

Mitchel Klein  
Department of Epidemiology  
Rollins School of Public Health at  
Emory University  
1518 Clifton Road NE  
Atlanta GA 30306  
Email: mklein@sph.emory.edu

*Series Editors*

M. Gail  
National Cancer Institute  
Rockville, MD 20892  
USA

K. Krickeberg  
Le Châtelet  
F-63270 Manglieu  
France

J. 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

Wing 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.

Library of Congress Control Number: 2005925181

ISBN-10: 0-387-23918-9  
ISBN-13: 978-0387-23918-7

Printed on acid-free paper.

© 2005, 1996 Springer Science+Business Media, Inc.

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, Inc., 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 in the United States of America. (TechBooks/HAM)

9 8 7 6 5 4 3 2 1

springeronline.com

Survival Analysis  
A Self-Learning Text

**Springer**

*New York*

*Berlin*

*Heidelberg*

*Hong Kong*

*London*

*Milan*

*Paris*

*Tokyo*

David G. Kleinbaum    Mitchel Klein

# Survival Analysis

A Self-Learning Text

Second Edition

 Springer

*To Rosa Parks*

*Nelson Mandela*

*Dean Smith*

*Sandy Koufax*

*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 second edition of this text on survival analysis, originally published in 1996. As in the first edition, 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 second edition has expanded the first edition by adding three new chapters and a revised computer appendix. The three new chapters are:

Chapter 7. Parametric Survival Models

Chapter 8. Recurrent Event Survival Analysis

Chapter 9. Competing Risks Survival Analysis

Chapter 7 extends survival analysis methods to 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 such parametric models are acceleration failure time models, which provide an alternative measure to the hazard ratio called the “acceleration factor”. The general form of the likelihood for a parametric model that allows for left, right, or interval censored data is also described. The chapter concludes with an introduction to frailty models.

Chapter 8 considers survival events that may occur more than once over the follow-up time for a given subject. Such events are called “recurrent events”. Analysis of such data can be carried out using a Cox PH model with the data layout augmented so that each subject has a line of data for each recurrent event. A variation of this approach uses a stratified Cox PH model, which stratifies on the order in which recurrent events occur. The use of “robust variance estimates” are recommended to adjust the variances of estimated model coefficients for correlation among recurrent events on the same subject.

Chapter 9 considers survival data in which each subject can experience only one of several different types of events (“competing risks”) over follow-up. Modeling such data can be carried out using a Cox model, a parametric survival model or a model which uses *cumulative incidence* (rather than survival).

The Computer Appendix in the first edition of this text has now been revised and extended to provide step-by-step instructions for using the computer packages STATA (version 7.0), SAS (version 8.2), and SPSS (version 11.5) to carry out the survival analyses presented in the main text. These computer packages are described in separate self-contained sections of the Computer Appendix, with the analysis of the same datasets illustrated in each section. The SPIDA package used in the first edition is no longer active and has therefore been omitted from the appendix and computer output in the main text.

In addition to the above new material, the original six chapters have been modified slightly to correct for errata in the first edition, to clarify certain issues, and to add theoretical background, particularly regarding the formulation of the (partial) likelihood functions for the Cox PH (Chapter 3) and extended Cox (Chapter 6) models.

The authors’ website for this textbook has the following web-link: <http://www.sph.emory.edu/~dkleinb/surv2.htm>

This website 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 nine 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 use 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, Muller K, and Nizam A, **Applied Regression Analysis and Other Multivariable Methods, Third Edition**, Duxbury Press, Pacific Grove, 1998, 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, Second Edition**, Springer-Verlag Publishers, New York, Chapters 4–7, 11.

Kleinbaum D, **ActivEpi-A CD Rom Electronic Textbook on Fundamentals of Epidemiology**, Springer-Verlag 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 wish to thank Allison Curry for carefully reviewing and editing the previous edition's first six chapters for clarity of content and errata. We also thank several faculty and MPH and PhD students in the Department of Epidemiology who have provided feedback on the first edition of this text as well as new chapters in various stages of development. This includes Michael Goodman, Mathew Strickland, and Sarah Tinker. 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 second edition.

# Contents

---

**Preface**    vii

**Acknowledgments**    xi

**Chapter 1**                      **Introduction to Survival Analysis**    1

Introduction    2  
Abbreviated Outline    2  
Objectives    3  
Presentation    4  
Detailed Outline    34  
Practice Exercises    38  
Test    40  
Answers to Practice Exercises    42

**Chapter 2**                      **Kaplan–Meier Survival Curves and the  
Log–Rank Test**    45

Introduction    46  
Abbreviated Outline    46  
Objectives    47  
Presentation    48  
Detailed Outline    70  
Practice Exercises    73  
Test    77  
Answers to Practice Exercises    79  
Appendix: Matrix Formula for the Log–Rank Statistic  
for Several Groups    82

**Chapter 3**                      **The Cox Proportional Hazards Model and Its  
Characteristics**    83

Introduction    84  
Abbreviated Outline    84  
Objectives    85  
Presentation    86  
Detailed Outline    117  
Practice Exercises    119  
Test    123  
Answers to Practice Exercises    127

<b>Chapter 4</b>	<b>Evaluating the Proportional Hazards Assumption</b>	<b>131</b>
	Introduction	132
	Abbreviated Outline	132
	Objectives	133
	Presentation	134
	Detailed Outline	158
	Practice Exercises	161
	Test	164
	Answers to Practice Exercises	167
<b>Chapter 5</b>	<b>The Stratified Cox Procedure</b>	<b>173</b>
	Introduction	174
	Abbreviated Outline	174
	Objectives	175
	Presentation	176
	Detailed Outline	198
	Practice Exercises	201
	Test	204
	Answers to Practice Exercises	207
<b>Chapter 6</b>	<b>Extension of the Cox Proportional Hazards Model for Time-Dependent Variables</b>	<b>211</b>
	Introduction	212
	Abbreviated Outline	212
	Objectives	213
	Presentation	214
	Detailed Outline	246
	Practice Exercises	249
	Test	253
	Answers to Practice Exercises	255
<b>Chapter 7</b>	<b>Parametric Survival Models</b>	<b>257</b>
	Introduction	258
	Abbreviated Outline	258
	Objectives	259
	Presentation	260
	Detailed Outline	313
	Practice Exercises	319
	Test	324
	Answers to Practice Exercises	327

**Chapter 8 Recurrent Event Survival Analysis 331**

Introduction	332
Abbreviated Outline	332
Objectives	333
Presentation	334
Detailed Outline	371
Practice Exercises	377
Test	381
Answers to Practice Exercises	389

**Chapter 9 Competing Risks Survival Analysis 391**

Introduction	392
Abbreviated Outline	394
Objectives	395
Presentation	396
Detailed Outline	440
Practice Exercises	447
Test	452
Answers to Practice Exercises	458

**Computer Appendix: Survival Analysis on the Computer 463**

A. STATA	465
B. SAS	508
C. SPSS	542

**Test Answers 557****References 581****Index 585**

# 1

# Introduction to Survival Analysis

## **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 8–14)**
- IV. Goals of survival analysis (page 15)**
- V. Basic data layout for computer (pages 15–19)**
- VI. Basic data layout for understanding analysis (pages 19–24)**
- VII. Descriptive measures of survival experience (pages 24–26)**
- VIII. Example: Extended remission data (pages 26–29)**
- IX. Multivariable example (pages 29–31)**
- X. Math models in survival analysis (pages 31–33)**

## Objectives

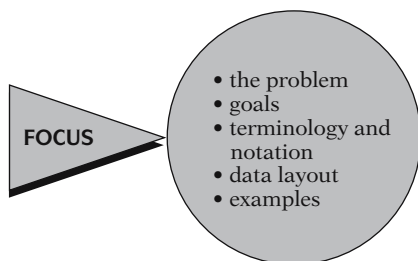
Upon completing the chapter, the learner should be able to:

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?

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



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*.

**Event:** death  
disease  
relapse  
recovery

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.

Assume 1 event



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.

Time  $\equiv$  survival time

Event  $\equiv$  failure

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 events or a **competing risk** problem, which are discussed in Chapters 8 and 9, respectively.

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

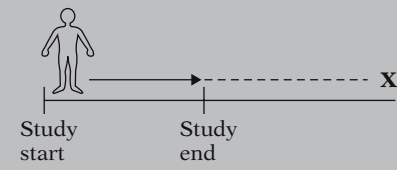
**Censoring:** don't know survival time exactly

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**.

## 6 1. Introduction to Survival Analysis

### 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.

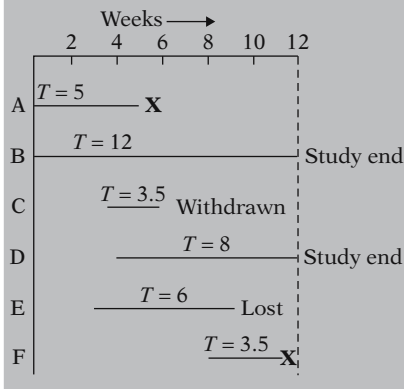
Why censor?

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



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 third 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.

**X** Event occurs

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.

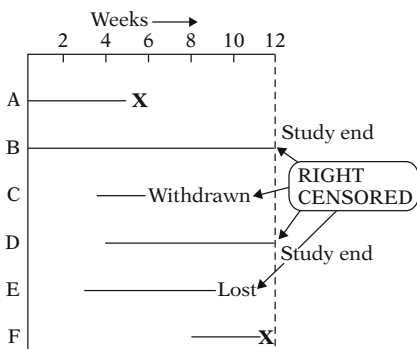
**SUMMARY**

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

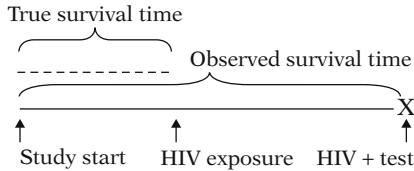
In **summary**, of the six persons observed, two get the event (persons A and F) and four are censored (B, C, D, and E).

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.



Notice in our example that for each of the four persons censored, we know that the person's exact 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 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 exactly the 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 tests positive.

### III. Terminology and Notation

$T =$  survival time ( $T \geq 0$ )  
 ↙ random variable

$t =$  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.

$\delta = (0, 1)$  random variable

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

- study ends
- lost to follow-up
- withdraws

Finally, we let the Greek letter delta ( $\delta$ ) denote a (0,1) random variable indicating either failure or censorship. That is,  $\delta = 1$  for failure if the event occurs during the study period, or  $\delta = 0$  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,  $\delta = 0$  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  $h(t)$ .

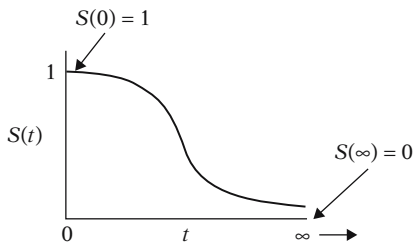
$$S(t) = P(T > t)$$

$t$	$S(t)$
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  $S(t)$ :

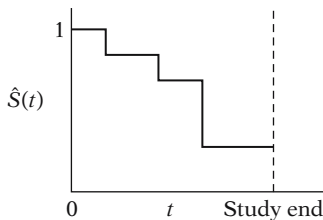


Theoretically, as  $t$  ranges from 0 up to infinity, the survivor function can be graphed as a smooth curve. As illustrated by the graph, where  $t$  identifies the X-axis, all survivor functions have the following characteristics:

- 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:

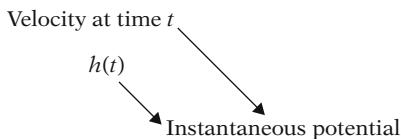
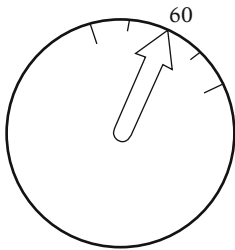
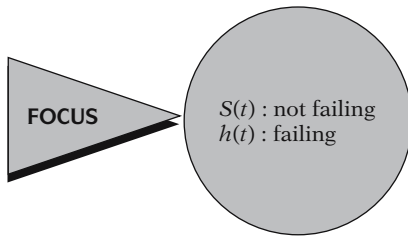


In practice, when using actual data, we usually obtain graphs that are **step functions**, as illustrated here, rather than smooth curves. Moreover, because the study period is never infinite in length and there may be competing risks for failure, it is possible that not everyone studied gets the event. 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.

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

The hazard function, denoted by  $h(t)$ , is given by the formula:  $h(t)$  equals the limit, as  $\Delta t$  approaches zero, of a probability statement about survival, divided by  $\Delta t$ , where  $\Delta t$  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.

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

Given

Conditional probabilities:  $P(A|B)$

$$P(t \leq T < t + \Delta t \mid T \geq t)$$

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

Hazard function  $\equiv$  conditional failure **rate**

$$\lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t \mid T \geq t)}{\Delta t}$$

Probability per unit time

Rate: 0 to  $\infty$

$$P = P(t \leq T < t + \Delta t \mid T \geq t)$$

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

In mathematical terms, the given part of the formula for the hazard function is found in the probability statement—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**.

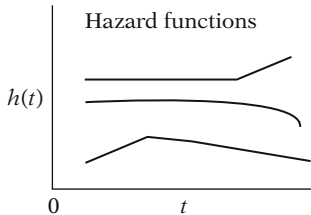
We now explain why the hazard is a **rate** rather than a probability. Note that in the hazard function formula, the expression to the right of the limit sign gives the ratio of two quantities. The numerator is the conditional probability we just discussed. The denominator is  $\Delta t$ , which denotes 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.

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  $\Delta t$  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.**



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

Gives  
instantaneous  
potential



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

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:

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

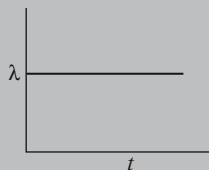
These two features follow from the ratio expression in the formula for  $h(t)$ , because both the probability in the numerator and the  $\Delta t$  in the denominator are nonnegative, and since  $\Delta t$  can range between 0 and  $\infty$ .

### EXAMPLE

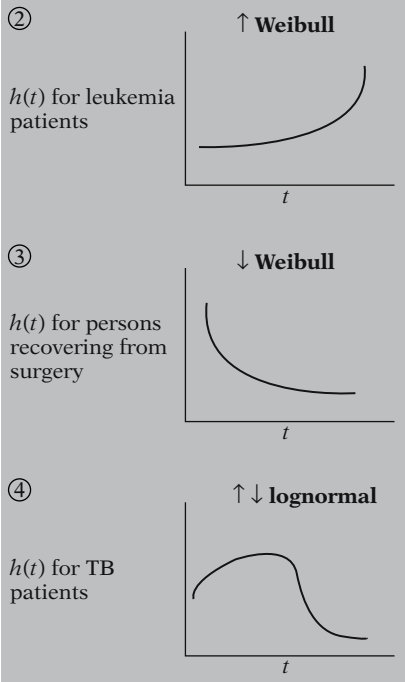
①

Constant hazard  
(exponential model)

$h(t)$  for healthy  
persons



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,  $\lambda$ . 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**

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

General formulae:

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

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



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  $\lambda$ —then it can be shown that the corresponding survival function is given by the following formula:  $S(t)$  equals  $e$  to the power minus  $\lambda$  times  $t$ .

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  $T$

$\delta$  = (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  $\delta$  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.

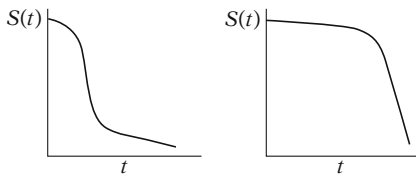
#### IV. Goals of Survival Analysis

We now state the basic goals of survival analysis.

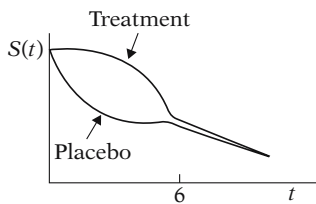
**Goal 1:** To estimate and interpret survivor and/or hazard functions from survival data.

**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 modules.

#### V. Basic Data Layout for Computer

Two types of data layouts:

- for computer use
- for understanding

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.

For computer:

Indiv. #	$t$	$\delta$	$X_1$	$X_2$	$\dots$	$X_p$
1	$t_1$	$\delta_1$	$X_{11}$	$X_{12}$	$\dots$	$X_{1p}$
2	$t_2$	$\delta_2$	$X_{21}$	$X_{22}$	$\dots$	$X_{2p}$
$\vdots$						
5	$t_5 = 3$	got event				
$\vdots$						
8	$t_8 = 3$	censored				
$\vdots$						
$n$	$t_n$	$\delta_n$	$X_{n1}$	$X_{n2}$	$\dots$	$X_{np}$

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.  $\delta$ ) the dichotomous variable that indicates censorship status.

Indiv. #	$t$	Failure status	Explanatory variables			
		$\delta$	$X_1$	$X_2$	$\dots$	$X_p$
1	$t_1$	$\delta_1$	$X_{11}$	$X_{12}$	$\dots$	$X_{1p}$
2	$t_2$	$\delta_2$	$X_{21}$	$X_{22}$	$\dots$	$X_{2p}$
$\vdots$						
5	$t_5 = 3$	$\delta_5 = 1$				
$\vdots$						
8	$t_8 = 3$	$\delta_8 = 0$				
$\vdots$						
$n$	$t_n$	$\delta_n$	$X_{n1}$	$X_{n2}$	$\dots$	$X_{np}$

$X_i = \text{Age, } E, \text{ or Age} \times \text{Race}$

Thus,  $\delta_1$  is 1 if person 1 gets the event or is 0 if person 1 is censored;  $\delta_2$  is 1 or 0 similarly, and so on, up through  $\delta_n$ . In the example just considered, person 5, who failed at 3 weeks, has a  $\delta$  of 1; that is,  $\delta_5$  equals 1. In contrast, person 8, who was censored at 3 weeks, has a  $\delta$  of 0; that is,  $\delta_8$  equals 0.

Note that if all of the  $\delta_j$  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,  $X_i$ , 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.

		Columns						
		#	$t$	$\delta$	$X_1$	$X_2$	$\dots$	$X_p$
Rows	1	$t_1$	$\delta_1$	$X_{11}$	$X_{12}$	$\dots$	$X_{1p}$	
	2	$t_2$	$\delta_2$	$X_{21}$	$X_{22}$	$\dots$	$X_{2p}$	
	$\vdots$							
	$\vdots$							
	$\vdots$							
	$\vdots$							
	$j$	$t_j$	$\delta_j$	$X_{j1}$	$X_{j2}$	$\dots$	$X_{jp}$	
	$\vdots$							
	$n$	$t_n$	$\delta_n$	$X_{n1}$	$X_{n2}$	$\dots$	$X_{np}$	

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  $j$ , the observed information is given by the values  $t_j, \delta_j, X_{j1}, X_{j2}$ , etc., up to  $X_{jp}$ . 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,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

+ denotes censored

- In remission at study end
- Lost to follow-up
- Withdraws

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,	1, 1, 2, 2, 3,
13, 16, 22, 23,	4, 4, 5, 5,
6+, 9+, 10+, 11+,	8, 8, 8, 8,
17+, 19+, 20+,	11, 11, 12, 12,
25+, 32+, 32+,	15, 17, 22, 23
34+, 35+	

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

	Indiv. #	$t$ (weeks)	$\delta$ (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
	9	23	1	1
GROUP	10	6	0	1
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 six 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 six weeks. The remaining persons in group one 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  $\delta = 1$ , 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  $\delta = 0$ . The  $X$  value is again 1 because person #14 is also in group 1.

EXAMPLE (continued)				
	Indiv. #	$t$ (weeks)	$\delta$ (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  $\delta = 1$ ; the  $X$  value is 0 because person #32 is in group 2.

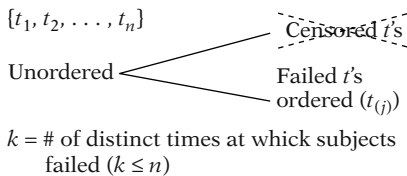
## VI. Basic Data Layout for Understanding Analysis

For analysis:

Ordered failure times ( $t_{(j)}$ )	# of failures ( $m_j$ )	# censored in [ $t_{(j)}, t_{(j+1)}$ ) ( $q_j$ )	Risk set $R(t_{(j)})$
$t_{(0)} = 0$	$m_0 = 0$	$q_0$	$R(t_{(0)})$
$t_{(1)}$	$m_1$	$q_1$	$R(t_{(1)})$
$t_{(2)}$	$m_2$	$q_2$	$R(t_{(2)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(k)}$	$m_k$	$q_k$	$R(t_{(k)})$

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.

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.



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,  $k = 7$ )

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	3	1	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	1	1	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	1	2	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ wks
Totals	9	12	

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

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive $\geq 0$ wks
$t_{(1)} = 1$	2	0	21 persons survive $\geq 1$ wk
$t_{(2)} = 2$	2	0	19 persons survive $\geq 2$ wks
$t_{(3)} = 3$	1	0	17 persons survive $\geq 3$ wks
$t_{(4)} = 4$	2	0	16 persons survive $\geq 4$ wks
$t_{(5)} = 5$	2	0	14 persons survive $\geq 5$ wks
$t_{(6)} = 8$	4	0	12 persons survive $\geq 8$ wks
$t_{(7)} = 11$	2	0	8 persons survive $\geq 11$ wks
$t_{(8)} = 12$	2	0	6 persons survive $\geq 12$ wks
$t_{(9)} = 15$	1	0	4 persons survive $\geq 15$ wks
$t_{(10)} = 17$	1	0	3 persons survive $\geq 17$ wks
$t_{(11)} = 22$	1	0	2 persons survive $\geq 22$ wks
$t_{(12)} = 23$	1	0	1 person survive $\geq 23$ wks
Totals	21	0	

For example, using the remission data for group 1, we find that nine of the 21 persons failed, including three persons at 6 weeks and one 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  $m_j$ , of those persons who failed at each distinct failure time. When there are no ties at a certain failure time, then  $m_j = 1$ . 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  $m_j$ '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**

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	3	1	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	1	1	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	1	2	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ wks
Totals	9	12	

**Remission Data: Group 1**

#	$t$ (weeks)	$\delta$	$X$ (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  $q_j$ , of those persons censored in the time interval starting with failure time  $t_{(j)}$  up to the next failure time denoted  $t_{(j+1)}$ . 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  $q_j$ '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 here person #12, who was censored at week 10, 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**

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	3	1	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	1	1	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	1	2	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ 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:**  $R(t_{(j)})$  is the set of individuals for whom  $T \geq t_{(j)}$ .

**Remission Data: Group 1**

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	0	0	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	3	1	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	1	1	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	1	2	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ 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  $R(t_{(j)})$  is the collection of individuals who have survived at least to time  $t_{(j)}$ ; that is, each person in  $R(t_{(j)})$  has a survival time that is  $t_{(j)}$  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)**

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
$t_{(0)} = 0$	<del>0</del>	<del>0</del>	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	<del>3</del>	<del>1</del>	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	1	1	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	1	2	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ wks
Totals	9	12	

$t_{(0)} = 0$	<del>0</del>	<del>0</del>	21 persons survive $\geq 0$ wks
$t_{(1)} = 6$	<del>3</del>	<del>1</del>	21 persons survive $\geq 6$ wks
$t_{(2)} = 7$	<del>1</del>	<del>1</del>	17 persons survive $\geq 7$ wks
$t_{(3)} = 10$	<del>1</del>	<del>2</del>	15 persons survive $\geq 10$ wks
$t_{(4)} = 13$	1	0	12 persons survive $\geq 13$ wks
$t_{(5)} = 16$	1	3	11 persons survive $\geq 16$ wks
$t_{(6)} = 22$	1	0	7 persons survive $\geq 22$ wks
$t_{(7)} = 23$	1	5	6 persons survive $\geq 23$ wks
Totals	9	12	

Now let's look at the risk set at 7 weeks. This set consists of seventeen 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 nine persons from 21, leaves twelve persons in group 1 still at risk for getting the event at 13 weeks. Thus, the risk set consists of these twelve persons.

**How we work with censored data:** Use all informaton 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 **non-informative**—censored subjects are not at increased risk for failure. See Chapter 9 on competing risks for further details.)

EXAMPLE			
$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
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 20th 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.

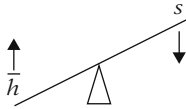
Having introduced the basic terminology and data layouts to this point, we now consider some data analysis issues and some additional applications.

## 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
$T_1$ (ignoring + 's) = 17.1	$T_2 = 8.6$
$\bar{h}_1 = \frac{9}{359} = .025$	$\bar{h}_2 = \frac{21}{182} = .115$
Average hazard rate $(\bar{h}) = \frac{\# \text{ failures}}{\sum_{i=1}^n t_i}$	

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  $T$  "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  $h$  "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.

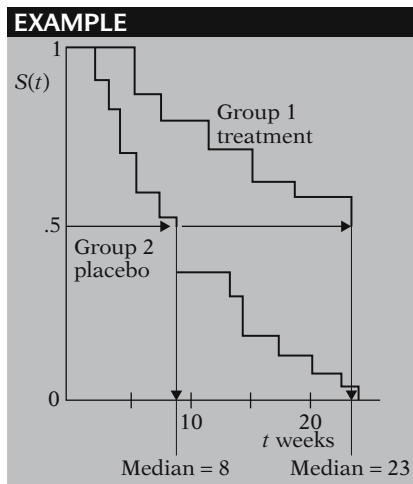
In our example, the average hazard for the treatment group is smaller than the average hazard for the placebo group.

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

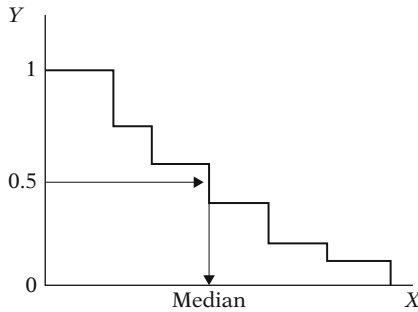
Thus, using average hazard rates, we again see that the treatment group appears to be doing better overall than the placebo group; that is, the treatment group is less prone to fail than the placebo group.

Descriptive measures ( $\bar{T}$  and  $\bar{h}$ ) 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.



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 Chapter 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.

---

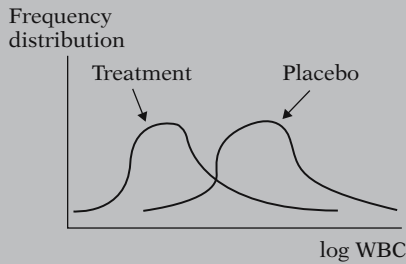
### VIII. Example: Extended Remission Data

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



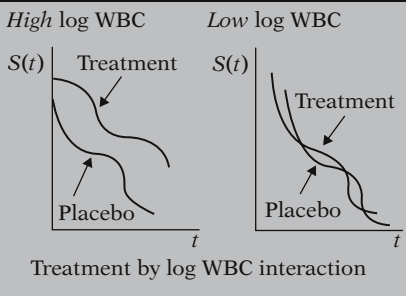
Need to adjust for imbalance in the distribution of log WBC

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.**

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**



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 problem:

Compare two groups after adjusting for confounding and interaction.

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.

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)	$\delta$	$X_1$ (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

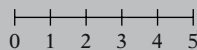
Event = death

Explanatory variables:

- exposure variable
- confounders
- interaction variables

*Exposure:*

Social Network Index (SNI)



Absence of social network                      Excellent social network

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.,  $\text{SNI} \nearrow \Rightarrow S(t) \nearrow$

*Explanatory variables:*

<b>SNI</b>	}	Exposure variable
<b>AGE</b>		
<b>SBP</b>		Potential confounders/ interaction variables
<b>CHR</b>		
<b>QUET</b>		
<b>SOCL</b>		

Note :  $\text{QUET} = \frac{\text{weight}}{(\text{height})^2} \times 100$

**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 ( $\text{QUET} = \text{weight over height squared times } 100$ ), 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

#	$t$	$\delta$	SNI	AGE	SBP	CHR	QUET	SOCL
1	$t_1$	$\delta_1$	SNI <sub>1</sub>	AGE <sub>1</sub>	SBP <sub>1</sub>	CHR <sub>1</sub>	QUET <sub>1</sub>	SOCL <sub>1</sub>
2	$t_2$	$\delta_2$	SNI <sub>2</sub>	AGE <sub>2</sub>	SBP <sub>2</sub>	CHR <sub>2</sub>	QUET <sub>2</sub>	SOCL <sub>2</sub>
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
170	$t_{170}$	$\delta_{170}$	SNI <sub>170</sub>	AGE <sub>170</sub>	SBP <sub>170</sub>	CHR <sub>170</sub>	QUET <sub>170</sub>	SOCL <sub>170</sub>

### X. Math Models in Survival Analysis

General framework



Controlling for  $C_1, C_2, \dots, C_p$ .

**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  $\beta$

Logistic regression:  
odds ratio  $e^\beta$

Survival analysis:  
hazard ratio  $e^\beta$

In linear regression modeling, the measure of effect is usually some regression coefficient  $\beta$ .

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  $\beta$ .

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):

$HR = 1 \Rightarrow$  no relationship

$HR = 10 \Rightarrow$  exposed hazard  
10 times unexposed

$HR = 1/10 \Rightarrow$  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.

---

## 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 Chapter 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 exact survival time.
- B. Reasons: study ends without subject getting event; lost to follow-up; withdraws.
- C. Examples of survival data for different persons; summary table.

### III. Terminology and notation (pages 8–14)

- A. Notation:  $T$  = survival time random variable:  
 $t$  = specific value for  $T$   
 $\delta$  = (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:

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

- 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”;
  - $h(t) \geq 0$ ; 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:

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

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

**IV. Goals of survival analysis** (page 15)

- 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** (15–19)

## A. General layout:

#	$t$	$\delta$	$X_1$	$X_2 \cdots X_p$
1	$t_1$	$\delta_1$	$X_{11}$	$X_{12} \cdots X_{1p}$
2	$t_2$	$\delta_2$	$X_{21}$	$X_{22} \cdots X_{2p}$
.	.	.		.
.	.	.		.
.	.	.		.
$j$	$t_j$	$\delta_j$	$X_{j1}$	$X_{j2} \cdots X_{jp}$
.	.	.		.
.	.	.		.
.	.	.		.
$n$	$t_n$	$\delta_n$	$X_{n1}$	$X_{n2} \cdots X_{np}$

## B. Example: Remission time data



## VI. Basic data layout for understanding analysis (pages 19–24)

### A. General layout:

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

*Note:*  $k$  = # of distinct times at which subjects failed;  $n$  = # of subjects ( $k \leq n$ );  $R(t_{(j)})$ , the risk set, is the set of individuals whose survival times are at least  $t_{(j)}$  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 24–26)

### A. Average survival time (ignoring censorship status):

$$\bar{T} = \frac{\sum_{j=1}^n t_j}{n}$$

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

### B. Average hazard rate:

$$\bar{h} = \frac{\# \text{ failures}}{\sum_{j=1}^n t_j}$$

### C. Descriptive measures $\bar{T}$ and $\bar{h}$ 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 26–29)

- 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 29–31)

- 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.

#	$t$	$\delta$	SNI	AGE	SBP	CHR	QUET	SOCL
1	$t_1$	$\delta_1$	$SNI_1$	$AGE_1$	$SBP_1$	$CHR_1$	$QUET_1$	$SOCL_1$
2	$t_2$	$\delta_2$	$SNI_2$	$AGE_2$	$SBP_2$	$CHR_2$	$QUET_2$	$SOCL_2$
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.
170	$t_{170}$	$\delta_{170}$	$SNI_{170}$	$AGE_{170}$	$SBP_{170}$	$CHR_{170}$	$QUET_{170}$	$SOCL_{170}$

**X. Math models in survival analysis** (pages 31–33)

- 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.

**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,

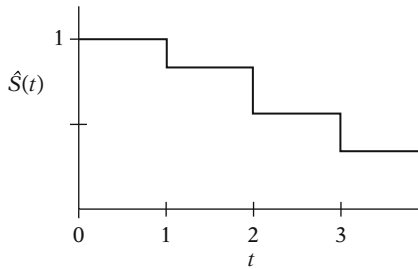
1, 1, 1+, 1+, 1+, 2, 2, 2, 2+, 2+, 3, 3, 3+, 4+, 5+

where + denotes censored data, complete the following table:

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
0	0	0	15 persons survive $\geq 0$ weeks
1			
2			
3			

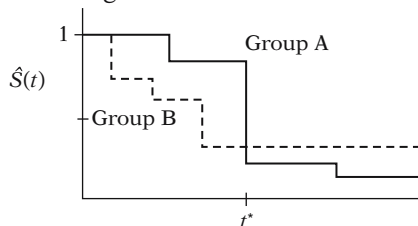
Also, compute the average survival time ( $\bar{T}$ ) and the average hazard rate ( $\bar{h}$ ) using the raw data (ignoring + signs for  $\bar{T}$ ).

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  $\infty$ .
- 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 six weeks is the set of individuals whose survival times are less than or equal to six weeks.
- T F 12. If the risk set at six weeks consists of 22 persons, and four persons fail and three persons are censored by the 7th week, then the risk set at seven 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 ( $\text{CHR} = 0$ ), and group 2 has a positive history of chronic disease ( $\text{CHR} = 1$ ):
 

Group 1 ( $\text{CHR} = 0$ ): 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 ( $\text{CHR} = 1$ ): 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:

	$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
Group 1:	0.0	0	0	25 persons survived $\geq 0$ years
	1.8	1	0	25 persons survived $\geq 1.8$ 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 ( $\bar{T}$ ) and the average hazard rate ( $\bar{h}$ ) for each group are given as follows:

	$\bar{T}$	$\bar{h}$
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.

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
0	0	0	15 persons survive $\geq 0$ weeks
1	2	3	15 persons survive $\geq 1$ weeks
2	3	2	10 persons survive $\geq 2$ weeks
3	2	3	5 persons survive $\geq 3$ weeks

$$\bar{T} = \frac{33}{15} = 2.2; \bar{h} = \frac{7}{33} = 0.2121$$

17. Median = 3 weeks

18. Group A

19. Group B

20. Group A



# 2

**Kaplan–**

---

**Meier**

---

**Survival**

---

**Curves and**

---

**the Log–**

---

**Rank Test**

---

## 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.

## 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.

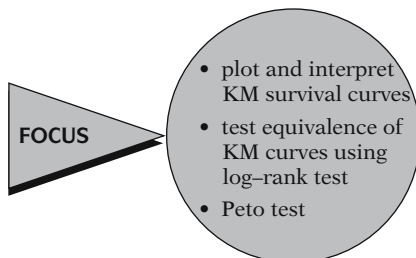
- I. **Review (pages 48–50)**
- II. **An example of Kaplan–Meier curves (pages 51–55)**
- III. **General features of KM curves (pages 56–57)**
- IV. **The log–rank test for two groups (pages 57–61)**
- V. **The log–rank test for several groups (pages 61–63)**
- VI. **Alternatives to the log–rank test (pages 63–68)**
- VII. **Summary (page 68)**

## 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.

## 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.

### I. Review

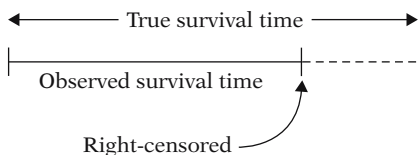


**Event:** death  
disease  
relapse

Time = survival time

Event = failure

**Censoring:** Don't know survival time exactly



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.

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**.

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  $T$ .

$$\delta = (0, 1) \text{ random variable}$$

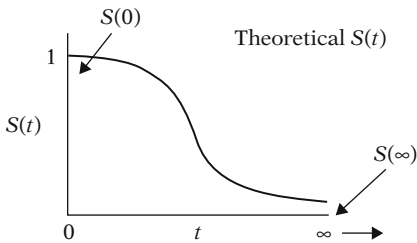
$$= \begin{cases} 1 & \text{if failure} \\ 0 & \text{if censored} \end{cases}$$

We let the Greek letter delta ( $\delta$ ) 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.

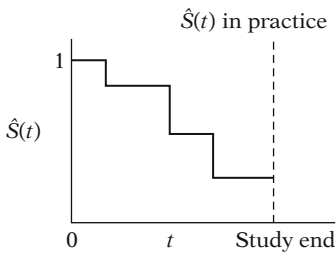
$$S(t) = \text{survivor function}$$

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

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.



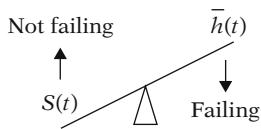
In practice, using data, we usually obtain estimated survivor curves that are **step functions**, as illustrated here, rather than smooth curves.

$$h(t) = \text{hazard function}$$

$$= \text{instantaneous potential}$$

$$\text{given survival up to time } t$$

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.

$$h(t) \text{ is a rate: } 0 \text{ to } \infty$$



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. #	$t$	$\delta$	$X_1$	$X_2 \cdots X_p$
1	$t_1$	$\delta_1$	$X_{11}$	$X_{12} \cdots X_{1p}$
2	$t_2$	$\delta_2$	$X_{21}$	$X_{22} \cdots X_{2p}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n$	$t_n$	$\delta_n$	$X_{n1}$	$X_{n2} \cdots X_{np}$

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  $\delta$ , 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, $t_{(j)}$	# of failures $m_j$	# censored in $[t_{(j)}, t_{(j+1)})$ , $q_j$	Risk set, $R(t_{(j)})$
$t_{(0)} = 0$	$m_0 = 0$	$q_0$	$R(t_{(0)})$
$t_{(1)}$	$m_1$	$q_1$	$R(t_{(1)})$
$t_{(2)}$	$m_2$	$q_2$	$R(t_{(2)})$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$t_{(k)}$	$m_k$	$q_k$	$R(t_{(k)})$

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  $q_j$ , of those persons censored in the time interval starting with failure time  $t_{(j)}$  up to but not including the next failure time, denoted by  $t_{(j+1)}$ . The last column gives the **risk set**, which denotes the collection of individuals who have survived at least to time  $t_{(j)}$ .

Table of ordered failures:

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

Survival probability:

Use **Kaplan–Meier (KM)** method.

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.

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:

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

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

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)

$t_{(j)}$	$n_j$	$m_j$	$q_j$
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)

$t_{(j)}$	$n_j$	$m_j$	$q_j$
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)

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\hat{S}(t_{(j)})$
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  $n_j$  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  $t_{(j)}$ , it is assumed that  $n_j$  includes those persons failing at time  $t_{(j)}$ . In other words,  $n_j$  counts those subjects at risk for failing instantaneously prior to time  $t_{(j)}$ .

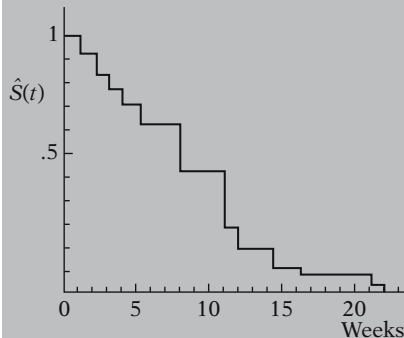
We now describe how to compute the KM curve for the table for group 2. The computations for group 2 are quite straightforward because there are no censored subjects for this group.

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)

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\hat{S}(t_{(j)})$
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	$0/21 = .00$

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

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 one week is given by  $19/21$  or (.90) because 2 people failed at one week, so that 19 people from the original 21 remain as survivors past one week.

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

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

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

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

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

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

17 = # in risk set at week 3

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

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

**KM** formula = product limit  
formula

Group 1 (treatment)

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\hat{S}(t_{(j)})$
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 four 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  $t_{(j)}$  given that a subject survives up to that failure time.

Thus, in the KM formula for survival past four weeks, the term 19/21 gives the probability of surviving past the first ordered failure time, one week, given survival up to the first week. Note that all 21 persons in group 2 survived up to one week, but that 2 failed at one week, leaving 19 persons surviving past one 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 one 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 four weeks stop at the fourth week with the component 14/16. Similarly, the KM formula for surviving past eight 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.

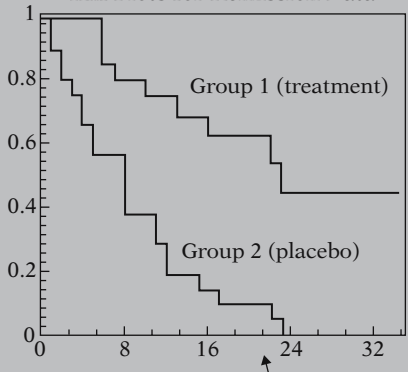
**EXAMPLE (continued)**

Group 1 (treatment)				
$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\hat{S}(t_{(j)})$
0	21	0	0	①
6	21	3	1	$1 \times \frac{18}{21} = .8571$
7	17	1	1	$.8571 \times \frac{16}{17} = .8067$
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  $t_{(j)}$ :  $\Pr(T > t_{(j)} | T \geq t_{(j)})$

Not available at  $t_{(j)}$ : failed prior to  $t_{(j)}$   
 or  
 censored prior to  $t_{(j)}$   
 ↙  
 group 1 only

**KM Plots for Remission Data**



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

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 one of these fails to survive past week 7. The other fractions are calculated similarly.

For a specified failure time  $t_{(j)}$ , the fraction may be generally expressed as the conditional probability of surviving past time  $t_{(j)}$ , given availability (i.e., in the risk set) at time  $t_{(j)}$ . 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  $t_{(j)}$  for one of two reasons: (1) either the subject has failed prior to  $t_{(j)}$ , or (2) the subject has been censored prior to  $t_{(j)}$ . 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  $t_{(j)}$ .

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, and SPSS. 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:

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

KM formula = product limit formula

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

#### EXAMPLE

$$\begin{aligned} \hat{S}(10) &= .8067 \times \frac{14}{15} = .7529 \\ &= \frac{18}{21} \times \frac{16}{17} \times \frac{14}{15} \end{aligned}$$

$$\begin{aligned} \hat{S}(16) &= .6902 \times \frac{10}{11} \\ &= \frac{18}{21} \times \frac{16}{17} \times \frac{14}{15} \times \frac{11}{12} \times \frac{10}{11} \end{aligned}$$

$$\begin{aligned} \hat{S}(t_{(j)}) &= \prod_{i=1}^j \hat{\Pr}[T > t_{(i)} | T \geq t_{(i)}] \\ &= \hat{S}(t_{(j-1)}) \\ &\quad \times \hat{\Pr}(T > t_{(j)} | T \geq t_{(j)}) \end{aligned}$$

Math proof:

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

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

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

For example, the probability of surviving past ten weeks is given in the table for group 1 (page 55) 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.

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

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.

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.

$$A = "T \geq t_{(j)}" \rightarrow A \text{ and } B = B$$

$$B = "T > t_{(j)}"$$

$$\Pr(A \text{ and } B) = \Pr(B) = \mathcal{S}(t_{(j)})$$

If we let A be the event that a subject survives to at least time  $t_{(j)}$  and we let B be the event that a subject survives past time  $t_{(j)}$ , 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  $t_{(j)}$ .

No failures during  $t_{(j-1)} < T < t_{(j)}$

$$\Pr(A) = \Pr(T > t_{(j-1)}) = \mathcal{S}(t_{(j-1)})$$

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

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

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

Thus, from  $\Pr(A \text{ and } B)$  formula,

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

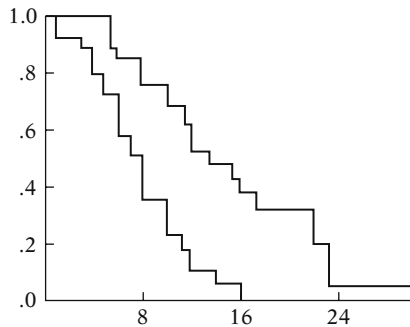
$$\Pr(A \text{ and } B) = \Pr(A) \times \Pr(B | A)$$

$$\mathcal{S}(t_{(j)}) = \mathcal{S}(t_{(j-1)})$$

$$\times \Pr(T > t_{(j)} | T \geq t_{(j)})$$

#### IV. The Log-Rank Test for Two Groups

Are KM curves statistically equivalent?



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.

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:  $n = 42$ 

$t_{(j)}$	# failures		# in risk set	
	$m_{1j}$	$m_{2j}$	$n_{1j}$	$n_{2j}$
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,  $t_{(j)}$ , in the entire set of data, we show the numbers of subjects ( $m_{ij}$ ) failing at that time, separately by group ( $i$ ), followed by the numbers of subjects ( $n_{ij}$ ) 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, one 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:**

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

$\uparrow$   
 Proportion  
in risk set

$\uparrow$   
 # of failures over  
both groups

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

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  $j$  (i.e.,  $e_{1j}$ ) as the proportion of the total subjects in both groups who are at risk at time  $j$ , that is,  $n_{1j}/(n_{1j} + n_{2j})$ , multiplied by the total number of failures at that time for both groups (i.e.,  $m_{1j} + m_{2j}$ ). For group 2,  $e_{2j}$  is computed similarly.

**EXAMPLE**

Expanded Table (Remission Data)

<i>j</i>	<i>t</i> ( <i>j</i> )	# failures		# in risk set		# expected		Observed–expected	
		<i>m</i> <sub>1<i>j</i></sub>	<i>m</i> <sub>2<i>j</i></sub>	<i>n</i> <sub>1<i>j</i></sub>	<i>n</i> <sub>2<i>j</i></sub>	<i>e</i> <sub>1<i>j</i></sub>	<i>e</i> <sub>2<i>j</i></sub>	<i>m</i> <sub>1<i>j</i></sub> – <i>e</i> <sub>1<i>j</i></sub>	<i>m</i> <sub>2<i>j</i></sub> – <i>e</i> <sub>2<i>j</i></sub>
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

$$O_i - E_j = \sum_{j=1}^{17} (m_{ij} - e_{ij}), \quad i = 1, 2$$

**EXAMPLE**

$$O_1 - E_1 = -10.26$$

$$O_2 - E_2 = 10.26$$

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.

**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.

$$\begin{aligned} & \text{Var}(O_i - E_i) \\ &= \sum_j \frac{n_{1j}n_{2j}(m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2(n_{1j} + n_{2j} - 1)} \\ & i = 1, 2 \end{aligned}$$

$H_0$ : no difference between survival curves

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

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 ( $n_{ij}$ ) and the number of failures in each group ( $m_{ij}$ ) at time  $j$ . The summation is over all distinct failure times.

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.

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, and SPSS 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**

$$O_2 - E_2 = 10.26$$

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

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

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**

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

$$\text{Log-rank statistic} = 16.793$$

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.

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

Approximation formula:

$$X^2 = \sum_i^{\text{\# of groups}} \frac{(O_i - E_i)^2}{E_i}$$

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

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.

If the number of groups being compared is  $G (\geq 2)$ , 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.

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).

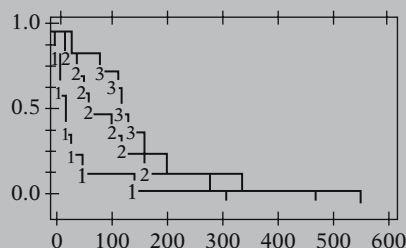
Among the variables listed, we now focus on the performance status variable (column 7). This variable is an continuous variable, so before we can obtain KM curves and the log–rank test, we need to categorize this variable.

**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 =  $\chi^2(2) = 29.18$ P-value =  $\Pr > \chi^2 = 0.0000$  $G = 3$  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

Wilcoxon

Tarone-Ware

Peto

Flemington-Harrington

There are several alternatives to the log rank test offered by Stata, SAS, and SPSS 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

$$O_i - E_i = \sum_j (m_{ij} - e_{ij})$$

i = group #

j = jth 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:

$$\frac{\left( \sum_j w(t_j)(m_{ij} - e_{ij}) \right)^2}{\text{var} \left( \sum_j w(t_j)(m_{ij} - e_{ij}) \right)}$$

i = 1, 2

j = jth failure time

$w(t_j)$  = weight at jth 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 jth failure time (as shown on the left for two groups).

*Wilcoxon Test*

- $w(t_j) = n_j$  (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  $t_j$  by the number at risk  $n_j$ , over all groups at time  $t_j$ . 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	$w(t_j)$
Log rank	1
Wilcoxon	$n_j$
Tarone–Ware	$\sqrt{n_j}$
Peto	$\tilde{s}(t_j)$
Flemington–Harrington	$\hat{s}(t_{j-1})^p [1 - \hat{s}(t_{j-1})]^q$

The Tarone–Ware test statistic also applies more weight to the early failure times by weighting the observed minus expected score at time  $t_j$  by the square root of the number at risk  $\sqrt{n_j}$ . The Peto test weights the jth 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–Harrington test uses the Kaplan–Meier survival estimate  $\hat{s}(t)$  over all groups to calculate its weights for the jth failure time,  $\hat{s}(t_{j-1})^p [1 - \hat{s}(t_{j-1})]^q$ . The weights for each of these test statistics are summarized on the left.

*Flemington–Harrington Test*

$$w(t) = \hat{s}(t_{j-1})^p [1 - \hat{s}(t_{j-1})]^q$$

if  $p = 1$  and  $q = 0$ ,  $w(t) = \hat{s}(t_{j-1})$

if  $p = 0$  and  $q = 1$ ,

$$w(t) = 1 - \hat{s}(t_{j-1})$$

if  $p = 0$  and  $q = 0$ ,

$$w(t) = 1 \text{ (log rank test)}$$

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  $w(t) = \hat{s}(t_{j-1})$  which gives more weight for the earlier survival times when  $\hat{s}(t_{j-1})$  is close to one. However, if  $p = 0$  and  $q = 1$  then  $w(t) = 1 - \hat{s}(t_{j-1})$  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	0.0005

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.

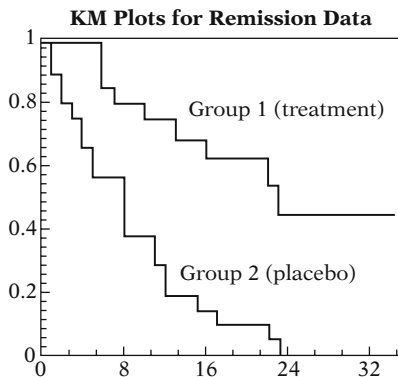
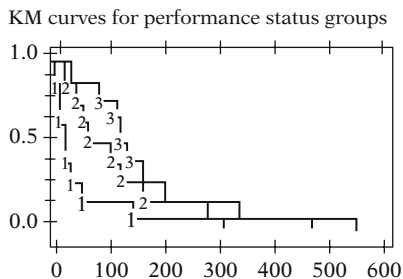
Vets Data, 3-Level Performance  
Status

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

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.

Remission Data, 2-Level Treatment

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



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 follow-up 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

-&gt;lwbc3 = 1

rx	Events observed	Events expected
0	0	<b>2.91</b>
1	4	<b>1.09</b>
Total	4	4.00

-&gt;lwbc3 = 2

rx	Events observed	Events expected
0	5	<b>7.36</b>
1	5	<b>2.64</b>
Total	10	10.00

-&gt;lwbc3 = 3

rx	Events observed	Events expected
0	4	<b>6.11</b>
1	12	<b>9.89</b>
Total	16	16.00

-&gt; Total

rx	Events observed	Events expected (*)
0	9	<b>16.38</b>
1	21	<b>13.62</b>
Total	30	30.00

(\*) sum over calculations within lwbc3 **chi2 (1) = 10.14, Pr > chi2 = 0.0014**

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

$$O_i - E_i = \sum_j (m_{ij} - e_{ij})$$

$i = \text{group \#}, \quad j = \text{jth failure time}$

Log rank stratified

$$O_i - E_i = \sum_s \sum_j (m_{ijs} - e_{ijs})$$

$i = \text{group \#}, \quad j = \text{jth failure time},$   
 $s = \text{stratum \#}$

Stratified or unstratified (G groups)

Under  $H_0$ :

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

Can stratify with other tests

Wilcoxon, Tarone–Ware,  
 Peto, Flemington–Harrington

Limitation

Sample-size may be small within  
 strata

Alternatively

Test associations using modeling

- Can simultaneously control covariates
- Shown in next chapter

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).

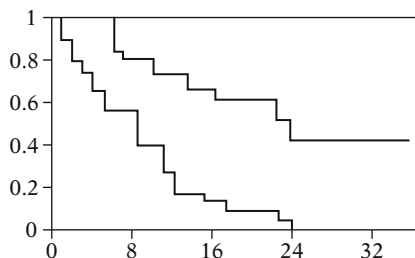
The stratified approach can also be applied to any of the weighted variations of the log rank test (e.g., Wilcoxon). A limitation of the stratified approach is the reduced sample size within each stratum. This is particularly problematic with the remission dataset, which has a small sample size to begin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

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. 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.



$t_{(j)}$ :  $j$ th ordered failure time

$$\begin{aligned}\hat{S}(t_{(j)}) &= \prod_{i=1}^j \hat{\Pr}[T > t_{(i)} | T \geq t_{(i)}] \\ &= \hat{S}(t_{(j-1)}) \\ &\quad \times \hat{\Pr}(T > t_{(j)} | T \geq t_{(j)})\end{aligned}$$

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

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.

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.

---

## Chapters

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

1. Introduction
- ✓ 2. Kaplan–Meier Survival Curves and the Log–Rank 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 48–50)
- 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  $T$
    - $\delta = (0, 1)$  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  $\infty$ .
  - 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 51–55)

- 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

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

- 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.,

$$\hat{S}_{(j)} = \hat{S}_{(j-1)} \hat{\Pr}[T > t_{(j)} | T \geq t_{(j)}].$$

### III. General Features of KM Curves (pages 56–57)

- A. Two alternative general formulae:

$$\mathcal{S}_{(j)} = \prod_{i=1}^j \Pr[T > t_{(i)} | T \geq t_{(i)}] \quad (\text{product limit formula})$$

$$\mathcal{S}_{(j)} = \mathcal{S}_{(j-1)} \Pr[T > t_{(j)} | T \geq t_{(j)}]$$

- 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 57–61)

- 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  $i$ th group at time  $j$ , where  $i = 1$  or  $2$ :

$$\text{observed counts} = m_{ij},$$

$$\text{expected counts} = e_{ij}, \text{ where}$$

$$\text{expected counts} = (\text{proportion in risk set}) \times (\# \text{ failures over both groups}),$$

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

- D. Log-rank statistic for two groups:

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

where  $i = 1, 2$ ,

$$O_i - E_i = \sum_j (m_{ij} - e_{ij}), \text{ and}$$

$$\begin{aligned} & \text{Var}(O_i - E_i) \\ &= \sum_j \frac{n_{1j}n_{2j}(m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2(n_{1j} + n_{2j} - 1)}, \\ & i = 1, 2 \end{aligned}$$

- 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 61–63)

- 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 63–68)

- A. Alternative tests supported by Stata: Wilcoxon, Tarone-Ware, Peto, and Fleming-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. Summary (page 68)

**Practice Exercises**

- 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

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

Group 1					Group 2				
$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\mathcal{S}(t_{(j)})$	$t_{(j)}$	$n_j$	$m_j$	$q_j$	$\mathcal{S}(t_{(j)})$
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				
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:

$t_{(j)}$	$m_{1j}$	$m_{2j}$	$n_{1j}$	$n_{2j}$	$e_{1j}$	$e_{2j}$	$m_{1j} - e_{1j}$	$m_{2j} - e_{2j}$
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

(Continued on next page)

$t_{(j)}$	$m_{1j}$	$m_{2j}$	$n_{1j}$	$n_{2j}$	$e_{1j}$	$e_{2j}$	$m_{1j} - e_{1j}$	$m_{2j} - e_{2j}$
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

(Continued on next page)

Subj	Survt	Relapse	Sex	log WBC	Rx
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
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  $\text{logwbc}$ . Because  $\text{logwbc}$  is continuous, we need to categorize this variable before we compute KM curves. Suppose we categorize  $\text{logwbc}$  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 =  $\text{chi}^2(2) = 26.39$

P-value =  $\text{Pr} > \text{chi}^2 = 0.0000$

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, 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/surv2.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)

- Compute and plot the KM plots for the two categories of the “clinic” variable and comment on the extent to which they differ.
- 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 =  $\text{chi2}(2) = 27.89$

P-value =  $\text{Pr} > \text{chi2} = 0.0000$

Wilcoxon =  $\text{chi2}(2) = 11.63$

P-value =  $\text{Pr} > \text{chi2} = 0.0007$

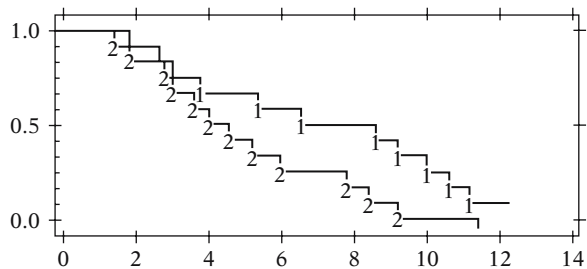
- Compute and evaluate KM curves and the log–rank test for comparing suitably chosen categories of the variable “Methadone dose.” Explain how you determined the categories for this variable.

**Answers to Practice Exercises**

1. a.

Group 1					Group 2				
$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)})$	$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)})$
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 four years, but are quite separate after four years, although they appear to come close again around twelve years.

c. Using the expanded table format, the following information is obtained:

$t_{(j)}$	$m_{1j}$	$m_{2j}$	$n_{1j}$	$n_{2j}$	$e_{1j}$	$e_{2j}$	$m_{1j} - e_{1j}$	$m_{2j} - e_{2j}$
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	-8.690	8.690

- d. The log-rank statistic can be computed from the totals of the expanded table using the formulae:

$$\begin{aligned} \text{log-rank statistic} &= \frac{(O_i - E_i)^2}{\widehat{\text{Var}}(O_i - E_i)} \\ \text{Var}(O_i - E_i) &= \sum_i \frac{n_{1j}n_{2j}(m_{1j} + m_{2j})(n_{1j} + n_{2j} - m_{1j} - m_{2j})}{(n_{1j} + n_{2j})^2(n_{1j} + n_{2j} - 1)} \end{aligned}$$

The variance turns out to be 9.448, so that the log-rank statistic is  $(8.69)^2/9.448 = 7.993$ .

Using Stata, the results for the log-rank test are given as follows:

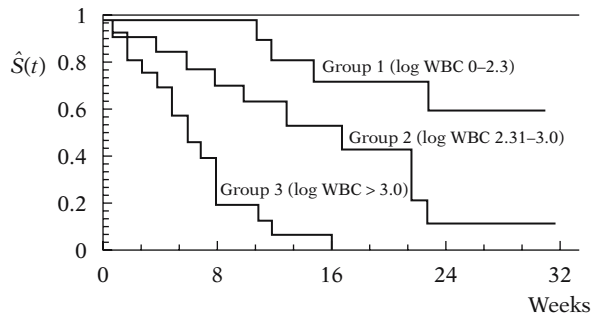
Group	Events observed	Events expected
1	22	30.79
2	25	16.21
Total	47	47.00

Log-rank =  $\text{chi2}(2) = 7.99$

P-value =  $\text{Pr} > \text{chi2} = 0.0047$

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  $i = 1, 2, \dots, G$  and  $j = 1, 2, \dots, k$ , where  $G = \#$  of groups and  $k = \#$  of distinct failure times,

$n_{ij} = \#$  at risk in  $i$ th group at  $j$ th ordered failure time

$m_{ij} =$  observed  $\#$  of failures in  $i$ th group at  $j$ th ordered failure time

$e_{ij} =$  expected  $\#$  of failures in  $i$ th group at  $j$ th ordered failure time

$$= \left( \frac{n_{ij}}{n_{1j} + n_{2j}} \right) (m_{1j} + m_{2j})$$

$$n_j = \sum_{i=1}^G n_{ij}$$

$$m_j = \sum_{i=1}^G m_{ij}$$

$$O_i - E_i = \sum_{j=1}^k (m_{ij} - e_{ij})$$

$$\text{Var}(O_i - E_i) = \sum_{j=1}^k \frac{n_{ij}(n_j - n_{ij})m_j(n_j - m_j)}{n_j^2(n_j - 1)}$$

$$\text{Cov}(O_i - E_i, O_l - E_l) = \sum_{j=1}^k \frac{-n_{ij}n_{lj}m_j(n_j - m_j)}{n_j^2(n_j - 1)}$$

$$\mathbf{d} = (O_1 - E_1, O_2 - E_2, \dots, O_{G-1} - E_{G-1})'$$

$$\mathbf{V} = ((v_{il}))$$

where  $v_{ii} = \text{Var}(O_i - E_i)$  and  $v_{il} = \text{Cov}(O_i - E_i, O_l - E_l)$  for  $i = 1, 2, \dots, G - 1$ ;  $l = 1, 2, \dots, G - 1$ .

Then, the log–rank statistic is given by the matrix product formula:

$$\text{Log–rank statistic} = \mathbf{d}'\mathbf{V}^{-1}\mathbf{d}$$

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**

## 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.

## 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 86–94)**
- II. **The formula for the Cox PH model (pages 94–96)**
- III. **Why the Cox PH model is popular (pages 96–98)**
- IV. **ML estimation of the Cox PH model (pages 98–100)**
- V. **Computing the hazard ratio (pages 100–103)**
- VI. **Adjusted survival curves using the Cox PH model (pages 103–107)**
- VII. **The meaning of the PH assumption (pages 107–111)**
- VIII. **The Cox likelihood (pages 111–115)**
- IX. **Summary (pages 115–116)**

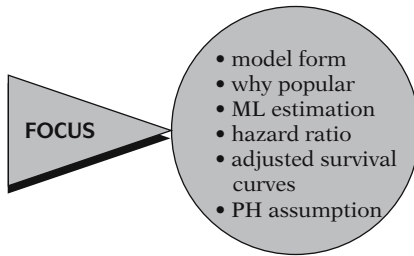


## Objectives

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.

## 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?  
 $X_3 = X_1 \times X_2 = \text{group status} \times \text{log WBC}$

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, and SPSS 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
$Rx \times \text{log WBC}$	-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				
<b>Model 1:</b>				
	Coef.	Std. Err.	p >  z	Haz. Ratio
<i>Rx</i>	1.509	0.410	0.000	4.523
No. of subjects = 42 Log likelihood = -86.380				
<b>Hazard ratios</b>				
↓				
<b>Model 2:</b>				
	Coef.	Std. Err.	p >  z	Haz. Ratio
<i>Rx</i>	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				
<b>Model 3:</b>				
	Coef.	Std. Err.	p >  z	Haz. Ratio
<i>Rx</i>	2.355	1.681	0.161	10.537
log WBC	1.803	0.447	0.000	6.067
<i>Rx</i> × 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 **hazard ratios** for the effect of each variable adjusted for the other variables in the model.

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  
 $Rx \times$  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	<u>0.510</u>	0.710
No. of subjects = 42 <u>Log likelihood = -72.066</u>				

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)**

$$P = 0.510: \frac{-0.342}{-0.520} = -0.66 = Z \quad \text{Wald statistic}$$

LR statistic: uses Log likelihood = -72.066

$$-2 \ln L (\text{log likelihood statistic}) = -2 \times (-72.066) = 144.132$$

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 <u>Log likelihood = -72.280</u>				

We now look at the printout 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.

**EXAMPLE (continued)**

$$LR (\text{interaction in model 3}) = -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}})$$

In general:

$$LR = -2 \ln L_R - (-2 \ln L_F)$$

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)**

$$\begin{aligned}
 LR \text{ (interaction in model 3)} \\
 &= -2 \ln L_{\text{model 2}} - (-2 \ln L_{\text{model 3}}) \\
 &= (-2 \times -72.280) - (-2 \times -72.066) \\
 &= 144.550 - 144.132 = 0.428
 \end{aligned}$$

(LR is  $\chi^2$  with 1 d.f. under  $H_0$ : no interaction.)  
 $0.40 < P < 0.50$ , **not significant**  
 Wald test  $P = 0.510$

LR  $\neq$  Wald

When in doubt, 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

Three statistical objectives.

1. **test for significance of effect**
2. **point estimate of effect**
3. **confidence interval for effect**

**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 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.

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.

We now focus on how to assess the effect of treatment status adjusting for log WBC using the model 2 output, again shown here.

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.

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.

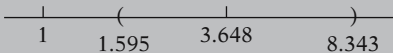
**OUTPUT**

**Model 2:**

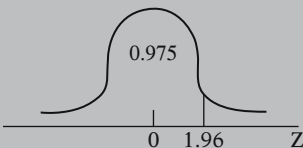
	Coef.	Std. Err.	z	p >  z	Haz. Ratio	[95% Conf. Interval]
<i>Rx</i>	1.294	0.422	3.07	0.002	3.648	1.595 8.343
log WBC	1.604	0.329	4.87	0.000	4.975	2.609 9.486
No. of subjects = 42		Log likelihood = -72.280		Prob > chi2 = 0.0000		

**EXAMPLE (continued)**

95% confidence interval for the *HR*:  
(1.595, 8.343)



95% CI for  $\beta_1$ :  $1.294 \pm (1.96)(0.422)$



95% CI for *HR* =  $e^{\beta_1}$ :

$$\exp[\hat{\beta}_1 \pm 1.96s_{\hat{\beta}_1}] = e^{1.294 \pm 1.96(0.422)}$$

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 ( $\beta_1$ ). 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  $\hat{\beta}$ 's and  $s_{\hat{\beta}}$ 's

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				
<b>Model 1:</b>				
	Coef.	Std. Err.	p >  z	Haz. Ratio
Rx	1.509	0.410	0.000	4.523
No. of subjects = 42 Log likelihood = -86.380				
<b>Model 2:</b>				
	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)
<i>HR</i> for model 1 (4.523) is higher than <i>HR</i> for model 2 (3.648).
<b>Confounding:</b> crude versus adjusted <i>HR</i> 's are meaningfully different.
Confounding due to log WBC ⇒ must control for log WBC, i.e., prefer model 2 to model 1.
If no confounding, then consider precision: e.g., if 95% CI is narrower for model 2 than model 1, we prefer model 2.

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.

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.

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 = <b>8.067</b>	
	width = <b>6.748</b>	
$Rx$ 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.

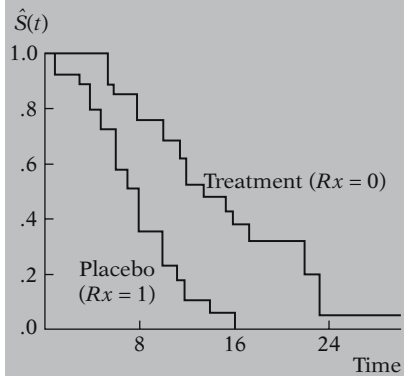
Cox model formulae not specified

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, 2002), and very similar to carrying out a classical linear regression analysis (see Kleinbaum et al., *Applied Regression Analysis*, 3rd ed., Chapter 16, 1997).

**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

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}$$

$\mathbf{X} = (X_1, X_2, \dots, X_p)$   
 explanatory/predictor variables

$h_0(t)$	×	$e^{\sum_{i=1}^p \beta_i X_i}$
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)

Time-dependent variables:  
Chapter 6

Time-independent variable:  
Values for a given individual  
do not change over time; e.g.,  
SEX and SMK

Assumed not to change  
once measured

AGE and WGT values do not change  
much, or effect on survival depends  
on one measurement.

$$\begin{aligned}
 X_1 = X_2 = \dots = X_p = 0 \\
 h(t, \mathbf{X}) &= h_0(t) e^{\sum_{i=1}^p \beta_i X_i} \\
 &= h_0(t) e^0 \\
 &= h_0(t)
 \end{aligned}$$

Baseline hazard

No  $X$ 's in model:  $h(t, \mathbf{X}) = h_0(t)$ .

$h_0(t)$  is unspecified.

Cox model: **semiparametric**

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**.

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.

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.

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.

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.

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.

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:

$$h(t, \mathbf{X}) = \lambda p t^{p-1}$$

$$\text{where } \lambda = \exp\left[\sum_{i=1}^p \beta_i X_i\right]$$

$$\text{and } h_0(t) = p t^{p-1}$$

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  $\lambda$ ,  $p$ , and the  $\beta_i$ '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  $\Rightarrow$  Cox model will approximate Weibull

Exponential  $\Rightarrow$  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.

$$h(t, \mathbf{X}) = h_0(t) \times e^{\sum_{i=1}^p \beta_i X_i}$$

Baseline hazard	Exponential
	↓
	$0 \leq h(t, \mathbf{X}) < \infty$ always

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.

$$h_0(t) \times \underbrace{\sum_{i=1}^p \beta_i X_i}_{\text{Linear}}$$

↓

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  $\beta$ 's.

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  $\beta$ 's in the exponential part of the model. As we will show later, all we need are estimates of the  $\beta$ '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.

Measure of effect: hazard ratio ( $HR$ ) involves only  $\beta$ 's, without estimating  $h_0(t)$ .

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  $S(t, \mathbf{X})$  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

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}$$

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		

Estimated model:

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t) e^{1.294 Rx + 1.604 \log WBC}$$

ML estimates: maximize likelihood function  $L$

$$L = \text{joint probability of observed data} = L(\beta)$$

We now describe how estimates are obtained for the parameters of the Cox model. The parameters are the  $\beta$ '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  $\hat{\beta}_i$  "hat."

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 eukemia patients.

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.

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  $\beta$ 's) in the model being considered.  $L$  is sometimes written notationally as  $L(\beta)$  where  $\beta$  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

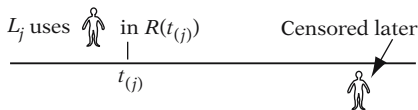
$$L = L_1 \times L_2 \times L_3 \times \dots \times L_k = \prod_{j=1}^k L_j$$

where

$L_j$  = portion of  $L$  for the  $j$ th failure time given the risk set  $R(t_{(j)})$

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  $j$ th failure time,  $L_j$  denotes the likelihood of failing at this time, given survival up to this time. Note that the set of individuals at risk at the  $j$ th failure time is called the “risk set,”  $R(t_{(j)})$ , 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  $j$ th failure time is part of the risk set used to compute  $L_j$ , even though this person is censored later.

Steps for obtaining ML estimates:

- form  $L$  from model
- maximize  $\ln L$  by solving

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.

$$\frac{\partial \ln L}{\partial \beta_i} = 0$$

$i = 1, \dots, p(\text{\# of parameters})$

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 86–94)

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

$$\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$$

and

$$\mathbf{X} = (X_1, X_2, \dots, X_p)$$

denote the set of  $X$ 's for two individuals

To interpret  $\widehat{HR}$ , want  $\widehat{HR} \geq 1$ , i.e.,

$$\hat{h}(t, \mathbf{X}^*) \geq \hat{h}(t, \mathbf{X}).$$

Typical coding:  $\mathbf{X}^*$ : group with larger  $h$   
 $\mathbf{X}$ : group with smaller  $h$

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.

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

$\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$ , where  $X_1^* = 1$  denotes **placebo** group.

$\mathbf{X} = (X_1, X_2, \dots, X_p)$ , where  $X_1 = 0$  denotes **treatment** group.



$$\widehat{HR} = \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i}}$$

$$\widehat{HR} = \frac{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i^*}}{\hat{h}_0(t) e^{\sum_{i=1}^p \hat{\beta}_i X_i}} = e^{\sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i)}$$

$$\widehat{HR} = \exp \left[ \sum_{i=1}^p \beta_i (X_i^* - X_i) \right]$$

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.

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  $\beta_i$  "hat" times the difference between  $X_i^*$  and  $X_i$ .

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**

$\mathbf{X} = (X_1, X_2, \dots, X_p) = (X_1)$ , where  $X_1$  denotes (0,1) exposure status ( $p = 1$ )

$X_1^* = 1, X_1 = 0$

$\widehat{HR} = \exp[\hat{\beta}_1(X_1^* - X_1)]$   
 $= \exp[\hat{\beta}_1(1 - 0)] = e^{\hat{\beta}_1}$

Model 1:

	Coef.	Std. Err.	P >  z	Haz. Ratio
Rx	(1.509)	0.410	0.000	(4.523)

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  $\beta_1$  "hat" times 1 minus 0, which simplifies to  $e$  to the  $\beta_1$  "hat."

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  $X^*$  and  $X$  though unspecified.

**EXAMPLE 2 (continued)**

$$\begin{aligned}\widehat{HR} &= \exp[\hat{\beta}_1(X_1^* - X_1) + \hat{\beta}_1(X_2^* - X_2)] \\ &= \exp[1.294(1-0) \\ &\quad + 1.604(\log \text{WBC} - \log \text{WBC})] \\ &= \exp[1.294(1) + 1.604(0)] = e^{1.294}\end{aligned}$$

**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.

**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
$Rx \times \log \text{WBC}$	-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:

$$\mathbf{X} = (X_1 = 0, X_2 = \log \text{WBC}, X_3 = 0 \times \log \text{WBC})$$

$$\widehat{HR} = \exp\left[\sum_{i=1}^3 \hat{\beta}_i (X_i^* - X_i)\right]$$

$$\begin{aligned}\widehat{HR} &= \exp[2.355(1 - 0) \\ &\quad + 1.803(\log \text{WBC} - \log \text{WBC}) \\ &\quad + (-0.342)(1 \times \log \text{WBC} \\ &\quad - 0 \times \log \text{WBC})] \\ &= \exp[2.355 - 0.342 \log \text{WBC}]\end{aligned}$$

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.

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.

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  $X_3^* = 1$  times log 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 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 3 (continued)**

$$\begin{aligned} \log \text{WBC} &= 2: \\ \widehat{HR} &= \exp[2.355 - 0.342 (2)] \\ &= e^{1.671} = 5.32 \\ \log \text{WBC} &= 4: \\ \widehat{HR} &= \exp[2.355 - 0.342 (4)] \\ &= e^{0.987} = 2.68 \end{aligned}$$

In order to get a numerical value for the hazard ratio, we must specify a value for log WBC. For instance, if  $\log \text{WBC} = 2$ , the estimated hazard ratio becomes 5.32, whereas if  $\log \text{WBC} = 4$ , 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:

$$\widehat{HR} = \exp\left[\hat{\beta} + \sum \hat{\delta}_j W_j\right]$$

where

$\hat{\beta}$  = coefficient of  $E$

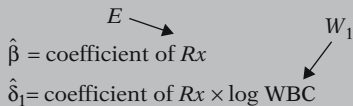
$\hat{\delta}_j$  = coefficient of  $E \times W_j$

( $\widehat{HR}$  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  $\hat{\beta}$  "hat" denotes the coefficient of the exposure variable and the  $\hat{\delta}$  "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:



$$\begin{aligned} \widehat{HR} \text{ (model 3)} &= \exp[\hat{\beta} + \hat{\delta}_1 \log \text{WBC}] \\ &= \exp[2.355 - 0.342 \log \text{WBC}] \end{aligned}$$

For model 3,  $\hat{\beta}$  "hat" is the coefficient of the  $Rx$  variable, and there is only one  $\hat{\delta}$  "hat" in the sum, which is the coefficient of the product term  $Rx \times \log \text{WBC}$ . Thus, there is only one  $W$ , namely  $W_1 = \log \text{WBC}$ . The hazard ratio formula for the effect of exposure is then given by exponentiating  $\hat{\beta}$  "hat" plus  $\hat{\delta}_1$  "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.

---

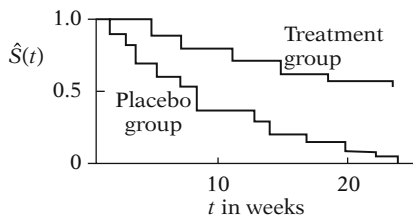
## VI. Adjusted Survival Curves Using the Cox PH Model

Two primary quantities:

1. estimated hazard ratios
2. estimated survival curves

The two primary quantities desired from a survival analysis point of view 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.

No model: use KM curves



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 functions).

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:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}$$

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  $\beta_i$  times  $X_i$ .

Cox model survival function:

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{-\sum_{i=1}^p \beta_i X_i}}$$

Estimated survival function:

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{e^{-\sum_{i=1}^p \hat{\beta}_i X_i}}$$

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**

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t) e^{1.294 Rx + 1.604 \log \text{WBC}}$$

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp(1.294 Rx + 1.604 \log \text{WBC})}$$

Specify values for  $\mathbf{X} = (Rx, \log \text{WBC})$ :

$Rx = 1, \log \text{WBC} = 2.93$ :

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp(1.294(1) + 1.604(2.93))}$$

$$= [\hat{S}_0(t)]^{\exp(5.99)} = \left( [\hat{S}_0(t)]^{400.9} \right)$$

$Rx = 0, \log \text{WBC} = 2.93$ :

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp(1.294(0) + 1.604(2.93))}$$

$$= [\hat{S}_0(t)]^{\exp(4.70)} = \left( [\hat{S}_0(t)]^{109.9} \right)$$

Adjusted Survival Curves

$Rx = 1, \log \text{WBC} = 2.93$ :

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{400.9}$$

$Rx = 0, \log \text{WBC} = 2.93$ :

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{109.9}$$

Typically, use  $X = \bar{X}$  or  $X_{\text{median}}$ .

Computer uses  $\bar{X}$ .

**EXAMPLE (continued)**

Remission data ( $n = 42$ ):

$\log \text{WBC} = 2.93$

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.

We can obtain a specific survival curve by specifying values for the vector  $\mathbf{X}$ , whose component variables are  $Rx$  and  $\log \text{WBC}$ .

For instance, if  $Rx = 1$  and  $\log \text{WBC} = 2.93$ , 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  $\log \text{WBC}$  for the entire dataset of 42 subjects.

Also, if  $Rx = 0$  and  $\log \text{WBC} = 2.93$ , 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  $t$ .

The two formulae just obtained, again shown here, allow us to compare survival curves for different treatment groups adjusted for the covariate  $\log \text{WBC}$ . Both curves describe estimated survival probabilities over time assuming the same value of  $\log \text{WBC}$ , in this case, the value 2.93.

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.

In our example, the mean  $\log \text{WBC}$  for all 42 subjects in the remission data set is 2.93. That is why we chose this value for  $\log \text{WBC}$  in the formulae for the adjusted survival curve.

General formulae for adjusted survival curves comparing two groups:

Exposed subjects:

$$\hat{S}(t, \mathbf{X}_1) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_1(1) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i]}$$

Unexposed subjects:

$$\hat{S}(t, \mathbf{X}_0) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_1(0) + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i]}$$

General formula for adjusted survival curve for all covariates in the model:

$$\hat{S}(t, \bar{\mathbf{X}}) = [\hat{S}_0(t)]^{\exp[\sum \hat{\beta}_i \bar{X}_i]}$$

#### EXAMPLE (continued)

Single survival curve for Cox model containing  $Rx$  and log WBC:

$$\overline{Rx} = 0.50$$

$$\overline{\log \text{WBC}} = 2.93$$

$$\begin{aligned} \hat{S}(t, \bar{\mathbf{X}}) &= [\hat{S}_0(t)]^{\exp(\hat{\beta}_1 \overline{Rx} + \hat{\beta}_2 \overline{\log \text{WBC}})} \\ &= [\hat{S}_0(t)]^{\exp(1.294(0.5) + 1.604(2.93))} \\ &= [\hat{S}_0(t)]^{\exp(5.35)} = \boxed{[\hat{S}_0(t)]^{210.6}} \end{aligned}$$

Compute survival probability by specifying value for  $t$  in

$$\hat{S}(t, \bar{\mathbf{X}}) = [\hat{S}_0(t)]^{210.6}$$

Computer uses  $t$ 's which are failure times.

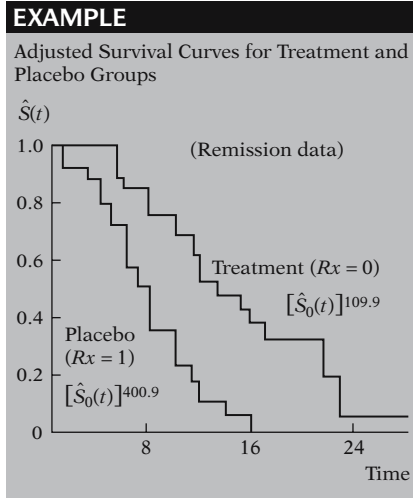
More generally, if we want to compare survival curves for two levels of an exposure variable, and we want to adjust for several covariates, we can write the formula for each curve as shown here. Note that we are assuming that the exposure variable is variable  $X_1$ , whose estimated coefficient is  $\hat{\beta}_1$  "hat," and the value of  $X_1$  is 1 for exposed and 0 for unexposed subjects.

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.

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.)

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.



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

Later presentations:

- how to evaluate PH
- analysis when PH not met

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.

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.

## VII. The Meaning of the PH Assumption

PH:  $HR$  is constant over time, i.e.,  
 $\hat{h}(t, \mathbf{X}^*) = \text{constant} \times \hat{h}(t, \mathbf{X})$

$$\begin{aligned} \widehat{HR} &= \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} \\ &= \frac{\hat{h}_0(t) \exp\left[\sum \hat{\beta}_i X_i^*\right]}{\hat{h}_0(t) \exp\left[\sum \hat{\beta}_i X_i\right]} \\ &= \exp\left[\sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i)\right] \end{aligned}$$

where  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$  and  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  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.

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

does not involve  $t$ .

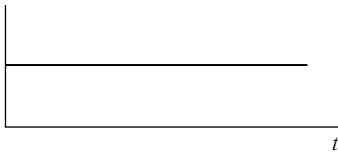
Let  $\overset{\text{Constant}}{\searrow}$

$$\hat{\theta} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

then

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta}$$

$\widehat{HR}(\mathbf{X}^* \text{ versus } \mathbf{X})$



$$\hat{h}(t, \mathbf{X}^*) = \hat{\theta} \hat{h}(t, \mathbf{X})$$

Proportionality constant  
(not dependent on time)

The final expression for the hazard ratio therefore involves the estimated coefficients  $\beta_1$  “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  $t$ .

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  $\theta$  “hat,” then we can write the hazard ratio as shown here. This is a mathematical expression which states the proportional hazards assumption.

Graphically, this expression says that the estimated hazard ratio comparing any two individuals plots as a constant over time.

Another way to write the proportional hazards assumption mathematically expresses the hazard function for individual  $\mathbf{X}^*$  as  $\theta$  “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  $\theta$  “hat,” which does not depend on time.

**EXAMPLE: Remission Data**

$$\hat{h}(t, \mathbf{X}) = \hat{h}_0(t) e^{1.294 Rx + 1.604 \log \text{WBC}}$$

$$\begin{aligned} \widehat{HR} &= \frac{\hat{h}(t, Rx=1, \log \text{WBC} = 2.93)}{\hat{h}(t, Rx=0, \log \text{WBC} = 2.93)} \\ &= \exp[1.294] = 3.65 \text{ Constant} \end{aligned}$$

Placebo  $\swarrow$

$$\hat{h}(t, Rx = 1, \log \text{WBC} = 2.93) = 3.65 \hat{h}(t, Rx = 0, \log \text{WBC} = 2.93)$$

$\nwarrow$  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  $\log \text{WBC}$ . For this model, the estimated hazard ratio that compares placebo ( $Rx = 1$ ) with treated ( $Rx = 0$ ) subjects controlling for  $\log \text{WBC}$  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.



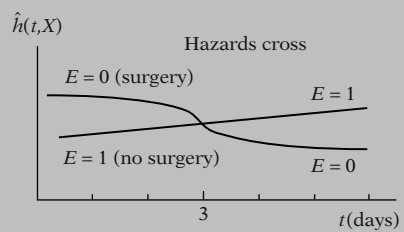
**EXAMPLE: PH Not Satisfied**



$$E = \begin{cases} 0 & \text{if surgery} \\ 1 & \text{if no surgery} \end{cases}$$

$$h(t, \mathbf{X}) = h_0(t)e^{\beta E}$$

Is the above Cox PH model appropriate?  
 Note:  
 Serious surgery  $\Rightarrow$  High risk for death early



2 days:  $\frac{\hat{h}(t=2, E=1)}{\hat{h}(t=2, E=0)} < 1$   
 but  
 5 days:  $\frac{\hat{h}(t=5, E=1)}{\hat{h}(t=5, E=0)} > 1$

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 three days, and that prior to three days the hazard for the surgery group is higher than the hazard for the no surgery group, whereas after three 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*:

$$h(t, \mathbf{X}) = h_0(t)e^{\beta E}$$

$$\widehat{HR} = \frac{\hat{h}(t, E = 1)}{\hat{h}(t, E = 0)} = e^{\hat{\beta}}$$

General rule:

If the hazards cross, then a Cox PH model is not appropriate.

Analysis when Cox PH model not appropriate? See Chapters 5 and 6.

**EXAMPLE (continued)**

Surgery study analysis options:

- stratify by exposure (use KM curves)
- start analysis at three 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

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 three days and greater than 1 after three 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  $\beta$  “hat,” which does not vary over time.

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.

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 brief reply with regard to the surgery study example just described.

Actually, for the surgery study there are several options available for the analysis. These include:

- analyze by stratifying on the exposure variable; that is, do not fit any model, and instead obtain 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 for greater than three days to get two different hazard ratio estimates, one for each of these two time 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.

Hazards  
cross  $\Rightarrow$  PH not met

but

?  $\Rightarrow$  PH met

See Chapter 4: Evaluating PH Assumption

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.

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."

---

## VIII. 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  $t_1, t_2, \dots$
- 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  $t_j$  ( $j = 1, 2, \dots$ ). 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 the 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 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.

For Cox model

Subject's pattern of covariates  
 affects likelihood of ordered  
 events

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

Cox PH model

$$h(t) = h_0(t)e^{\beta_1 \text{SMOKE}}$$

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

Cox Likelihood

$$L = \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right] \times \left[ \frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right] \times \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}} \right]$$

Likelihood is product of 3 terms

$$L = L_1 \times L_2 \times L_3$$

$$L_1 = \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right]$$

$$L_2 = \left[ \frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right]$$

$$L_3 = \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}} \right]$$

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.

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.

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<sub>1</sub>) 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<sub>2</sub>) 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<sub>3</sub>) just has Larry's hazard in the denominator and the numerator.

$t_1$ , 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

$$L = L_1 \times L_2 \times L_3 \times \cdots \times L_k$$

$$= \prod_{j=1}^k L_j$$

Obtaining maximum likelihood estimates

Solve system of equations

$$\frac{\partial \ln L}{\partial \beta_i} = 0, \quad i = 1, 2, 3, \dots, p$$

$p = \# \text{ of parameters}$

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  $L_j$  denote the contribution to the likelihood corresponding to the jth 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  $L_j$  is constructed in a similar manner as with the data for Gary, Larry, and Barry.

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).

---

## IX. Summary

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”

In this section we briefly summarize the content covered in this presentation.

- 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.

5. ML estimation: maximize a partial likelihood  
 $L =$  joint probability of observed data  $= L(\beta)$

6. Hazard ratio formula:

$$\widehat{HR} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

7. Adjusted survival curves: 0 or 1  
 Comparing  $E$  groups:

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_1 E + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i]}$$

Single curve:

$$\hat{S}(t, \bar{\mathbf{X}}) = [\hat{S}_0(t)]^{\exp[\sum \hat{\beta}_i \bar{X}_i]}$$

8. PH assumption:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta} \text{ (a constant over } t)$$

$$\text{i.e., } \hat{h}(t, \mathbf{X}^*) = \hat{\theta} \hat{h}(t, \mathbf{X})$$

Hazards cross  $\Rightarrow$  PH not met

9. Cox PH likelihood

- 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 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.
- Finally, we describe how the Cox likelihood is developed using the ordered failure times from the data.

---

## 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 86–94)

- 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 94–96)

$$A. \quad h(t, \mathbf{X}) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i \right]$$

- B.  $h_0(t)$  is called the **baseline hazard function**.
- C.  $\mathbf{X}$  denotes a collection of  $p$  explanatory variables  $X_1, X_2, \dots, X_p$ .
- 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 96–98)

- 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 98–100)

- 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 100–103)

- A. Formula for hazard ratio comparing two individuals,  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$  and  $\mathbf{X} = (X_1, X_2, \dots, X_p)$ :

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

- 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. Adjusted survival curves using the Cox PH model** (pages 103–107)

- 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  $\mathbf{X} = (X_1, X_2, \dots, X_p)$ .
- 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.

**VII. The meaning of the PH assumption** (pages 107–111)

- 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

**VIII. Cox likelihood** (pages 111–115)

- A. Lottery Example
- B. Likelihood based on order of events

**IX. Summary** (pages 115–116)

**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.

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
Four indicator variables for cell type	1	Treatment	Standard = 1, test = 2
	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

- State the Cox PH model used to obtain the above computer results.
- 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  $\times$  performance status; treatment  $\times$  disease duration; treatment  $\times$  age; and treatment  $\times$  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: *SPIDA manual*, Sydney, Australia, 1991; and 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)

122 3. The Cox Proportional Hazards Model and Its Characteristics

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

- For model 1, give an expression for the hazard ratio for the effect of the platelet variable adjusted for age and sex.
- 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.
- Carry out an appropriate test of hypothesis to evaluate whether there is any significant interaction in model 1. What is your conclusion?
- Considering models 2–5, evaluate whether age and sex need to be controlled as confounders?
- Which of the five models do you think is the best model and why?
- 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.

**Test**

1. Consider a hypothetical two-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 two 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  $\times$  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 $\times$ 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  $\text{CHR} = 1$  crosses the hazard function for persons with  $\text{CHR} = 0$ , 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  $\text{CHR} = 1$ , adjusted for AGE. Also, give an expression for the estimated survival curve for persons with  $\text{CHR} = 0$ , 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]	
<i>Rx</i>	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
<i>Rx</i> × Sex	1.952	0.907	.031	7.046	1.191	41.702
<i>Rx</i> × log WBC	-0.151	0.531	.776	0.860	0.304	2.433

Log likelihood = -69.515

**Model 2:**

<i>Rx</i>	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
<i>Rx</i> × Sex	2.013	0.883	.023	7.483	1.325	42.261

Log likelihood = -69.555

**Model 3:**

<i>Rx</i>	0.587	0.542	.279	1.798	0.621	5.202
Sex	-1.073	0.701	.126	0.342	0.087	1.353
<i>Rx</i> × Sex	1.906	0.815	.019	6.726	1.362	33.213

Log likelihood = -83.475

**Model 4:**

<i>Rx</i>	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

- 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?
- Evaluate whether you would prefer model 1 or model 2. Explain your answer.
- Using model 2, give an expression for the hazard ratio for the effect of the *Rx* variable adjusted for SEX and log WBC.
- Using your answer in part 3c, compute the hazard ratio for the effect of *Rx* for males and for females separately.
- By considering the potential confounding of log WBC, determine which of models 2 and 3 you prefer. Explain.
- 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  $\chi_3^2$  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 1a above).
  - d. 95% CI for adjusted  $HR$ :  
$$\exp\left[2\hat{\beta}_1 \pm 1.96 \times 2\sqrt{\text{var}(\hat{\beta}_1)}\right]$$
  - e.  $\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[4\hat{\beta}_1 + (\text{AGE})\hat{\beta}_2 + (\text{RACE})\hat{\beta}_3 + (\text{SEX})\hat{\beta}_4]}$
  - 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.
  
2. a.  $h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_7 X_7 + \dots + \beta_{10} X_{10}]$ 
  - b. Adeno cell type:  $\mathbf{X}^* = (\text{treatment}, 1, 0, 0, \text{perfstat}, \text{disdur}, \text{age}, \text{prther})$   
Large cell type:  $\mathbf{X} = (\text{treatment}, 0, 0, 0, \text{perfstat}, \text{disdur}, \text{age}, \text{prther})$   
$$HR = \frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp\left[\sum_{i=1}^p \beta_i (X_i^* - X_i)\right]$$
$$= \exp[0 + \hat{\beta}_3(1 - 0) + \hat{\beta}_4(0 - 0) + \hat{\beta}_5(0 - 0) + 0 + \dots + 0]$$
$$= \exp[\hat{\beta}_3] = \exp[0.789] = 2.20$$
  - c. Adeno cell type:  $\mathbf{X}^* = (\text{treatment}, 1, 0, 0, \text{perfstat}, \text{disdur}, \text{age}, \text{prther})$   
Squamous cell type:  $\mathbf{X} = (\text{treatment}, 0, 0, 1, \text{perfstat}, \text{disdur}, \text{age}, \text{prther})$   
$$HR = \frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp\left[\sum_{i=1}^p \beta_i (X_i^* - X_i)\right]$$
$$= \exp[0 + \hat{\beta}_3(1 - 0) + \hat{\beta}_4(0 - 0) + \hat{\beta}_5(0 - 1) + 0 + \dots + 0]$$
$$= \exp[\hat{\beta}_3 - \hat{\beta}_5] = \exp[0.789 - (-0.400)] = \exp[1.189] = 3.28$$

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.

$$e. \hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[2\hat{\beta}_1 + \hat{\beta}_5 + (\overline{\text{perfstat}})\hat{\beta}_7 + (\overline{\text{disdur}})\hat{\beta}_8 + (\overline{\text{age}})\hat{\beta}_9 + (\overline{\text{prther}})\hat{\beta}_{10}]}$$

$$f. HR = \frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp[\beta_1 + (\text{perfstat})\beta_{11} + (\text{disdur})\beta_{12} + (\text{age})\beta_{13} + (\text{prther})\beta_{14}]$$

where  $\beta_1$  is the coefficient of the treatment variable and  $\beta_{11}$ ,  $\beta_{12}$ ,  $\beta_{13}$ , and  $\beta_{14}$  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

# **Evaluating the Propor- tional Hazards Assumption**

## 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 134–135)**
- II. **Checking the PH assumption: Overview (pages 135–137)**
- III. **Graphical approach 1: log–log plots (pages 137–145)**
- IV. **Graphical approach 2: observed versus expected plots (pages 145–150)**
- V. **The goodness-of-fit (GOF) testing approach (pages 151–153)**
- VI. **Assessing the PH assumption using time-dependent covariates (pages 153–157)**

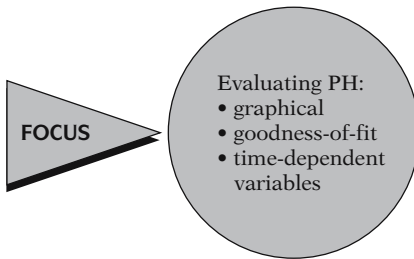
## Objectives

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:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{i=1}^p \beta_i X_i}$$

$\mathbf{X} = (X_1, X_2, \dots, X_p)$  explanatory/  
predictor variables

$h_0(t)$	×	$e^{\sum_{i=1}^p \beta_i X_i}$
Baseline hazard		Exponential
Involves $t$ but not $X$ 's		Involves $X$ 's but not $t$ ( $X$ 's are time- independent)

$X$ '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:

$$\widehat{HR} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

where  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$   
 and  $\mathbf{X} = (X_1, X_2, \dots, X_p)$   
 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: 0 or 1  
 Comparing  $E$  groups:

$$\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[\beta_1 E + \sum_{i \neq 1} \hat{\beta}_i \bar{X}_i]}$$

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:

$$\hat{S}(t, \bar{\mathbf{X}}) = [\hat{S}_0(t)]^{\exp[\hat{\beta}_i \bar{X}_i]}$$

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

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.

Hazards don't cross  $\nRightarrow$  PH met

---

## II. Checking the Proportional Hazards Assumption: Overview

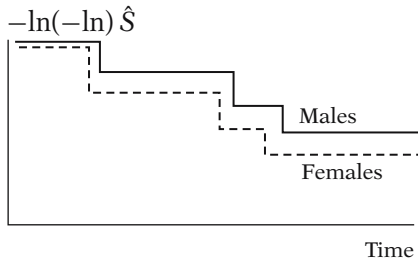
There are three general approaches for assessing the PH assumption, again listed here. We now briefly overview each approach, starting with graphical techniques.

Three approaches:

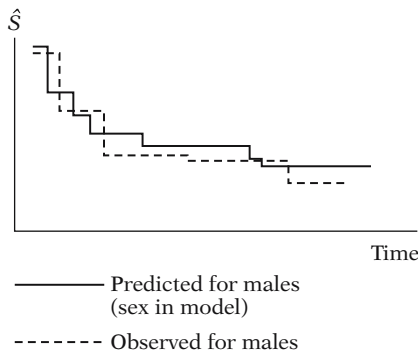
- graphical
- goodness-of-fit test
- time-dependent variables

Graphical techniques:

$-\ln(-\ln) S$  curves parallel?



Observed vs. predicted: Close?



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  $\Rightarrow$  PH satisfied  
(e.g.  $P > 0.10$ )

$p$ -value small  $\Rightarrow$  PH not satisfied  
(e.g.  $P < 0.05$ )

Time-dependent covariates:

Extended Cox model:

Add product term involving some function of time.

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.

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.

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, suggest 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.

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**

$h(t, X) = h_0(t) \exp[\beta \text{Sex} + \delta(\text{Sex} \times t)]$   
 $\delta \neq 0 \Rightarrow \text{PH assumption violated}$

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable “Sex × 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.

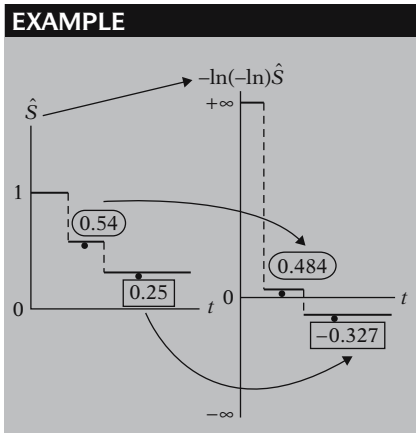
$\log - \log \hat{S} = \text{transformation of } \hat{S}$   
 $= -\ln(-\ln \hat{S})$

- $\ln \hat{S}$  is negative  $\Rightarrow -(\ln \hat{S})$  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.<sup>1</sup>

---

<sup>1</sup>An equivalent way to write  $-\ln(-\ln \hat{S})$  is  $-\ln(\int_0^t h(u)du)$ , where  $\int_0^t h(u)du$  is called the “cumulative hazard” function. This result follows from the formula  $S(t) = \exp[-\int_0^t h(u)du]$ , which relates the survivor function to the hazard function (see p. 14 in Chapter 1).



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,  $-\ln(-\ln 0.54) = 0.484$

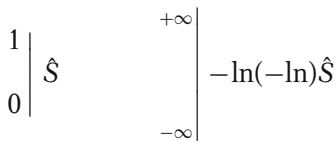
To illustrate the computation of a log-log value, suppose we start with an estimated survival probability of 0.54. Then the log-log transformation of this value is  $-\ln(-\ln 0.54)$ , which is  $-\ln(0.616)$ , because  $\ln(0.54)$  equals  $-0.616$ . Now, continuing further,  $-\ln(0.616)$  equals 0.484, 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:



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:

$$h(t, \mathbf{X}) = h_0(t) e^{\sum_{j=1}^p \beta_j X_j}$$

↓ From math

Cox PH survival function:

$$S(t, \mathbf{X}) = [S_0(t)]^{e^{-\sum_{j=1}^p \beta_j X_j}}$$

↖ Baseline survival function.

log-log ⇒ takes logs twice

log #1:

$$\ln S(t, \mathbf{X}) = e^{-\sum_{j=1}^p \beta_j X_j} \times \ln S_0(t)$$

$$0 \leq S(t, \mathbf{X}) \leq 1$$

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:

$$\begin{aligned} \ln[-\ln S(t, \mathbf{X})] &= \ln \left[ -e^{-\sum_{j=1}^p \beta_j X_j} \times \ln S_0(t) \right] \\ &= \ln \left[ e^{\sum_{j=1}^p \beta_j X_j} \right] + \ln[-\ln S_0(t)] \\ &= \sum_{i=1}^p \beta_i X_i + \ln[-\ln S_0(t)] \end{aligned}$$

$$\begin{aligned} -\ln[-\ln S(t, \mathbf{X})] &= -\sum_{i=1}^p \beta_i X_i - \ln[-\ln S_0(t)] \end{aligned}$$

or

$$\begin{aligned} \ln[-\ln S(t, \mathbf{X})] &= +\sum_{i=1}^p \beta_i X_i + \ln[-\ln S(t)] \end{aligned}$$

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  $h_0(t)$ .

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.

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.

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**.

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:

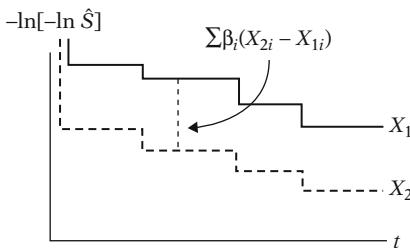
$$\mathbf{X}_1 = (X_{11}, X_{12}, \dots, X_{1p})$$

$$\mathbf{X}_2 = (X_{21}, X_{22}, \dots, X_{2p})$$

$$\begin{cases} -\ln[-\ln S(t, \mathbf{X}_1)] \\ = -\sum_{i=1}^p \beta_i X_{1i} - \ln[-\ln S_0(t)] \\ -\ln[-\ln S(t, \mathbf{X}_2)] \\ = -\sum_{i=1}^p \beta_i X_{2i} - \ln[-\ln S_0(t)] \end{cases}$$

$$\begin{aligned} &-\ln[-\ln S(t, \mathbf{X}_1)] \\ &- (-\ln[-\ln S(t, \mathbf{X}_2)]) \\ &= \sum_{i=1}^p \beta_i (X_{2i} - X_{1i}) \\ &\text{does not involve } t \end{aligned}$$

$$\begin{aligned} &-\ln[-\ln S(t, \mathbf{X}_1)] \\ &= -\ln[-\ln S(t, \mathbf{X}_2)] \\ &\quad + \sum_{i=1}^p \beta_i (X_{2i} - X_{1i}) \end{aligned}$$



Graphical approach using log-log plots: PH model is appropriate if “empirical” plots of log-log survival curves are parallel.

Now suppose we consider two different specifications of the  $\mathbf{X}$  vector, corresponding to two different individuals,  $\mathbf{X}_1$  and  $\mathbf{X}_2$ .

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  $\mathbf{X}$ .

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  $t$ .

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.

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

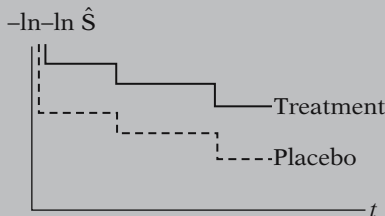
Predictors ( $X$ 's):  
 $Rx$  (= 1 if placebo, 0 if treatment)  
 log WBC

Cox PH model:

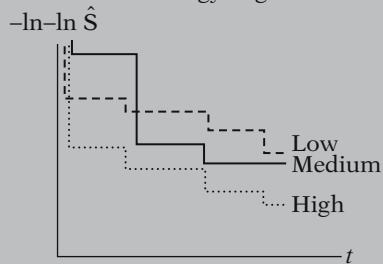
$$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 Rx + \beta_2 \log \text{WBC}]$$

Assessing PH assumption:  
 compare log-log survival curves for categories of  $Rx$  and log WBC

One-at-a-time strategy:  $Rx$  variable



One-at-a-time strategy: log WBC



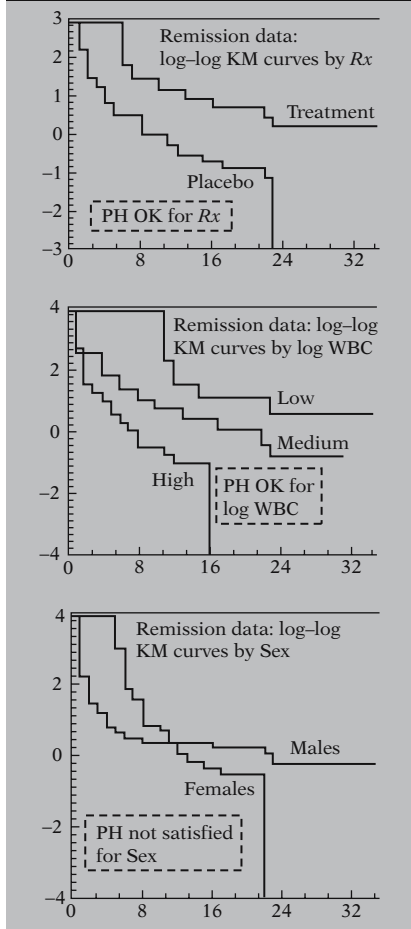
As an example, suppose we consider the comparison of treatment and placebo groups in a clinical trial of leukemia patients, where survival time is time, in weeks, until a patient goes out of remission. Two predictors of interest in this study are treatment group status (1 = placebo, 0 = treatment), denoted 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 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.

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, and SPSS 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 to categorize a continuous variable?

- many categories  $\Rightarrow$  data “thins out”
- different categorizations may give different graphical pictures

Recommend:

- small # of categories (2 or 3)
- meaningful choice
- reasonable balance (e.g., terciles)

How to evaluate several variables simultaneously?

Strategy:

- categorize variables separately
- form combinations of categories
- compare log-log curves on same graph

Drawback:

- data “thins out”
- difficult to identify variables responsible for nonparallelism

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.]

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).

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.

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.

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.

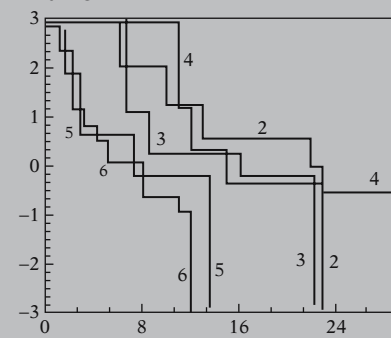
**EXAMPLE**

		Remission Data		
		log WBC	Low	Medium
<i>Rx</i>	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  $\widehat{KM}$  curves by six combinations of  $Rx$  by log WBC



Plots suggest PH not satisfied. However, the study is small, i.e., plots are unreliable.

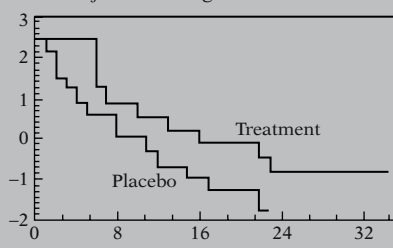
Alternative strategy:  
Adjust for predictors already satisfying PH assumption, i.e., use adjusted log-log  $\widehat{S}$  curves

**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  $\widehat{S}$  curves for  $Rx$  groups using PH model adjusted for log WBC



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 twelve 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.

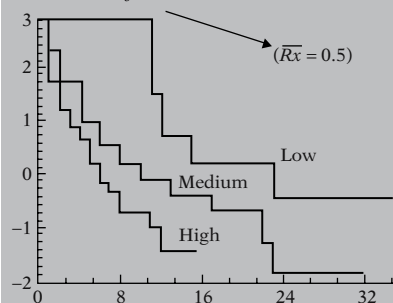
An alternative graphical strategy for considering several predictors together is to assess the PH assumption for one predictor adjusted for other predictors that are assumed to satisfy the PH assumption. Rather than using Kaplan-Meier curves, this involves a comparison of adjusted log-log survival curves.

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  $Rx$

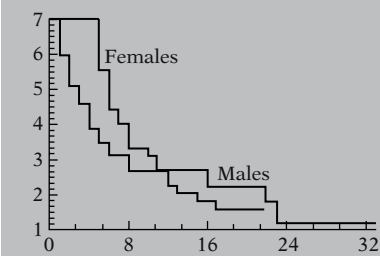


Remission data:

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)

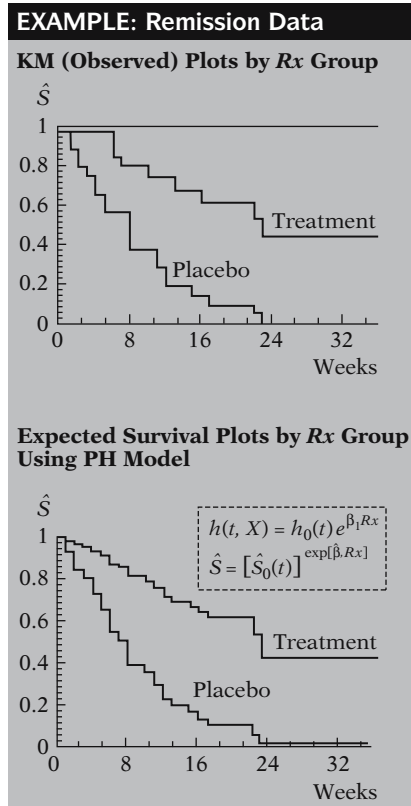
As with the log–log approach, the observed versus expected approach may be carried out using either or both of two strategies—(1) assessing the PH assumption for variables one-at-a-time, or (2) assessing the PH assumption after adjusting for other variables. The strategy which adjusts for other variables 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.

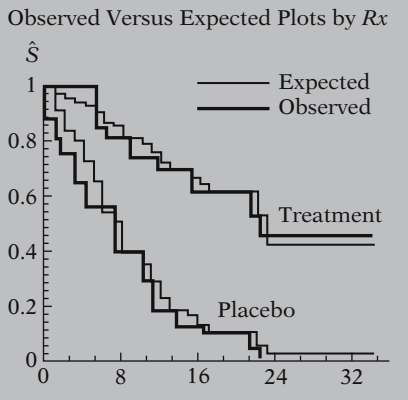


As an example, for the remission data on 42 leukemia patients we have illustrated earlier, the KM plots for the treatment and placebo groups, with 21 subjects in each group, are shown here. These are the “observed” plots.

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)**



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)**

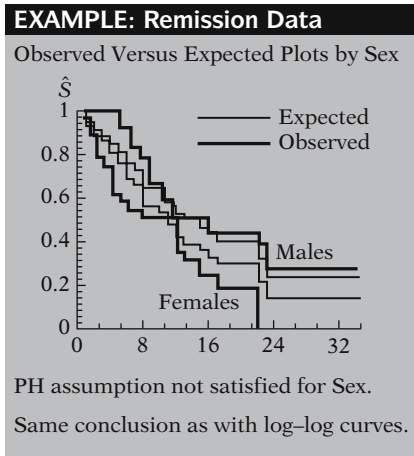
Observed and expected plots are close for each treatment group.  
Conclude PH assumption not violated.

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.



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 nonparallel.

Continuous variable:

- form strata from categories
- observed plots are KM curves for each category
- two options for expected plots
  1. Use PH model with  $k - 1$  dummy variables  $X_i$  for  $k$  categories, i.e.,

$$h(t, \mathbf{X}) = h_0(t) \exp\left(\sum_{i=1}^{k-1} \beta_i X_i\right)$$

Obtain adjusted survival curve:

$$\hat{S}(t, \mathbf{X}_c) = [\hat{S}_0(t)]^{\exp(\sum \hat{\beta}_i X_{ci})}$$

where

$\mathbf{X}_c = (X_{c1}, X_{c2}, \dots, X_{c,k-1})$   
gives values of dummy variables for category  $c$ .

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.

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:

$$h(t, X) = h_0(t) \exp(\beta X)$$

↙  
Continuous

Obtain adjusted survival curve:

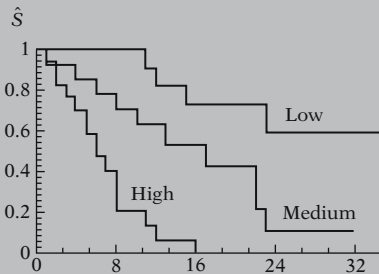
$$\hat{S}(t, \bar{X}_c) = [\hat{S}_0(t)]^{\exp(\hat{\beta} \bar{X}_c)}$$

where  $\bar{X}_c$  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 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.

**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:

$$X_1 = 1, X_2 = 0: \hat{S}(t, X_{\text{high}}) = [\hat{S}_0(t)]^{\exp(\hat{\beta}_1)}$$

$$X_1 = 0, X_2 = 1: \hat{S}(t, X_{\text{medium}}) = [\hat{S}_0(t)]^{\exp(\hat{\beta}_2)}$$

$$X_1 = 0, X_2 = 0: \hat{S}(t, X_{\text{low}}) = [\hat{S}_0(t)]$$

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.

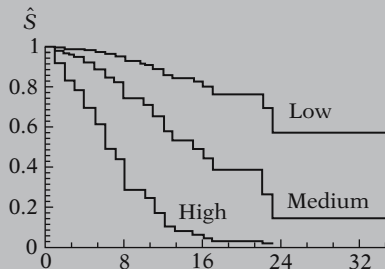
Using option 1, expected plots would be obtained by fitting a Cox PH model containing two dummy variables  $X_1$  and  $X_2$ , as shown here, where  $X_1$  takes the values 1 if high or 0 if other and  $X_2$  takes the values 1 if medium or 0 if other. Thus, when log WBC is high, the values of  $X_1$  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.

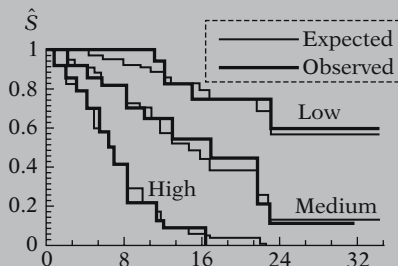


**EXAMPLE (continued)**

Expected Plots for log WBC Using Option 1 (Dummy Variables)



Observed Versus Expected Plots Using Option 1



Option 2: Treat log WBC as continuous  
 $h(t, \mathbf{X}) = h_0(t) \exp[\beta(\log \text{WBC})]$

$\overline{\log \text{WBC}}_{\text{high}} = 3.83:$

$$\hat{S}(t, X_{\text{high}}) = [\hat{S}_0(t)]^{\exp[3.83\hat{\beta}]}$$

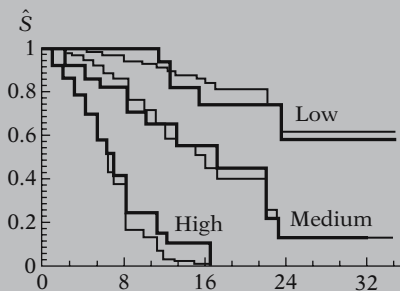
$\overline{\log \text{WBC}}_{\text{med}} = 2.64:$

$$\hat{S}(t, X_{\text{med}}) = [\hat{S}_0(t)]^{\exp[2.64\hat{\beta}]}$$

$\overline{\log \text{WBC}}_{\text{low}} = 1.71:$

$$\hat{S}(t, X_{\text{low}}) = [\hat{S}_0(t)]^{\exp[1.71\hat{\beta}]}$$

Observed Versus Expected Plots for log WBC Using Option 2



The expected plots using option 1 (the dummy variable approach) are shown here for the three categories of log WBC.

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.

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.

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

Test of Harrel and Lee (1986)

- Variation of test of Schoenfeld
- Uses Schoenfeld residuals

Schoenfeld residuals defined for

- Each predictor in model
- Every subject who has event

Consider Cox PH model

$$h(t) = h_0(t) \exp(\beta_1 RX + \beta_2 \log WBC + \beta_3 SEX)$$

3 predictors → 3 Schoenfeld residuals for each subject who has event

Schoenfeld residual for *i*th subject for LOGWBC

$$\text{Observed LOGWBC} - \text{LOGWBC weighted average}$$

Weights are other subjects' hazard (from subjects still at risk)

Underlying idea of test  
If PH holds then Schoenfeld residuals uncorrelated with time

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.

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.

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.

Suppose subject *i* has an event at time  $t_j$ . 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  $t_j$ . The weights are each subject's hazard.

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  
shown in Computer Appendix

The implementation of the test for the PH assumption in Stata, SAS, and, SPSS 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.	$P(PH)$
$R_x$	1.391	0.457	0.935
log WBC	1.594	0.330	0.828
Sex	0.263	0.449	0.038

log WBC and  $R_x$  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

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 objective, but can detect specific violations

Recommend—Use both graphs and tests

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.

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**

Extended Cox model: contains product terms of the form  $X \times g(t)$ , where  $g(t)$  is a 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.

One-at-a-time model:

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta X + \delta X \times g(t)]$$

Some choices for  $g(t)$ :

$$g(t) = t$$

$$g(t) = \log t$$

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases} \quad (\text{heaviside function})$$

$$H_0: \delta = 0$$

Under  $H_0$ , the model reduces to:

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta X]$$

Use either Wald statistic or likelihood ratio statistic:

$\chi^2$  with 1 df under  $H_0$

#### EXAMPLE

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 \text{Sex} + \beta_2 (\text{Sex} \times t)]$$

$\beta_2 \neq 0 \Rightarrow$  PH assumption violated

Strategies for assessing PH:

- one-at-a-time
- several predictors simultaneously
- for a given predictor adjusted for other predictors

When assessing predictors one-at-a-time, the extended Cox model takes the general form shown here for the predictor  $X$ .

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,  $\log t$ .

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 “ $\delta$  equal to zero.” Note that if the null hypothesis is true, the model reduces to a Cox PH model containing the single variable  $X$ .

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.

For example, if the PH assumption is being assessed for Sex, a Cox model might be extended to include the variable  $\text{Sex} \times t$  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.<sup>2</sup>

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.

<sup>2</sup>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  $i$ th predictor

$$\begin{aligned} H_0: \delta_1 = \delta_2 = \dots = \delta_p = 0 \\ LR = -2 \ln L_{\text{PH model}} \\ \quad - (-2 \ln L_{\text{ext. Cox model}}) \\ \quad \sim \chi_p^2 \text{ under } H_0 \end{aligned}$$

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  $i$ th predictor.

With the above model, we test for the PH assumption simultaneously by assessing the null hypothesis that all the  $\delta_i$  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— $-2 \ln L$ —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

$$\begin{aligned} h(t, \mathbf{X}) = h_0(t) \exp [ & \beta_1 (Rx) \\ & + \beta_2 (\log \text{ WBC}) + \beta_3 (\text{Sex}) \\ & + \delta_1 (Rx) \times g(t) + \delta_2 (\log \text{ WBC}) \\ & \times g(t) + \delta_3 (\text{Sex}) \times g(t) ] \end{aligned}$$

$$\text{where } g(t) = \begin{cases} 1 & \text{if } t \geq 7 \\ 0 & \text{if } t < 7 \end{cases}$$

$$H_0: \delta_1 = \delta_2 = \delta_3 = 0$$

$$LR \sim \chi^2 \text{ with 3 df}$$

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  $\delta$  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.

Assessing PH for a given predictor adjusted for other predictors:

$$h(t, \mathbf{X}) = h_0(t) \exp \left[ \sum_{i=1}^{p-1} \beta_i X_i + \beta^* X^* + \delta^* X^* \times g(t) \right]$$

$X^*$  = Predictor of interest

$H_0: \delta^* = 0$

Wald or LR statistic  $\sim \chi^2$  with 1 df

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.

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  $\delta^*$  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:

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 (Rx) + \beta_2 (\log WBC) + \beta^* (\text{Sex}) + \delta^* (\text{Sex}) \times g(t)]$$

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 Stata, SAS, and SPSS

Drawback: choice of  $g_i(t)$

Different choices may lead to different conclusions about PH assumption.

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, SPSS, and Stata.

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

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.

Recommend using at least two methods.

---

## 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

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.

Next:

5. The Stratified Cox Procedure
6. Extension of the Cox Proportional Hazards Model for Time-Dependent Variables



## Detailed Outline

### I. Background (pages 134–135)

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,

$$\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*) \text{ and } \mathbf{X} = (X_1, X_2, \dots, X_p):$$

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp \left[ \sum_{i=1}^p \beta_i (X_i^* - X_i) \right]$$

C. Adjusted survival curves using the Cox PH model:

$$\mathcal{S}(t, \mathbf{X}) = [\mathcal{S}_0(t)]^{\exp[\sum \beta_i X_i]}$$

i. To graph  $\mathcal{S}(t, \mathbf{X})$ , must specify values for

$$\mathbf{X} = (X_1, X_2, \dots, X_p).$$

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 PH assumption: Overview (pages 135–137)

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 137–145)

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

$$-\ln[-\ln \mathcal{S}(t, \mathbf{X})] = -\sum_{i=1}^p \beta_i X_i - \ln[-\ln \mathcal{S}_0(t)]$$

- 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

$$\begin{aligned} &-\ln[-\ln \mathcal{S}(t, \mathbf{X}_1)] \\ &= -\ln[-\ln \mathcal{S}(t, \mathbf{X}_2)] + \sum_{i=1}^p \beta_i (X_{2i} - X_{1i}) \end{aligned}$$

- 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  $-\ln -\ln \hat{S}$ :

- i.  $\hat{S}$  is a KM curve
- ii.  $\hat{S}$  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 145–150)

- 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 151–153)
- 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 153–157)

- 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:  
 $h(t, \mathbf{X}) = h_0(t) \exp[\beta X + \delta g(t)]$ .  
 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  $i$ th predictor. Test  $H_0: \delta_1 = \delta_2 = \dots = \delta_p = 0$  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, pp. 223–224, 1980). 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,  $\dots$ , 100 = best)

Column 8: Disease duration (months)

Column 9: Age

Column 10: Prior therapy (none = 0, some = 10)

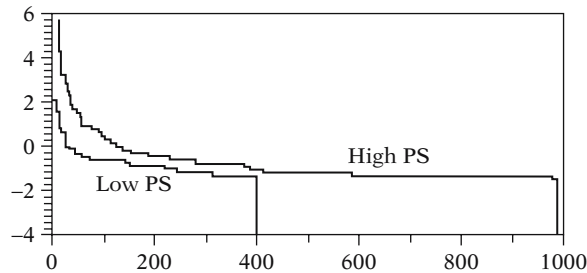
Column 11: Status (0 = censored, 1 = died)

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.

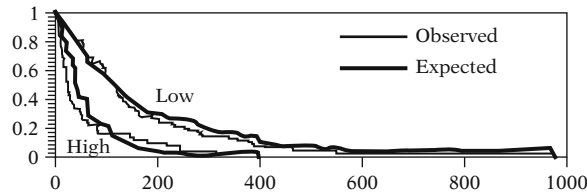
Cox regression	Coef.	Std. Err.	$p >  z $	Haz. Ratio	[95% Conf. Interval]		$P(PH)$
Treatment	0.290	0.207	0.162	1.336	0.890	2.006	0.628
Large cell	0.400	0.283	0.157	1.491	0.857	2.594	0.033
Adeno cell	1.188	0.301	0.000	3.281	1.820	5.915	0.081
Small cell	0.856	0.275	0.002	2.355	1.374	4.037	0.078
Performance 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 3, 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 3, 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)

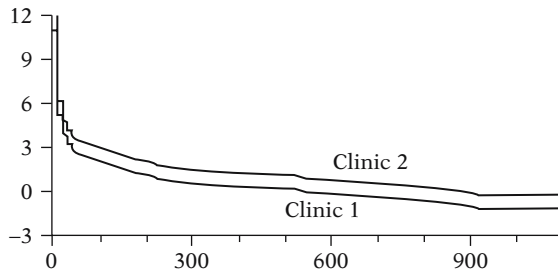
Column 6: Maximum methadone dose (mg/day)

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.

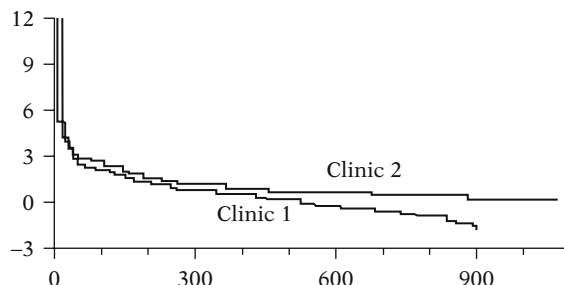


3. The following printout was obtained from fitting a stratified Cox PH model to these data, where the variable being stratified is clinic:

Stratified					
Cox regression					
Analysis time.t:					[95% Conf.
survt (in days)	Coef.	Std. Err.	p >  z	Haz. Ratio	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

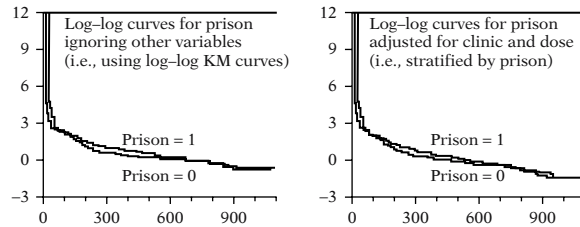
No. of subjects = 238    Log likelihood = -597.714    Stratified by clinic

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  $CT_i$  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:
- graphical, using either log–log plots or observed versus expected plots;
  - statistical test;
  - 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 obtain 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 heavyside function. To test the PH assumption simultaneously for all variables, the null hypothesis is stated as  $H_0: \delta_1 = \delta_2 = \dots = \delta_8 = 0$ . The test statistic is a likelihood-ratio statistic of the form

$$LR = -2 \ln L_R - (-2 \ln L_F)$$

where  $R$  denotes the reduced (PH) model obtained when all  $\delta$ '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**

---

## 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 176)**
- II. An Example (pages 176–180)**
- III. The General Stratified Cox (SC) Model (pages 180–181)**
- IV. The No-Interaction Assumption and How to Test It (pages 182–188)**
- V. A Second Example Involving Several Stratification Variables (pages 188–193)**
- VI. A Graphical View of the Stratified Cox Approach (pages 193–194)**
- VII. Summary (pages 195–196)**



## Objectives

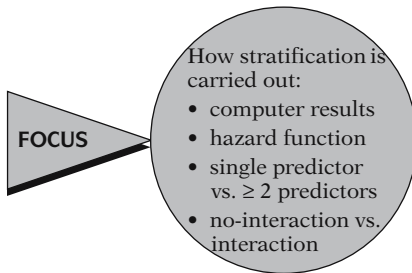
Upon completing the chapter, the learner should be able to:

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.

## 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

Appendix A illustrates SC procedures using Stata, SAS, and SPSS.

- Log WBC and Rx are included in SC model.
- SC model is stratified by SEX.

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 ( $Rx = 0$ )

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:  $\exp(0.931 \pm 1.96 \times 0.472)$

Wald test:  $P = 0.048$  (two-tailed), significant at the 0.05 level.

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, or SPSS).

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.

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  $Rx = 0$ ).

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.

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 just 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 =  $\overline{-59.648}$  Stratified by sex

$$\begin{aligned} LR &= (-2 \times -59.648) - (-2 \times -57.560) \\ &= 119.296 - 115.120 = 4.179 \quad (P < 0.05) \end{aligned}$$

*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  $\beta_1$  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  $g$ th 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  $\beta_1$  "hat," where  $\beta_1$  is the coefficient of the  $Rx$  variable.

**EXAMPLE (continued)**

Cannot estimate  $HR$  for SEX variable (SEX doesn't satisfy PH).

**Different** baseline hazard functions:  $h_{01}(t)$  for females and  $h_{02}(t)$  for males.

Same coefficients  $\beta_1$  and  $\beta_2$  for both female and male models.

Different baselines  $\left\{ \begin{array}{l} h_{01}(t) \Rightarrow \text{Survival curve} \\ \text{for females} \\ h_{02}(t) \Rightarrow \text{Survival curve} \\ \text{for males} \end{array} \right.$

Females and males:  
same  $\beta_1$  and  $\beta_2 \Rightarrow$  same  $\widehat{HR}$ 's, e.g.,  $e^{\hat{\beta}_1}$

No interaction assumption  
(see Section IV)

Estimates of  $\beta_1$  and  $\beta_2$ :

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  $h_2(t)$ .

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.

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  $\beta_1$  and  $\beta_2$  are the same for both female and male models.

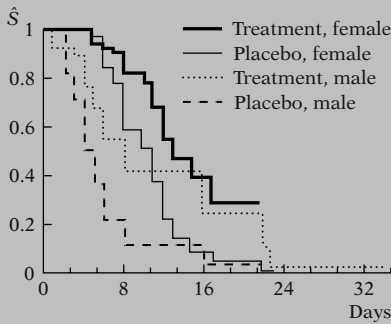
Because there are different baseline hazard functions, the fitted stratified Cox model will yield different estimated survival curves for females and males. These curves will be described shortly.

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  $\beta_1$  "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.

To obtain estimates of  $\beta_1$  and  $\beta_2$ , 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  $R_x$   
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

Example: one binary predictor

↓

General: several predictors, several strata

$Z_1, Z_2, \dots, Z_k$ , do not satisfy PH

$X_1, X_2, \dots, X_p$ , 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  $Z^*$

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.

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  $Z_1, Z_2, \dots, Z_k$ ; the variables satisfying the PH assumption we denote as  $X_1, X_2, \dots, X_p$ .

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.

The general SC model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p]$$

$g = 1, 2, \dots, k^*$ , strata defined from  $Z^*$

$Z^*$  not included in the model

$X_1, X_2, \dots, X_p$  included in the model

**Different** baseline hazard functions:

$$h_{0g}(t), g = 1, 2, \dots, k^*$$

**Same** coefficients:  $\beta_1, \beta_2, \dots, \beta_p$

$$\text{Different baselines} \left\{ \begin{array}{l} \hat{h}_{01}(t) \Rightarrow \hat{S}_1(t) \\ \hat{h}_{02}(t) \Rightarrow \hat{S}_2(t) \\ \vdots \\ \hat{h}_{0k}(t) \Rightarrow \hat{S}_k(t) \end{array} \right\} \text{Different survival curves}$$

$\widehat{HR}$  same for each stratum

(no-interaction assumption, Section IV)

(Partial) likelihood function:

$$L = L_1 \times L_2, \times \dots \times L_{k^*}$$

Strata:	1	2	...	$k^*$
Likelihood:	$L_1$	$L_2$	...	$L_{k^*}$
Hazard:	$h_1(t, \mathbf{X})$	$h_2(t, \mathbf{X})$	...	$h_{k^*}(t, \mathbf{X})$

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.

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  $g$ th stratum. The strata are defined as the different categories of the stratification variable  $Z^*$ , and the number of strata equals  $k^*$ .

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.

Note also that the baseline hazard function  $h_{0g}(t)$  is allowed to be different for each stratum. However, the coefficients  $\beta_1, \beta_2, \dots, \beta_p$  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.

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.

To obtain estimates of the regression coefficients  $\beta_1, \beta_2, \dots, \beta_p$ , 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

Stratified Cox model

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p]$$

$\beta$  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	Coef	StErr.	p-value	HR	0.95	CI	P(PH)
4 log WBC	1.639	0.519	0.002	5.150	1.862	14.242	0.228
5 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	Coef	StErr.	p-value	HR	0.95	CI	P(PH)
4 log WBC	1.170	0.499	0.019	3.222	1.213	8.562	0.674
5 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 previously pointed out that the SC model contains regression coefficients, denoted as  $\beta$ '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.

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:

$$\begin{aligned}
 (\diamond) h_g(t, \mathbf{X}) \\
 = h_{0g}(t) \exp[\beta_{1g} \log \text{WBC} + \beta_{2g} Rx] \\
 \text{where } g = 1 \text{ (females)}, g = 2 \text{ (males)}
 \end{aligned}$$

No-interaction model:

$$\begin{aligned}
 h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \log \text{WBC} + \beta_2 Rx] \\
 \text{where } g = 1 \text{ (females)}, g = 2 \text{ (males)}
 \end{aligned}$$

Alternative interaction model:

$$\begin{aligned}
 (\star) h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1^* \log \text{WBC} \\
 + \beta_2^* Rx + \beta_3^* (\text{SEX} \times \log \text{WBC}) + \beta_4^* \\
 \times (\text{SEX} \times Rx)]
 \end{aligned}$$

$$\text{where SEX} = \begin{cases} 1 & \text{if female} \\ 0 & \text{if male} \end{cases}$$

$h_{0g}(t)$  are different for  $g = 1, 2$   
 $\beta^*$  coefficients do not involve  $g$

Equivalence of models  $(\diamond)$  and  $(\star)$ :  
 $g = 1$  (females), so that sex = 1:

$$\begin{aligned}
 h_1(t, \mathbf{X}) &= h_{01}(t) \exp[\beta_1^* \log \text{WBC} + \beta_2^* Rx \\
 &\quad + \beta_3^* (1 \times \log \text{WBC}) + \beta_4^* (1 \times Rx)] \\
 &= h_{01}(t) \exp\left[\left(\beta_1^* + \beta_3^*\right) \log \text{WBC} \right. \\
 &\quad \left. + \left(\beta_2^* + \beta_4^*\right) Rx\right]
 \end{aligned}$$

$g = 2$  (males), so that sex = 0:

$$\begin{aligned}
 h_2(t, \mathbf{X}) &= h_{02}(t) \exp[\beta_1^* \log \text{WBC} + \beta_2^* Rx \\
 &\quad + \beta_3^* (0 \times \log \text{WBC}) + \beta_4^* (0 \times Rx)] \\
 &= h_{02}(t) \exp\left[\beta_1^* \log \text{WBC} + \beta_2^* Rx\right]
 \end{aligned}$$

Interaction models in same format:

Females ( $g = 1$ ):  $h_1(t, \mathbf{X})$ 

$$(\diamond) = h_{01}(t) \exp[\beta_{11} \log \text{WBC} + \beta_{21} Rx]$$

$$(\star) = h_{01}(t) \exp[(\beta_1^* + \beta_3^*) \log \text{WBC} + (\beta_2^* + \beta_4^*) Rx]$$

Males ( $g = 2$ ):  $h_2(t, \mathbf{X})$ 

$$(\diamond) = h_{02}(t) \exp[\beta_{12} \log \text{WBC} + \beta_{22} Rx]$$

$$(\star) = h_{02}(t) \exp[\beta_1^* \log \text{WBC} + \beta_2^* Rx]$$

One way to state the hazard model formula when there is interaction is shown here  $(\diamond)$ . 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  $\beta_{1g}$  and  $\beta_{2g}$ .

In contrast, in the no-interaction model, the coefficient ( $\beta_1$ ) of log WBC is the same for females and for males; also, the coefficient ( $\beta_2$ ) of  $Rx$  is the same for females and for males.

An alternative way to write the interaction model is shown here  $(\star)$ . This alternative form contains two product terms—SEX  $\times$  log WBC and SEX  $\times$   $Rx$ —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  $\beta^*$  coefficients do not involve the subscript  $g$  and therefore are the same for each sex.

Nevertheless, this alternative formula  $(\star)$  is equivalent to the interaction formula  $(\diamond)$  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  $\beta_1^*$  for males.

Similarly, the coefficients of  $Rx$  are different, namely,  $(\beta_2^* + \beta_4^*)$  for females versus  $\beta_2^*$  for males.

The preceding formulae indicate that two seemingly different formulae for the interaction model— $(\diamond)$  versus  $(\star)$ , shown earlier—can be written in the same format. We show these formulae here separately for females and males.

**EXAMPLE (continued)**

$$\begin{aligned} & (\diamond) \quad (\star) \\ \text{Females } (g = 1): & \beta_{11} = \beta_1^* + \beta_3^* \\ & \beta_{21} = \beta_2^* + \beta_4^* \end{aligned}$$

$$\begin{aligned} & (\diamond) \quad (\star) \\ \text{Males } (g = 2): & \beta_{12} = \beta_1^* \\ & \beta_{22} = \beta_2^* \end{aligned}$$

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:

$$\begin{aligned} \log \text{ WBC} & \begin{cases} \hat{\beta}_{11} = \underline{1.639} \\ \hat{\beta}_1^* + \hat{\beta}_3^* = 1.170 + 0.469 = \underline{1.639} \end{cases} \\ Rx & \begin{cases} \hat{\beta}_{21} = \underline{1.859} \\ \hat{\beta}_2^* + \hat{\beta}_4^* = 0.267 + 1.592 = \underline{1.859} \end{cases} \end{aligned}$$

Males:

$$\begin{aligned} \log \text{ WBC} & \hat{\beta}_{12} = \underline{1.170} = \hat{\beta}_1^* \\ Rx & \hat{\beta}_{22} = \underline{0.267} = \hat{\beta}_2^* \end{aligned}$$

Interaction model:

$$\begin{aligned} h_g(t, \mathbf{X}) = & h_{0g}(t) \exp[\beta_1^* \log \text{ WBC} + \beta_2^* Rx \\ & + \beta_3^* (\text{SEX} \times \log \text{ WBC}) \\ & + \beta_4^* (\text{SEX} \times Rx)] \end{aligned}$$

Notice that for females, the coefficient  $\beta_{11}$  in model  $(\diamond)$  must be equivalent to  $(\beta_1^* + \beta_3^*)$  in model  $(\star)$  because both models have the same format, and both  $\beta_{11}$  and  $(\beta_1^* + \beta_3^*)$  are coefficients of the same variable, log WBC. Similarly,  $\beta_{21}$  in model  $(\diamond)$  is equivalent to  $(\beta_2^* + \beta_4^*)$  in model  $(\star)$  because both are coefficients of the same variable,  $Rx$ .

For males, it follows in an analogous way, that the coefficient  $\beta_{12}$  is equivalent to  $\beta_1^*$ , and, similarly,  $\beta_{22}$  equals  $\beta_2^*$ .

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 shown 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 shown 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

$LR \sim \chi_{2df}^2$  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

Now, we generalize this process.

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.

No-interaction SC model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p]$$

$g = 1, 2, \dots, k^*$ , strata defined from  $Z^*$

SC model allowing interaction:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} X_1 + \beta_{2g} X_2 + \cdots + \beta_{pg} X_p]$$

$g = 1, 2, \dots, k^*$ , strata defined from  $Z^*$

Alternative SC interaction model:

- uses product terms involving  $Z^*$
- define  $k^* - 1$  dummy variables  $Z_1^*, Z_2^*, \dots, Z_{k^*-1}^*$ , from  $Z^*$
- products of the form  $Z_i^* \times X_j$ , where  $i = 1, \dots, k^* - 1$  and  $j = 1, \dots, p$ .

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \cdots + \beta_p X_p + \beta_{11}(Z_1^* \times X_1) + \cdots + \beta_{p1}(Z_1^* \times X_p) + \beta_{12}(Z_2^* \times X_1) + \cdots + \beta_{p2}(Z_2^* \times X_p) + \cdots + \beta_{1,k^*-1}(Z_{k^*-1}^* \times X_1) + \cdots + \beta_{p,k^*-1}(Z_{k^*-1}^* \times X_p)]$$

$g = 1, 2, \dots, k^*$ , 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.

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  $g$ th stratum and indicates that the regression coefficients are different for different strata of  $Z^*$ .

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  $Z_1^*, Z_2^*, \dots, Z_{k^*-1}^*$ , needs to be involved in a product term with each of the  $X$ 's.

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  $\beta$  coefficients.

**EXAMPLE (Remission Data)**

$$\begin{aligned}
 Z^* &= \text{sex}, k^* = 2, \\
 Z_1^* &= \text{sex}(0,1), \\
 X_1 &= \log \text{WBC}, X_2 = Rx \quad (p = 2) \\
 h_g(t, \mathbf{X}) &= h_{0g}(t) \exp[\beta_1^* X_1 + \beta_2^* X_2 \\
 &\quad + \beta_{11}^* (Z_1^* \times X_1) \\
 &\quad + \beta_{21}^* (Z_1^* \times X_2)] \\
 &= h_{0g}(t) \exp[\beta_1^* \log \text{WBC} \\
 &\quad + \beta_2^* Rx + \beta_3^* (\text{sex} \times \log \text{WBC}) \\
 &\quad + \beta_4^* (\text{sex} \times Rx)] \\
 g &= 1, 2 \\
 \beta_1 &= \beta_1^*, \beta_2 = \beta_2^*, \beta_{11} = \beta_3^*, \text{ and } \beta_{21} = \beta_4^*
 \end{aligned}$$

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:

$LR = -2 \ln L_R - (-2 \ln L_F)$   
 $R =$  reduced (no-interaction) model  
 $F =$  full (interaction) model  
 contains product terms

$$H_0: \begin{cases} \beta_{11} = \dots = \beta_{p1} = 0 \\ \beta_{12} = \dots = \beta_{p2} = 0 \\ \vdots \\ \beta_{1,k^*-1} = \dots = \beta_{p,k^*-1} = 0 \end{cases}$$

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  $-2 \ln L_R$  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.

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.

$$\begin{aligned}
 LR &\sim \chi_{p(k^*-1)}^2 \text{ df} \\
 &\text{under } H_0: \text{ no interaction}
 \end{aligned}$$

$p(k^* - 1)$  gives number of product terms being tested in interaction model

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.

**EXAMPLE (Remission Data)**

$Z^* = \text{sex}, k^* = 2,$   
 $Z_1^* = \text{sex}(0,1),$   
 $X_1 = \log \text{WBC}, X_2 = Rx \ (p = 2)$   
 $p(k^* - 1) = 2,$  so  
 $LR \sim \chi_{2df}^2$  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)

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

Variables not satisfying PH:

- cell type (3 dummy variables)
- performance status
- prior therapy (possibly)

SC model: stratifies on cell type and performance status

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.

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.

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.

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:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \text{Treatment} + \beta_2 \text{Age}]$$

$g = 1, 2, \dots, 8$  (= # of strata defined from  $Z^*$ )

Interaction model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} \text{Treatment} + \beta_{2g} \text{Age}]$$

$g = 1, 2, \dots, 8$

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  $4 \times 2$ , 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:

$$\begin{aligned}
 h_g(t, \mathbf{X}) &= h_{0g}(t) \exp[\beta_1 \text{Treatment} \\
 &\quad + \beta_2 \text{Age} \\
 &\quad + \beta_{11}(Z_1^* \times \text{Treatment}) + \cdots \\
 &\quad + \beta_{17}(Z_7^* \times \text{Treatment}) \\
 &\quad + \beta_{21}(Z_1^* \times \text{Age}) + \cdots + \beta_{27}(Z_7^* \times \text{Age})] \\
 g &= 1, 2, \dots, 8
 \end{aligned}$$

Another version of interaction model:

$$\begin{aligned}
 &\text{Replace } Z_1^*, \dots, Z_7^* \text{ by} \\
 &Z_1^* = \text{large cell (binary)} \\
 &Z_2^* = \text{adeno cell (binary)} \\
 &Z_3^* = \text{small cell (binary)} \\
 &Z_4^* = \text{PSbin (binary)} \\
 &Z_5^* = Z_1^* \times Z_4^* \\
 &Z_6^* = Z_2^* \times Z_4^* \\
 &Z_7^* = Z_3^* \times Z_4^*
 \end{aligned}$$

$$\begin{aligned}
 h_g(t, \mathbf{X}) &= h_{0g}(t) \exp[\beta_1 \text{Treatment} + \beta_2 \text{Age} \\
 &\quad + \beta_{11}(\text{tr } Z_1^*) + \beta_{12}(\text{tr } Z_2^*) + \beta_{13}(\text{tr } Z_3^*) \\
 &\quad + \beta_{14}(\text{tr } Z_4^*) + \beta_{15}(\text{tr } Z_1^* Z_4^*) \\
 &\quad + \beta_{16}(\text{tr } Z_2^* Z_4^*) + \beta_{17}(\text{tr } Z_3^* Z_4^*) \\
 &\quad + \beta_{21}(\text{AGE } Z_1^*) + \beta_{22}(\text{AGE } Z_2^*) \\
 &\quad + \beta_{23}(\text{AGE } Z_3^*) + \beta_{24}(\text{AGE } Z_4^*) \\
 &\quad + \beta_{25}(\text{AGE } Z_1^* Z_4^*) + \beta_{26}(\text{AGE } Z_2^* Z_4^*) \\
 &\quad + \beta_{27}(\text{AGE } Z_3^* Z_4^*)]
 \end{aligned}$$

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.

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 ( $Z_4^*$ ).

The latter interaction model is shown here. In this model, the variable  $\text{tr } Z_1^*$  denotes the product of treatment status with the large cell dummy  $Z_1^*$ , the variable  $\text{tr } Z_2^*$  denotes the product of treatment status with the adeno cell variable  $Z_2^*$ , and so on. Also, the variable  $\text{tr } Z_1^* Z_4^*$  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  $\text{AGE } Z_1^*$  denotes the product of age with  $Z_1^*$ , and the variable  $\text{AGE } Z_1^* Z_4^*$  denotes the triple product of age times  $Z_1^*$  times  $Z_4^*$ .

Note that we are just 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	0.187	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  
 $\hat{h}_1(t, \mathbf{X}) = \hat{h}_{01}(t) \exp[(0.286) \text{Treatment}]$

$g = 2: Z_1^* = 1, Z_2^* = Z_3^* = Z_4^* = 0$   
(Large cell type and PSbin = 0)

Nonzero product terms	Coefficients
Age $Z_1^* = \text{Age}$	$\beta_{21}$
tr $Z_1^* = \text{Treatment}$	$\beta_{11}$

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  $\beta_{21}$  and  $\beta_{11}$ , respectively.

**EXAMPLE (continued)**

$$h_2(t, \mathbf{X}) = h_{02}(t) \exp[(\beta_1 + \beta_{11})\text{Treatment} + (\beta_2 + \beta_{21})\text{Age}]$$

$$\hat{\beta}_1 = 0.286, \hat{\beta}_2 = 0.000$$

$$\hat{\beta}_{11} = 2.351, \hat{\beta}_{21} = 0.078$$

Hazard functions for interaction model:

$$g = 1: (Z_1^* = Z_2^* = Z_3^* = Z_4^* = 0):$$

$$\hat{h}_1(t, \mathbf{X}) = \hat{h}_{01}(t) \exp[(0.286)\text{Treatment}]$$

$$g = 2: (Z_1^* = 1, Z_2^* = Z_3^* = Z_4^* = 0):$$

$$\hat{h}_2(t, \mathbf{X}) = \hat{h}_{02}(t) \exp[(2.637)\text{Treatment} + (0.078)\text{Age}]$$

$$g = 3: (Z_2^* = 1, Z_1^* = Z_3^* = Z_4^* = 0):$$

$$\hat{h}_3(t, \mathbf{X}) = \hat{h}_{03}(t) \exp[(-0.872)\text{Treatment} + (-0.047)\text{Age}]$$

$$g = 4: (Z_3^* = 1, Z_1^* = Z_2^* = Z_4^* = 0):$$

$$\hat{h}_4(t, \mathbf{X}) = \hat{h}_{04}(t) \exp[(0.868)\text{Treatment} + (-0.059)\text{Age}]$$

$$g = 5: (Z_1^* = Z_2^* = Z_3^* = 0, Z_4^* = 1):$$

$$\hat{h}_5(t, \mathbf{X}) = \hat{h}_{05}(t) \exp[(-0.747)\text{Treatment} + (0.051)\text{Age}]$$

$$g = 6: (Z_1^* = 1, Z_2^* = Z_3^* = 0, Z_4^* = 1):$$

$$\hat{h}_6(t, \mathbf{X}) = \hat{h}_{06}(t) \exp[(0.810)\text{Treatment} + (-0.038)\text{Age}]$$

$$g = 7: (Z_2^* = 1, Z_1^* = Z_3^* = 0, Z_4^* = 1):$$

$$\hat{h}_7(t, \mathbf{X}) = \hat{h}_{07}(t) \exp[(0.880)\text{Treatment} + (-0.041)\text{Age}]$$

$$g = 8: (Z_3^* = 1, Z_1^* = Z_2^* = 0, Z_4^* = 1):$$

$$\hat{h}_8(t, \mathbf{X}) = \hat{h}_{08}(t) \exp[(0.297)\text{Treatment} + (0.033)\text{Age}]$$

LR test to compare no-interaction model with interaction model:

$H_0$ : no-interaction model acceptable, i.e.,

Treatment:  $\beta_{11} = \beta_{12} = \dots = \beta_{17} = 0$

and Age:  $\beta_{21} = \beta_{22} = \dots = \beta_{27} = 0$

$$14 \text{ coefficients} \Rightarrow \text{df} = 14$$

$$LR = -2 \ln L_R - (2 \ln L_F)$$

$R$  = reduced (no-interaction) model

$F$  = full (interaction) model

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.

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.

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.

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.

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.

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)**

$LR \sim \chi^2_{14df}$  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.

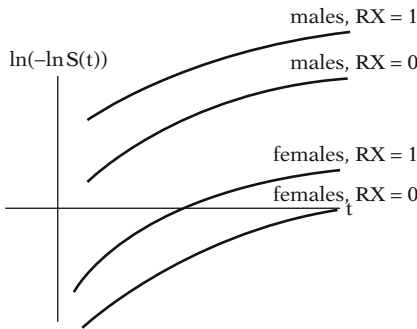
Thus, under the null hypothesis, the  $LR$  statistic is approximately chi-square with 14 degrees of freedom.

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**

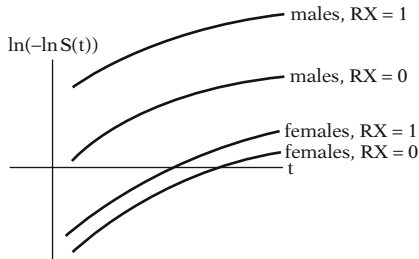
a.  $h(t) = h_0(t)\exp(\beta_1RX + \beta_2SEX)$   
 $\ln(-\ln S(t)) = \ln(-\ln S_0(t)) + \beta_1RX + \beta_2SEX$



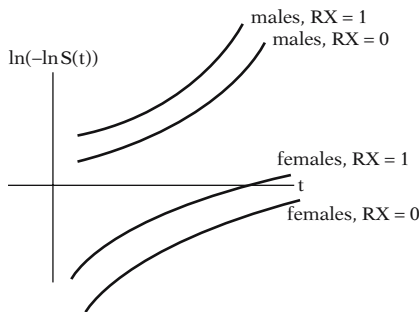
In this section we examine four log-log survival plots illustrating the assumptions underlying a stratified Cox model with or without interaction. Each of the four models considers two dichotomous predictors: treatment (coded  $RX = 1$  for placebo and  $RX = 0$  for new treatment) and  $SEX$  (coded 0 for females and 1 for males). The four models are as follows (see left).

- a.  $h_0(t)\exp(\beta_1RX + \beta_2SEX)$ . 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.

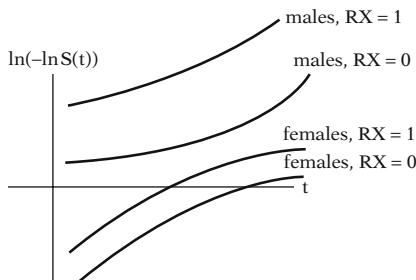
b.  $h(t) = h_0(t) \exp(\beta_1 RX + \beta_2 SEX + \beta_3 RX \times SEX)$   
 $\ln(-\ln S(t)) = \ln(-\ln S_0(t)) + \beta_1 RX + \beta_2 SEX + \beta_3 RX \times SEX$



c.  $h(t) = h_{0g}(t) \exp(\beta_1 RX)$   
 ( $g = 1$  for males,  $g = 0$  for females)  
 $\ln(-\ln S(t)) = \ln(-\ln S_{0g}(t)) + \beta_1 RX$



d.  $h(t) = h_{0g}(t) \exp(\beta_1 RX + \beta_2 RX \times SEX)$   
 ( $g = 1$  for males,  $g = 0$  for females)  
 $\ln(-\ln S(t)) = \ln(-\ln S_{0g}(t)) + \beta_1 RX + \beta_2 RX \times SEX$



b.  $h(t) = h_0(t) \exp(\beta_1 RX + \beta_2 SEX + \beta_3 RX \times SEX)$ . 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.  $h(t) = h_{0g}(t) \exp(\beta_1 RX)$ , where  $g = 1$  for males,  $g = 0$  for females. 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.  $h(t) = h_{0g}(t) \exp(\beta_1 RX + \beta_2 RX \times SEX)$ , where  $g = 1$  for males,  $g = 0$  for females. 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. Summary

Stratified Cox (SC) model:

- stratification of predictors not satisfying PH assumption
- includes predictors satisfying PH
- does not include stratified variables

### Computer Results

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 stratified Cox model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p]$$

$g = 1, 2, \dots, k^*$ , strata defined from  $Z^*$

$Z^*$  has  $k^*$  categories

$X_1, X_2, \dots, X_p$  satisfy PH

Stratification variable  $Z^*$ :

- identify  $Z_1, Z_2, \dots, Z_k$  not satisfying PH
- categorize each  $Z$
- form combinations of categories (strata)
- each combination is a stratum of  $Z^*$

We now summarize the most important features of the stratified Cox (SC) model described in this presentation.

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.

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.

The general hazard function form for the stratified Cox model is shown here. This formula contains a subscript  $g$  that indicates the  $g$ th 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.

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  $\beta_1, \beta_2, \dots, \beta_p$  for each  $g$ , i.e.,  $Z^*$  does not interact with the  $X$ 's.

Different baselines  $\left\{ \begin{array}{l} h_{01}(t) \Rightarrow \hat{S}_1(t) \\ h_{02}(t) \Rightarrow \hat{S}_2(t) \\ \vdots \\ h_{0k}(t) \Rightarrow \hat{S}_{k^*}(t) \end{array} \right\}$  Different survival curves

$\widehat{HR}$  same for each stratum

(Partial) likelihood function:

$$L = L_1 \times L_2 \times \dots \times L_{k^*}$$

Stratified Cox model allowing interaction:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} X_1 + \beta_{2g} X_2 + \dots + \beta_{pg} X_p]$$

$g = 1, 2, \dots, k^*$ , strata defined from  $Z^*$ .

Alternative stratified Cox interaction model:

- uses product terms involving  $Z^*$
- define  $k^* - 1$  dummy variables from  $Z^*$
- products of the form  $Z_i^* \times X_j$

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

$$LR \sim \chi_{p(k^*-1)\text{df}}^2 \text{ under } H_0: \text{no interaction}$$

The above model is designated as a “no-interaction” model because the  $\beta$ '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.

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.

However, because the coefficients of the  $X$ 's are the same for each stratum, estimates of hazard ratios are the same for each stratum.

Regression coefficients in the SC model are estimated by maximizing a partial likelihood function that is obtained by multiplying likelihood functions for each stratum.

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  $\beta$  coefficient has a subscript  $g$ .

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.

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 176)

- 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 176–180)

- 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  $\beta_1$  and  $\beta_2$  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 180–181)

- A. 
$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p],$$

$$g = 1, 2, \dots, k^*$$

where the strata are defined from the stratification variable  $Z^*$ .

- B.  $Z^*$  defined from  $Z_1, Z_2, \dots, Z_k$  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  $\beta_1, \beta_2, \dots, \beta_p$  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  $L = L_1 \times L_2 \times \dots \times L_{k^*}$ , where  $L_i$  is likelihood for  $i$ th stratum.

#### IV. The No-Interaction Assumption and How to Test It (pages 182–188)

- A. Assumes same coefficients  $\beta_1, \beta_2, \dots, \beta_p$  for each  $g$ .
- B. Interaction model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g}X_1 + \beta_{2g}X_2 + \dots + \beta_{pg}X_p],$$

$g = 1, 2, \dots, k^*$  strata defined from  $Z^*$ .

- C. Alternative stratified Cox interaction model:

- uses product terms involving  $Z^*$
- define  $k^* - 1$  dummy variables  $Z_1^*, Z_2^*, \dots, Z_{k^*-1}^*$  from  $Z^*$
- products of the form  $Z_i^* \times X_j$ , where  $i = 1, \dots, k^* - 1; j = 1, \dots, p$
- hazard function:  $g = 1, 2, \dots, k^*$  strata defined from  $Z^*$

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 X_1 + \dots + \beta_p X_p + \beta_{11}(Z_1^* \times X_1) + \dots + \beta_{p1}(Z_1^* \times X_p) + \beta_{12}(Z_2^* \times X_1) + \dots + \beta_{p2}(Z_2^* \times X_p) + \dots + \beta_{1,k^*-1}(Z_{k^*-1}^* \times X_1) + \dots + \beta_{p,k^*-1}(Z_{k^*-1}^* \times X_p)]$$

- 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
- $LR \sim \chi_{p(k^*-1)}^2$  under  $H_0$ : no interaction, i.e.,  $\beta_{11} = \beta_{21} = \dots = \beta_{p,k^*-1} = 0$

#### V. A Second Example Involving Several Stratification Variables (pages 188–193)

- 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:  
$$h_g(t, \mathbf{X}) = h_{0g}(t) \mathbf{exp}[\beta_1 \text{Treatment} + \beta_2 \text{Age}],$$
  
 $g = 1, 2, \dots, 8$
- H. Hazard function for interaction model:  
$$h_g(t, \mathbf{X}) = h_{0g}(t) \mathbf{exp}[\beta_{1g} \text{Treatment} + \beta_{2g} \text{Age}],$$
  
 $g = 1, 2, \dots, 8$
- I. Alternative version of interaction model:  
$$h_g(t, \mathbf{X}) = h_{0g}(t) \mathbf{exp}[\beta_1 \text{Treatment} + \beta_2 \text{Age}$$
  
 $+ \beta_{11}(Z_1^* \times \text{Treatment}) + \dots + \beta_{17}(Z_7^* \times \text{Treatment})$   
 $+ \beta_{21}(Z_1^* \times \text{Age}) + \dots + \beta_{27}(Z_7^* \times \text{Age})],$   
 $g = 1, 2, \dots, 8$   
where  $Z_1^* = \text{large cell (binary)}$ ,  $Z_2^* = \text{adeno cell (binary)}$ ,  $Z_3^* = \text{small cell (binary)}$ ,  $Z_4^* = \text{PSbin (binary)}$ ,  $Z_5^* = Z_1^* \times Z_4^*$ ,  $Z_6^* = Z_2^* \times Z_4^*$ ,  $Z_7^* = Z_3^* \times Z_4^*$
- 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:  $\beta_{11} = \beta_{12} = \dots = \beta_{17} = 0$   
and  $\beta_{21} = \beta_{22} = \dots = \beta_{27} = 0$
  - $LR \sim \chi^2_{14 \text{ df}}$  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 193–194)

Comparison of log–log survival curves

1. Describe interaction of Rx and Sex.
2. Describe violation of PH assumption for Sex.

## VII. Summary (pages 195–196)

## Practice Exercises

The following questions derive from the dataset **vets.dat** concerning the Veteran's Administration Lung Cancer Trial that we previously considered in the presentation on the stratified Cox model. 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]		<i>P(PH)</i>
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]		<i>P(PH)</i>
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  $Z_1 = \text{"small cell"}$  and  $Z_2 = \text{"performance status."}$  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.

- 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?
- 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.
- 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.
- 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 print-out 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 1.1t:

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 <sub>1</sub> <sup>*</sup>	-0.029	0.024	0.234	0.971	0.926	1.019
AgeZ <sub>1</sub> <sup>*</sup>	-0.055	0.037	0.135	0.946	0.880	1.018
PTZ <sub>1</sub> <sup>*</sup>	0.043	0.075	0.564	1.044	0.901	1.211
DDZ <sub>2</sub> <sup>*</sup>	0.025	0.032	0.425	1.026	0.964	1.092
AgeZ <sub>2</sub> <sup>*</sup>	0.001	0.024	0.956	1.001	0.956	1.049
PTZ <sub>2</sub> <sup>*</sup>	-0.078	0.054	0.152	0.925	0.831	1.029
DDZ <sub>1</sub> Z <sub>2</sub> <sup>*</sup>	-0.071	0.059	0.225	0.931	0.830	1.045
AgeZ <sub>1</sub> Z <sub>2</sub> <sup>*</sup>	0.084	0.049	0.084	1.088	0.989	1.196
PTZ <sub>1</sub> Z <sub>2</sub> <sup>*</sup>	-0.005	0.117	0.963	0.995	0.791	1.250
trZ <sub>1</sub> <sup>*</sup>	0.560	0.732	0.444	1.751	0.417	7.351
trZ <sub>2</sub> <sup>*</sup>	-0.591	0.523	0.258	0.554	0.199	1.543
trZ <sub>1</sub> Z <sub>2</sub> <sup>*</sup>	-0.324	0.942	0.731	0.723	0.114	4.583

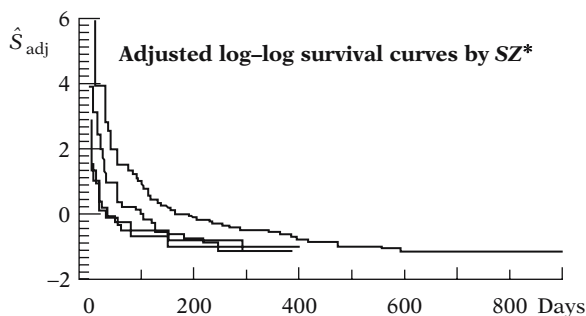
No. of subjects = 137

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]	<i>P(PH)</i>
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?

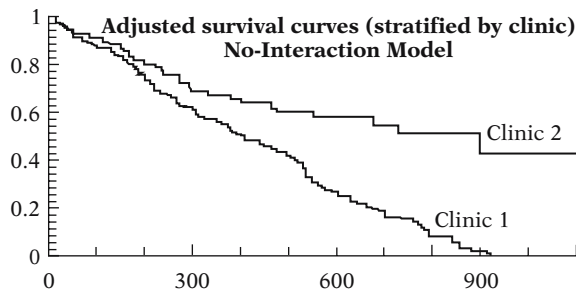
2. The following printout was obtained from fitting a SC PH model to these data, where the variable being stratified is clinic:

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  $\text{clinpr} = \text{clinic} \times \text{prison}$  and  $\text{clindos} = \text{clinic} \times \text{dose}$ .

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

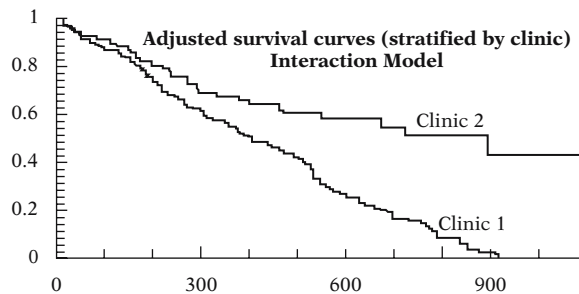
Stratified by clinic

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 curves 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 or 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.  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \text{Treatment} + \beta_2 DD + \beta_3 \text{Age} + \beta_4 PT]$ ,  $g = 1, \dots, 4$ , 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:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} \text{Treatment} + \beta_2 DD + \beta_3 \text{Age} + \beta_4 PT]$ ,  $g = 1, \dots, 4$ .  
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  $g = 1, \dots, 4$ .

6. Version 1:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} \text{Treatment} + \beta_{2g} DD + \beta_{3g} \text{Age} + \beta_{4g} PT]$ ,  $g = 1, \dots, 4$ .

Version 2:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \text{Treatment} + \beta_2 DD + \beta_3 \text{Age} + \beta_4 PT + \beta_5(Z_1^* \times \text{Treatment}) + \beta_6(Z_1^* \times DD) + \beta_7(Z_1^* \times \text{Age}) + \beta_8(Z_1^* \times PT) + \beta_9(Z_2^* \times \text{Treatment}) + \beta_{10}(Z_2^* \times DD) + \beta_{11}(Z_2^* \times \text{Age}) + \beta_{12}(Z_2^* \times PT) + \beta_{13}(Z_1^* \times Z_2^* \times \text{Treatment}) + \beta_{14}(Z_1^* \times Z_2^* \times DD) + \beta_{15}(Z_1^* \times Z_2^* \times \text{Age}) + \beta_{16}(Z_1^* \times Z_2^* \times PT)]$ ,  
 $g = 1, \dots, 4$ .

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.

$$H_0: \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = \beta_{12} \\ = \beta_{13} = \beta_{14} = \beta_{15} = \beta_{16} = 0 \text{ from version 2.}$$

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:

$$g = 1; Z_1^* = Z_2^* = 0:$$

$$\hat{h}_1(t, \mathbf{X}) = \hat{h}_{01}(t) \exp[(0.381)\text{Treatment} + (0.015)DD \\ + (0.000)\text{Age} + (0.023)PT]$$

$$g = 2; Z_1^* = 1, Z_2^* = 0:$$

$$\hat{h}_2(t, \mathbf{X}) = \hat{h}_{02}(t) \exp[(0.941)\text{Treatment} + (-0.014)DD \\ + (-0.055)\text{Age} + (0.066)PT]$$

$$g = 3; Z_1^* = 0, Z_2^* = 1:$$

$$\hat{h}_3(t, \mathbf{X}) = \hat{h}_{03}(t) \exp[(-0.210)\text{Treatment} + (0.040)DD \\ + (0.001)\text{Age} + (-0.055)PT]$$

$$g = 4; Z_1^* = 1, Z_2^* = 1:$$

$$\hat{h}_4(t, \mathbf{X}) = \hat{h}_{04}(t) \exp[(0.026)\text{Treatment} + (-0.060)DD \\ + (0.030)\text{Age} + (-0.017)PT]$$

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 Propor- tional Hazards Model for Time- Dependent Variables**

## 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 214)**
- II. **Review of the Cox PH Model (pages 214–216)**
- III. **Definition and Examples of Time-Dependent Variables (pages 216–219)**
- IV. **The Extended Cox Model for Time-Dependent Variables (pages 219–221)**
- V. **The Hazard Ratio Formula for the Extended Cox Model (pages 221–223)**
- VI. **Assessing Time-Independent Variables That Do Not Satisfy the PH Assumption (pages 224–229)**
- VII. **An Application of the Extended Cox Model to an Epidemiologic Study on the Treatment of Heroin Addiction (pages 230–234)**
- VIII. **An Application of the Extended Cox Model to the Analysis of the Stanford Heart Transplant Data (pages 235–239)**
- IX. **The Extended Cox Likelihood (pages 239–242)**
- X. **Summary (pages 242–245)**

## Objectives

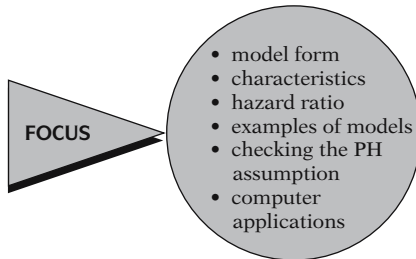
Upon completing the chapter, the learner should be able to:

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 the heaviside function(s), for a model involving heaviside function(s).
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

$$h(t, \mathbf{X}) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i \right]$$

$\mathbf{X} = (X_1, X_2, \dots, X_p)$   
Explanatory/predictor variables

$$h_0(t) \times \exp \left[ \sum_{i=1}^p \beta_i X_i \right]$$

Baseline hazard	Exponential
Involves $t$ but not $X$ 's	Involves $X$ 's but not $t$ ( $X$ 's are time-independent)

The 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.

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)

Hazard ratio formula:

$$\widehat{HR} = \exp \left[ \sum_{i=1}^p \hat{\beta}_i (X_i^* - X_i) \right]$$

where  $\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*)$  and  $\mathbf{X} = (X_1, X_2, \dots, X_p)$  denote the two sets of  $X$ 's.

PH assumption:

$$\frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \hat{\theta} \text{ (a constant over } t\text{)}$$

$$\text{i.e., } \hat{h}(t, \mathbf{X}^*) = \hat{\theta} \hat{h}(t, \mathbf{X})$$

Hazards cross  $\Rightarrow$  PH not met

Hazards don't cross  $\nRightarrow$  PH met

Three approaches:

- graphical
- time-dependent variables
- goodness-of-fit test

Time-dependent covariates:

Extend Cox model: add product term(s) involving some function of time

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.

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  $\mathbf{X}$ .

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.

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.

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.

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 gender, a Cox model might be extended to include the variable  $\text{sex} \times t$  in addition to  $\text{sex}$ . If the coefficient of the product term turns out to be non-significant, we can conclude that the PH assumption is satisfied for  $\text{sex}$  provided that the variable  $\text{sex} \times t$  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:

Time-dependent	Time-independent
Value of variable differs over time	Value of variable is constant over time
Example: Race $\times t$	Race

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.

As a simple example, the variable RACE is a time-independent variable, whereas the variable RACE  $\times$  time is a time-dependent variable.

**EXAMPLES OF DEFINED VARIABLES**

**Defined** variable:  $RACE \times t$

Time-independent

$$Race = 1 \Rightarrow Race \times t = t$$

$$Race = 0 \Rightarrow Race \times t = 0 \text{ (at any } t\text{)}$$

$$E \times (\log t - 3)$$

Function of  $t$

[ $E$  denotes a (0,1) exposure variable].

$$E \times g(t) \text{ where } g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

Heaviside function

$$t \geq t_0: E \times g(t) = E$$

$$t < t_0: E \times g(t) = 0$$

Heaviside functions used when PH assumptions not met.

**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.

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—that is,  $\log t - 3$ —rather than time 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.

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.

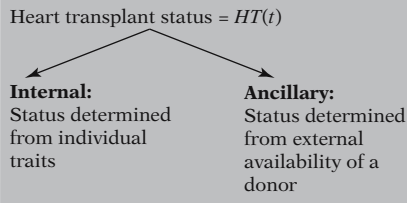
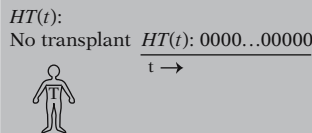
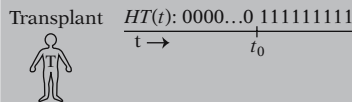
**EXAMPLES OF ANCILLARY VARIABLES**

Air pollution index at time  $t$ ;  $EMP(t)$

**ANOTHER EXAMPLE**

Heart transplant status at time  $t$ :

$$HT(t) = \begin{cases} 1 & \text{if received transplant at some time } t_0 \leq t \\ 0 & \text{if did not receive transplant by time } t \end{cases}$$



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

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t) \right]$$

$$\mathbf{X}(t) = \underbrace{(X_1, X_2, \dots, X_{p_1})}_{\text{Time-independent}}, \underbrace{(X_1(t), X_2(t), \dots, X_{p_2}(t))}_{\text{Time-dependent}}$$

**EXAMPLE**

$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta(E \times t)],$   
 $p_1 = 1, p_2 = 1,$   
 $\mathbf{X}(t) = (X_1 = E, X_1(t) = E \times t)$

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 hazards 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  $\mathbf{X}(t)$ .

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)	}	Computer Appendix
SAS (PHREG)		
SPSS (COXREG)		

Computer packages that include programs for fitting the extended Cox model include Stata, SAS, and SPSS. See the Computer Appendix at the end of this text for a comparison of the Stata, SAS, and SPSS 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:**

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t - L_j) \right]$$

$X_j(t - L_j)$  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.

$$\widehat{HR}(t) = \frac{\hat{h}(t, \mathbf{X}^*(t))}{\hat{h}(t, \mathbf{X}(t))} = \exp \left[ \sum_{i=1}^{p_1} \hat{\beta}_i [X_i^* - X_i] + \sum_{j=1}^{p_2} \delta_j [X_j^*(t) - X_j(t)] \right]$$

Two sets of predictors:

$$\mathbf{X}^*(t) = (X_1^*, X_2^*, \dots, X_{p_1}^*, X_1^*(t), X_2^*(t), \dots, X_{p_2}^*(t))$$

$$\mathbf{X}(t) = (X_1, X_2, \dots, X_{p_1}, X_1(t), X_2(t), \dots, X_{p_2}(t))$$

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.

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  $\mathbf{X}(t)$ .

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**

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta(E \times t)]$$

$$E = \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if unexposed} \end{cases}$$

$$\mathbf{X}^*(t) = (E = 1, E \times t = t)$$

$$\mathbf{X}(t) = (E = 0, E \times t = 0)$$

$$\widehat{HR}(t) = \frac{\hat{h}(t, E = 1)}{\hat{h}(t, E = 0)}$$

$$= \exp[\hat{\beta}(1 - 0) + \hat{\delta}((1 \times t) - (0 \times t))]$$


$$= \exp[\hat{\beta} + \hat{\delta}t]$$

$\hat{\delta} > 0 \Rightarrow \widehat{HR}(t) \uparrow$  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  $\beta$  “hat” plus  $\delta$  “hat” times  $t$ . This formula says that the hazard ratio is a function of time; in particular, if  $\delta$  “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.

$$\widehat{HR}(t) = \exp \left[ \sum_{i=1}^{p_1} \hat{\beta}_i [X_i^* - X_i] + \sum_{j=1}^{p_2} \hat{\delta}_j [X_j^*(t) - X_j(t)] \right]$$


  
 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  $\delta_j$  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  $\delta_j$  “hat” of the difference in values of the  $j$ th 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**

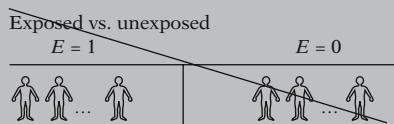
$E(t)$  = chemical exposure status at time  $t$  (weekly)  
 $= \begin{cases} 0 & \text{if unexposed at time } t \\ 1 & \text{if exposed at time } t \end{cases}$

 :

$E(t)$	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

$\delta$  represents the overall effect of  $E(t)$ .

$$\begin{aligned} \widehat{HR}(t) &= \frac{\hat{h}(t, E(t) = 1)}{\hat{h}(t, E(t) = 0)} \\ &= \exp[\delta[1 - 0]] \\ &= e^{\delta}, \text{ a fixed number} \end{aligned}$$

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 five-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,  $\delta$ , corresponding to the one variable in the model. Thus,  $\delta$  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  $\delta$  "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.

Three methods for checking PH assumption:

1. graphical
2. extended Cox model
3. GOF test

Cox PH model for  $p$  time-independent  $X$ 's:

$$h(t, \mathbf{X}) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i \right]$$

### Extended Cox model:

Add product terms of the form:

$$X_i \times g_i(t)$$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \delta_i X_i g_i(t) \right]$$

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.

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.

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.

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  $i$ th time-independent variable is denoted as  $X_i$ , then we can define the  $i$ th product term as  $X_i \times g_i(t)$  where  $g_i(t)$  is some function of time for the  $i$ th variable.

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**

$g_i(t) = t \Rightarrow X_i g_i(t) = X_i \times t$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \delta_i (X_i \times t) \right]$$

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  $X_i \times t$ . The extended Cox model in this case takes the form shown here.

**EXAMPLE 3: one variable at a time**

$X_L$  only  $\Rightarrow \begin{cases} g_L(t) = t, \\ g_i(t) = 0 \text{ for other } i \end{cases}$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i + \delta_L (X_L \times t) \right]$$

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  $X_L \times t$ , as shown here.

**EXAMPLE 4**

$g_i(t) = \ln t \Rightarrow X_i g_i(t) = X_i \times \ln t$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \delta_i (X_i \times \ln t) \right]$$

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**

$g_i(t) = \begin{cases} 0 & \text{if } t \geq t_0 \\ 1 & \text{if } t < t_0 \end{cases}$

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.

$H_0 : \delta_1 = \delta_2 = \dots = \delta_p = 0$

To check the PH assumption using a statistical test, we consider the null hypothesis that all the  $\delta$  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:

$$h(t, \mathbf{X}) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i \right]$$

$$\begin{aligned} LR &= -2 \ln L_{\text{PH model}} \\ &\quad - (-2 \ln L_{\text{ext. Cox model}}) \\ &\sim \chi_p^2 \text{ under } H_0 \end{aligned}$$

#### 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:  
 $h(t, \mathbf{X}) = h_0(t) \exp[\beta E]$

$LR = -2 \ln L_R - (-2 \ln L_F)$   
 $\sim \chi^2$  with 1 df under  $H_0$

$F = \text{full (extended), } R = \text{reduced (PH)}$

SAS: **PHREG** fits both PH and extended Cox models.

Stata: **Stcox** fits both PH and extended Cox models.

If PH test significant: Extended Cox model is preferred; HR is time-dependent.

Under this null hypothesis, the model reduces to the PH model.

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  $H_0$ .

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.

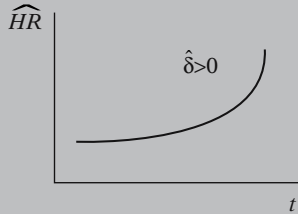
Note that to carry out the computations for this test, two different types of models, a PH model and an extended Cox model, need to be fit.

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**

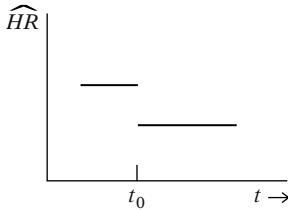
$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta(E \times t)]$$

$$\widehat{HR} = \exp[\hat{\beta} + \hat{\delta}t]$$



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  $\beta$  “hat” plus  $\delta$  “hat” times  $t$ . Thus, depending on whether  $\delta$  “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  $\delta$  “hat” is positive.

**Heaviside function:**



$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta E g(t)]$$

$$\begin{aligned} t \geq t_0: & \quad g(t) = 1 \Rightarrow E \times g(t) = E \\ & \quad h(t, \mathbf{X}) = h_0(t) \exp[(\beta + \delta)E] \\ & \quad \widehat{HR} = \exp[\hat{\beta} + \hat{\delta}] \end{aligned}$$

$$\begin{aligned} t < t_0: & \quad g(t) = 0 \Rightarrow E \times g(t) = 0 \\ & \quad h(t, \mathbf{X}) = h_0(t) \exp[\beta E] \\ & \quad \widehat{HR} = \exp[\hat{\beta}] \end{aligned}$$

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.

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.

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  $\beta$  “hat” plus  $\delta$  “hat.”

If  $t < t_0$ ,  $g(t) = 0$ , the corresponding hazard ratio is simplified to  $e$  to the  $\beta$  “hat.”

A single heaviside function in the model

$$\begin{aligned} h(t, \mathbf{X}) & \\ & = h_0(t) \exp[\beta E + \delta(E \times g(t))] \end{aligned}$$

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.

yields two hazard ratios:

$$\begin{aligned} t \geq t_0: & \quad \widehat{HR} = \exp(\hat{\beta} + \hat{\delta}) \\ t < t_0: & \quad \widehat{HR} = \exp(\hat{\beta}) \end{aligned}$$

Alternative model with two heaviside functions:

$$h(t, \mathbf{X}) = h_0(t) \exp[\delta_1(E \times g_1(t)) + \delta_2(E \times g_2(t))]$$

$$g_1(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{if } t \geq t_0 \end{cases}$$

**Note:** Main effect for  $E$  not in model.

Two  $HR$ 's from the alternative model:

$$\begin{aligned} t \geq t_0: g_1(t) &= 1, g_2(t) = 0 \\ h(t, \mathbf{X}) &= h_0(t) \exp[\delta_1(E \times 1) + \delta_2(E \times 0)] \\ &= h_0(t) \exp[\delta_1 E] \end{aligned}$$

so that  $\widehat{HR} = \exp(\hat{\delta}_1)$

$$\begin{aligned} t < t_0: g_1(t) &= 0, g_2(t) = 1 \\ h(t, \mathbf{X}) &= h_0(t) \exp[\delta_1(E \times 0) + \delta_2(E \times 1)] \\ &= h_0(t) \exp[\delta_2 E] \end{aligned}$$

so that  $\widehat{HR} = \exp(\hat{\delta}_2)$

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:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta(E \times g(t))]$$

$$t \geq t_0: \widehat{HR} = \exp(\hat{\delta}_1) = \exp(\hat{\beta} + \hat{\delta})$$

$$t < t_0: \widehat{HR} = \exp(\hat{\delta}_2) = \exp(\hat{\beta})$$

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.

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  $\delta_1$  "hat."

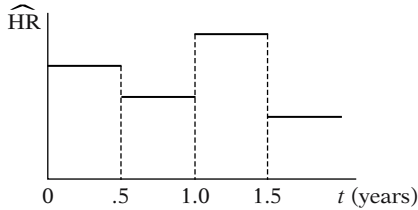
When  $t < t_0$ , the value of  $g_1(t)$  is 0 and the value of  $g_2(t)$  is 1. Then, the exponential part of the model becomes  $\delta_2 \times E$ , and the corresponding hazard ratio formula is  $e$  to the  $\delta_2$  "hat."

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,  $\delta_1$  "hat" in the alternative model equals  $\beta$  "hat" plus  $\delta$  "hat" in the original model (containing one heaviside function), and  $\delta_2$  "hat" in the alternative model equals  $\beta$  "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

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
- $E \times g_1(t), E \times g_2(t), E \times g_3(t), E \times g_4(t)$

1	2	3	4
0	0.5	1.0	1.5
$t(\text{years})$			

$h(t, \mathbf{X}(t))$

$$= h_0(t) \exp[\delta_1 E g_1(t) + \delta_2 E g_2(t) + \delta_3 E g_3(t) + \delta_4 E g_4(t)]$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } 0 \leq t < 0.5 \text{ year} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } 0.5 \text{ year} \leq t < 1.0 \text{ year} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_3(t) = \begin{cases} 1 & \text{if } 1.0 \text{ year} \leq t < 1.5 \text{ years} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_4(t) = \begin{cases} 1 & \text{if } t \geq 1.5 \text{ years} \\ 0 & \text{if otherwise} \end{cases}$$

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.

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.

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.

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:

$$4 \widehat{HR}'s \begin{cases} 0 \leq t < 0.5: \widehat{HR} = \exp(\hat{\delta}_1) \\ 0.5 \leq t < 1.0: \widehat{HR} = \exp(\hat{\delta}_2) \\ 1.0 \leq t < 1.5: \widehat{HR} = \exp(\hat{\delta}_3) \\ t \geq 1.5: \widehat{HR} = \exp(\hat{\delta}_4) \end{cases}$$

## VII. An Application of the Extended Cox Model to An Epidemiologic Study on the Treatment of Heroin Addiction

### EXAMPLE

#### 1991 Australian study (Capehorn 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)

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) ↘ 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

$P(PH)$  for the variables prison and dose are nonsignificant  $\Rightarrow$  remain in model

A 1991 Australian study by Capehorn 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.

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.

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)**

**Adjusted Survival Curves Stratified by Clinic**  
(prison and dose in the model)

**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

$$\text{clinic} = \begin{cases} 1 & \text{if clinic 1} \\ 0 & \text{if clinic 2} \end{cases}$$

*Note:*  
Previously clinic = 2 for clinic 2

$t \geq 365 \text{ days: } HR = \exp(\hat{\beta}_1 + \hat{\delta})$   
 $t < 365 \text{ days: } HR = \exp(\hat{\beta}_1)$

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 one year (i.e., 365 days), but the two curves diverge greatly after one 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)**

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta_1(\text{clinic})g_1(t) + \delta_2(\text{clinic})g_2(t)]$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } t < 365 \text{ days} \\ 0 & \text{if } t \geq 365 \text{ days} \end{cases}$$

and

$$g_2(t) = \begin{cases} 1 & \text{if } t \geq 365 \text{ days} \\ 0 & \text{if } t < 365 \text{ days} \end{cases}$$

$$t < 365 \text{ days: } \widehat{HR} = \exp(\hat{\delta}_1)$$

$$t \geq 365 \text{ days: } \widehat{HR} = \exp(\hat{\delta}_2)$$

	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 $\times$ $g_1$	0.460	0.255	0.072	1.583	0.960	2.611
Clinic $\times$ $g_2$	1.828	0.386	0.000	6.223	2.921	13.259

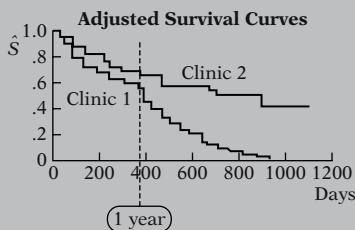
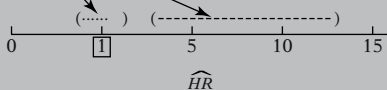
$$t < 365 \text{ days: } \widehat{HR} = e^{0.460} = 1.583$$

$$t \geq 365 \text{ days: } \widehat{HR} = e^{1.828} = 6.223$$

95% confidence intervals for clinic effect:

$t < 365$  days: (0.960, 2.611)

$t \geq 365$  days: (2.921, 13.259)



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  $\delta_1$  “hat” and  $e$  to the  $\delta_2$  “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  $\times$   $g_1$ ” and that the estimated hazard ratio of 6.223 is computed by exponentiating the estimated coefficient 1.828 of the product term “clinic  $\times$   $g_2$ ”.

Note also that the 95% confidence interval for the clinic effect prior to 365 days—that is, for the product term “clinic  $\times$   $g_1(t)$ ”—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  $\times$   $g_2$ ”—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 one year in contrast to a small difference in clinic survival times prior to one 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

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta(\text{clinic} \times t)]$$

$$\widehat{HR} = \exp(\hat{\beta}_1 + \hat{\delta}t)$$

$\widehat{HR}$  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

$$\widehat{HR} = \exp(\hat{\beta}_1 + 91\hat{\delta})$$

$t = 274$ :

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{clinic}) + \beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta(\text{clinic})(274)]$$

$$\widehat{HR} = \exp(\hat{\beta}_1 + 274\hat{\delta})$$

$t = 458.5$ :

$$\widehat{HR} = \exp(\hat{\beta}_1 + 458.5\hat{\delta})$$

$t = 639$ :

$$\widehat{HR} = \exp(\hat{\beta}_1 + 639\hat{\delta})$$

$t = 821.5$ :

$$\widehat{HR} = \exp(\hat{\beta}_1 + 821.5\hat{\delta})$$

$$\hat{\delta} > 0 \Rightarrow \widehat{HR} \uparrow \text{ as time } \uparrow$$

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 one 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 one 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  $\beta_1$  times the clinic variable plus  $\delta$  times the clinic variable times 91: the corresponding estimated hazard ratio for the clinic effect is then  $e$  to the power  $\beta_1$  "hat" plus  $\delta$  "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  $\delta$  times the clinic variable times 274: the corresponding hazard ratio for the clinic effect is then  $e$  to  $\beta_1$  "hat" plus 274  $\delta$  "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  $\delta$  "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

$$\widehat{\text{cov}}(\hat{\beta}_1, \hat{\delta}) = -0.000259 \text{ Log likelihood} = -667.642$$

$$\hat{\beta}_1 = -0.0183 \quad \hat{\delta} = 0.003$$

$\widehat{\text{HR}}$  depends on  $\hat{\beta}_1$  and  $\hat{\delta}$ .

$$t = 91.5: \widehat{\text{HR}} = \exp(\hat{\beta}_1 + \hat{\delta}t) = 1.292$$

$$t = 274: \widehat{\text{HR}} = \exp(\hat{\beta}_1 + \hat{\delta}t) = 2.233$$

$$t = 458.5: \widehat{\text{HR}} = \exp(\hat{\beta}_1 + \hat{\delta}t) = 3.862$$

$$t = 639: \widehat{\text{HR}} = \exp(\hat{\beta}_1 + \hat{\delta}t) = 6.677$$

$$t = 821.5: \widehat{\text{HR}} = \exp(\hat{\beta}_1 + \hat{\delta}t) = 11.544$$

$$\exp\left[\hat{\beta}_1 + \hat{\delta}t \pm 1.96\sqrt{\widehat{\text{Var}}(\hat{\beta}_1 + \hat{\delta}t)}\right]$$

$$\text{Var}(\hat{\beta}_1 + \hat{\delta}t) = s_{\hat{\beta}_1}^2 + t^2 s_{\hat{\delta}}^2 + 2t \widehat{\text{cov}}(\hat{\beta}_1, \hat{\delta})$$

$$\begin{matrix} \uparrow & \uparrow & \uparrow \\ (0.347)^2 & (0.001)^2 & (-0.000259) \end{matrix}$$

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  $\beta_1$  “hat” equals  $-0.0183$ , and the estimated coefficient  $\delta$  “hat” obtained for the product term equals  $0.003$ . For the model being fit, the hazard ratio depends on the values of both  $\beta_1$  “hat” and  $\delta$  “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  $\beta_1$  “hat” and  $\delta$  “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

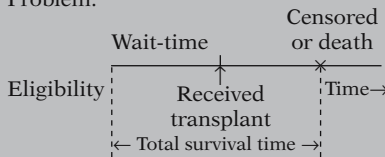
### EXAMPLE

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?

One approach:  
 Compare two separate groups: 65 transplants vs. 38 nontransplants

Problem:



**Note:** Wait-time contributes to survival time for nontransplants.

Covariates:

Tissue mismatch score } prognostic only  
 Age at transplant } for transplants

Age at eligibility: not considered prognostic for nontransplants

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 to the analysis of this 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.

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.

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

$$HT(t) = \begin{cases} 0 & \text{if did not receive transplant} \\ & \text{by time } t, \text{ i.e., if } t < \text{wait-} \\ & \text{time} \\ 1 & \text{if received transplant prior} \\ & \text{to time } t, \text{ i.e., if } t \geq \text{wait-} \\ & \text{time} \end{cases}$$

No transplant


 $HT(t)$   
 0000...00000  
 $t \rightarrow$ 

Transplant


 $HT(t)$   
 0000...0111111111  
 $t \rightarrow$ 

Time of transplant

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:

$$TMS(t) = \begin{cases} 0 & \text{if } t < \text{wait-time} \\ TMS & \text{if } t \geq \text{wait-time} \end{cases}$$

$$AGE(t) = \begin{cases} 0 & \text{if } t < \text{wait-time} \\ AGE & \text{if } t \geq \text{wait-time} \end{cases}$$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\delta_1 HT(t) + \delta_2 TMS(t) + \delta_3 AGE(t)]$$

Focus:

Assessing the effect of  $HT(t)$  adjusted for  $TMS(t)$  and  $AGE(t)$ .

**Note:**  $HT(t)$  does not satisfy PH assumption.

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

$$\widehat{HR} = e^{-3.1718} = 0.0417 = \frac{1}{23.98}$$

$$\widehat{HR} = \frac{\hat{h}(\text{transplants})}{\hat{h}(\text{nontransplants})} \approx \frac{1}{24} ?$$

Not appropriate!

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.

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.

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.

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.

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

Alternative interpretation:

At time  $t$ ,

$\hat{h}$  ("not yet received transplant")

$\approx 24 \hat{h}$  ("already received transplant")

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

$i$  denotes  $i$ th transplant patient

$\mathbf{X}^*(t) = (HT(t) = 1, TMS(t) = TMS_i, AGE(t) = AGE_i)$

$\mathbf{X}(t) = (HT(t) = 0, TMS(t) = 0, AGE(t) = 0)$

$$\begin{aligned}\widehat{HR}(t) &= \exp[\hat{\delta}_1(1 - 0) + \hat{\delta}_2(TMS_i - 0) \\ &\quad + \hat{\delta}_3(AGE_i - 0)] \\ &= \exp[\hat{\delta}_1 + \hat{\delta}_2 TMS_i + \hat{\delta}_3 AGE_i]\end{aligned}$$

$$= \exp[-3.1718 + 0.4442 TMS_i + 0.0552 AGE_i]$$

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 non-transplants.

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*.

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.

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.

The hazard ratio formula then reduces to  $e$  to the sum of  $\delta_1$  "hat" plus  $\delta_2$  "hat" times  $TMS_i$  plus  $\delta_3$  "hat" times  $AGE_i$ , where the  $\delta$  "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**

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

Cox PH model:  $h(t) = h_0(t)e^{\beta_1 SMOKE}$

*Cox PH Likelihood*

$$L = \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right] \times \left[ \frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1}} \right] \times \left[ \frac{h_0(t)e^{\beta_1}}{h_0(t)e^{\beta_1}} \right]$$

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.

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 whereas Gary and Harry were nonsmokers.

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  $t_j$  (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  $t_j$ , including the censored subject Harry. The numerator of each term is the hazard of the subject who got the event at  $t_j$ . The reader may wish to reread Section VIII of Chapter 3.

Cox extended model

$$h(t) = h_0(t)e^{\beta_1 \text{SMOKE} + \beta_2 \text{SMOKE} \times \text{TIME}}$$

Time-dependent covariate  
(its value changes over 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	$h_0(t)e^{\beta_1+8\beta_2}$

Cox extended model

$L =$

$$\left[ \frac{h_0(t)e^{\beta_1+2\beta_2}}{h_0(t)e^{\beta_1+2\beta_2} + h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1+2\beta_2}} \right] \times \left[ \frac{h_0(t)e^0}{h_0(t)e^0 + h_0(t)e^0 + h_0(t)e^{\beta_1+3\beta_2}} \right] \times \left[ \frac{h_0(t)e^{\beta_1+8\beta_2}}{h_0(t)e^{\beta_1+8\beta_2}} \right]$$

Likelihood is product of 3 terms:

$$L = L_1 \times L_2 \times L_3$$

Barry (t = 2)      Gary (t = 3)      Larry (t = 8)

SMOKE × TIME = 0 for nonsmokers

SMOKE × TIME changes over time for smokers

Larry's hazard changes over  $L_1, L_2, L_3$ .

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, a smoker, got the event at TIME = 8. However at TIME = 2, 3, and 8, the covariate SMOKE × TIME changes values, thus affecting 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.

The likelihood for the extended Cox model is constructed in a similar manner to that of the likelihood for the Cox PH model. The difference is that the expression for the subject's hazard is allowed to vary over time. The extended Cox likelihood for these data is shown on the left.

Just as with the Cox PH likelihood shown previously, 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 Gary got the event. Therefore, Harry's hazard is still in the denominator of  $L_1$  and  $L_2$ .

The inclusion of the time-varying covariate SMOKE × TIME does not change the expression for the hazard for the nonsmokers (Gary and Harry) because 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  $L_3$  (see dashed arrows above).

$h_0(t)$  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.

Incorrent coding of SMOKE × TIME

ID	TIME	STATUS	SMOKE	SMOKE × TIME
Barry	2	1	1	2
Gary	3	1	0	0
Harry	5	0	0	0
Larry	8	1	1	8

Coded as time-independent, not time-dependent

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 his survival time. In other words, SMOKE × TIME should not be coded like the typical interaction term. In fact, if SMOKE × TIME were 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<sub>1</sub>, L<sub>2</sub>, and L<sub>3</sub>.

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 were included in a Cox model it would not be surprising if the coefficient estimate were highly significant even if the PH assumption were not violated. It would be expected that a product term with each individual's survival time would predict the outcome (his survival time), but it would not be meaningful. Nevertheless, this is a common mistake.

Correctly coding SMOKE × TIME

- Time dependent
- Computer packages allow definition in the analytic procedure
- See Computer Appendix for details

To obtain a correctly defined SMOKE × TIME time-dependent variable, computer packages typically allow the variable to be defined within the analytic procedure. See Computer Appendix to see how time-dependent variables are defined in Stata, SAS, and SPSS.

Coding SMOKE × TIME as time-dependent

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

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: 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

Review Cox PH model.

Define time-dependent variable: defined, internal, ancillary.

### Extended Cox model:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t) \right]$$

$$\widehat{HR}(t) = \exp \left[ \sum_{i=1}^{p_1} \hat{\beta}_i [X_i^* - X_i] + \sum_{j=1}^{p_2} \hat{\delta}_j [X_j^*(t) - X_j(t)] \right]$$

Function of time

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.

Next, we gave the form of the “extended Cox model,” shown here again, which allows for time-dependent as well as time-independent variables.

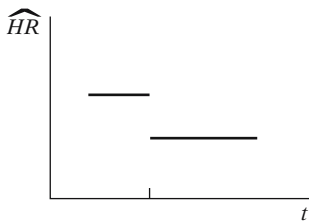
We then described various characteristics of this extended Cox model, including the formula for the hazard ratio. The latter formula is time-dependent so that the PH assumption is not satisfied.

### Model for assessing PH assumption:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp \left[ \sum_{i=1}^p \beta_i X_i + \sum_{i=1}^p \delta_i X_i g_i(t) \right]$$

Examples of  $g_i(t)$ :  
 $t$ ,  $\log t$ , heaviside function

### Heaviside functions:



$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta E g(t)]$$

where

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1 E g_1(t) + \beta_2 E g_2(t)]$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0 \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } t < t_0 \\ 0 & \text{if } t \geq t_0 \end{cases}$$

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.

The functions  $g_i(t)$  denote functions of time for the  $i$ th variable that are to be determined by the investigator. Examples of such functions are  $g_i(t) = t$ ,  $\log t$ , or a heaviside function.

The use of heaviside functions were described and illustrated. Such functions allow for the hazard ratio to be constant within different time intervals.

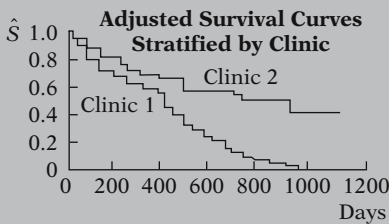
For two time intervals, the model can take either one of two equivalent forms as shown here. The first model contains a main effect of exposure and only one heaviside function. The second model contains two heaviside functions without a main effect of exposure. Both models yield two distinct and equivalent values for the hazard ratio.

#### 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)**

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \delta_1(\text{clinic})g_1(t) + \delta_2(\text{clinic})g_2(t)]$$

Heaviside functions

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_2(\text{prison}) + \beta_3(\text{dose}) + \beta_1(\text{clinic}) + \delta(\text{clinic} \times t)]$$

where

$T(t)$  1, 3, 5, 7, 9 in half-year intervals

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

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.

Next:

7. Parametric models

## Detailed Outline

- I. Preview** (page 214)  
**II. Review of the Cox PH Model** (pages 214–216)

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:

$$\mathbf{X}^* = (X_1^*, X_2^*, \dots, X_p^*) \text{ and } \mathbf{X} = (X_1, X_2, \dots, X_p):$$

$$\frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp \left[ \sum_{i=1}^p \beta_i (X_i^* - X_i) \right]$$

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:

- Graphical*: Compare ln–ln survival curves or observed versus predicted curves
- Time-dependent covariates*: Use product (i.e., interaction) terms of the form  $X \times g(t)$ .
- Goodness-of-fit test*: Use a large sample  $Z$  statistic.

E. Options when the PH assumption is not met:

- Use a stratified Cox procedure.
- 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 216–219)

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 219–221)

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  $\mathbf{X}(t) = (X_1, X_2, \dots, X_{p_1}, X_1(t), X_2(t), \dots, X_{p_2}(t))$  denotes the entire collection of predictors at time  $t$ ,  $X_i$  denotes the  $i$ th time-independent variable, and  $X_j(t)$  denotes the  $j$ th 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 221–223)

A. 
$$HR(t) = \exp \left[ \sum_{i=1}^{p_1} \hat{\beta}_i [X_i^* - X_i] + \sum_{j=1}^{p_2} \hat{\delta}_j [X_j^*(t) - X_j(t)] \right]$$

- 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 224–229)

- 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  $H_0: \delta_1 = \delta_2 = \dots = \delta_p = 0$
- 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:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta E + \delta E g(t)]$$

F. The model with two heaviside functions:

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\delta_1 E g_1(t) + \delta_2 E g_2(t)]$$

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:

$$t \geq t_0: \widehat{HR} = \exp(\hat{\beta} + \hat{\delta}) = \exp(\hat{\delta}_1)$$

$$t < t_0: \widehat{HR} = \exp(\hat{\beta}) = \exp(\hat{\delta}_2)$$

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  $E$ :

$$h(t, \mathbf{X}(t)) = h_0(t) \exp[\delta_1 E g_1(t) + \delta_2 E g_2(t) + \delta_3 E g_3(t) + \delta_4 E g_4(t)]$$

where

$$g_i(t) = \begin{cases} 1 & \text{if } t \text{ is within interval } i \\ 0 & \text{if otherwise} \end{cases}$$

## VII. An Application of the Extended Cox Model to an Epidemiologic Study on the Treatment of Heroin Addiction (pages 230–234)

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 one year of treatment.

C. Two extended Cox models were considered:

- Use heaviside functions to obtain two distinct hazard ratios, one for less than one year and the other for greater than one 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 235–239)

- 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. Extended Cox Likelihood** (pages 239–242)

- A. Review of PH likelihood (Chapter 3).
- B. Barry, Gary, Larry, example of Cox likelihood.

**X. Summary** (pages 242–245)

**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)

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	[95% Conf.					
Analysis time.t: survt	Coef.	Std. Err.	p >  z	Haz. Ratio	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 subjees = 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  $\times$  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 $\times$ 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 heavy-side 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

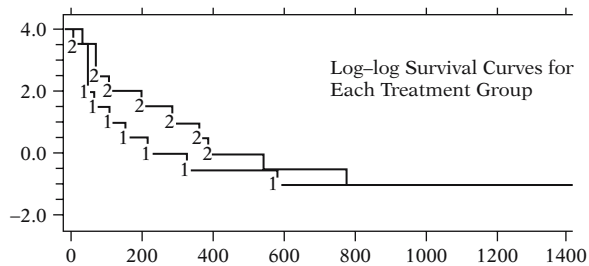
Log likelihood = -57.560 Stratified by sex

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). A listing of the dataset (called chemo) is given at the end of the presentation.

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.



Number at risk

	45	26	20	11	10	7	5	2
	45	40	25	17	10	7	6	2

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.

$$3. \quad h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{sex}) + \beta_2(\log \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex})g_1(t) + \delta_2(\log \text{WBC})g_2(t) + \delta_3(Rx)g_3(t)]$$

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.

$$6. \quad h(t, \mathbf{X}(t)) = h_0(t) \exp[\beta_1(\text{sex}) + \beta_2(\log \text{WBC}) + \beta_3(Rx) + \delta_1(\text{sex} \times t)]$$

7. The hazard ratio for the effect of Sex in each time interval, controlling for  $Rx$  and log WBC is given as follows:

$$\begin{aligned} t = 8 \text{ weeks} & \quad \widehat{HR} = \exp[\hat{\beta}_1 + 8\hat{\delta}_1] \\ t = 16 \text{ weeks} & \quad \widehat{HR} = \exp[\hat{\beta}_1 + 16\hat{\delta}_1] \end{aligned}$$

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**

---

## 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 260–262)**
- II. **Probability Density Function in Relation to the Hazard and Survival Function (pages 262–263)**
- III. **Exponential Example (pages 263–265)**
- IV. **Accelerated Failure Time Assumption (pages 266–268)**
- V. **Exponential Example Revisited (pages 268–272)**
- VI. **Weibull Example (pages 272–277)**
- VII. **Log-Logistic Example (pages 277–282)**
- VIII. **A More General Form of the AFT Model (pages 282–284)**
- IX. **Other Parametric Models (pages 284–286)**
- X. **The Parametric Likelihood (pages 286–289)**
- XI. **Interval-Censored Data (pages 289–294)**
- XII. **Frailty Models (pages 294–308)**
- XIII. **Summary (pages 309–312)**

## Objectives

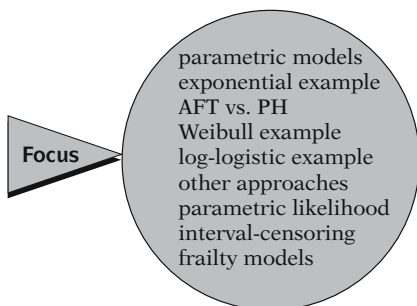
Upon completing this chapter, the learner should be able to:

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*

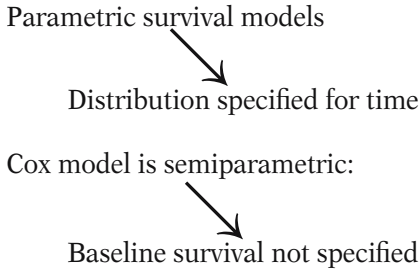
- Outcome assumed to follow some family of distributions
- Exact distribution is unknown if parameters are unknown
- Data used to estimate parameters
- Examples of parametric models:
  - Linear regression
  - Logistic regression
  - Poisson regression

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 three and another distribution is normal with a mean of seven, 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**, all of which are supported by SAS or Stata software.

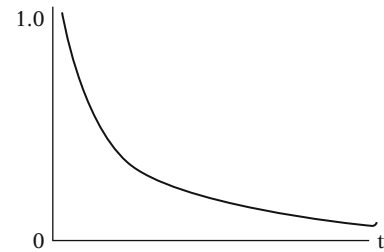


The Cox proportional hazards model, by contrast, is not a fully parametric model. Rather it is a **semi-parametric model** because even if the regression parameters (the betas) are known, the distribution of the outcome remains unknown. The baseline survival (or hazard) function is not specified in a Cox model.

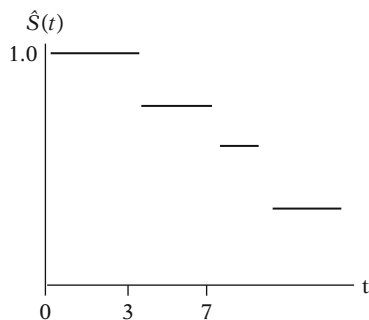
Cox model widely popular:

- 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, and SPSS 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 function known then  
Survival and hazard can be found

$$S(t) = P(T > t) = \int_t^{\infty} f(u)du$$

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

Survival in terms of hazard

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

Cumulative hazard:  $\int_0^t h(u)du$

$$f(t) = h(t)S(t)$$

*Key Point*

Specifying one of  $f(t)$ ,  $S(t)$ , or  $h(t)$  specifies all three 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).

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  $t$ .

Finally, the probability function can be expressed as the product of the hazard and the survival functions,  $f(t) = h(t)S(t)$ .

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	$\exp(-\lambda t)$	$\lambda$
Weibull	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$
Log-logistic	$\frac{1}{1 + \lambda t^p}$	$\frac{\lambda p t^{p-1}}{1 + \lambda t^p}$

$$f(t) = h(t)S(t)$$

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:

- $\lambda$  reparameterized for regression
- $p$  held fixed

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  $\lambda$ . The Weibull and log-logistic distributions have two parameters  $\lambda$  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.

Typically for parametric survival models, the parameter  $\lambda$  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  $\lambda$  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.

$$h(t) = \lambda = \exp(\beta_0 + \beta_1 \text{TRT})$$

$$\text{TRT} = 1: h(t) = \exp(\beta_0 + \beta_1)$$

$$\text{TRT} = 0: h(t) = \exp(\beta_0)$$

$$\text{HR}(\text{TRT} = 1 \text{ vs. } \text{TRT} = 0)$$

$$= \frac{\exp(\beta_0 + \beta_1)}{\exp(\beta_0)} = \exp(\beta_1)$$

Constant Hazards

⇒ Proportional Hazards

Proportional Hazards

⇒ Constant Hazards

Exponential Model—Hazards are constant

Cox PH Model—Hazards are proportional not necessarily constant

### Remission Data

Exponential regression  
log relative-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

Coefficient estimates obtained by MLE

↙  
asymptotically normal

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  $\lambda$  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.

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.

Output from running the exponential model is shown on the left. The model was run using Stata software (version 7.0). The parameter estimates are listed under the column called Coef. The parameter estimate for the coefficient of TRT ( $\beta_1$ ) 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).

The regression coefficients are estimated **using maximum likelihood estimation (MLE)**, and are **asymptotically normally** distributed.

$$\text{TRT} = 1: \hat{h}(t) = \exp(-2.159 \\ + (-1.527)) = 0.025$$

$$\text{TRT} = 0: \hat{h}(t) = \exp(-2.159) \\ = 0.115$$

$$\widehat{HR} (\text{TRT} = 1 \text{ vs. } 0) = \exp(-1.527) \\ = 0.22$$

$$95\% \text{ CI} = \exp[-1.527 \pm 1.96(0.398)] \\ = (0.10, 0.47)$$

Results: suggest treatment lowers hazard

Parametric models

- Need not be PH models
- Many are AFT models

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

*AFT vs. PH*

- Different interpretation of parameters
- AFT applies to comparison of survival times
- PH applies to comparison of hazards

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  $\exp[-1.527 \pm 1.96(0.398)]$  yielding a CI of (0.10, 0.47). These results suggest that the experimental treatment delays remission.

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 models rather than PH models.** The exponential and Weibull distributions can accommodate both the PH and AFT assumptions.

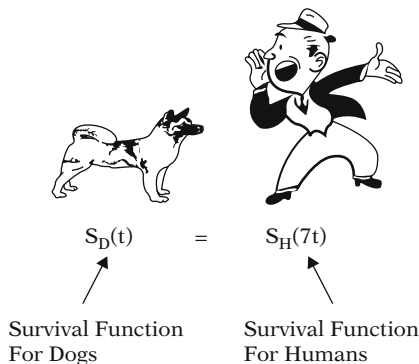
On the left is Stata output from the AFT form of the exponential model with TRT as the only predictor. Stata can output both the PH or AFT form of an exponential or Weibull model (see Computer Appendix). SAS (version 8.2) only runs the AFT form of parametric models and SPSS (version 11.5) does not yet provide commands to run parametric models.

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

##### Second Illustration

$S_1(t)$ —Survival function for smokers

$S_2(t)$ —Survival function for nonsmokers

AFT assumption:

$$S_2(t) = S_1(\gamma t) \text{ for } t \geq 0$$

$\gamma$  is the acceleration factor

If  $\gamma = \exp(\beta)$

$$S_2(t) = S_1([\exp(\alpha)]t)$$

or

$$S_2([\exp(-\alpha)]t) = S_1(t)$$

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  $S_D(t) = S_H(7t)$ , where  $S_D(t)$  and  $S_H(t)$  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.**

For a second illustration of the accelerated failure time assumption consider a comparison of survival functions among smokers  $S_1(t)$  and nonsmokers  $S_2(t)$ . The AFT assumption can be expressed as  $S_2(t) = S_1(\gamma t)$  for  $t \geq 0$ , where  $\gamma$  is a constant called the **acceleration factor** comparing smokers to nonsmokers. In a regression framework the acceleration factor  $\gamma$  could be parameterized as  $\exp(\alpha)$  where  $\alpha$  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

$$S_2(t) = S_1(0.75t)$$

$T_1$ —Survival time for smokers

$T_2$ —Survival time for nonsmokers

AFT assumption in terms of random variables:

$$T_1 = \gamma T_2$$

Acceleration factor

Measure of association  
on survival time

Hazard ratio

Measure of association on the  
hazard

Acceleration factor ( $\gamma$ )

- 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)

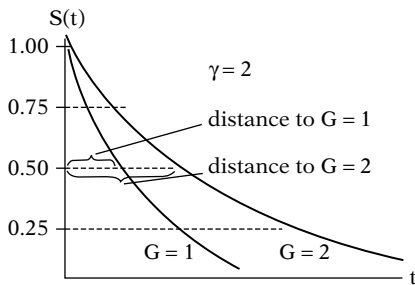
- Time to  $S(t) = 0.50$  (median) is double for Group 2
- Time to  $S(t) = 0.20$  is double for Group 2
- Time to  $S(t) = 0.83$  is double for Group 2
- Time to  $S(t) = 0.98$  is double for Group 2
- Time to  $S(t) = \mathbf{q}$  is double for Group 2 (**generalization**)

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.,  $S_2(t) = S_1(0.75t)$ ).

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  $T_1 = \gamma T_2$ .

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.

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 ( $\text{TRT} = 1$ )  
21 patients given placebo ( $\text{TRT} = 0$ )

Previously discussed PH form of model

Now discuss AFT form of model

Exponential survival and hazard functions:

$$S(t) = \exp(-\lambda t)$$

$$h(t) = \lambda$$

Recall for PH model:

$$h(t) = \lambda = \exp(\beta_0 + \beta_1 \text{TRT})$$

We return to the exponential example applied to the remission data with treatment status ( $\text{TRT}$ ) 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.

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.

AFT assumption  
(comparing 2 levels)

- Ratio of times is constant to all fixed  $S(t)$

Strategy for developing the model:

- Solve for  $t$  in terms of  $S(t)$
- Scale  $t$  in terms of the predictors

$$S(t) = \exp(-\lambda t)$$

$$t = [-\ln(S(t))] \times \frac{1}{\lambda}$$

$$\text{let } \frac{1}{\lambda} = \exp(\alpha_0 + \alpha_1 \text{TRT})$$

$$t = [-\ln(S(t))] \times \mathbf{\exp(\alpha_0 + \alpha_1 \text{TRT})}$$

Scaling of  $t$  

Median survival time,  $S(t) = 0.5$ :

$$t_m = [-\ln(0.5)] \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

The underlying AFT assumption, for comparing two levels of covariates, is that the ratio of times to any fixed value of  $S(t) = q$  is constant for any probability  $q$ . We develop the model with the survival function and solve for  $t$  in terms of  $S(t)$ . We then scale  $t$  in terms of the predictors.

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  $\lambda$ , 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:

$$\gamma (\text{TRT} = 1 \text{ vs. } \text{TRT} = 0)$$

$$\begin{aligned} \gamma &= \frac{[-\ln(q)] \exp(\alpha_0 + \alpha_1)}{[-\ln(q)] \exp(\alpha_0)} \\ &= \exp(\alpha_1) \end{aligned}$$

The expression for  $t$  is restated on the left in terms of any fixed probability  $S(t) = q$ . The acceleration factor  $\gamma$  is found by taking the ratio of the times to  $S(t) = q$  for  $\text{TRT} = 1$  and  $\text{TRT} = 0$ . After canceling,  $\gamma$  reduces to  $\exp(\alpha_1)$ .

*Remission Data*

Exponential regression accelerated failure-time form

<i>_t</i>	Coef.	Std. Err.	z	p >  z
trt	1.527	.398	3.83	0.00
_cons	2.159	.218	9.90	0.00

$$\hat{\gamma} = \exp(1.527) = 4.60$$

$$95\% \text{ CI: } \exp[1.527 \pm 1.96(0.398)] \\ = (2.11, 10.05)$$

$$t = [-\ln(q)] \times \exp(\alpha_0 + \alpha_1 \text{TRT}) \\ \hat{t} = [-\ln(q)] \\ \times \exp(2.159 + 1.527(\text{TRT}))$$

Estimated Survival Times by S(t)  
Quartiles for TRT = 1 and  
TRT = 0 (Exponential Model)

S(t) = q	$\hat{t}_{\text{TRT}=0}$	$\hat{t}_{\text{TRT}=1}$
0.25	12.0	55.3
0.50	6.0	27.6
0.75	2.5	11.5

$$\hat{\gamma} = 4.60 \text{ (for TRT = 1 vs. TRT = 0)}$$

Ratio of survival times:

$$\frac{55.3}{12.0} = \frac{27.6}{6.0} = \frac{11.5}{2.5} = 4.60$$

Effect of treatment:

- Stretches survival by a factor of 4.6
- Interpretation of  $\gamma$  has intuitive appeal

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  $\gamma$  is calculated as  $\exp[1.527 \pm 1.96(0.398)]$  yielding a CI of (2.11, 10.05).

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.

The ratio of survival times for each row in the table comparing TRT = 1 vs. TRT = 0 is 4.60, which not coincidentally 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.



HR and  $\gamma$  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

$\text{HR} > 1 \Rightarrow$  exposure harmful to survival

$\gamma < 1 \Rightarrow$  exposure harmful to survival

$\text{HR} < 1 \Rightarrow$  exposure benefits survival

$\gamma = \text{HR} = 1 \Rightarrow$  no effect from exposure

Exponential PH and AFT models:

- Same model
- Different parameterization
- Same estimates for
  - Survival function
  - Hazard function
  - Median survival

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.,  $\text{TRT} = 1$  vs.  $\text{TRT} = 0$ ) 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 the exposure is beneficial to survival whereas a hazard ratio greater than one implies the exposure is harmful to survival (and vice versa).**

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

Exponential model is special case of Weibull 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.

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,  $\lambda$  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  $\Rightarrow$  PH and PH  $\Rightarrow$  AFT  
 Holds if  $p$  is fixed

HR vs. AFT

Hazard ratio  $\Rightarrow$  Comparison of rates

Acceleration factor  $\Rightarrow$  Effect on survival

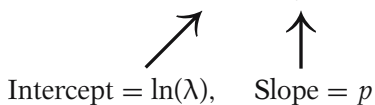
Useful Weibull property:

- $\ln[-\ln S(t)]$  is linear with  $\ln(t)$
- Enables graphical evaluation using KM survival estimates

Linearity of  $\ln(t)$

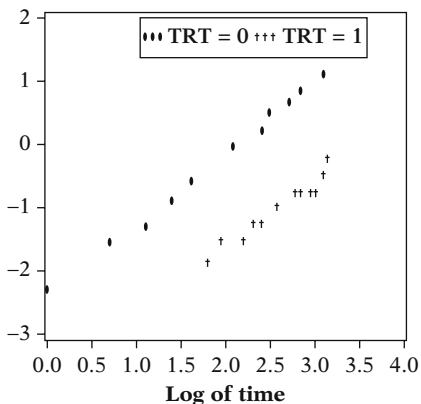
$$S(t) = \exp(-\lambda t^p)$$

$$\Rightarrow \ln[-\ln S(t)] = \ln(\lambda) + p \ln(t)$$



Remission data: evaluate Weibull assumption for TRT = 1 and TRT = 0

$\ln[-\ln \hat{S}(t)]$  plotted against  $\ln(t)$



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. 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.

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**.

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.

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  $\Rightarrow$  Weibull, PH, and AFT assumptions hold
2. Parallel straight lines with slope of 1  $\Rightarrow$  Exponential. PH and AFT
3. Parallel but not straight lines  $\Rightarrow$  PH but not Weibull, not AFT (can use Cox model)
4. Not parallel and not straight  $\Rightarrow$  Not Weibull, PH violated
5. Not parallel but straight lines  $\Rightarrow$  Weibull holds, but PH and AFT violated, different  $p$

Previous plot suggests Weibull and PH assumption reasonable for TRT

Weibull PH model:

$$h(t) = \lambda p t^{p-1}$$

where  $\lambda = \exp(\beta_0 + \beta_1 \text{TRT})$ .

Hazard ratio (TRT = 1 vs. TRT = 0)

$$\begin{aligned} \text{HR} &= \frac{\exp(\beta_0 + \beta_1) p t^{p-1}}{\exp(\beta_0) p t^{p-1}} \\ &= \exp(\beta_1) \end{aligned}$$

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.

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.

The Weibull hazard function is  $h(t) = \lambda p t^{p-1}$ . A Weibull PH model is defined by reparameterizing lambda  $\lambda$  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).

## 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		

Weibull PH

$$\widehat{HR} (\text{TRT} = 1 \text{ vs. } 0) = \exp(-1.731) = 0.18$$

$$95\% \text{ CI} = \exp[-1.731 \pm 1.96(0.413)] = (0.08, 0.40)$$

Weibull:  $\widehat{HR} = 0.18$ Exponential:  $\widehat{HR} = 0.22$ 

Suggests preventive effect of TRT

Comparing Cox and Weibull PH models

Cox: estimate  $\beta_1$ 


$$h(t) = h_0(t) \exp(\beta_1 \text{TRT})$$


 baseline hazard unspecified
Weibull: estimate  $\beta_0, \beta_1, p$ 

$$h(t) = \lambda p t^{p-1} \text{ where}$$

$$\lambda = \exp(\beta_0 + \beta_1 \text{TRT}).$$

$$h(t) = [\exp(\beta_0) p t^{p-1}] \exp(\beta_1 \text{TRT}).$$


 baseline hazard specified parametrically

On the left is output (Stata version 7.0) 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.

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.

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 ( $\beta_1$ ) and the distribution of the baseline hazard ( $h_0(t)$ ) remains unspecified.

With some manipulation, the Weibull PH model can also be expressed as a product of a baseline hazard and  $\exp(\beta_1 \text{TRT})$  (see left). There are three parameters to estimate  $\beta_0, \beta_1$ , and  $p$  that fully specify the hazard.

$$S(t) = \exp(-\lambda t^p)$$

solve for t

$$t = [-\ln S(t)]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

$$\text{let } \frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{RX})$$

$$t = [-\ln S(t)]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

Scaling of t



Let  $S(t) = q$

$$t = [-\ln(q)]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

Median survival time ( $q = 0.5$ )

$$t_m = [-\ln(0.5)]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{TRT})$$

Acceleration factor,  $\gamma$  (TRT = 1 vs. TRT = 0)

$$\begin{aligned} \gamma &= \frac{[-\ln(q)]^{1/p} \exp(\alpha_0 + \alpha_1)}{[-\ln(q)]^{1/p} \exp(\alpha_0)} \\ &= \exp(\alpha_1) \end{aligned}$$

*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		

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  $1/\lambda^{1/p} = \exp(\alpha_0 + \alpha_1 \text{TRT})$ , it can be seen that the predictor variable TRT is used to scale the 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).

The acceleration factor  $\gamma$  is obtained as the ratio of the times to  $S(t) = q$  for TRT = 1 and for TRT = 0. After canceling,  $\gamma$  reduces to  $\exp(\alpha_1)$ . As with the PH form of the model, this result depends on p not varying by treatment status; otherwise  $\gamma$  would depend on q.

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  $\gamma$  is calculated to be (1.93, 6.53).

Weibull AFT:

$$\hat{\gamma}(\text{TRT} = 1 \text{ vs. } 0) = \exp(1.267) \\ = 3.55$$

$$95\% \text{ CI} = \exp[1.267 \pm 1.96(0.311)] \\ = (1.93, 6.53)$$

Weibull:  $\hat{\gamma} = 3.55$

Exponential:  $\hat{\gamma} = 4.60$  (assumes  $h(t) = \lambda$ )

*Relating Weibull AFT and PH coefficients*

$$\text{AFT: } \lambda^{1/p} = \exp[-(\alpha_0 + \alpha_1 \text{TRT})] \\ (1/p) \ln \lambda = -(\alpha_0 + \alpha_1 \text{TRT}) \\ \ln \lambda = -p(\alpha_0 + \alpha_1 \text{TRT})$$

$$\text{PH: } \lambda = \exp(\beta_0 + \beta_1 \text{TRT}) \\ \ln \lambda = \beta_0 + \beta_1 \text{TRT}$$

Relationship of coefficients:

$$\beta_j = -\alpha_j p \text{ so that} \\ \beta = -\alpha \text{ for exponential } (p = 1)$$

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

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.

Corresponding coefficients obtained from the PH and AFT forms of the Weibull models are related as follows:  $\beta_j = -\alpha_j p$  for the  $j$ th 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.

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,  $\beta = -\alpha$ .

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

$$\text{Log-logistic hazard: } h(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p} \\ (\text{where } p > 0 \text{ and } \lambda > 0)$$

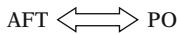
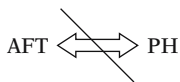
The log-logistic distribution accommodates an AFT model but not a PH model. Its hazard function is shown on the left. The shape parameter is  $p(>0)$ .

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: 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 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

$$S(t) = \frac{1}{1 + \lambda t^p} \quad 1 - S(t) = \frac{\lambda t^p}{1 + \lambda t^p}$$

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  $\lambda t^p$  (see left).

Log-logistic PO model:

- Reparameterize  $\lambda$  in terms of  $X_s$  and  $\beta_s$

A log-logistic proportional odds model can be formulated by reparameterizing  $\lambda$  in terms of predictor variables and regression parameters. We come back to this point later in this section.



Log Odds Is Linear with  $\ln(t)$

$$\begin{aligned} \log(\text{failure odds}) &= \ln(\lambda t^p) \\ &= \ln(\lambda) + p[\ln(t)] \end{aligned}$$

$\nearrow$  Intercept =  $\ln(\lambda)$        $\uparrow$  slope =  $p$

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 =  $p$

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 =  $-p$

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.

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  $p$ .

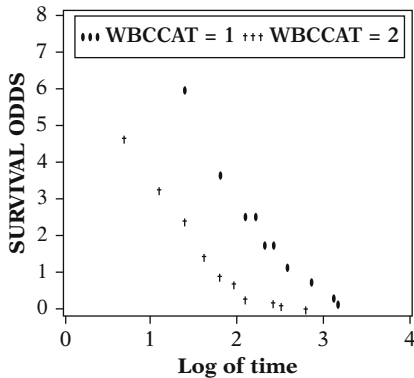
We could alternatively plot the log of the survival odds,  $\ln(\hat{S}(t))/(1 - \hat{S}(t))$ , against  $\ln(t)$ . If the log-logistic assumption is correct the resulting plots should be a straight line of slope  $-p$ .

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  $\implies$  Log-logistic

Parallel plots  $\implies$  PO

Log-logistic and PO  $\implies$  AFT

Log-logistic and Weibull graphical approach analogous

- Check PH for Weibull
- Check PO for log-logistic

AFT log-logistic model

$$S(t) = \frac{1}{1 + \lambda t^p} = \frac{1}{1 + (\lambda^{1/p} t)^p}$$

solve for t to obtain

$$t = \left[ \frac{1}{S(t)} - 1 \right]^{1/p} \times \frac{1}{\lambda^{1/p}}$$

$$\text{let } \frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{WBCCAT})$$

$$t = \left[ \frac{1}{S(t)} - 1 \right]^{1/p} \times \exp(\alpha_0 + \alpha_1 \text{WBCCAT})$$

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:

- straight lines support the log-logistic assumption,**
- parallel curves support the PO assumption, and**
- 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  $\text{WBCCAT} = 2$  (high count) and  $\text{WBCCAT} = 1$  (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 the predictor variable  $\text{WBCCAT}$  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,  
 $\gamma$  (WBCCAT = 2 vs. WBCCAT = 1)

$$= \frac{[q^{-1} - 1]^{1/p} \exp(\alpha_0 + 2\alpha_1)}{[q^{-1} - 1]^{1/p} \exp(\alpha_0 + 1\alpha_1)} = \exp(\alpha_1)$$

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_gamma	-.779	.164	-4.73	0.000
gamma	.459	.0756		

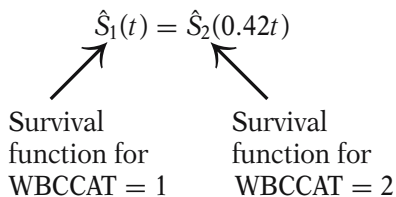
$$p = 1/(0.459) = 2.18$$

WBCCAT = 2 vs. WBCCAT = 1 (log-logistic):

$$\hat{\gamma} = \exp(-0.871) = 0.42$$

$$95\% \text{ CI for } \gamma = \exp[-0.871 \pm 1.96(0.296)] = (0.23, 0.75)$$

Comparing estimated survival



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  $\lambda = \exp(\beta_0 + \beta_1 \text{WBCCAT})$

OR (WBCCAT = 2 vs. WBCCAT = 1)

$$= \frac{t^p \exp(\beta_0 + 2\beta_1)}{t^p \exp(\beta_0 + 1\beta_1)} = \exp(\beta_1)$$

The acceleration factor  $\gamma$  is found by taking the ratio of the times to  $S(t) = q$  for WBCCAT = 2 and for WBCCAT = 1. After canceling,  $\gamma$  reduces to  $\exp(\alpha_1)$ .

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  $p$  ( $\text{gamma} = 1/p$ ) rather than for  $p$ . The estimate for gamma is 0.459. Therefore, the estimate for  $p$  is  $1/(0.459) = 2.18$ .

An estimate of the acceleration factor  $\hat{\gamma}$  comparing WBCCAT = 2 to WBCCAT = 1 is found by exponentiating the estimate  $-0.871$  of  $\alpha_1$  to obtain 0.42. The 95% confidence interval for  $\gamma$  is calculated to be (0.23, 0.75).

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.

The proportional odds form of the log-logistic model can also be formulated by reparameterizing  $\lambda$ . Recall that the log-logistic failure odds is  $\lambda t^p$ .

By setting  $\lambda = \exp(\beta_0 + \beta_1 \text{WBCCAT})$ , an odds ratio comparing WBCCAT = 2 to WBCCAT = 1 can be calculated (see left). After canceling, the odds ratio reduces to  $\exp(\beta_1)$ .

*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  $j$ th covariate. This result is obtained using a similar argument to that presented for the Weibull example in the previous section.

The estimate for  $\alpha_1$  in the AFT model is  $-0.871$  and the estimate for  $p$  is  $2.18$ . Therefore, an estimate for  $\beta_1$  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)

$$\ln(T) = \alpha_0 + \alpha_1 \text{TRT} + \epsilon$$


  
random error

With additional parameter

$$\ln(T) = \alpha_0 + \alpha_1 \text{TRT} + \sigma \epsilon$$


  
 $\sigma$  scales the error

If  $\epsilon \sim N(0, 1)$ , then  
 $\ln(T) \sim N(\mu = \alpha_0 + \alpha_1 \text{TRT}, \text{sd} = \sigma)$

Similar to linear regression (except for inclusion of censorships)

In general,

$$\mu_{\ln(T)} \neq (\alpha_0 + \alpha_1 \text{TRT}), \quad \text{sd} \neq \sigma$$

Interpretation of parameters depends on distribution

Let  $\sigma = \frac{1}{p}$ , then

$$\ln(T) = \alpha_0 + \alpha_1 \text{TRT} + \frac{1}{p} \epsilon$$

Additive model in terms of  $\ln(T)$   
 but  
 multiplicative model in terms of T

$$\begin{aligned} T &= \exp\left(\alpha_0 + \alpha_1 \text{TRT} + \frac{1}{p} \epsilon\right) \\ &= \exp[\alpha_0 + \alpha_1 \text{TRT}] \times \exp\left(\frac{1}{p} \epsilon\right) \end{aligned}$$

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  $\epsilon$  is random error following some distribution.

Some distributions have an additional parameter ( $\sigma$ ) scaling  $\epsilon$ . The model including this additional parameter can be expressed as shown on the left, where the random error  $\epsilon$  is multiplied by a scale parameter  $\sigma$ .

If  $\epsilon$  follows a standard normal distribution and  $\ln(T) = \alpha_0 + \alpha_1 \text{TRT} + \sigma \epsilon$ , then  $\ln(T)$  would follow a normal distribution with mean  $\mu = \alpha_0 + \alpha_1 \text{TRT}$  and standard deviation  $\sigma$ . 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, for other distributions, the mean of  $\ln(T)$  is not  $\alpha_0 + \alpha_1 \text{TRT}$  and its standard deviation is not  $\sigma$ . In other words, it should not be assumed that the mean of  $\epsilon$  is 0 and the standard deviation is 1. The interpretation of the parameters depends on the underlying distribution.

Sometimes the model is parameterized using  $\sigma = 1/p$ . The model can then be restated by replacing  $\sigma \epsilon$  with  $(1/p)\epsilon$ .

The AFT model is **additive on the log scale but a multiplicative model with respect to T**.

In particular, the model can be expressed in terms of T by exponentiating  $\ln(T)$ , as shown on the left.

Collapse  $\alpha_0$  into baseline term

$$T_0 = \exp(\alpha_0) \exp\left(\frac{1}{p}\epsilon\right)$$

so that  $T = \exp(\alpha_1 \text{TRT}) \times T_0$   
 where  $T_0$  is a random variable for  
 $\text{TRT} = 0$

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

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 ( $\text{TRT} = 0$ ).

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:

$$h(t) = [\exp(\gamma t)] \times \exp(\beta_0 + \beta_1 \text{TRT})$$



$$h_0(t) = \exp(\gamma t) \text{ (parametrically specified)}$$

- $\gamma > 0$  hazard exponentially increases with  $t$
- $\gamma < 0$  hazard exponentially decreases with  $t$
- $\gamma = 0$  constant hazard (exponential model)

AFT model: multiplicative


$$\begin{aligned} T &= \exp(\alpha_0 + \alpha_1 \text{TRT} + \epsilon) \\ &= \exp(\alpha_0) \times \exp(\alpha_1) \times \exp(\epsilon) \end{aligned}$$

but

additive on log scale:

$$\ln(T) = \alpha_0 + \alpha_1 \text{TRT} + \epsilon$$

Additive failure time model



$$T = \alpha_0 + \alpha_1 \text{TRT} + \epsilon$$

T rather than  $\log(T)$  is linear with TRT

*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

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  $\gamma$  (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.

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).

An alternative parametric model is to define an **additive failure time model** in terms of  $T$ . Consider the model:  $T = \alpha_0 + \alpha_1 \text{TRT} + \epsilon$ . Now  $T$ , rather than  $\ln(T)$ , is expressed as a linear function of the regression parameters. Stated in statistical language: the log link function is omitted from this failure time model. SAS supports such an additive failure time model (see Computer Appendix).

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  $\lambda$  and  $p$  are modeled as functions of treatment status (TRT). If  $\delta_1$  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:  $\text{-----} \frac{|}{10} \text{X}$  time

Left-censored:  $\text{X} \frac{|}{10} \text{-----}$  time

Interval-censored:  $\text{-----} \frac{|}{8} \text{X} \frac{|}{10} \text{-----}$  time

**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

$$L = f(2) \times \int_8^{\infty} f(t)dt \times f(6) \times \int_0^2 f(t)dt \times \int_4^8 f(t)dt$$

(Barry  $\times$  Gary  $\times$  Harry  $\times$  Carrie  $\times$  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
  
- Subjects independent
  - Allows L to be formulated as product of subjects' contributions

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.

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 subject's 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

Weibull:  $h(t) = \lambda p t^{p-1}$ ,

$S(t) = \exp(-\lambda t^p)$

$f(t) = h(t)S(t)$

$f(t) = \lambda p t^{p-1} \exp(-\lambda t^p)$

where  $\lambda = \exp(\beta_0 + \beta_1 \text{SMOKE})$

(PH form of the model)

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).

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:  $\lambda = \beta_0 + \beta_1 \text{SMOKE}$ .

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 LOWER and 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

$$L = f(2) \times \int_8^{\infty} f(t)dt \times f(6) \times \int_0^2 f(t)dt$$

$$\times \int_4^8 f(t)dt$$

$$L = \exp(\beta_0 + \beta_1)p(2)^{p-1} \exp(-\exp(\beta_0 + \beta_1)2^p)$$

$$\times \int_8^{\infty} \exp(\beta_0)p(t)^{p-1} \exp(-\exp(\beta_0)t^p)dt$$

$$\times \exp(\beta_0)p(6)^{p-1} \exp(-\exp(\beta_0)6^p)$$

$$\times \int_0^2 \exp(\beta_0)p(t)^{p-1} \exp(-\exp(\beta_0)t^p)dt$$

$$\times \int_4^8 \exp(\beta_0 + \beta_1)p(t)^{p-1} \exp(-\exp(\beta_0 + \beta_1)t^p)dt$$

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 (5 subjects) for ease of illustration but the process can be generalized for any number of subjects.

Obtaining maximum likelihood estimates

Solve system of equations:

$$\frac{\partial L_n(L)}{\partial \beta_j} = 0 \quad j = 1, 2, \dots, N$$

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 <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub>	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

EVENT: dichotomous outcome coded 1 if event, 0 for no event during the interval

D<sub>1</sub>, D<sub>2</sub>, D<sub>3</sub>: 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

*Logistic Model*

$$\text{Logit } P(Y = 1) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 \text{TRT}$$

where  $P(Y = 1)$  is the probability of event for a given interval conditioned on survival of previous intervals

*Interpretation of Parameters*

$\beta_1$ : Log odds of event in 1st interval among TRT = 0

$\beta_2$ : Log odds of event in 2nd interval given survival of 1st interval among TRT = 0

$\beta_3$ : Log odds of event in 3rd interval given survival of first two intervals among TRT = 0

$\beta_4$ : Log odds ratio for TRT

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<sub>1</sub> is a dummy variable coded 1 if the observation represents the first interval and 0 otherwise. Similarly, D<sub>2</sub> is coded 1 for the second interval and D<sub>3</sub> is coded 1 for the third interval.

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.

A logistic model (shown at left) containing the three dummy variables and TRT can be formulated with the data in this form.

Care must be taken with the interpretation of the parameters:  $\beta_1$  is the log odds of the event occurring in the first interval among the placebo group;  $\beta_2$  is the log odds of the event occurring in the second interval conditioned on survival of the first interval among the placebo group;  $\beta_3$  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  $\beta_4$  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

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

*Alternative Binary Model*

$$\log(-\log(1 - P(Y = 1))) = \beta_1 D_1 + \beta_2 D_2 + \beta_3 D_3 + \beta_4 \text{TRT}$$

where  $1 - P(Y = 1)$  is the probability of surviving a given interval conditioned on survival of previous intervals

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

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.

The odds ratio comparing TRT = 1 to TRT = 0 is obtained by exponentiating  $\beta_4$ . 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.

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.

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

Log–log survival curves:  
parallel  $\implies$  additive effects  
 $\implies$  PH

Complementary log–log link:  
additive effects on log–log scale  
 $\implies$  PH

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

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

**The complementary log–log binary model is a proportional hazards model.** The hazard ratio comparing TRT = 1 to TRT = 0 is obtained by exponentiating  $\beta_4$ .

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, 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** 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  $\alpha$  ( $\alpha > 0$ )**

- Unobserved multiplicative effect on hazard
- Follows distribution  $g(\alpha)$  with  $\mu = 1$
- $\text{Var}(g(\alpha)) = \theta$ , parameter to be estimated

Hazard and survival conditioned on frailty

$$h(t|\alpha) = \alpha h(t)$$

$$S(t|\alpha) = S(t)^\alpha$$

 $\alpha > 1$ 

- 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):  $\alpha h(t) = h(t)$

Survival functions  
(with frailty models)

1. Conditional,  $S(t|\alpha)$ , individual level
2. Unconditional,  $S_U(t)$ , population level

Unconditional survival function  $S_U(t)$

$$S_U(t) = \int_0^\infty S(t|\alpha)g(\alpha)d\alpha$$

$$h_U(t) = \frac{-d[S_U(t)]/dt}{S_U(t)}$$

**The frailty  $\alpha$**  is an unobserved multiplicative effect on the hazard function assumed to follow some distribution  $g(\alpha)$  with  $\alpha > 0$  and the mean of  $g(\alpha)$  equal to 1. The variance of  $g(\alpha)$  is a parameter  $\theta$  (theta) that is typically estimated from the data.

An individual's hazard function conditional on the frailty can be expressed as  $\alpha$  **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  $\alpha$  power**.

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.

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  $\alpha$ . The corresponding unconditional hazard  $h_U(t)$  can then be found using the relationship between the survival and hazard functions (see left).

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  $\theta$

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  $\theta$  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:  $h(t) = \lambda pt^{p-1}$  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&gt;=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 ( $\theta$ ). 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 3 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  $j$ th individual:

$$h_j(t|\alpha_j) = \alpha_j h(t) \quad j = 1, 2, \dots, n$$

where  $h(t) = \lambda p t^{p-1}$

$$\begin{aligned} \text{with } \lambda = \exp(\beta_0 + \beta_1 \text{TX} \\ + \beta_2 \text{PERF} + \beta_3 \text{DD} \\ + \beta_4 \text{AGE} + \beta_5 \text{PRIORTX}) \end{aligned}$$

and where  $\alpha \sim \text{gamma} (\mu = 1, \text{variance} = \theta)$

$\alpha_j$  not estimable

- An  $\alpha_j$  associated with each subject
- Too many parameters

Rather,  $\text{var}[g(\alpha)]$  is estimated

- Gamma is 2-parameter distribution
  - Mean set at 1.0
  - $\theta = \text{Var}[g(\alpha)]$  is estimated

Interpreting coefficients in Model 2

$$\widehat{HR} = \exp(\hat{\beta}_1) = 1.11$$

Estimates HR comparing two individuals

- With same  $\alpha$
- One with TX = 2, other with TX = 1
- With same levels of other predictors

For Model 2 we can express the Weibull model with a gamma distributed frailty in terms of the individual level hazard for the  $j$ th subject.

If  $\alpha_j$  denotes the frailty for the  $j$ th subject, then that subject's hazard  $h_j(t|\alpha_j)$  can be expressed as  $\alpha_j$  **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).

The values for each  $\alpha_j$  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, 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.

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 covariates 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  $\alpha_1 = \alpha_2$

Another interpretation for  $\exp(\beta_1)$

- Ratio of conditional hazards from the same individual
- Effect for individual taking test rather than 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

For Model 2

Conditional hazard increases

but

unconditional hazard unimodal

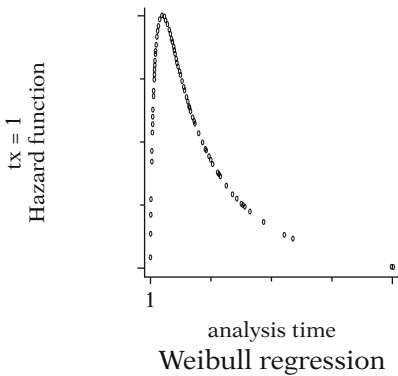
To clarify, recall that the individual level or conditional hazard function can be expressed as a multiplied by  $h(t)$ . Suppose  $h_1(t|\alpha_1)$  and  $h_2(t|\alpha_2)$  are the conditional hazard functions for individuals who use the standard and test treatments, respectively, at the mean levels 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 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.

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.

Although the estimated individual level or **conditional** hazard is estimated to increase from Model 2, 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.

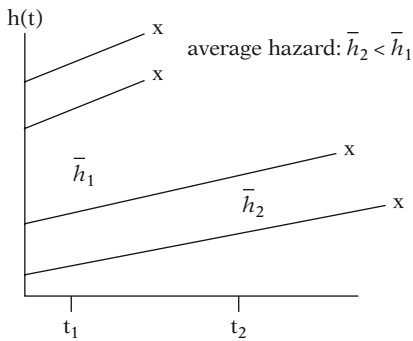
Estimated unconditional hazard Model 2 (TX = 1, mean level for other covariates,  $\hat{p} = 1.54$ )



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)

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  $h_U(t) \rightarrow 0$
- So  $h_U(t)$  increases and then decreases (unimodal)

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)

$$\frac{h_{U2}(t)}{h_{U1}(t)} = \frac{h_2(t)}{h_1(t)} \times \frac{1 - \theta \ln[S_1(t)]}{1 - \theta \ln[S_2(t)]}$$

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.

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.

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:

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.

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

Model 2 (Weibull with frailty)

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.

- Used PH parameterization
- Can equivalently use AFT parameterization



Unconditional survival function  $S_U(t)$  with gamma frailty  $g(\alpha)$

$$S_U(t) = \int_0^\infty S(t | \alpha)g(\alpha)d\alpha$$

$$= [[1 - \theta \ln S(t)]^{-1/\theta}$$

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

$$\frac{1}{\lambda^{1/p}} = \exp(\alpha_0 + \alpha_1 \text{TX} + \alpha_2 \text{PERF} + \alpha_3 \text{DD} + \alpha_4 \text{AGE} + \alpha_5 \text{PRIORTX})$$

*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

$$\hat{\gamma}(\text{TX} = 2 \text{ vs. } 1) = \exp(-0.068)$$

$$= 0.93$$

Comparing individuals with same  $\alpha$

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 (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).

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 (multiply by  $-p$  to get the PH coefficient estimates in Model 2).

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.

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 ( $p = 0.721$ )

PH assumption

Individual level PH  $\Rightarrow$  Population level PH

AFT assumption

Individual level AFT  $\Rightarrow$  Population level AFT

Population level survival (with gamma frailty)

$$S_{U1}(t) = [[1 - \theta \ln S_1(t)]^{-1/\theta}$$

$$S_{U2}(t) = [[1 - \theta \ln S_2(t)]^{-1/\theta}$$

If  $S_1(t) = S_2(\gamma t)$

then

$$S_{U1}(t) = [[1 - \theta \ln S_1(t)]^{-1/\theta}$$

$$= [[1 - \theta \ln S_2(\gamma t)]^{-1/\theta}$$

$$= S_{U2}(\gamma t)$$

Thus,

Individual level AFT

$\Rightarrow$  Population level AFT

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.

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.

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  $\gamma$  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).

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

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).

## Log-logistic model

- Accommodates unimodal hazard without a frailty component

## 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  $\theta$

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

*Unshared Frailty*

- The type of frailty we have described previous to this point
- Frailty distributed independently among subjects

*Shared Frailty Models*

- Similar to random effect regression models
- Accounts for within-cluster correlation
- $\theta$  is a measure of the degree of correlation

Hazard conditional on shared frailty (for  $j$ th subject in  $k$ th cluster)

$$h_{jk}(t|\alpha_k) = \alpha_k h_{jk}(t)$$

where

$$h_{jk}(t) = h(t|\mathbf{X}_{jk})$$

for  $j = 1, 2, \dots, n_k$

and total  $n_k$  subjects in  $k^{\text{th}}$  cluster

If family is the cluster variable,  
then

subjects of same family have same

$\alpha_k$

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.

By contrast, the frailty described previous to this point (**unshared frailty**) has been assumed to be distributed independently among subjects.

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 2002). The estimate for the variance parameter  $\theta$  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.

For a shared frailty model, the conditional hazard function for the  $j$ th subject from the  $k$ th cluster can be expressed as  $\alpha_k$  **multiplied by**  $h_{jk}(t)$  where  $h_{jk}(t)$  depends on the subject's covariates  $\mathbf{X}_{jk}$ . Notice that the frailty  $\alpha_k$  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

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.

## Shared frailty models

- Accounts for dependence among subjects who share frailty

 $S_{\cup}(t)$  and  $h_{\cup}(t)$ 

- Population averages in unshared frailty models
- Population averages in shared frailty models provided that cluster size is uncorrelated with frailty

Recall, with unshared frailty models, we interpreted the unconditional survival and hazard functions as representing population averages. With shared frailty models, however, these unconditional functions may not strictly represent population averages unless the number of subjects in a cluster is uncorrelated with the level of frailty.

## 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

## Cox shared frailty model

$$h_{jk}(t|\alpha_j) = \alpha_k h_0(t) \exp(\beta \mathbf{X}_{jk})$$

for  $j = 1, 2, \dots, n_k$   
total of  $n_k$  subjects in  $k$ th cluster

PH violation of  $h_{\cup}(t)$  in Cox model

- If gamma-distributed frailty included
- Interpreting coefficient estimates
  - Only used for HR estimates among those who share same  $\alpha$

## 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

## Recurrent events:

- Topic of next chapter (Chapter 8)

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  $j$ th subject from the  $k$ th cluster can be expressed as  $\alpha_k$  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.

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.

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.

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

$f(t)$  specified  $\Rightarrow$  corresponding  $S(t)$ ,  
 $h(t)$  also determined

Moreover,  
Specifying one of  $f(t)$ ,  $S(t)$ , or  $h(t)$   
determines all three functions

#### Parametric models

- Need not be PH models
- Many are AFT models

#### Acceleration factor ( $\gamma$ )

- Key measure of association in AFT models
- Describes stretching or contraction of  $S(t)$

#### AFT assumption

$$S_2(t) = S_1(\gamma t)$$

$\uparrow$              $\uparrow$   
 Group 2    Group 1

Detailed examples presented:

- Exponential model
- Weibull model
- Log-logistic model

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.

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.

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.

The acceleration factor ( $\gamma$ ) 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  $S_2(t) = S_1(\gamma t)$ .

We presented detailed examples of the exponential, Weibull, and log-logistic model using the remission dataset.

## Exponential Model

- $h(t) = \lambda$  (constant hazard)
- Special case of Weibull model

## Weibull Model

- AFT  $\longleftrightarrow$  PH

## Log-logistic Model

- Not a PH model
- AFT  $\longleftrightarrow$  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

*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

## Presented other parametric models

- Generalized gamma model
- Lognormal model
- Gompertz 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.,  $h(t) = \lambda$ ). 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 accommodate 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).

The idea underlying the proportional odds assumption is that the survival (or failure) odds ratio comparing two specifications of covariates remains constant over time.

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  $\ln(t)$ .

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  $f(t)$

$$\int_0^{t_1} f(t)dt: \text{ left-censored at } t_1$$

$$\int_{t_1}^{\infty} f(t)dt: \text{ right-censored at } t_1$$

$$\int_{t_1}^{t_2} f(t)dt: \text{ interval-censored from } t_1 \text{ to } t_2$$

Full likelihood ( $L$ )

$$L = \prod_{j=1}^N L_j \quad j = 1, 2, \dots, N$$

where  $L_j$  is the contribution from  $j$ th subject

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

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

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).

Assuming independence among subjects, the full likelihood can be formulated as a product of each subject's contribution.

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 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,  $\alpha$

$$h(t|\alpha) = \alpha h(t)$$



multiplicative effect on  $h(t)$   
mean = 1, variance =  $\theta$

$\theta$  estimated from data

We concluded with a discussion of frailty models. The frailty  $\alpha$  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  $\theta$  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 260–262)
  - 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 262–263)
  - 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  $S(t) = \exp(-\int_0^t h(u)du)$  and  $f(t) = h(t)S(t)$
- III. Exponential Example** (pages 263–265)
  - 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:  $HR(X_1 = 1 \text{ vs. } X_1 = 0) = \exp(\beta_1)$
- IV. Acceleration Failure Time Assumption** (pages 266–268)
  - 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 ( $\gamma$ ) 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  $\gamma$  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 268–272)

- 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  $\beta_j$  and  $\alpha_j$  are PH and AFT parameterization for the  $j$ th 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 272–277)

- 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  $\beta_j$  and  $\alpha_j$  are PH and AFT parameterization, respectively, for the  $j$ th covariate

## VII. Log-Logistic Example (pages 277–282)

- 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:  

$$S(t) = 1/(1 + \lambda t^p) = 1/(1 + (\lambda^{1/p} t)^p)$$
 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:  
 $HR (X_1 = 1 \text{ vs. } X_1 = 0) = \exp(\beta_1)$ 
    - a. PO assumption is that the odds ratio is constant over time
  - v. Survival odds ratio:  
 $HR (X_1 = 1 \text{ vs. } X_1 = 0) = \exp(-\beta_1)$ 
    - 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  $\beta_j$  and  $\alpha_j$  are PO and AFT parameterization for the  $j$ th covariate
    - a. Shape parameter with Stata is parameterized as  $\gamma = 1/p$

**VIII. A More General Form of the AFT Model**

(pages 282–284)

A. General form with one predictor ( $X_1$ ):  $\ln(T) = \alpha_0 + \alpha_1 X_1 + \epsilon$

B. Include additional parameter,  $\sigma$ :  $\ln(T) = \alpha_0 + \alpha_1 X_1 + \sigma \epsilon$

C. Let  $\sigma = 1/p \implies \ln(T) = \alpha_0 + \alpha_1 X_1 + (1/p)\epsilon$

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  $\alpha_0$  into baseline term,

$$\text{let } T_0 = \exp(\alpha_0) \exp\left(\frac{1}{p}\epsilon\right):$$

$$\text{so } T = \exp(\alpha_1 X_1) \times T_0$$

**IX. Other Parametric Models** (pages 284–286)

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)

D. 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 286–289)

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 289–294)

- 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 294–308)

- A. The frailty  $\alpha$  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  $\theta$ 
  - i. The variance  $\theta$  is a parameter estimated by the data
  - ii. Gamma and inverse-Gaussian are distributions offered by Stata 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
- Play similar role as adding a random effect to a linear regression
  - Can account for within-group correlation.

**XIII. Summary** (pages 309–312)

**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

		LR chi2(5) = 61.31		
Log likelihood = -200.196		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  $\gamma$  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:  $-2 \log \text{likelihood} + 2p$**  (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

LR  $\chi^2(5) = 52.86$

Log likelihood =  $-200.626$  Prob >  $\chi^2 = 0.0000$

.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  $\gamma$  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  $\lambda$  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		Prob > chi2 = 0.1941	
Log likelihood = -47.063396					
	.t	Coef.	Std. Err.	z	p >  z
<hr/>					
.t					
	trt	-1.682	1.374	-1.22	0.221
	_cons	-3.083	.646	-4.77	0.000
<hr/>					
ln.p					
	trt	-.012	.328	-0.04	0.970
	_cons	.315	.174	1.82	0.069
<hr/>					

15. Even though  $\lambda$  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  $\gamma$  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 five 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  $j$ th 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. 
$$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2 \text{PERF} + \alpha_3 \text{DD} + \alpha_4 \text{AGE} + \alpha_5 \text{PRIORTX}]}{\exp[\alpha_0 + \alpha_1(1) + \alpha_2 \text{PERF} + \alpha_3 \text{DD} + \alpha_4 \text{AGE} + \alpha_5 \text{PRIORTX}]}$$

$$= \exp(\alpha_1)$$

$$\hat{\gamma} = \exp(-0.054087) = 0.95$$

$$95\% \text{ CI} = \exp[-0.054087 \pm 1.96(0.1863349)] = (0.66, 1.36)$$

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  $\beta_1 = -\alpha_1 p = -\beta_1 \div \text{gamma}$ , where  $\beta_1$  is the coefficient for TX in a PO model.  
 $\text{OR} = \exp(-\alpha_1 \div \text{gamma})$   
 estimated OR =  $\exp(-0.054087 \div 0.6168149) = 0.92$   
 estimated survival OR =  $1/[\exp(-0.054087 \div 0.6168149)] = 1.09$ .

11. The AIC statistic is calculated as  $-2 \log \text{likelihood} + 2p$  (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. 
$$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2PERF + \alpha_3DD + \alpha_4AGE + \alpha_5PRIORTX]}{\exp[\alpha_0 + \alpha_1(1) + \alpha_2PERF + \alpha_3DD + \alpha_4AGE + \alpha_5PRIORTX]}$$

$$= \exp(\alpha_1)$$

$$\hat{\gamma} = \exp(-0.131) = 0.88$$

$$95\% \text{ CI} = \exp[(-0.131 \pm 1.96(0.1908))] = (0.60, 1.28)$$

13. The generalized gamma distribution reduces to a lognormal distribution if  $\kappa = 0$ .

$$H_0: \kappa = 0$$

$$\text{Wald test statistic: } z = \frac{0.2376}{0.2193} = 1.08 \text{ (from output)}$$

$$\text{p-value: } 0.279 \text{ (from output)}$$

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  $\lambda = \exp(\beta_0 + \beta_1 \text{TRT})$  and  $p = \exp(\delta_0 + \delta_1 \text{TRT})$   
 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)
- $$\text{HR (TRT} = 0 \text{ vs. TRT} = 1) = \frac{\lambda_0 p_0 t^{p_0-1}}{\lambda_1 p_1 t^{p_1-1}}$$
- $$\widehat{\text{HR}}(\text{as a function of } t) = \frac{(0.0458)(1.3703)t^{0.3703}}{(0.0085)(1.3539)t^{0.3539}}$$
- $$\widehat{\text{HR}}(t = 10) = \frac{(0.0458)(1.3703)(10^{0.3703})}{(0.0085)(1.3539)(10^{0.3539})} = 5.66$$
- $$\widehat{\text{HR}}(t = 20) = \frac{(0.0458)(1.3703)(20^{0.3703})}{(0.0085)(1.3539)(20^{0.3539})} = 5.73$$

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  $\text{TRT} = 1$  and  $\text{TRT} = 0$  and the lines are parallel.

# **8**

---

**Recurrent**

---

**Event**

---

**Survival**

---

**Analysis**

---

## 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 334)
- II. Examples of Recurrent Event Data (pages 334–336)
- III. Counting Process Example (pages 336–337)
- IV. General Data Layout for Counting Process Approach (pages 338–339)
- V. The Counting Process Model and Method (pages 340–344)
- VI. Robust Estimation (pages 344–346)
- VII. Results for CP Example (pages 346–347)
- VIII. Other Approaches—Stratified Cox (pages 347–353)
- IX. Bladder Cancer Study Example (Continued) (pages 353–357)
- X. A Parametric Approach Using Shared Frailty (pages 357–359)
- XI. A Second Example (pages 359–364)
- XII. Survival Curves with Recurrent Events (pages 364–367)
- XIII. Summary (pages 367–370)

**Objectives**

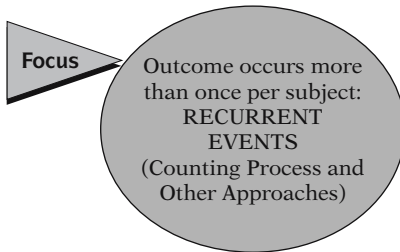
Upon completing this chapter, the learner should be able to:

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 marginal model approach for analyzing correlated data.
5. Given recurrent event data, outline the steps needed to analyze such data using the marginal model approach.

## 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

Recurrent events identical



**Counting Process Approach**  
(Andersen et al., 1993)

Recurrent events: different disease categories or event order important



**Stratified Cox (SC) Model Approaches**

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.

The approach to analysis typically used when recurrent events are treated as identical is called the **Counting Process Approach** (Andersen et al., 1993).

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

**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

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.

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 3 events experienced by Al occurred at different survival times (from the start of initial follow-up) from the 3 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.

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**

$\left\{ \begin{array}{l} N \text{ subjects} \\ r_i \text{ time intervals for subject } i \\ \delta_{ij} \text{ event status (0 or 1) for subject } \\ \quad i \text{ in interval } j \\ t_{ij0} \text{ start time for subject } i \text{ in} \\ \quad \text{interval } j \\ t_{ij1} \text{ stop time for subject } i \text{ in} \\ \quad \text{interval } j \\ X_{ijk} \text{ value of } k\text{th predictor for} \\ \quad \text{subject } i \text{ in interval } j \\ i = 1, 2, \dots, N; j = 1, 2, \dots, n_i; \\ \quad k = 1, 2, \dots, p \end{array} \right.$

The general data layout for the counting process approach is presented in Table 8.3 for a dataset involving  $N$  subjects.

The  $i$ th subject has  $r_i$  recurrent events.  $\delta_{ij}$  denotes the event status ( $1 = \text{failure}$ ,  $0 = \text{censored}$ ) for the  $i$ th subject in the  $j$ th time interval.  $t_{ij0}$  and  $t_{ij1}$  denote the start and stop times, respectively, for the  $i$ th subject in the  $j$ th interval.  $X_{ijk}$  denotes the value of the  $k$ th predictor for the the  $i$ th subject in the  $j$ th interval.

**Table 8.3.** General Data Layout: CP Approach

		I		S		Predictors	
u	t	S		S			
b	e	t	S				
j	r	a	t	S			
e	v	t	a	t			
c	a	u	r	o	⏟		
t	l	s	t	p			
i	j	$\delta_{ij}$	$t_{ij0}$	$t_{ij1}$	$X_{ij1}$	$\dots$	$X_{ijp}$
1	1	$\delta_{11}$	$t_{110}$	$t_{111}$	$X_{111}$	$\dots$	$X_{11p}$
1	2	$\delta_{12}$	$t_{120}$	$t_{121}$	$X_{121}$	$\dots$	$X_{12p}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
1	$r_1$	$\delta_{1r_1}$	$t_{1r_10}$	$t_{1r_11}$	$X_{1r_11}$	$\dots$	$X_{1r_1p}$
.	.	.	.	.	.	.	.
i	1	$\delta_{i1}$	$t_{i10}$	$t_{i11}$	$X_{i11}$	$\dots$	$X_{i1p}$
i	2	$\delta_{i2}$	$t_{i20}$	$t_{i21}$	$X_{i21}$	$\dots$	$X_{i2p}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
i	$r_i$	$\delta_{ir_i}$	$t_{ir_i0}$	$t_{ir_i1}$	$X_{ir_i1}$	$\dots$	$X_{ir_i p}$
.	.	.	.	.	.	.	.
N	1	$\delta_{N1}$	$t_{N10}$	$t_{N11}$	$X_{N11}$	$\dots$	$X_{N1p}$
N	2	$\delta_{N2}$	$t_{N20}$	$t_{N21}$	$X_{N21}$	$\dots$	$X_{N2p}$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
N	$r_N$	$\delta_{Nr_N}$	$t_{Nr_N0}$	$t_{Nr_N1}$	$X_{Nr_N1}$	$\dots$	$X_{Nr_N p}$

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 ( $\delta_{ij} = 0$ ), then the number of recurrent events for this subject is  $r_i - 1$ ; previous time intervals, however, usually end with a failure ( $\delta_{ij} = 1$ ).

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.,  $X$ s) may be time-independent or time-dependent for a given subject. For example, if one of the  $X$ s is “gender” ( $1 = \text{female}$ ,  $0 = \text{male}$ ), 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

*Cox PH Model*

$$h(t, \mathbf{X}) = h_0(t) \exp[\sum \beta_i X_i]$$

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

The model typically used to carry out the **Counting Process** approach is the standard Cox PH model, once again shown here at the left.

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 <b>independent</b> even though several outcomes on the <b>same</b> subject	Different lines of data are treated as <b>independent</b> because they come from <b>different</b> 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

$$h(t, \mathbf{X}) = h_0(t)\exp[\beta \mathbf{tx} + \gamma_1 \mathbf{num} + \gamma_2 \mathbf{size}]$$

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**

**Table 8.5.** Ordered Failure Time and Risk Set Information for First 26 Subjects in Bladder Cancer Study

Ordered failure times $t_{(j)}$	# in risk set $n_j$	# failed $m_j$	# censored in $[t_{(j)}, t_{(j+1)})$	Subject ID #s for outcomes in $[t_{(j)}, t_{(j+1)})$
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

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 confounders. 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 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  $j$ th to the  $j + 1$ th ordered failure time provided that no subject who fails at time  $t_{(j)}$  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  $t_{(j)}$  does not drop out of the risk set after  $t_{(j)}$ .*

**Table 8.6.** Focus on Subject #s **19** and **25**

$t_{(j)}$	$n_j$	$m_j$	$q_{(j)}$	Subject ID #s
0	26	—	1	1
1	25	1	1	2, 18
<b>2</b>	<b>24</b>	2	0	<b>19, 25</b>
3	<b>24</b>	4	1	3, 13, 14, 16, 26
			•	
			•	
			•	
<b>17</b>	19	1	3	8, 9, 10, <b>25</b>
21	16	1	0	14
<b>22</b>	16	1	0	<b>25</b>
23	16	1	3	11, 12, 13, 14
24	12	1	0	15
25	11	2	0	16, 20
<b>26</b>	10	1	2	17, 18, <b>19</b>
28	7	1	4	20, 21, 22, 23, 24
<b>30</b>	3	1	2	24, <b>25</b> , 26

For example, at month  $t_{(j)} = 2$ , subject #s **19** and **25** fail, but the number in the risk set at that time ( $n_j = 24$ ) 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  $t_{(j)} = 26$  and subject #**25** has two recurrent events at months  $t_{(j)} = 17$  and  $t_{(j)} = 22$  and has additional follow-up time until censored month 30.

**Table 8.7.** Focus on Subject #s **3**, **13**, **14**, **16**, **26**

$t_{(j)}$	$n_j$	$m_j$	$q_{(j)}$	Subject ID #s
0	26	—	1	1
1	25	1	1	2, 18
2	24	2	0	19, 25
<b>3</b>	<b>24</b>	4	1	<b>3, 13, 14, 16, 26</b>
5	<b>23</b>	1	0	9
<b>6</b>	23	2	0	6, <b>26</b>
7	23	1	1	4, 15
<b>8</b>	22	1	0	<b>26</b>
<b>9</b>	22	1	0	<b>14</b>
10	22	2	2	5, 6, 12, 15
<b>12</b>	20	2	1	7, 10, <b>26</b>
<b>15</b>	19	2	0	12, <b>16</b>
<b>16</b>	19	3	0	10, <b>13</b> , 15
17	19	1	3	8, 9, 10, 25
21	16	1	0	14
22	16	1	0	25
<b>23</b>	16	1	3	11, 12, <b>13</b> , <b>14</b>
24	12	1	0	15
<b>25</b>	11	2	0	<b>16</b> , 20
26	10	1	2	17, 18, 19
28	7	1	4	20, 21, 22, 23, 24
<b>30</b>	3	1	2	24, 25, <b>26</b>

As another example from Table 8.5, subject #s **3**, **13**, **14**, **16**, **26** contribute information at ordered failure time  $t_{(j)} = 3$ , but the number in the risk set only drops from 24 to 23 even though the last four of these subjects all fail at  $t_{(j)} = 3$ . Subject #**3** is censored at month 4 (see Table 8.4), so this subject is removed from the risk set after failure time  $t_{(j)} = 3$ . However, subjects **13**, **14**, **16**, and **26** all have recurrent events after  $t_{(j)} = 3$ , so they are not removed from the risk set after  $t_{(j)} = 3$ .

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	

No Interaction Cox PH Model

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta \mathbf{tx} + \gamma_1 \mathbf{num} + \gamma_2 \mathbf{size}]$$

Partial likelihood function:

$$\mathbf{L} = \mathbf{L}_1 \times \mathbf{L}_2 \times \cdots \times \mathbf{L}_{22}$$

$L_j$  = individual likelihood at  $t_{(j)}$   
 = Pr[*failing at  $t_{(j)}$  | survival up to  $t_{(j)}$* ]  
 $j = 1, 2, \dots, 22$

$$L_j = \frac{\exp(\beta \mathbf{tx}_{(j)} + \gamma_1 \mathbf{num}_{(j)} + \gamma_2 \mathbf{size}_{(j)})}{\sum_{s \in R(t_{(j)})} \exp(\beta \mathbf{tx}_{s(j)} + \gamma_1 \mathbf{num}_{s(j)} + \gamma_2 \mathbf{size}_{s(j)})}$$

$\mathbf{tx}_{(j)}$ ,  $\mathbf{num}_{(j)}$ , and  $\mathbf{size}_{(j)}$  values of **tx**, **num**, and **size** at  $t_{(j)}$

$\mathbf{tx}_{s(j)}$ ,  $\mathbf{num}_{s(j)}$ , and  $\mathbf{size}_{s(j)}$  values of **tx**, **num**, and **size** for subject  $s$  in  $R(t_{(j)})$

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	<b>22</b>	0	1	5
25	4	0	22	30	0	1	5

$j = 15$ th ordered failure time  
 $n_{15} = 16$  subjects in risk set at  
 $t_{(15)} = 22$ :

$$R(t_{(15)} = 22) = \{\text{subject \#s } 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26\}$$

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.

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.

Each individual likelihood  $L_j$  essentially gives the conditional probability of failing at time  $t_{(j)}$ , given survival (i.e., remaining in the risk set) at  $t_{(j)}$ .

If there is only one failure at the  $j$ th ordered failure time,  $L_j$  is expressed as shown at the left for the above no-interaction model. In this expression  $\mathbf{tx}_{(j)}$ ,  $\mathbf{num}_{(j)}$ , and  $\mathbf{size}_{(j)}$  denote the values of the variables **tx**, **num**, and **size** for the subject failing at month  $t_{(j)}$ .

The terms  $\mathbf{tx}_{s(j)}$ ,  $\mathbf{num}_{s(j)}$ , and  $\mathbf{size}_{s(j)}$  denote the values of the variables **tx**, **num**, and **size** for the subject  $s$  in the risk set  $R(t_{(j)})$ . Recall that  $R(t_{(j)})$  consists of all subjects remaining at risk at failure time  $t_{(j)}$ .

For example, subject #25 from Table 8.4 failed for the third time at month 22, which is the  $j = 15$ th ordered failure time listed in Table 8.5. It can be seen that  $n_j = 16$  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.



$$\mathbf{L}_{15} = \frac{\exp(\beta(0) + \gamma_1(1) + \gamma_1(5))}{\sum_{s \text{ in } \mathbf{R}(t_{(15)})} \exp(\beta \mathbf{t}_{s(15)} + \gamma_1 \mathbf{num}_{s(15)} + \gamma_1 \mathbf{size}_{s(15)})}$$

Computer program formulates partial likelihood  $\mathbf{L}$   
(See Computer Appendix)

The corresponding likelihood  $\mathbf{L}_{15}$  at  $t_{(15)} = 22$  is shown at the left. Subject #25's values  $\mathbf{tx}_{25(15)} = 0$ ,  $\mathbf{num}_{25(15)} = 1$ , and  $\mathbf{size}_{25(15)} = 5$ , 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  $t_{(15)} = 22$ .

The overall partial likelihood  $\mathbf{L}$  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

### 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

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.

Nevertheless,

- Observations of the same subject are correlated
- Makes sense to adjust for such correlation in the analysis

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, 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  $\widehat{\mathbf{Var}}(\widehat{\beta}_k)$  where  $\widehat{\beta}_k$  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, 2002).

CP approach: assumes independence

Goal of robust estimation: adjust for correlation within subjects

Same goal for other approaches for analyzing recurrent event data

Do not adjust

$$\widehat{\beta}_k$$

Only adjust

$$\widehat{\text{Var}}(\widehat{\beta}_k)$$

*Robust (Empirical) Variance*

allows

tests of hypotheses and confidence intervals

that account for correlated data

Matrix formula:

derived from ML estimation

Formula not essential for using computer packages

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.

This is the same goal for other approaches for analyzing recurrent event data that we describe later in this chapter.

Note that the estimated regression coefficients themselves are not adjusted; only the estimated variances of these coefficients are adjusted.

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.

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, 2002).

However, the information below is not essential for an understanding of how to obtain robust estimators using computer packages. (See Computer Appendix for the code using SAS and Stata.)

Extension (Lin and Wei, 1989) of **information sandwich estimator** (Zeger and Liang, 1986)

Matrix formula

$$\widehat{\mathbf{R}}(\widehat{\beta}) = \widehat{\mathbf{Var}}(\widehat{\beta})[\mathbf{R}'_s\mathbf{R}_s]\widehat{\mathbf{Var}}(\widehat{\beta})$$

where

$$\widehat{\mathbf{Var}}(\widehat{\beta})$$

is the **information matrix**, and

$$\mathbf{R}_s$$

is matrix of **score residuals**

Formula applies to other approaches for analyzing recurrent event data

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.

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  $\mathbf{R}_s$  expression in the middle of the formula denotes the matrix of **score residuals** obtained from ML estimation.

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

**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 8.9.** 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

We now describe the results from using the **CP** approach on the Bladder Cancer Study data involving all 85 subjects.

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 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**

Nonrobust standard error for **tx**  
 = **0.2001**

*Summary of Results from  
 CP Approach*

Hazard Ratio **tx**:  $\exp(-0.407) = 0.667$   
 (= 1/1.5)

Wald Chi-Square **tx**:   robust   nonrobust  
   2.83    4.14  
 P-value **tx**:                   .09     .04

(H<sub>0</sub>: no effect of **tx**, H<sub>A</sub>: two sided)

95% CI for HR **tx** (robust): (0.414,  
 1.069)

H<sub>A</sub>: one-sided, both p-values < .05

We return to the analysis of these data when we discuss other approaches for analysis of recurrent event data.

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.

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 .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.

---

**VIII. Other Approaches—  
 Stratified Cox**

3 stratified Cox (SC) approaches:

**Conditional 1** (Prentice,  
 Williams and

**Conditional 2** Peterson,1981)

**Marginal** (Wei, Lin, and  
 Weissfeld,1989)

Goal: distinguish order of recurrent events

Strata variable: time interval #  
 treated as  
 categorical

We now describe three other approaches for analyzing recurrent event data, each of which uses a Stratified Cox (SC) PH model. They are called **conditional 1**, **conditional 2**, and **marginal**. These approaches are often used to distinguish the order in which recurrent events occur.

The “strata” variable for each approach treats the time interval number as a categorical variable.

Example:

maximum of 4 failures per subject

↓

Strata = 1 for time interval #1  
variable 2 for time interval #2  
3 for time interval #3  
4 for time interval #4

**Conditional:** time **between** two events

**Conditional 1**  $\frac{0 \quad 50 \rightarrow 80}{\text{entry}}$

**Conditional 2**  $\frac{0 \rightarrow 30}{\text{ev1} \quad \text{ev2}}$

**Marginal**

- Total survival time from study entry until kth event
- Recurrent events of different types

**Conditional 1** 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

**Conditional 2** 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

**Marginal** approach

Standard (nonrecurrent event) layout (i.e., without (start, stop) columns)

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.

Both **conditional** approaches focus on survival time **between** two events. However, **conditional 1** uses the actual times of the two events from study entry, whereas **conditional 2** starts survival time at 0 for the earlier event and stops at the later event.

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.

The **conditional 1** approach uses the exact same (start, stop) data layout format used for the **CP** approach, except that for **conditional 1**, an **SC** model is used rather than a standard (unstratified) **PH** model. The strata variable here is **int** in this listing.

The **conditional 2** 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.

The **marginal** approach uses the standard (non-recurrent 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**, **conditional 1**, and **conditional 2** approaches

**Marginal** approach

Each subject at risk for all failures that might occur

# 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

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.

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 (**conditional 1**, **conditional 2**, and **marginal**) fundamentally differ in the way the risk set is determined for strata corresponding to events after the first event.

**Conditional 2:** time until 1st event does not influence risk set for later events (i.e., clock reset to 0 after event occurs)

With **conditional 2**, 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.

**Conditional 1:** time until 1st event influences risk set for later events

In contrast, with **conditional 1**, 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  $k$ th event ( $k = 1, 2, \dots$ ) identifies those at risk for the  $k$ th event since entry into the study.

EXAMPLE					
ID	Status	Stratum	Days		tx
			Start	Stop	
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.

### Conditional 1

Stratum 1			Stratum 2		
$t_{(j)}$	$n_j$	$R(t_{(j)})$	$t_{(j)}$	$n_j$	$R(t_{(j)})$
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 **conditional 1** 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  $t = 105$ .

### Conditional 2

Stratum 1			Stratum 2		
$t_{(j)}$	$n_j$	$R(t_{(j)})$	$t_{(j)}$	$n_j$	$R(t_{(j)})$
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 **conditional 2** approach is used. Notice that the data for **stratum 1** are identical to those for **conditional 1**. 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		
$t_{(j)}$	$n_j$	$R(t_{(j)})$	$t_{(j)}$	$n_j$	$R(t_{(j)})$
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}

Stratum 3 for Marginal approach follows

**Marginal  
Stratum 3**

$t_{(j)}$	$n_j$	$R(t_{(j)})$
0	3	{M, H, P}
85	2	{M, P}

Note: H censored by  $t = 85$

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)

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)

No-interaction SC model for bladder cancer data

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta \mathbf{t}\mathbf{x} + \gamma_1 \text{num} + \gamma_2 \text{size}]$$

where  $g = 1, 2, 3, 4$

We next consider the **marginal** approach. For **stratum 1**, the data are identical again to those for **conditional 1**. 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  $t = 60$ .

Because Polly experienced three events, there is also a third stratum for this example, which we describe for the marginal approach only.

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  $t = 105$ .

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.

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.

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*

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_g \mathbf{tx} + \gamma_{1g} \mathbf{num} + \gamma_{2g} \mathbf{size}]$$

$$g = 1, 2, 3, 4$$

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*

$$\begin{aligned} h_g(t, \mathbf{X}) = h_{0g}(t) \exp[ & \beta \mathbf{tx} + \gamma_1 \mathbf{num} \\ & + \gamma_2 \mathbf{size} + \delta_{11}(Z_1^* \times \mathbf{tx}) \\ & + \delta_{12}(Z_2^* \times \mathbf{tx}) + \delta_{13}(Z_3^* \times \mathbf{tx}) \\ & + \delta_{21}(Z_1^* \times \mathbf{num}) + \delta_{22}(Z_2^* \times \mathbf{num}) \\ & + \delta_{23}(Z_3^* \times \mathbf{num}) + \delta_{31}(Z_1^* \times \mathbf{size}) \\ & + \delta_{32}(Z_2^* \times \mathbf{size}) + \delta_{33}(Z_3^* \times \mathbf{size})] \end{aligned}$$

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  $Z_3^*$ .

where  $Z_1^*$ ,  $Z_2^*$ , and  $Z_3^*$  are 3 dummy variables for the 4 strata.

$H_0(\text{Version 1})$

$$\begin{aligned} \beta_1 = \beta_2 = \beta_3 = \beta_4 & \equiv \beta, \\ \gamma_{11} = \gamma_{12} = \gamma_{13} = \gamma_{14} & \equiv \gamma_1, \\ \gamma_{21} = \gamma_{22} = \gamma_{23} = \gamma_{24} & \equiv \gamma_2 \end{aligned}$$

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.

$H_0(\text{Version 2})$

$$\begin{aligned} \delta_{11} = \delta_{12} = \delta_{13} = \delta_{21} = \delta_{22} = \delta_{23} \\ = \delta_{31} = \delta_{32} = \delta_{33} = 0 \end{aligned}$$

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

Recommend using

### robust estimation

$$\widehat{\mathbf{R}}(\widehat{\boldsymbol{\beta}}) = \widehat{\mathbf{Var}}(\widehat{\boldsymbol{\beta}})[\mathbf{R}'_s \mathbf{R}_s] \widehat{\mathbf{Var}}(\widehat{\boldsymbol{\beta}})$$

to adjust for correlation of observations on the same subject

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.

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)

**Table 8.10.** Estimated  $\beta$ s and HRs for **tx** from Bladder Cancer Data

Model	$\hat{\beta}$	$\hat{\text{HR}} = \exp(\hat{\beta})$
CP	-0.407	0.666 (=1/1.50)
C1	-0.334	0.716 (=1/1.40)
C2	-0.270	0.763 (=1/1.31)
M	-0.580	0.560 (=1/1.79)

CP = Counting Process, C1 = Conditional 1, C2 = Conditional 2, M = Marginal

HR for **M: 0.560** (=1/1.79) differs from

HRs for **CP: 0.666** (=1/1.50),  
**C1: 0.716** (=1/1.40),  
**C2: 0.763** (=1/1.31)

We now present and compare SAS results from using all four methods we have described—**CP**, **conditional 1**, **conditional 2**, and **marginal**—for analyzing the recurrent event data from the bladder cancer study.

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.

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  $\beta$ s,  $SE(\beta)$ 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
C1	-0.334	0.216	0.197	.122	.090
C2	-0.270	0.208	0.208	.195	.194
M	-0.580	0.201	0.303	.004	.056

CP = Counting Process, C1 = Conditional 1, C2 = Conditional 2, M = Marginal, NR = Nonrobust, R = Robust, P = Wald P-value

**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**  
**C1: P(NR) = .122 > P(R) = .090**  
**C2: P(NR) = .195 = P(R) = .194**

Wald test statistic(s):

$$Z = \hat{\beta}/SE(\hat{\beta}) \Leftrightarrow Z^2 = [\hat{\beta}/SE(\hat{\beta})]^2 \sim N(0,1) \text{ under } H_0: \beta = 0 \sim \chi^2_{1 \text{ df}}$$

**Table 8.12.** Estimated  $\beta$ s and Robust  $SE(\beta)$ s for **tx** from Interaction SC Model for Bladder Cancer Data

Model	Interaction SC Model				No Interaction
	Str1 $\hat{\beta}_1$ (SE)	Str2 $\hat{\beta}_2$ (SE)	Str3 $\hat{\beta}_3$ (SE)	Str4 $\hat{\beta}_4$ (SE)	$\hat{\beta}$ (SE)
CP	—	—	—	—	-.407 (.242)
C1	-.518 (.308)	-.459 (.441)	.117 (.466)	-.041 (.515)	-.334 (.197)
C2	-.518 (.308)	-.259 (.402)	.221 (.620)	-.195 (.628)	-.270 (.208)
M	-.518 (.308)	-.619 (.364)	-.700 (.415)	-.651 (.490)	-.580 (.303)

CP = Counting Process, C1 = Conditional 1, C2 = Conditional 2, M = Marginal

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.

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.

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 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 3 approaches (other than the CP approach) that use a SC model.

Notice that for each of the 3 SC modeling approaches, the estimated  $\beta$ s and corresponding standard errors are different over the four strata as well as for the no-interaction model. For example, using the **conditional 1** 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*

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_g \mathbf{tx} + \gamma_{1g} \mathbf{num} + \gamma_{2g} \mathbf{size}]$$

$g = 1, 2, 3, 4$

Note: subscript  $g$  allows for different regression coefficients for each stratum

**Conditional 1** for Subject 10

id	int	event	start	stop	tx	num	size
10	1	1	0	12	0	1	1

**Conditional 2** 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

**Marginal** Approach for Subject 10

id	int	event	stime	tx	num	size
10	1	1	12	0	1	1

**Note:** int = stratum #

Marginal approach

**start** time = 0 always

**stop** time = **stime**

Subject #10: (start, stop) = (0, 12)

*Bladder Cancer Study*

1. Which approach is best?
2. Conclusion about **tx**?

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.

Notice also that for stratum 1 the estimated  $\beta$  and its standard error are identical ( $-0.518$  and  $0.308$ , resp.) for the **conditional 1**, **conditional 2**, 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 3 SC approaches and does not start to differ until stratum 2.

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 **conditional 1** and **2** data for this subject.

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

If event order important:

Choose from the 3 SC approaches.

**Conditional 1:** time of recurrent  
 event from entry  
 into the study

**Conditional 2:** Use time from  
 previous event to  
 next recurrent  
 event

**Marginal:** Consider strata as  
 representing different  
 event types

**Conditional 1** versus **Marginal**  
 (subtle choice)

Recommend: Choose conditional 1  
 unless strata  
 represent different  
 event types

What do we conclude about **tx**?

Conclusions based on results from  
**CP** and **conditional 1** approaches

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, 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?

The **conditional 1** 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.

The **conditional 2** 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.

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.

We (the authors) consider the choice between the **conditional 1** and **marginal** approaches to be quite subtle. We prefer **conditional 1**, 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.

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 **conditional 1** 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	Conditional 1
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 **conditional 1** 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 **conditional 1** 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:  $\widehat{HR}(\mathbf{CP}) = 0.667 = 1/1.50$ ,  $\widehat{HR}(\mathbf{C1}) = 0.716 = 1/1.40$ . This effect is borderline nonsignificant (2-sided tests):  $P(\mathbf{CP}) = .09 = P(\mathbf{C1})$ . 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

We now present a parametric approach

- Weibull PH model
- Gamma shared frailty component
- Bladder Cancer dataset
  - Data layout for the counting process approach

Can review Chapter 7  
Weibull model (Section VI)  
Frailty models (Section XII)

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.

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).

Hazard conditioned on frailty  $\alpha_k$

$$h_k(t|\alpha, \mathbf{X}_{jk}) = \alpha_k h(t|\mathbf{X}_{jk})$$

where  $\alpha \sim \text{gamma}(\mu = 1, \text{var} = \theta)$   
 and where  $h(t|\mathbf{X}_{jk}) = \lambda_{jk} p t^{p-1}$   
 (Weibull) with  $\lambda_{jk} = \exp(\beta_0 + \beta_1 \mathbf{tx}_{jk} + \beta_2 \mathbf{num}_{jk} + \beta_3 \mathbf{size}_{jk})$

We define the model in terms of the hazard of the  $j$ th outcome occurrence for the  $k$ th subject conditional on his or her frailty  $\alpha_k$ . The frailty is a multiplicative random effect on the hazard function  $h(t|\mathbf{X}_{jk})$ , assumed to follow a gamma distribution of mean 1 and variance theta  $\theta$ . We assume  $h(t|\mathbf{X}_{jk})$  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

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		

The model output (obtained using Stata version 7) 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.

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.

Likelihood ratio test of theta = 0:  
 chibar(01) = 7.34  
 Prob >= chibar2 = 0.003

*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  $\mathbf{tx}$  (robust): (0.414, 1.069)

Interpretations of HR from frailty model

- Compares 2 individuals with same  $\alpha$
- Compares individual with himself
  - What is effect if individual had used treatment rather than placebo?

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  $\mathbf{tx}$ . 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).

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

$\mathbf{tx} = 1$  if treatment, 0 if placebo

8 years of follow-up

The exposure variable of interest was treatment group ( $\mathbf{tx}$ ), 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 eight 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 meters, 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

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**.

Each model contains **tx**, **age**, and **sex**.

### CP model

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta \mathbf{tx} + \gamma_1 \mathbf{age} + \gamma_2 \mathbf{sex}]$$

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.

### No-interaction SC model

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta \mathbf{tx} + \gamma_1 \mathbf{age} + \gamma_2 \mathbf{sex}]$$

where  $g = 1, 2$

### Interaction SC model:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_g \mathbf{tx} + \gamma_{1g} \mathbf{age} + \gamma_{2g} \mathbf{sex}]$$

where  $g = 1, 2$

**Table 8.14.** Comparison of Parameter Estimates and Robust Standard Errors for Treatment Variable (tx) Controlling for Age and Sex (Macular Degeneration Data)

Model	"Interaction" Cox stratified model		"No-interaction" SC model
	Stratum 1 $\hat{\beta}_1$ (SE)	Stratum 2 $\hat{\beta}_2$ (SE)	$\hat{\beta}$ (SE)
Counting process	n/a	n/a	-0.174 (0.104)
Cond 1	-0.055 (0.286)	-0.955 (0.443)	-0.306 (0.253)
Cond 2	-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

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 "no-interaction" stratified Cox models for the three approaches other than the counting process approach.

Notice that the estimated coefficients for  $\beta_1$  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  $\beta_2$  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.

**Table 8.15.** Comparison of Results for the Treatment Variable (tx) Obtained for Conditional 1 and Marginal Approaches (Macular Degeneration Data)

		Cond 1	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
	$H_0: \beta_2 = 0$	<b>4.6395</b>	<b>3.4281</b>
	$H_0: \beta = 0$	1.4569	1.0609
P-value	$H_0: \beta_1 = 0$	0.8458	0.8478
	$H_0: \beta_2 = 0$	<b>0.0312</b>	<b>0.0641</b>
	$H_0: \beta = 0$	0.2274	0.3030
Hazard ratio	exp( $\hat{\beta}_1$ )	0.946	0.946
	exp( $\hat{\beta}_2$ )	<b>0.385</b>	<b>0.423</b>
	exp( $\hat{\beta}$ )	0.736	0.742
95% Conf. interval	exp( $\hat{\beta}_1$ )	(0.540, 1.656)	(0.540, 1.656)
	exp( $\hat{\beta}_2$ )	<b>(0.161, 0.918)</b>	<b>(0.170, 1.052)</b>
	exp( $\hat{\beta}$ )	(0.448, 1.210)	(0.420, 1.310)

In Table 8.15, we summarize the statistical inference results for the effect of the treatment variable (**tx**) for the **conditional 1** 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 **conditional 2** 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.

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 **conditional 1** and **marginal** approaches give the identical result that the estimated treatment effect ( $\widehat{HR} = 0.946 = 1/1.06$ ) is neither meaningful nor significant ( $P = 0.85$ ).

For the *second event*, indicating a clinically severe stage of macular degeneration, the Wald statistic P-value for the **conditional 1** 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 **conditional 1** 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 **conditional 1** 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 **conditional 1** or **marginal** approaches to the analysis.

Conclusions regarding 2nd event:

- Clinically moderate and statistically significant treatment effect
- Similar for **conditional 1** and **marginal** approaches, but more support from **conditional 1** 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 **conditional 1** analysis than from the **marginal** analysis.

Comparison of **conditional 1** with **marginal** Approach

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.

What if results had been different?

Recommend **conditional 1** if

Can assume 2nd event cannot occur without 1st event previously occurring

↓

Should consider survival time to 2nd event **conditional on** experiencing 1st event

Nevertheless, we authors find it difficult to make such a decision, even for this example. The **conditional 1** 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.

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

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.

Macular degeneration data: recommend **marginal** approach

**In general:** carefully consider interpretation of each 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?

Focus on one ordered event at a time

$S_1(t)$  : 1st event

$S_2(t)$  : 2nd event

...

$S_k(t)$  : kth event

### 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)

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?

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.

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

Two versions

**Conditional:**

$T_{2c}$  = 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

**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

Two versions

**Conditional:**

$T_{kc}$  = 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

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.

There are two possible versions for such a plot.

**Conditional:** 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.

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.

As with survival to the second event, there are two possible versions, **conditional** or **marginal**, for such a plot, as shown on the left.

ID	Status	Stratum	Days		tx
			Start	Stop	
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**

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$
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

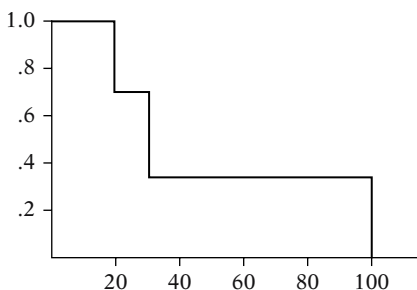
Deriving  $S_{2c}(t)$ : **Stratum 2 (Conditional)**

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_{2c}(t_{(j)})$
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)**

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_{2m}(t_{(j)})$
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

*Survival Plots for Molly, Holly  
and Polly Recurrent Event  
Data ( $n = 3$ )*



**Figure 8.1.**  $S_1(t)$ : Survival to 1st Event

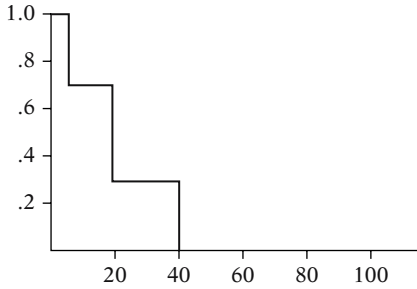
The survival plot for survival to the first event  $S_1(t)$  is derived from the stratum 1 data layout for any of the three alternative SC analysis approaches. Recall that  $m_j$  and  $q_j$  denote the number of failures and censored observations at time  $t_{(j)}$ . The survival probabilities in the last column use the KM product limit formula.

The **conditional** survival plot for survival to the second event is derived from the stratum 2 data layout for the **conditional 2** approach. We denote this survival curve as  $S_{2c}(t)$ . Notice that the survival probabilities here are identical to those in the previous table; however, the failure times  $t_{(j)}$  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  $S_{2m}(t)$ . Again, the last column here is identical to those in the previous two tables, but, once again, the failure times  $t_{(j)}$  in each table are different.

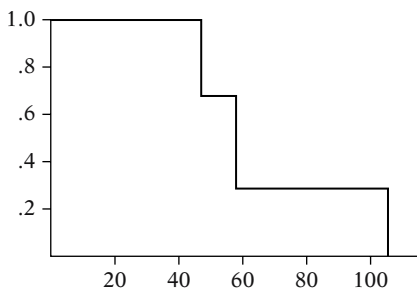
The survival plots that correspond to the above three data layouts are shown in Figures 8.1 to 8.3.

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 (Conditional)

Figure 8.2 shows **conditional** 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),  
conditional 1, conditional 2  
marginal**

The 4 approaches

- Differ in how risk set is determined
- Differ in data layout
- All involve standard Cox model program
- Latter 3 approaches use a SC model

Identical recurrent events



**CP approach**

We have described four approaches for analyzing recurrent event survival data.

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.

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

**CP** approach: **Start** and **Stop** times

Standard layout: only **Stop** (survival) times (no recurrent events)

**Conditional 1:** same **Start** and **Stop** Times as **CP**, but uses **SC** model

**Conditional 2:** **Start** and **Stop** Times  
**Start** = 0 always  
**Stop** = time since previous event  
**SC** model

**Marginal** approach:

Standard layout (nonrecurrent event), that is, without (**Start**, **Stop**) columns

Each failure is a separate process

Recommend using **robust estimation** to adjust for correlation of observations on the same subject.

**Application 1:** Bladder Cancer study  
 n = 86  
 64 months of follow-up

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.

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.

The **conditional 1** approach uses the exact same (start, stop) data layout format used for the **CP** approach, except that for **conditional 1**, the model used is a SC PH model rather than an unstratified PH model.

The **conditional 2** 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.

The **marginal** approach uses the standard (non-recurrent 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.

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.

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

**tx** = 1 if thiotepa, 0 if placebo  
**num** = initial # of tumors  
**size** = initial size of tumors

**CP** results: no strong evidence for **tx** (HR = 0.67, P = .09, 95% CI: 0.414, 1.069)

Alternative parametric approach

- Weibull PH model
- Gamma shared frailty component
- Bladder cancer dataset
- Similar HR and confidence interval as for counting process approach

**Application 2:** Clinical trial

n = 43  
 8 years of follow-up  
 High doses of antioxidants and zinc  
 Age-related macular degeneration

Exposure: **tx** = 1 if treatment, 0 if placebo  
 Covariates: **age**, **sex**

Two possible events:

1st event: visual acuity score <50 (i.e., poor vision)  
 2nd event: clinically advanced severe stage of macular degeneration

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.

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**).

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** (HR = 0.67, P = .09, 95% CI: 0.414, 1.069)

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.

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 eight years.

The exposure variable of interest was treatment group (**tx**). Covariates considered were **age** and **sex**.

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 <b>conditional 1</b> vs. <b>marginal</b> (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 <b>conditional 1</b> and <b>marginal</b> 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 <ul style="list-style-type: none"> <li>• No treatment effect</li> <li>• Same for <b>conditional 1</b> and <b>marginal</b> approaches</li> </ul>	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 <b>conditional 1</b> or <b>marginal</b> approaches to the analysis.
Conclusions regarding 2nd event <ul style="list-style-type: none"> <li>• Clinically moderate and statistically significant treatment effect</li> </ul>	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 <b>marginal</b> approach (but not clear-cut)	The choice between the conditional 1 and marginal approaches for these data was not clear-cut, although the <b>Marginal</b> approach was perhaps more appropriate because the two events were of very different types.
<b>In general:</b> 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: <b>Conditional:</b> only subjects with $k - 1$ events <b>Marginal:</b> 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 <b>conditional</b> or <b>marginal</b> plot, which typically differ.

## Detailed Outline

- I. **Overview** (page 334)
  - 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.
- II. **Examples of Recurrent Event Data** (pages 334–336)
  - 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 336–337)
  - 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 for Counting Process

##### Approach (pages 338–339)

- A.  $r_i$  time intervals for subject  $i$ .  
 $\delta_{ij}$  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  $k$ th predictor for subject  $i$  in interval  $j$ .  
 $i = 1, 2, \dots, N$ ;  $j = 1, 2, \dots, n_i$ ;  $k = 1, 2, \dots, p$ .
- B. Layout for subject  $i$ :

$i$	$j$	$\delta_{ij}$	$t_{ij0}$	$t_{ij1}$	$X_{ij1} \dots X_{ijp}$
$i$	1	$\delta_{i1}$	$t_{i10}$	$t_{i11}$	$X_{i11} \dots X_{i1p}$
$i$	2	$\delta_{i2}$	$t_{i20}$	$t_{i21}$	$X_{i21} \dots X_{i2p}$
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
$i$	$r_i$	$\delta_{ir_i}$	$t_{ir_i0}$	$t_{ir_i1}$	$X_{ir_i1} \dots X_{ir_i p}$

- 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 ( $\mathbf{tx}$ , 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 340–344)

- A. The model typically used to carry out the **Counting Process (CP)** approach is the standard Cox PH model:  $h(t, \mathbf{X}) = h_0(t) \exp[\sum \beta_i X_i]$ .
- 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  $t_{(j)}$  does not drop out of the risk set after  $t_{(j)}$  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

$$h(t, \mathbf{X}) = h_0(t)\exp[\beta \mathbf{tx} + \gamma_1 \mathbf{num} + \gamma_2 \mathbf{size}].$$

- 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 344–346)

- 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 346–347)

- 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** (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 347–353)

- 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. **Conditional 1** approach:
  1. Same **Start** and **Stop** Times as **CP**.
  2. **SC** model.
- E. **Conditional 2** 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:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta \mathbf{tx} + \gamma_1 \mathbf{num} + \gamma_2 \mathbf{size}]$ . where  $g = 1, 2, 3, 4$ .

Interaction model:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_g \mathbf{tx} + \gamma_{1g} \mathbf{num} + \gamma_{2g} \mathbf{size}]$ . 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 353–357)

- A. Results from using all four methods—**CP**, **conditional 1**, **conditional 2**, 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 C1: 0.716 C2: 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  $\beta$ s and corresponding standard errors are different over the four strata (i.e., four events) for each model separately.
- E. The estimated  $\beta$ '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 **conditional 1** and **marginal** is difficult, but prefer **conditional 1** 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 357–359)

- 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 359–364)

- A. Clinical trial ( $n = 43$ , 8-year study) on effect of using high doses of antioxidants and zinc (i.e.,  $\mathbf{tx} = 1$  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 **conditional 1** 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 **conditional 1** and **marginal** approaches.
- G. Conclusions regarding 2nd event:
  1. **Conditional 1**:  $HR = .385 = 1/2.60$ , two-sided  $P = .03$ .
  2. **Marginal**:  $HR = .423 = 1/2.36$ , two-sided  $P = .06$ .
  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 364–367)

- 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  $k$ th event, the survival curve is given by  $S_k(t) = \Pr(T_k > t)$  where  $T_k =$  survival time up to occurrence of the  $k$ th event).

- D. Two versions for  $S_k(t)$ :
- i.  **$S_{kc}(t)$  conditional:**  $T_{kc}$  = time from the  $k - 1$ st to  $k$ th event, restricting data to subjects with  $k - 1$  events.
  - ii.  **$S_{km}(t)$  marginal:**  $T_{km}$  = time from study entry to  $k$ th 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 367–370)
- A. Four approaches for analyzing recurrent event survival data: the **counting process (CP)**, **conditional 1**, **conditional 2**, 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  $t_{(j)}$  does not drop out of the risk set after  $t_{(j)}$ .
- 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  $t_{(j)} = 10$ , 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 **conditional 1** 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 **conditional 1** 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 **conditional 1** and the **marginal** approaches, the **marginal** approach would be preferred provided the different strata clearly represent different event types.

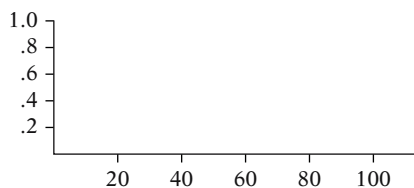
- T F 14. When using an interaction SC model to analyze recurrent event data, the estimated regression coefficients and corresponding standard errors for the first stratum always will be identical for the **conditional 1**, **conditional 2**, and **marginal** approaches.
- T F 15. The choice among the **CP**, **conditional 1**, **conditional 2**, 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  $m_j$  and  $q_j$  denote the number of failures and censored observations at time  $t_{(j)}$ . The survival probabilities in the last column use the KM product limit formula.

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

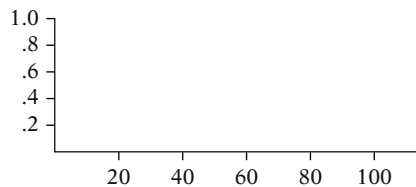
17. Plot the survival curve that corresponds to the data layout obtained for Question 16.



18. Fill in the following data layout describing survival (in weeks) **from the first to second event** using the **conditional** approach:

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$
0	3	0	0	{A, S, C}	1.00
10	-	-	-	-	-
-	-	-	-	-	-
-	-	-	-	-	-

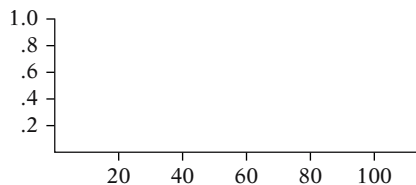
19. Plot the survival curve that corresponds to the data layout obtained for Question 18.



20. Fill in the following data layout describing survival (in weeks) **to the second event** using the **marginal** approach:

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$
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**

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 ( $\mathbf{tx}$ , = 1 if treatment A, = 0 if treatment B) to reduce their risk for future heart attacks over a four-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

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 (**conditional 1**, **conditional 2**, **marginal** approaches) that controlled for the covariate **smoking**.

**Table T.1.** Comparison of Results for the Treatment Variable (**tx**) Obtained from No-Interaction Models<sup>a</sup> Across Four Methods (Defibrillator Study)

Model	<b>CP</b>	<b>Conditional 1</b>	<b>Conditional 2</b>	<b>Marginal</b>
Parameter estimate <sup>b</sup>	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)

<sup>a</sup> No-interaction SC model fitted with PROC PHREG for the **conditional 1**, **conditional 2** and **marginal** methods; no-interaction standard Cox PH model fitted for **CP** approach.

<sup>b</sup> Estimated coefficient of **tx** variable.

1. State the hazard function formula for the no-interaction model used to fit the **CP** approach.
2. Based on the **CP** approach, what do you conclude about the effect of treatment (tx)? Explain briefly using the results in Table T.1.
3. 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.
4. 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).
5. 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?
6. Based on the study description given earlier, why does it make sense to recommend the **CP** approach over the other alternative approaches?
7. 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 $t_{(j)}$	# in risk set $n_j$	# failed $m_j$	# censored in $[t_{(j)}, t_{(j+1)})$	Subject ID #s for outcomes in $[t_{(j)}, t_{(j+1)})$
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)



**Table T.2.** (Continued)

Ordered failure times $t_{(j)}$	# in risk set $n_j$	# failed $m_j$	# censored in $[t_{(j)}, t_{(j+1)})$	Subject ID #s for outcomes in $[t_{(j)}, t_{(j+1)})$
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

8. In Table T.2, why does the number in the risk set ( $n_j$ ) remain unchanged through failure time (i.e., day) 68, even though 50 events occur up to that time?
9. Why does the number in the risk set change from 31 to 26 when going from time 96 to 97?
10. Why is the number of failures ( $m_j$ ) equal to 3 and the number of censored subjects equal to 1 in the interval between failure times 79 and 80?
11. What 5 subjects were censored in the interval between failure times 111 and 112?
12. 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

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)})\Pr(T > t_{(j)}   T \geq t_{(j)})$
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	$.22 \times 34/35 = .22$

(Continued)

**Table T.3.** (Continued)

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)})\Pr(T > t_{(j)} T \geq t_{(j)})$
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	$.17 \times 33/35 = .16$
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	$.03 \times 1/1 = .03$

13. Suppose the survival probabilities shown in Table T.3 are plotted on the y-axis versus corresponding ordered failure times on the x-axis.
- What is being plotted by such a curve? (Circle one or more choices.)
    - $\Pr(T_1 > t)$  where  $T_1 =$  time to first event from study entry.
    - $\Pr(T > t)$  where  $T =$  time from any event to the next recurrent event.
    - $\Pr(T > t)$  where  $T =$  time to any event from study entry.
    - $\Pr(\text{not failing prior to time } t)$ .
    - None of the above.
  - Can you criticize the use of the product limit formula for  $S(t_{(j)})$  in Table T.3? Explain briefly.

14. Use Table T.2 to complete the data layouts for plotting the following survival curves.

a.  $S_1(t) = \Pr(T_1 > t)$  where  $T_1 =$  time to first event from study entry

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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	-	-	-	-

b. **Conditional**  $S_{2c}(t) = \Pr(T_{2c} > t)$  where  $T_{2c} =$  time to second event from first event.

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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)

*(Continued)*

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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	-	-	-	-

- c. **Marginal**  $S_{2m}(t) = \Pr(T_{2m} > t)$  where  $T_{2m}$  = time to second event from study entry.

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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	-	-	-	-

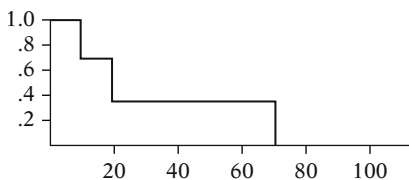
15. The survival curves corresponding to each of the data layouts (a, b, c) described in Question 14 will be different. Why?

## 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**, **conditional 1**, and **conditional 2** 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 **conditional 1** or **conditional 2**. 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**, **conditional 1**, **conditional 2**, and **marginal** approaches depends on carefully considering the interpretation of each approach.

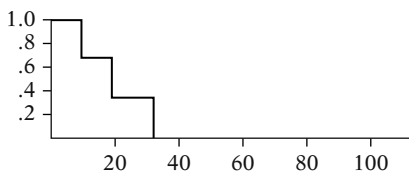
16. $t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$
0	3	0	0	{A, S, C}	1.00
10	3	1	0	{A, S, C}	0.67
20	2	1	0	{A, S}	0.33
70	1	1	0	{A}	0.00

17.  $S_1(t)$



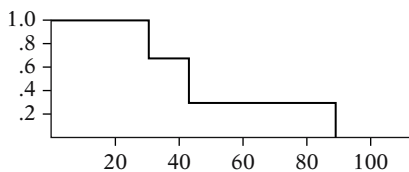
18. $t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$ Conditional
0	3	0	0	{A, S, C}	1.00
10	3	1	0	{A, S, C}	0.67
20	2	1	0	{A, C}	0.33
30	1	1	0	{C}	0.00

19.  $S_{2c}(t)$  Conditional



20. $t_{(j)}$	$n_j$	$m_j$	$q_j$	$R(t_{(j)})$	$S_1(t_{(j)})$ 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.  $S_{2m}(t)$  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**

---



## 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 396)**
- II. **Examples of Competing Risks Data (pages 396–398)**
- III. **Byar Data (pages 399–400)**
- IV. **Method 1—Separate Models for Different Event Types (pages 400–403)**
- V. **The Independence Assumption (pages 403–411)**
- VI. **Cumulative Incidence Curves (CIC) (pages 412–420)**
- VII. **Conditional Probability Curves (CPC) (pages 420–421)**
- VIII. **Method 2—Lunn–McNeil (LM) Approach (pages 421–427)**
- IX. **Method 2a—Alternative Lunn–McNeil (LM<sub>alt</sub>) Approach (pages 427–430)**
- X. **Method 1—Separate Models versus Method 2—LM Approach (pages 431–434)**
- XI. **Summary (pages 434–439)**

## Objectives

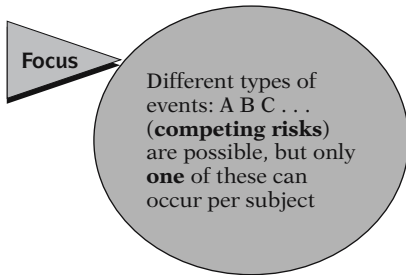
Upon completing this chapter, the learner should be able to:

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 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

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

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.

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

$X_1, X_2, \dots, X_p \Rightarrow$  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)*

$HR_{LC}(E \text{ vs. not } E) = 1?$   
(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.

$HR(LC \text{ vs. Stroke}) = 1?$   
(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)*

$HR_{HOSPINF}(E \text{ vs. not } E) = 1?$   
(allowing for competing risk from surgery)

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.

Note: death from surgery reduces number of hospital infections to be treated

*Accidental Death vs. Combat Death (3)*

$HR_{COMBAT}(E \text{ vs. not } E)$   
(allowing competing risk of accidental 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.

Suppose entire company dies at accident time  $t$  before entering combat

↓

$$S_{COMBAT}(t) = P(T_{COMBAT} > t) = 1$$

where  $T_{COMBAT}$  = time to combat death

However,

$T_{C+A}$  = combat or accidental death

↓

“event free”  $S_{C+A}(t) = P(T_{C+A} > t) = 0$

Moreover,

$S_{KM}(T_{COMBAT} > t)$

is undefined because the risk set is empty at time  $t$

Competing Risks Data Survival Curve Interpretation?

*Limb sarcoma patients (4)*

Competing risks

1 = local recurrence, 2 = lung metastasis, or 3 = other metastasis

$HR_c(E \text{ vs. not } E)$ ,  $c = 1, 2, 3$   
(allowing for competing risk from other two failure types)

$HR(\text{Lung Metastasis vs. Local Recurrence})$ ? Controlling for Predictors

No failure types involve death

↓

Recurrent events possible

But can use classical competing risk methods if focus on **only first failure**

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, the KM survival probability for combat death at time  $t$  is undefined because no one was at risk for a combat death at time  $t$ .

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).

In the fourth example involving limb sarcoma patients, the competing risks are the three failure types shown at the left.

In this study, the investigators wanted hazard ratios for each failure type, allowing for competing risks from the other two failure types.

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.

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.

### 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

Competing risks: deaths from

**Cancer** (main focus)

**CVD**

**Other**

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	≤74 years
	1	75–79 years
	2	≥80 years
Standardized <sup>a</sup> weight (Wt)	0	≥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	≥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 <sup>2</sup>
	1	≥30 cm <sup>2</sup>
Gleason Score <sup>+</sup> (SG)	0	≤10
	1	>10

<sup>a</sup> weight (kg) – height (cm) + 200

<sup>+</sup> index of tumor invasiveness/aggressiveness

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).

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.

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

If only one failure type of interest

↓

Estimate only one hazard or HR

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 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

$$h_{\mathbf{c}}(t) = \lim_{\Delta t \rightarrow 0} P(t \leq T_{\mathbf{c}} < t + \Delta t | T_{\mathbf{c}} \geq t) / \Delta t$$

where  $T_{\mathbf{c}}$  = time-to-failure from event  $\mathbf{c}$

$\mathbf{c} = 1, 2, \dots, C$  (# 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  $\mathbf{c}$ . Thus,  $h_{\mathbf{c}}(t)$  gives the instantaneous failure rate at time  $t$  for event type  $\mathbf{c}$ , given not failing from event  $\mathbf{c}$  by time  $t$ .

**Cox PH cause-specific model**  
(event-type  $\mathbf{c}$ ):

$$h_{\mathbf{c}}(t, \mathbf{X}) = h_{0\mathbf{c}}(t) \exp\left[\sum_{i=1}^p \beta_{i\mathbf{c}} X_i\right],$$

$\mathbf{c} = 1, \dots, C$

Using a Cox PH model that considers predictors  $\mathbf{X} = (X_1, X_2, \dots, X_p)$ , the **cause-specific hazard model** for event type  $\mathbf{c}$  has the form shown at the left. Note that  $\beta_{i\mathbf{c}}$ , the regression coefficient for the  $i$ th predictor, is subscripted by  $\mathbf{c}$  to indicate that the effects of the predictors may be different for different event types.

$\beta_{i\mathbf{c}}$  allows effect of  $X_i$  to differ by event-type

BYAR DATA EXAMPLE					
Competing Risks: <b>Cancer, CVD, Other</b>					
Cause-specific model: <b>Cancer</b>					
No-interaction model:					
$h_{Ca}(t, \mathbf{X}) = h_{0Ca}(t) \exp[\beta_{1Ca}Rx + \beta_{2Ca}Age + \beta_{3Ca}Wt + \beta_{4Ca}PF + \beta_{5Ca}Hx + \beta_{6Ca}HG + \beta_{7Ca}SZ + \beta_{8Ca}SG]$					
$HR_{Ca}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1Ca}]$					
<b>CVD</b> and <b>Other</b> deaths are censored					
Cause-specific model: <b>CVD</b>					
$h_{CVD}(t, \mathbf{X}) = h_{0CVD}(t) \exp[\beta_{1CVD}Rx + \beta_{2CVD}Age + \beta_{3CVD}Wt + \beta_{4CVD}PF + \beta_{5CVD}HX + \beta_{6CVD}HG + \beta_{7CVD}SZ + \beta_{8CVD}SG]$					
$HR_{CVD}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1CVD}]$					
<b>Cancer</b> and <b>Other</b> are censored					
Cause-specific model: <b>Other</b>					
$H_{OTH}(t, \mathbf{X}) = h_{0OTH}(t) \exp[\beta_{1OTH}Rx + \beta_{2OTH}Age + \beta_{3OTH}Wt + \beta_{4OTH}PF + \beta_{5OTH}HX + \beta_{6OTH}HG + \beta_{7OTH}SZ + \beta_{8OTH}SG]$					
<b>Cancer</b> and <b>CVD</b> are 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

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  $\exp[\beta_{1Ca}]$ .

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.

$$\widehat{HR}_{Ca}(RX = 1 \text{ vs. } RX = 0)$$

$$= \exp(-0.550) = 0.577$$

$$\text{Wald ChiSq} = (-.550/.170)^2$$

$$= 10.345 \text{ (} P = 0.001 \text{)}$$

**Signif. below .01 level**

95% CI for  $\exp[\beta_{1Ca}]$ :

$$\exp[-0.550 \pm 1.96(0.170)]$$

$$= (0.413, 0.807)$$

**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

$$\widehat{HR}_{CVD}(RX = 1 \text{ vs. } RX = 0)$$

$$= \exp(0.354) = 1.425$$

$$\text{Wald ChiSq} = (.354/.174)^2$$

$$= 4.220 \text{ (} P = 0.042 \text{)}$$

**Signif. at .05 level**

95% CI for  $\exp[\beta_{1CVD}]$ :

$$\exp.[0.354 \pm 1.96(0.174)]$$

$$= (1.013, 2.004)$$

**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

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.

Also, the 95% confidence interval for this HR is (0.413, 0.807) = (1/2.43, 1/1.24).

We next provide edited output when the event-type is **CVD**, treating **Cancer** and **Other** as censored.

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.

The 95% confidence interval for this HR is (1.013, 2.004).

Last, we provide edited output when the event-type is **Other**, treating **Cancer** and **CVD** as censored.

$$\widehat{HR}_{\text{COTH}}(\text{RX} = 1 \text{ vs. } \text{RX} = 0) \\ = \exp(-0.580) = 0.561$$

$$\text{Wald ChiSq} = (-.578/.279)^2 \\ = 4.29 \text{ (P} = 0.038)$$

**Signif. at .05 level**

$$\text{95\% CI for } \exp[\beta_{1\text{OTH}}]: \\ \exp[-0.578 \pm 1.96(0.279)] \\ = (0.325, 0.969)$$

Not assessed in the above analysis:

PH assumption

Interaction of Rx with control variables

Here, the adjusted  $\widehat{HR}$  for the effect of Rx is 0.561 (= 1/1.78).

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.

The 95% confidence interval for this HR is (0.325, 0.969), which is somewhat imprecise.

We have thus completed a competing risk analysis of 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

Right Censoring vs. left censoring

↓

- More often
- Our focus

Important assumption

- Required for all approaches/models described to this point
- Relevant for competing risks

*Censoring Is Noninformative*  
(Synonym: Independent)

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.

We also briefly introduced in Chapter 1 an important assumption about censoring that is required for all approaches/models for analyzing survival data described up to this point, including data with competing risks. This assumption is typically stated as follows: censoring is **noninformative or independent**.

Typical context

- No competing risks
- Homogeneous sample

Definition: noninformative censoring:

**Probability of being censored at time  $t$  does not depend on prognosis for failure at time  $t$**

#### EXAMPLE

Harry in homogeneous risk set at time  $t$

Event: death from any cause

3 possible outcomes at time  $t$   
Fail, not fail, unknown (censored) status

Harry in poorer health than other subjects in risk set at time  $t$

↓

Harry's prognosis for failing higher than other subjects in the risk set at time  $t$

*Noninformative censoring*

↓↑

$\Pr(C | PH) = \Pr(C | GH)$   
where

C = censoring  
PH = poor health  
GH = good health

Noninformative censoring can be defined in a context that assumes the absence of competing risks and a homogeneous study sample; that is, all subjects have the same values for covariate predictors.

In the above context, we define noninformative censoring to mean that the probability of being censored for any subject in the risk set at time  $t$  does not depend on that subject's prognosis for failure at time  $t$ .

Suppose, for example, that Harry is in a homogeneous risk set at time  $t$  (e.g., the entire risk set is, say, white, male, age 60) and the event of interest is death from any cause, so there are no competing risks.

Given the above scenario, one of the following three outcomes can be observed on each subject, including Harry, in the risk set at time  $t$ : he can fail (i.e., die), not fail, or have unknown (censored) outcome from withdrawal or loss to follow-up.

Now, suppose Harry is in much poorer health than other subjects in the risk set at time  $t$ . Then Harry's potential for failing at time  $t$  is likely to be higher than for other subjects in the risk set at time  $t$ .

Now, if censoring is noninformative, then despite Harry's being in poorer health than other subjects at time  $t$ , a subject like Harry would just as likely be censored as any other subject in the risk set, including subjects healthier than Harry who have a lower prognosis for failure.

Harry more likely to be censored than other subjects in risk set



$$\Pr(C | PH) > \Pr(C | GH)$$



*Informative Censoring*

(Synonym: **DEPENDENT**)



**Biased results**

$\hat{S}(t)$  overestimates  $S(t)$   
if

large proportion of censored subjects actually failed at time  $t$

**Competing risks**



Different types of censoring

- Failure from competing risks
- Lost to follow-up
- Withdrawal
- End of study

*Noninformative (i.e., Independent) Censoring with Competing Risks*

Harry in risk set at time  $t$



Harry just as likely to be censored as any other subject in risk set regardless of reason for censoring or prognosis for event-type **A**

Byar data: 3 competing risks  
(**Cancer, CVD, Other** deaths)

Noninformative censoring?

Nevertheless, because Harry has poor health, he might be more likely to drop out of the study (i.e., be censored) at time  $t$  because of his poor health than other healthier subjects in the risk set. If so, Harry's censoring would be **informative** (or **dependent**).

Informative censoring unfortunately can lead to **biased results** in a survival analysis. A bias can result because people who get censored are more likely to have failed than those not censored. Thus, the estimated survival probability at any time  $t$  may overestimate the true survival probability at time  $t$  if a large proportion of those with unknown status (i.e., censored) actually failed.

When the survival analysis problem involves competing risks, the requirement of **noninformative** or **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.

Thus, for competing risks, censoring is **noninformative** or **independent** if for a subject like Harry in the risk set at time  $t$ , Harry is just as likely to be censored at time  $t$  as any other subject in the risk set at  $t$ , regardless of the reason for censoring, including failure from a competing risk, or Harry's prognosis for failing from event-type **A**.

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 noninformative?

Cause-specific focus: **Cancer**  
**Noninformative censoring**



Harry *just as likely* to be censored  
 as any other subject in risk set  
 regardless of type of censoring

**Types of censoring—competing risks:**  
**CVD** or **Other** death or usual censoring

*Informative (Dependent)*  
*Censoring*

Harry in poorer health than other  
 subjects in risk set at time  $t$



Harry *more likely* to be censored  
 than other subjects in risk set

Initial context: subjects are homoge-  
 neous

More general context: each subject  
 representative of subjects in the risk  
 set with the same values of predic-  
 tors

#### EXAMPLE

E = exposed, Age = 45, Gender = male

**Noninformative censoring**



All subjects in risk set for which  
 E = exposed, Age = 45, Gender = male  
 are equally *likely* to be censored  
 regardless of type of censoring

Suppose **censoring is noninformative** 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 of covariates is just as likely to be censored at time  $t$  as any other subject in the risk set with the same set of covariates regardless of whether the reason for censoring is a **CVD** or **Other** death, withdrawal from study, or loss to follow-up.

On the other hand, if **censoring is informative (or dependent)** and Harry was in poorer health than other subjects in the risk set at time  $t$ , he might be more likely to be censored, including dying from **CVD** or **Other** cause, at time  $t$  than other subjects in the risk set.

Recall that the context in which we initially defined noninformative censoring assumed subjects in the risk set at time  $t$  to be “homogeneous,” that is, having the same values for the predictors of interest. Actually, because predictors are typically included in one’s model, the more general context assumes that each subject in the risk set at time  $t$  is representative of all subjects with the same values of the predictors who survive to time  $t$ .

For example, if the predictors are exposure (E), Age, and Gender, then noninformative censoring requires that a subject in the risk set at time  $t$  who is exposed, 45 years old, and male, is just as likely to be censored at time  $t$  as any other subject in the risk set at time  $t$  who is also exposed, 45 years old, and male.

Important assumption for competing risks

**Censoring is noninformative (i.e., independent)**

regardless of different types of censoring possible

Synonym: **Competing risks are independent**

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?

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**.

**CVD** death

↓

**Cancer** death unobservable

In general

Failure from competing risk A

↓

Failure from competing risk B unobservable

Answer to 2:

Alternative strategies available but no strategy is always best

The important message at this point when analyzing competing risks survival data is that it is typically assumed that **censoring is noninformative** or **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.

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?

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**.

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  $t$ .

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****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**.

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.

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****EXAMPLE OF STRATEGY  
2—CANCER VS. CVD**

Include age smoking in model to remove the common effects of these variables on competing risks

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.

### 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 9.4** 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.

**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

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
<b>Worst-Case (1):</b>					
Rx	1	-0.185	0.110	0.092	0.831
<b>Worst-Case (2):</b>					
Rx	1	-0.411	0.169	0.015	0.663
<b>Independent competing risks:</b>					
Rx	1	-0.550	0.171	0.001	0.577

	WC(1)	WC(2)	Independent
$\widehat{HR}$ s	0.831	0.663	0.577
P-values	0.092 (N.S.)	0.015 (<.05)	0.001 (<<.01)

Independence	Nonindependence
x	[ ]
<b>.577</b>	.663 .831

If competing risks not independent then conclusions about the effect of Rx could be very different

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.

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.

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 nonindependence.

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 (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.

Worst-case (2)

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**

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.

Evaluates how badly biased the results can get if independence 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)

Competing risks: **KM** may not be as informative as when only one risk

### *Hypothetical Study*

- $n = 100$  subjects
- All subjects with prostate cancer

Surv <sub>t</sub> (months)	# Died	Cause
3	99	CVD
5	1	Cancer

Study goal: cause-specific cancer survival

Censored: CVD deaths

**Table 9.6.** Hypothetical Survival Data

$j$	$t_j$	$n_j$	$m_j$	$q_j$	$S_{Ca}(t_j) \leftrightarrow \text{KM}$
0	0	100	0	0	1
1	3	100	0	99	1
2	5	1	1	—	0

We have previously discussed (Chapter 1 and beyond) the use of survival curves to provide summary information over time of the survival experience of (sub) 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 curves. 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.

When competing risks are being considered, the KM survival curve may not be as informative as with only one risk.

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.

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** summarizes the survival experience in this hypothetical study. The first five columns of this table show the ordered failure-time interval number ( $j$ ), the time of failure ( $t_j$ ), the number in the risk set ( $n_j$ ), the number who fail ( $m_j$ ), and the number who are censored at each failure time ( $q_j$ ), 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  $S_{Ca}(t_j)$  for cause-specific cancer at each failure time.

Risk set at  $t_j = 5$ : 1 subject

$$\Pr(T > 5 | T \geq 5) = (1 - 1)/2 = 0$$

$$\begin{aligned} \mathbf{KM}_{Ca}: S_{Ca}(t = 5) \\ &= S(t = 4) \times \Pr(T > 5 | T \geq 5) \\ &= 1 \times 0 \\ &= \mathbf{0} \end{aligned}$$

$$\begin{aligned} \mathbf{KM}_{Ca} \Rightarrow \text{Risk}_{Ca}(T \leq 5) \\ &= 1 - 0 = \mathbf{1} \end{aligned}$$

Nevertheless,

$$\frac{1 \text{ cancer death}}{100 \text{ initial subjects}} = \mathbf{0.01} \text{ (small)}$$

Question:

How many of the 99 CVD deaths would have died of cancer at  $t = 5$  if they hadn't died of CVD at  $t = 3$ ?

Cannot answer: unobservable

**Table 9.7.** Hypothetical Survival Data Sensitivity Analysis A (99 CVD Deaths of Cancer at  $t = 5$ )

j	$t_j$	$n_j$	$m_j$	$q_j$	$S_{Ca}(t_j) \leftrightarrow \mathbf{KM}$
0	0	100	0	0	1
1	3	100	0	0	1
2	5	100	100	0	<b>0</b>

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.

Thus, use of the  $\mathbf{KM}_{Ca}$  curve in the presence of competing risks (for CVD), suggests that the 5-month risk for cancer death is 1; that is,  $1 - S_{Ca}(t = 5)$ . 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.

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?

Unfortunately, we cannot ever answer this question because those dying of CVD cannot be observed further once they have died.

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

$$\begin{aligned} &\downarrow \\ &\Pr(T > 5 | \text{censored at month 3}) \\ = & \\ &\Pr(T > 5 | \text{survived to month 5}) = 0 \\ &\downarrow \\ &99 \text{ CVDs deaths would have been} \\ &\text{cancer deaths at month 5} \end{aligned}$$

The reason why **Tables 9.6** and **9.7** give the same 5-month survival probability (=0) is that the KM method assumes noninformative (i.e., independent) censoring. For the original data (**Table 9.6**), noninformative 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)

j	t <sub>j</sub>	n <sub>j</sub>	m <sub>j</sub>	q <sub>j</sub>	S <sub>Ca</sub> (t <sub>j</sub> ) ↔ KM
0	0	100	0	0	1
1	3	100	0	0	1
2	5	100	1	99	<b>0.99</b>

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<sub>Ca</sub>(t = 5) = 0.99  
different from

**Table 9.6:** S<sub>Ca</sub>(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:  
Risk<sub>Ca</sub>(T ≤ 5) = 1 – 0.99 = **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:** Risk<sub>Ca</sub>(T ≤ 5) = 1  
derived from the data

**Table 9.8:** Risk<sub>Ca</sub>(T ≤ 5) = **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,

$$\text{Risk}_{\text{Ca}}(T \leq 5) = 1 \quad \text{or} \quad 0.01?$$

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

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.

- Requires independence assumption about competing risks
- Independence assumption 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: **CIC** = 1 – **KM**

**CIC** with competing risks

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.

- Derived from cause-specific hazard function
- Estimates **marginal probability** when competing risks are present
- Does not require independence assumption



## Marginal probabilities

Useful to assess treatment utility  
in cost-effectiveness analyses

for 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 .01 (5-month) marginal probability for cancer derived from hypothetical data in Table 9.6 illustrates small treatment utility for cancer.

Steps to Construct **CIC**

1. Estimate hazard at ordered failure times  $t_j$  for event type (**c**) of interest

$$\hat{h}_c(t_j) = \frac{m_{cj}}{n_j}$$

where

$m_{cj}$  = # of events for event type  
**c** at time  $t_j$

$n_j$  = # of subjects at risk at time  
 $t_j$

How does one construct a **CIC**? We first estimate the hazard at ordered time points  $t_j$  when the event of interest occurs. This hazard estimate is simply the number of events that occur at  $t_j$  divided by the number at risk at  $t_j$  (analogous to the KM estimate). We can write this as  $\hat{h}_c(t_j) = m_{cj}/n_j$  where the  $m_{cj}$  denotes the number of events for risk **c** at time  $t_j$  and  $n_j$  is the number of subjects at that time. Thus, at any particular time,  $m_{cj}/n_j$  is the estimated proportion of subjects failing from risk **c**.

2. Estimate

$S(t_{j-1})$  = **overall** survival  
probability of  
surviving previous  
time ( $t_{j-1}$ )

overall  $\Rightarrow$  subject survives  
all other competing  
events

To be able to fail at time  $t_j$ , the subject needs to be “around to fail”; that is, he must have survived the previous time when a failure occurred. The probability of surviving the previous time  $t_{j-1}$  is denoted  $S(t_{j-1})$ , where  $S(t)$  denotes the **overall survival curve** rather than the cause-specific survival curve  $S_c(t)$ . We must consider “overall” survival here, because the subject must have survived all other competing events.

3. Compute estimated incidence of failing from event-type **c** at time  $t_j$

$$\hat{I}_c(t_j) = \hat{S}(t_{j-1}) \times \hat{h}_c(t_j)$$

The probability (i.e., incidence) of failing from event-type **c** at time  $t_j$  is then simply the probability of surviving the previous time period multiplied by  $\hat{h}_c(t_j)$ .

4.  $\mathbf{CIC}(t_j) = \sum_{j'=1}^j \hat{I}_c(t_{j'})$   
 $= \sum_{j'=1}^j \hat{S}(t_{j'-1}) \hat{h}_c(t_{j'})$

The cumulative incidence at time  $t_j$  is then the *cumulative* sum up to time  $t_j$ , (i.e., from  $j' = 1$  to  $j' = j$ ) of these incidence values over all event-type **c** failure times.

**EXAMPLE OF CIC CALCULATION  
(ANOTHER HYPOTHETICAL STUDY)**

- 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**

$t_j$	$n_j$	$m_j$	$\hat{h}_{ca}(t_j)$	$\hat{S}_{(t_j-1)}$	$\hat{I}_{ca}(t_j)$	CIC( $t_j$ )
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	0.084
3.2	20	0	0	0.833	0	0.084
3.8	19	0	0	0.833	0	0.084
4.7	18	0	0	0.789	0	0.084
4.9	17	1	0.059	0.745	0.044	0.128
6	16	2	0.125	0.702	0.088	0.216
6.9	14	1	0.071	0.614	0.044	0.260
7.0	13	0	0	0.570	0	0.260
7.6	12	0	0	0.526	0	0.260
10	11	1	0.091	0.526	0.048	0.308
10.8	7	1	0.143	0.383	0.055	0.363
11.0	6	0	0	0.328	0	0.363
11.2	5	0	0	0.328	0	0.363
15	4	0	0	0.262	0	0.363
17.1	3	1	0.333	0.262	0.087	0.450
<b>20.3</b>	<b>2</b>	<b>1</b>	<b>0.5</b>	<b>0.175</b>	<b>0.088</b>	<b>0.537</b>
24.4	1	0	0	0.087	0	0.537

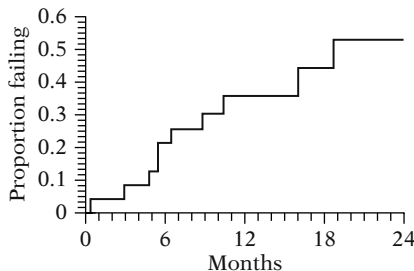
We illustrate the calculation of a **CIC** through an example.

Consider another hypothetical study involving 24 individuals receiving radiotherapy (XRT) for the treatment of head and neck cancer. Patients may die of the disease (cancer), other causes, or still be alive at the time of analysis.

The data are shown at the left.

The calculation of the **CIC** for these data is shown in **Table 9.9**.

From the table, we can see that the highest **CIC** probability of 0.537 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 53.7% when allowing for the presence of competing risks for CVD and Other Deaths.



Cumulative Incidence Curve

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 ( $t = 24.4$ ).

**CIC** Summary

- Gives marginal probability
- Does not use product limit formulation
- Not included in mainstream commercially available statistical packages (e.g., SAS, STATA, SPSS)

Independence of competing risks not required for **CIC** approach

Nevertheless, **CIC** requires

$$h(t) = h_{c1}(t) + h_{c2}(t) + \dots + h_{ck}(t)$$

where

$h(t)$  = overall hazard

$h_c(t)$  = hazard for event-type  $c$

Note: satisfied if

- Mutually exclusive event-types
- Non-recurrent events

Comparing **CICs** for 2 or more groups:

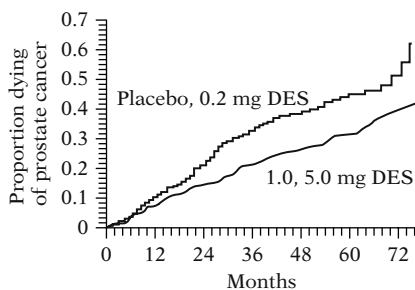
- Statistical test available (Gray, 1988)
- Analogous to log rank test
- No independence assumption
- Does not adjust for covariates

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.

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, 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; that is, one and only one event can occur at any one time and only once over time.

Gray (1988) developed a test to compare two or more **CICs**. 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 at the left gives the **CICs** for the two treatments for the Byar data.

Cumulative Incidence Curves –  
Byar Data

Gray's test results:  $\chi^2 = 6.6$ ,  $df = 1$   
 p-value: 0.01

Using Gray's test to compare the two **CICs** 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

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.

(but **CIC** meaningful for treatment utility)

Modeling **CIC** with covariates using PH model: Fine and Gray (1999)

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. The mathematical details of these methods are beyond the scope of this text but software is available that allows for such models to be fitted (Gebski, 1997)

(**CIC** also called **subdistribution** function)

Software available (Gebski, 1997)

Fine and Gray model analogous to Cox PH model

The **CIC** models developed by Fine and Gray are analogous to the Cox PH model but, for any failure type, they model 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.

Effects of predictors (e.g., HRs) have similar interpretation

**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

For the Byar data, the fitted **CIC** model that focuses on cancer deaths as the event-type of interest is shown in **Table 9.10** below which we have repeated **Table 9.1**, which uses the standard competing risks Cox PH model approach.

-2 LOG L = 1662.766546

**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

	Fine and Gray <b>CIC</b> (Table 9.10)	Standard Cox PH (Table 9.1)
$\hat{\beta}_{Rx}$ :	-0.414	-0.550
$\widehat{HR}_{Rx}$ :	0.661 (= $\frac{1}{1.513}$ )	0.577 (= $\frac{1}{1.73}$ )
P-value:	0.008	0.001

Although corresponding coefficient estimates and standard errors are different in the two outputs, both outputs are reasonably similar.

For example, the estimated coefficient of Rx is -0.414 in **Table 9.10** versus -0.550 in **Table 9.1**. The corresponding hazard ratio estimates ( $e^\beta$ ) are .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**)

$$CPC_c = \Pr(T_c \leq t | T \geq t)$$

where,  $T_c$  = time until event c occurs

$T$  = time until any competing risk event occurs

for example, Byar data

$$CPC_{pc} = \Pr(T_{pc} \leq 24 | T \geq 24)$$

where pc = prostate cancer

$$CPC_c = CIC_c / (1 - CIC_{c'})$$

where  $CIC_{c'}$  = CIC from risks other than c

Another approach to competing risks is called the **Cumulative Conditional Probability** or **CPC**. CPCs provide a third summary measure, in addition to (1 minus KM) and CIC, of the risk of failure of an event in the presence of competing risks. Put simply, the **CPC<sub>c</sub>** 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*.

Thus, with the Byar dataset, we may be interested in the risk of dying of prostate cancer at 24 months, given that the subject is alive at 24 months to experience this event.

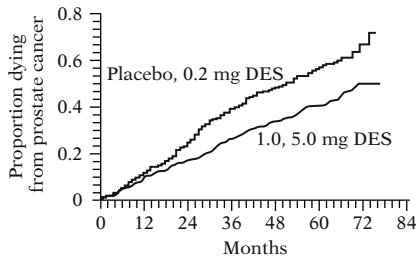
For risk type c, the **CPC** is defined by  $CPC_c = CIC_c / (1 - CIC_{c'})$ , where  $CIC_{c'}$  is the cumulative incidence of failure from risks other than risk c (i.e., all other risks considered together).

Graphs of CPCs obtained from CICs

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 allows for strata.

Tests to compare CPCs:

Pepe and Mori (1993)—2 curves  
Lunn (1998)—g curves



Byar Data: Cumulative Conditional Probability

For the Byar data, 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 **CPCs** in a regression framework.)

Test for equality: p-value = .01 (Pepe–Mori)

The Pepe–Mori test shows a significant difference between these two CPC curves.

### VIII. Method 2—The Lunn–McNeil (LM) Approach

Method 1: separate estimates for each failure type, treating the competing failure types as censored

We have previously (Section IV) described an approach (called Method 1) for analyzing competing risks data that 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 not failing from any event-type.

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 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 1 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 flexibility to perform statistical inferences about various features of the competing risk models that cannot be conveniently assessed using Method 1.

**Table 9.11.** Augmented Data for  $i$ th Subject at Time  $t_i$  Using LM Approach

Subj	Stime	Status	$D_1$	$D_2$	$D_3 \dots D_C$	$X_1 \dots X_p$
$i$	$t_i$	$e_1$	1	0	0 ... 0	$X_{i1} \dots X_{ip}$
$i$	$t_i$	$e_2$	0	1	0 ... 0	$X_{i1} \dots X_{ip}$
$i$	$t_i$	$e_3$	0	0	1 ... 0	$X_{i1} \dots X_{ip}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots \vdots \vdots$	$\vdots \vdots \vdots$
$i$	$t_i$	$e_c$	0	0	0 ... 1	$X_{i1} \dots X_{ip}$

$D_1, D_2, D_3, \dots, D_C$ : 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 9.11** for the  $i$ th subject with survival time  $t_i$  in the table. Also,  $C$  dummy variables  $D_1, D_2, D_3, \dots, D_C$  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  $X$ s 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  $D_1, D_2, D_3, \dots, D_C$  are indicators that distinguish the  $C$  competing risks (i.e., event-types).

**Definition**

$D_c$  equals 1 for event-type  $c$  and 0 otherwise,  $c = 1, 2, \dots, C$

Thus, the dummy variable  $D_c$  equals 1 for event-type  $c$  and 0 otherwise.

for example,

- Event-type 1:  $D_1 = 1,$   
 $D_2 = 0, D_3 = 0, \dots, D_C = 0$
- Event-type 2:  $D_1 = 0, D_2 = 1,$   
 $D_3 = 0, \dots, D_C = 0$
- Event-type 3:  $D_1 = 0, D_2 = 0,$   
 $D_3 = 1, \dots, D_C = 0$

For example, for event type 1, the  $D$ s are  $D_1 = 1, D_2 = 0, D_3 = 0, \dots, D_C = 0$ ; for event-type 2, the  $D$ s are  $D_1 = 0, D_2 = 1, D_3 = 0, \dots, D_C = 0$ ; and for event-type 3, the  $D$ s are  $D_1 = 0, D_2 = 0, D_3 = 1, \dots, D_C = 0$ .

**Table 9.12.** 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 9.12** 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.

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.

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*

$$\begin{aligned}
 g &= 1, 2, \dots, C \\
 h_g^*(t, \mathbf{X}) &= h_{0g}^*(t) \\
 &\times \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \\
 &+ \delta_{21} D_2 X_1 + \delta_{22} D_2 X_2 + \dots + \delta_{2p} D_2 X_p \\
 &+ \delta_{31} D_3 X_1 + \delta_{32} D_3 X_2 + \dots + \delta_{3p} D_3 X_p \\
 &+ \dots \\
 &+ \delta_{c1} D_c X_1 + \delta_{c2} D_c X_2 + \dots + \delta_{cp} D_c X_p]
 \end{aligned}$$

- 1st row: predictors  $X_1, X_2, \dots, X_p$
- 2nd row: product terms  $D_2 X_1, D_2 X_2, \dots, D_2 X_p$
- ...
- Cth row: product terms  $D_c X_1, D_c X_2, \dots, D_c X_p$

*LM Hazard Model for Event-Type 1*

$$\begin{aligned}
 h_1(t, \mathbf{X}) &= h_{01}^*(t) \\
 &\times \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p] \\
 &\quad (D_2 = D_3 = \dots = D_c = 0)
 \end{aligned}$$

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 9.11** is shown at the left.

Recall that the  $X_1, X_2, \dots, X_p$  denote the  $p$  predictors of interest.  $D_2, D_3, \dots, D_c$  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  $X$ s, the second row contains product terms involving  $D_2$  with each of the  $X$ s, and so on, with the last ( $C$ th) row containing product terms of  $D_c$  with each of the  $X$ s. The strata ( $g = 1, \dots, c$ ) are the  $C$  event-types.

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  $D_2, D_3, \dots, D_c$  are  $D_2 = D_3 = \dots = D_c = 0$ .



No product terms in model

$$HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\beta_1]$$

Product terms  $X_j X_1$  in model

$$HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) \\ = \exp[\beta_1 + \sum \beta_j X_j]$$

(product terms  $X_j X_1$  in the model)

*LM Hazard Model for  
Event-Type  $g$  ( $> 1$ )*

$$h_g(t, \mathbf{X}) = h_{0g}^*(t) \\ \times \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \\ + \delta_{g1} X_1 + \delta_{g2} X_2 + \dots + \delta_{gp} X_p] \\ = h_{0g}^*(t) \exp[(\beta_1 + \delta_{g1}) X_1 + (\beta_2 + \delta_{g2}) X_2 \\ + \dots + (\beta_p + \delta_{gp}) X_p]$$

No product terms  $X_j X_1$  in the model

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) \\ = \exp[(\beta_1 + \delta_{g1})]$$

Product terms  $X_j X_1$  in the model

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) \\ = \exp[(\beta_1 + \delta_{g1}) \\ + \sum (\beta_j + \delta_{gj} X_j)]$$

#### EXAMPLE OF LM MODEL FOR BYAR DATA

Separate models approach (Method 1):  
Cause-specific model: **Cancer**  
**CVD** and **Other** deaths censored

No-interaction model  
 $h_{Ca}(t, X) = h_{0Ca}(t) \exp[\beta_{1Ca} Rx + \beta_{2Ca} Age \\ + \beta_{3Ca} Wt + \beta_{4Ca} PF + \beta_{5Ca} Hx \\ + \beta_{6Ca} HG + \beta_{7Ca} SZ + \beta_{8Ca} SG]$

$$HR_{Ca}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_{1Ca}]$$

Thus, for  $g = 1$ , if  $X_1$  is a (0,1) variable, the other  $X_s$  are covariates, and there are no product terms  $X_j X_1$  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  $X_j X_1$ .

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.

Thus, for  $g > 1$ , if  $X_1$  is a (0,1) variable, the other  $X_s$  are covariates, and there are no product terms  $X_j X_1$  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  $X_j X_1$ .

We now illustrate the above general LM model formation using the Byar data.

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.

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

$g = 1, 2, 3$

$$h_g^*(t, \mathbf{X}) = h_{0g}(t) \times \exp[\beta_1 Rx + \beta_2 \text{Age} + \dots + \beta_8 \text{SG} + \delta_{21} D_2 Rx + \delta_{22} D_2 \text{Age} + \dots + \delta_{28} D_2 \text{SG} + \delta_{31} D_3 Rx + \delta_{32} D_3 \text{Age} + \dots + \delta_{38} D_3 \text{SG}]$$

1st row: predictors

Rx, Age, Wt, PF, . . . , SG

2nd row: products

$D_2 Rx, D_2 \text{Age}, \dots, D_2 \text{SG}$

3rd row: products

$D_3 Rx, D_3 \text{Age}, \dots, D_3 \text{SG}$

$D_2 = \text{CVD}$  and  $D_3 = \text{OTH}$  are (0,1) dummy variables that distinguish the 3 event-types

$$\text{HR}_{\text{Ca}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1]$$

$$\text{HR}_{\text{CVD}}(Rx = 1 \text{ vs. } Rx = 0)$$

$$= \exp[(\beta_1 + \delta_{21})]$$

$$\text{HR}_{\text{OTH}}(Rx = 1 \text{ vs. } Rx = 0)$$

$$= \exp[(\beta_1 + \delta_{31})]$$

**Table 9.13.** Edited Output for LM Model (No-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

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.

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  $\delta_{g1}$  of the product term  $D_g Rx, g = 2, 3$ , is added to the coefficient  $\beta_1$  of Rx in the exponential term.

**Table 9.13** 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  $\delta_{gj}$  coefficients in the **LM** model rather than the sum  $(\beta_1 + \delta_{gj})$  required for computing the HR when  $g = 2$  and 3.

$$\begin{aligned}\widehat{HR}_{Ca}(Rx = 1 \text{ vs. } Rx = 0) &= \exp[-0.550] = 0.577 \\ &= (1/1.733) \\ \text{Wald ChiSq} &= (-.550/.171)^2 \\ &= 10.345(P = 0.001)\end{aligned}$$

$$\begin{aligned}95\% \text{ CI for } \exp[\beta_{1Ca}]: \\ \exp[-0.550 \pm 1.96(0.171)] \\ = (0.413, 0.807)\end{aligned}$$

**LM** results for **Cancer** identical to Method 1 results for **Cancer**

$$\begin{aligned}\widehat{HR}_{CVD}(Rx = 1 \text{ vs. } Rx = 0) &= \exp(\hat{\beta}_1 + \hat{\delta}_{11}) \\ &= \exp(-0.550 + 0.905) \\ &= \exp(0.355) = 1.426\end{aligned}$$

$$\begin{aligned}\widehat{HR}_{OTH}(Rx = 1 \text{ vs. } Rx = 0) &= \exp(\hat{\beta}_1 + \hat{\delta}_{21}) \\ &= \exp(-0.550 - 0.028) \\ &= \exp(-0.578) = 0.561\end{aligned}$$

**LM** results for **CVD** and **Other** identical to Method 1 results for **CVD** and **Other**

Wald test statistics for **CVD** and **Other**

$$\begin{aligned}\text{Wald}_{CVD} &= \left[ \frac{\hat{\beta}_1 + \hat{\delta}_{11}}{\text{SE}_{\hat{\beta}_1 + \hat{\delta}_{11}}} \right]^2 \\ \text{Wald}_{OTH} &= \left[ \frac{\hat{\beta}_1 + \hat{\delta}_{21}}{\text{SE}_{\hat{\beta}_1 + \hat{\delta}_{21}}} \right]^2\end{aligned}$$

Computer packages provide for computation of the above formulas

SAS: **test** statement

STATA: **lincom** command

From **Table 9.13**, 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).

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.

Using **Table 9.13** 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.

These results are shown at the left, and they are identical to those obtained in **Tables 9.2 and 9.3** using Method 1.

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 9.13** gives only **individual standard errors** of  $\hat{\beta}_1$ ,  $\hat{\delta}_{11}$ , and  $\hat{\delta}_{21}$ .

SAS and STATA provide special syntax to specify the computer code for such computations: SAS's PHREG allows a "test" statement; STATA allows a "lincom" command.

Alternative **LM** formulation ( $\mathbf{LM}_{\text{alt}}$  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  $\mathbf{LM}_{\text{alt}}$  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 ( $\mathbf{LM}_{\text{alt}}$ ) Approach**

- Uses same data layout as **Table 9.11**
- Column headings:
  - Dummy variables  $D_1, D_2, \dots, D_C$
  - Predictor variables  $X_1, X_2, \dots, X_p$
- Above variables are transformed into product terms

The data layout required to fit the  $\mathbf{LM}_{\text{alt}}$  model is the same as shown earlier in **Table 9.11**. However, the variables listed in the columns of this table, namely, the dummy variables  $D_1, D_2, \dots, D_C$  and the predictor variables  $X_1, X_2, \dots, X_p$ , serve as basic variables that are transformed into product terms that define the  $\mathbf{LM}_{\text{alt}}$  model.

1st row of  $\mathbf{LM}_{\text{alt}}$  model:  
 product terms  $D_1X_1, D_1X_2, \dots, D_1X_p$   
 coefficients  $\delta'_{11}, \dots, \delta'_{1p}$

The primary difference in the two formulas is that the first row of the exponential term in the  $\mathbf{LM}_{\text{alt}}$  model contains product terms  $D_1X_1, D_1X_2, \dots, D_1X_p$  with coefficients denoted  $\delta'_{11}, \dots, \delta'_{1p}$  whereas the first row in the **LM** model contains the predictors  $X_1, X_2, \dots, X_p$  without product terms and coefficients denoted  $\beta_1, \dots, \beta_p$ .

1st row of **LM** model  
 predictors  $X_1, X_2, \dots, X_p$   
 coefficients  $\beta_1, \dots, \beta_p$

*General Stratified Cox  $\mathbf{LM}_{\text{alt}}$  Model*

The general form of the  $\mathbf{LM}_{\text{alt}}$  model is shown at the left. We have used a superscript prime (') to distinguish the hazard model formula for the  $\mathbf{LM}_{\text{alt}}$  model from the corresponding formula for the **LM** model given earlier.

$$\begin{aligned}
 &g = 1, \dots, C \\
 &h'_g(t, \mathbf{X}) = h'_{0g}(t) \\
 &\quad \times \exp[\delta'_{11}D_1X_1 + \delta'_{12}D_1X_2 + \dots + \delta'_{1p}D_1X_p \\
 &\quad + \delta'_{21}D_2X_1 + \delta'_{22}D_2X_2 + \dots + \delta'_{2p}D_2X_p \\
 &\quad + \delta'_{31}D_3X_1 + \delta'_{32}D_3X_2 + \dots + \delta'_{3p}D_3X_p \\
 &\quad + \dots \\
 &\quad + \delta'_{C1}D_CX_1 + \delta'_{C2}D_CX_2 + \dots + \delta'_{Cp}D_CX_p]
 \end{aligned}$$

- **LM<sub>alt</sub>** 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

**LM<sub>alt</sub>** Hazard Model for Event-Type 1

$$h'_1(t, \mathbf{X}) = h'_{01}(t) \times \exp[\delta'_{11}X_1 + \delta'_{12}X_2 + \cdots + \delta'_{1p}X_p]$$

( $D_1 = 1, D_2 = D_3 = \cdots = D_C = 0$ )  
 $HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\delta'_{11}]$   
 (no products  $X_jX_1$  in model)

**LM** HR =  $\exp[\beta_1]$

**LM<sub>alt</sub>** Hazard Model for Event-Type  $g (> 1)$

$$h'_1(t, \mathbf{X}) = h'_{0g}(t) \times \exp[\delta'_{g1}X_1 + \delta'_{g2}X_2 + \cdots + \delta'_{gp}X_p]$$

( $D_g = 1$ , and  $D_{g'} = 0$  for  $g' \neq g$ )  
 $HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\delta'_{g1}]$   
 (no products  $X_jX_1$  in model)

**LM** HR =  $\exp[\beta_1 + \delta_{g1}]$

Because the **LM<sub>alt</sub>** 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.

For  $g = 1$  (i.e., event-type 1), the **LM<sub>alt</sub>** 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  $D_2 = D_3 = \cdots = D_C = 0$ .

Thus, for  $g = 1$ , if  $X_1$  is a (0,1) variable, the other  $X_s$  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  $\exp[\delta_{11}]$ .

Recall that for the **LM** model, the corresponding HR formula also involved the coefficient of the  $X_1$  variable, denoted as  $\beta_1$ .

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  $X_s$  are covariates, and there are no products  $X_jX_1$  in model, the formula for the HR for the effect of  $X_1$  adjusted for the covariates is  $\exp[\delta'_{g1}]$ .

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)

$\mathbf{LM}_{\text{alt}}$  model: need standard error for  $\hat{\delta}'_{g1}$  (directly provided by output)

$\mathbf{LM}$  model: standard error of  $(\hat{\beta}_1 + \hat{\delta}_{g1})$ . (more complicated computation)

Next: Byar data example of  $\mathbf{LM}_{\text{alt}}$  model

$\mathbf{LM}_{\text{alt}}$  SC Model for Byar Data

$g = 1, 2, 3$

$$h'_g(t, \mathbf{X}) = h_{0g'}(t) \times \exp[\delta'_{11}D_1Rx + \delta'_{12}D_1Age + \dots + \delta'_{1p}D_1SG + \delta'_{21}D_2Rx + \delta'_{22}D_2Age + \dots + \delta'_{28}D_2SG + \delta'_{31}D_3Rx + \delta'_{32}D_3Age + \dots + \delta'_{38}D_3SG]$$

$D_1 = \mathbf{CA}$ ,  $D_2 = \mathbf{CVD}$ , and  $D_3 = \mathbf{OTH}$  are (0,1) dummy variables for the 3 event-types

1st row: products

$D_1Rx, D_1Age, \dots, D_1SG$   
( $\mathbf{LM}$  predictors, Rx, Age, ..., SG)

2nd row: products

$D_2Rx, D_2Age, \dots, D_2SG$

3rd row: products

$D_3Rx, D_3Age, \dots, D_3SG$

$$HR_{\mathbf{Ca}}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\delta'_{11}]$$

$$HR_{\mathbf{CVD}}(Rx = 1 \text{ vs. } Rx = 0)$$

$$= \exp[\delta'_{21}]$$

$$HR_{\mathbf{OTH}}(Rx = 1 \text{ vs. } Rx = 0)$$

$$= \exp[\delta'_{31}]$$

Thus for  $g > 1$ , statistical inferences about HRs using the  $\mathbf{LM}_{\text{alt}}$  model only require use of the standard error for  $\hat{\delta}'_{g1}$  that is directly provided in the output.

In contrast, the  $\mathbf{LM}$  model requires computing the more complicated standard error of the sum  $(\hat{\beta}_1 + \hat{\delta}_{g1})$ .

We now illustrate the above general  $\mathbf{LM}_{\text{alt}}$  model formation using the Byar data.

The stratified Cox (SC)  $\mathbf{LM}_{\text{alt}}$  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**.

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  $\mathbf{LM}$  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  $\mathbf{LM}$  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.

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  $\delta'_{g1}$  is the coefficient of the product term  $D_gRx$  in the  $\mathbf{LM}_{\text{alt}}$  model.

$$\text{Wald}_g = \left[ \frac{\hat{\delta}'_{g1}}{\text{SE}_{\hat{\delta}'_{g1}}} \right]^2$$

$g = 1$  (CA),  $2$  (CVD),  $3$  (OTH)

Statistical inference information

**LM<sub>alt</sub>** model: directly provided by output

**LM** model: not directly provided by output (requires additional computer code)

**Table 9.14.** Edited Output for SC **LM<sub>alt</sub>** 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 9.14 (LM<sub>alt</sub>)** output identical to **Tables 9.1, 9.2, 9.3 (Method 1)** output combined

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 **LM<sub>alt</sub>** model.

Thus, the **LM<sub>alt</sub>** 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 9.14** shows edited output obtained from fitting the above **LM<sub>alt</sub>** 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.

Thus, the output in **Table 9.14** using the single **LM<sub>alt</sub>** 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 **LM<sub>alt</sub>** models when you can simply fit 3 separate models?

Answer: Can perform statistical inferences that cannot be done when fitting 3 separate models

#### **LM** Model for Byar Data

$g = 1, 2, 3$

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp[\beta_1 Rx + \beta_2 Age + \dots + \beta_8 SG + \delta'_{21} D_2 Rx + \delta'_{22} D_2 Age + \dots + \delta'_{28} D_2 SG + \delta'_{31} D_3 Rx + \delta'_{32} D_3 Age + \dots + \delta'_{38} D_3 SG]$$

Inference question: Byar data

No-interaction **SC LM** model versus interaction **SC LM** model

No-interaction **SC** model

$g = 1, 2, 3$

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp[\beta_1 Rx + \beta_2 Age + \dots + \beta_8 SG]$$

Assumes

$$\begin{aligned} HR_{CA}(X_i) &= HR_{CVD}(X_i) \\ &= HR_{OTH}(X_i) \\ &\equiv HR(X_i) \text{ for any } X_i \text{ variable} \end{aligned}$$

for example,  $Rx = 0$  vs  $Rx = 1$ :

$$\begin{aligned} HR_{CA}(Rx) &= HR_{CVD}(Rx) \\ &= HR_{OTH}(Rx) \\ &= \exp[\beta_1] \end{aligned}$$

The reader may have the following question at this point: Why bother with the **LM** or **LM<sub>alt</sub>** 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 **LM<sub>alt</sub>** 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.

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 **LM<sub>alt</sub>** model (see Exercises at end of this chapter).

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.

The no-interaction **SC** model is shown here at the left.

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, in the above no-interaction **SC LM** model the hazard ratio for the effect of  $Rx$  is  $\exp[\beta_1]$  for each  $g$ , where  $\beta_1$  is the coefficient of  $Rx$ .



$H_0$ : all  $\delta_{gj} = 0$ ,  
 $g = 2, 3; j = 1, 2, \dots, 8$   
 where  $\delta_{gj}$  is coefficient of  $D_g X_j$  in the  
 interaction SC LM model

*Likelihood Ratio Test*

$$\begin{aligned} \text{LR} &= -2 \log L_R - (-2 \log L_F) \\ &\text{approx } \chi_{16}^2 \text{ under } H_0 \\ R &= \text{no interaction SC (reduced)} \\ &\text{model} \\ F &= \text{interaction SC (full) model} \end{aligned}$$

**Table 9.15.** 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 9.15:** Log likelihood<sub>R</sub>  
 = -1892.091

**Table 9.13:** Log likelihood<sub>F</sub>  
 = -1831.916

$$\begin{aligned} \text{LR} &= -2 \log L_R - (-2 \log L_F) \\ &= -2(-1892.091) \\ &\quad - (-2(-1831.916)) \\ &= 120.35 \text{ approx } \chi_{16}^2 \text{ under } H_0 \\ &\quad (P < 0.001) \end{aligned}$$

Reject  $H_0$ : interaction SC model  
 more appropriate than  
 no-interaction SC 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 ( $\delta_{gj}$ ) in the interaction SC model are equal to zero.

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  $\delta_{gj}$  coefficients being set equal to zero under  $H_0$ .

**Table 9.15** 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  $g$ .

From **Table 9.15**, we find that the log-likelihood statistic for the reduced (no-interaction SC) model is -1892.091. From **Table 9.13** (or **9.14**), the log-likelihood statistic for the full (interaction SC) model is -1831.916.

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  $H_0$ .

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



$$HR_{Ca}(Rx = 1 \text{ vs. } 0) \neq HR_{CVD}(Rx = 1 \text{ vs. } 0)$$

**DIFFERENT STUDY EXAMPLE**

Competing risks: Stage 1 vs. Stage 2 Breast Cancer  
 ⇓  
 $HR_{stg1}(Rx = 0 \text{ vs. } 1) = HR_{stg2}(Rx = 0 \text{ vs. } 1)$   
 Plausible  
 ⇓  
 Non-interaction SC Cox reasonable depending on similarity of competing risks

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.

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 (**LM<sub>U</sub>**):

$$h^*(t, \mathbf{X}) = h_0^*(t) \times \exp[\gamma_1 CVD + \gamma_2 OTH + \beta_1^{\bullet} Rx + \beta_2^{\bullet} Age + \dots + \beta_8^{\bullet} SG + \delta_{21}^{\bullet} D_2 Rx + \delta_{22}^{\bullet} D_2 Age + \dots + \delta_{28}^{\bullet} SG + \delta_{31}^{\bullet} D_3 Rx + \delta_{32}^{\bullet} D_3 Age + \dots + \delta_{38}^{\bullet} SG]$$

Returning again to the Byar data example, another variation of the **LM** model is shown at the left and denoted **LM<sub>U</sub>**. This is a Cox PH model applied to the augmented data of **Table 9.11** 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 (•) to distinguish the **LM<sub>U</sub>** model from the **LM** and **LM<sub>alt</sub>** models.

**LM<sub>U</sub>** model: **CVD** and **OTH** included in model

**LM** model: **CVD** and **OTH** not included in model  
 (Both **LM<sub>U</sub>** and **LM** models use augmented dataset)

The **LM<sub>U</sub>** 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 **LM<sub>U</sub>** model is based on the augmented dataset given in **Table 9.11**.

**LM<sub>U</sub>** model: need to check PH assumption (Chapter 4)  
 PH assumption not satisfied



Use **LM** instead of **LM<sub>U</sub>** model

Because the **LM<sub>U</sub>** 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 ( $\mathbf{CVD} = \mathbf{OTH} = 0$ ):

$$HR_{Ca}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\delta_{11}^{\bullet}]$$

**CVD** survival ( $\mathbf{CVD} = 1, \mathbf{OTH} = 0$ ):

$HR_{CVD}(Rx = 1 \text{ vs. } Rx = 0)$

$$= \exp[\gamma_1 + \delta_{21}^{\bullet}]$$

**Other** survival ( $\mathbf{CVD} = 0, \mathbf{OTH} = 1$ ):

$HR_{OTH}(Rx = 1 \text{ vs. } Rx = 0)$

$$= \exp[\gamma_1 + \delta_{31}^{\bullet}]$$

In particular, to obtain the hazard ratio for the effect of Rx on **Cancer** survival, we would specify  $\mathbf{CVD} = \mathbf{OTH} = 0$  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  $\gamma_1$  and  $\gamma_2$  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  $\mathbf{LM}$ -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  $\mathbf{LM}$ -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 all subjects in risk set
2. Product-limit (e.g., KM) curve has questionable interpretation

Several alternative strategies regarding independence assumption: 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
- Independence assumption not easily verifiable
- Typical analysis assumes independence assumption is satisfied

### **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

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.

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.

Regarding the independence assumption, several alternative strategies for addressing this issue are described, although no single strategy is always best.

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.

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.

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.

$$\begin{aligned} \text{CIC}(t_j) &= \sum_{j'=1}^j \hat{I}_c(t_{j'}) \\ &= \sum_{j'=1}^j \hat{S}(t_{j'-1}) \hat{h}_c(t_{j'}) \end{aligned}$$

$\hat{h}_c(t_j)$  = estimated hazard at ordered failure time  $t_j$  for the event-type (**c**)

$S(t_{j'-1})$  = **overall** survival probability of previous time ( $t_{j'-1}$ )

### CIC

- Does not use product limit formulation
- Not included in mainstream commercially available statistical packages (e.g., SAS, STATA, SPSS)

PH model used to obtain **CIC**

↓

Independence of competing risks 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

Alternative to **CIC**

$$\text{CPC}_c = \Pr(T_c \leq t \mid T \geq t)$$

where  $T_c$  = time until event **c** occurs

$T$  = time until any competing risk event occurs

$$\text{CPC}_c = \text{CIC}_c / (1 - \text{CIC}_{c'})$$

where  $\text{CIC}_{c'}$  = CIC from risks other than **c**

The formula for the calculating the **CIC** is shown at the left. The  $\hat{h}_c(t_j)$  in the formula is the estimated hazard at survival time  $t_j$  for the event-type (**c**) of interest. The term  $S(t_{j'-1})$  denotes the **overall** survival probability of previous time ( $t_{j'-1}$ ), where “overall survival” indicates a subject that survives all competing events.

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.

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.

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.

An alternative to the **CIC** is the Conditional Probability Curve (**CPC**). For risk type **c**,  $\text{CPC}_c$  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  $t$ .*

The **CPC** can be computed from the **CIC** through the formula  $\text{CPC}_c = \text{CIC}_c / (1 - \text{CIC}_{c'})$ , where  $\text{CIC}_{c'}$  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 <sub>1</sub>	D <sub>2</sub>	D <sub>3</sub> ... D <sub>C</sub>	X <sub>1</sub> ... X <sub>p</sub>
i	t <sub>i</sub>	e <sub>1</sub>	1	0	0 ... 0	X <sub>i1</sub> ... X <sub>ip</sub>
i	t <sub>i</sub>	e <sub>2</sub>	0	1	0 ... 0	X <sub>i1</sub> ... X <sub>ip</sub>
i	t <sub>i</sub>	e <sub>3</sub>	0	0	1 ... 0	X <sub>i1</sub> ... X <sub>ip</sub>
⋮	⋮	⋮	⋮	⋮	⋮	⋮
i	t <sub>i</sub>	e <sub>c</sub>	0	0	0 ... 1	X <sub>i1</sub> ... X <sub>ip</sub>

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.

$g = 1, 2, \dots, C$

$$\begin{aligned}
 h_g^*(t, \mathbf{X}) &= h_{0g}^*(t) \\
 &\times \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p \\
 &+ \delta_{21} D_2 X_1 + \delta_{22} D_2 X_2 + \dots + \delta_{2p} D_2 X_p \\
 &+ \delta_{31} D_3 X_1 + \delta_{32} D_3 X_2 + \dots + \delta_{3p} D_3 X_p \\
 &+ \dots \\
 &+ \delta_{c1} D_c X_1 + \delta_{c2} D_c X_2 + \dots + \delta_{cp} D_c X_p]
 \end{aligned}$$

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

$$(\hat{\beta}_1 + \hat{\delta}_{g1})$$

(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, and SPSS.

Alternative **LM** formulation: **LM<sub>alt</sub>** model

**LM<sub>alt</sub>** yields output identical to Method 1

1st row of **LM<sub>alt</sub>** model

Product terms  $D_1X_1, D_1X_2, \dots, D_1X_p$   
Coefficients  $\delta'_{11}, \dots, \delta'_{1p}$

1st row of **LM** model

Predictors  $X_1, X_2, \dots, X_p$   
Coefficients  $\beta_1, \dots, \beta_p$

**LM<sub>alt</sub>** model:  $\text{Wald}_g = \left[ \frac{\delta'_{g1}}{\text{SE}_{\delta'_{g1}}} \right]^2$

directly obtained from output

Statistical inference information

**LM<sub>alt</sub>** model: directly provided by output

**LM** model: not directly provided by output (requires additional computer code)

Advantage of **LM** (Method 2) over method 1:

**LM** offers flexibility for statistical inferences to consider simpler models

Nevertheless, there is an alternative formation of the **LM** model that avoids the need for special syntax. This alternative formulation, called the **LM<sub>alt</sub>** model, yields output that is identical to the output from the separate models (Method 1) approach for analyzing competing risk data.

The primary difference in the two formulas is that the first row of the exponential term in the **LM<sub>alt</sub>** model contains product terms  $D_1X_1, D_1X_2, \dots, D_1X_{p1}$  with coefficients denoted  $\delta'_{11}, \dots, \delta'_{1p}$  whereas the first row in the **LM** model contains the predictors  $X_1, X_2, \dots, X_p$  without product terms and coefficients denoted  $\beta_1, \dots, \beta_p$ .

Using the **LM<sub>alt</sub>** 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 **LM<sub>alt</sub>** model.

Thus, the **LM<sub>alt</sub>** 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.

An advantage of using either the **LM** or **LM<sub>alt</sub>** 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

Overall,

- Can use standard computer packages
- Independence assumption required

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, 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 396)
  - 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 396–398)
  - 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 399–400)
  - 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 Gleason 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 400–403)

- 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:

$$h_{\mathbf{c}}(t) = \lim_{\Delta t \rightarrow 0} P(t \leq T_{\mathbf{c}} < t + \Delta t | T_{\mathbf{c}} \geq t) / \Delta t$$

where  $T_{\mathbf{c}}$  = time-to-failure from event  $\mathbf{c}$ ,  $\mathbf{c} = 1, 2, \dots, C$  (# of event types).

- C. Cox PH cause-specific model (event-type  $\mathbf{c}$ ):

$$h_{\mathbf{c}}(t, \mathbf{X}) = h_{0\mathbf{c}}(t) \exp\left[\sum_{i=1}^p \beta_{i\mathbf{c}} X_i\right]$$

where  $\mathbf{c} = 1, \dots, C$ , 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**:

$$h_{\mathbf{Ca}}(t, \mathbf{X}) = h_{0\mathbf{Ca}}(t) \exp[\beta_{1\mathbf{Ca}}RX + \beta_{2\mathbf{Ca}}Age + \beta_{3\mathbf{Ca}}Wt + \beta_{4\mathbf{Ca}}PF + \beta_{5\mathbf{Ca}}Hx + \beta_{6\mathbf{Ca}}HG + \beta_{7\mathbf{Ca}}SZ + \beta_{8\mathbf{Ca}}SG]$$

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):

$$\widehat{HR}_{\mathbf{Ca}}(RX = 1 \text{ vs. } RX = 0) = 0.575 \text{ (} P = 0.001\text{)}$$

- b. Cause-specific **CVD** results for RX (with **Cancer** and **Other** censored):

$$\widehat{HR}_{\text{CVD}}(\text{RX} = 1 \text{ vs. } \text{RX} = 0) = 1.429 \text{ (P} = 0.040\text{)}$$

- c. Cause-specific **Other** results for RX (with **Cancer** and **CVD** censored):

$$\widehat{HR}_{\text{OTH}}(\text{RX} = 1 \text{ vs. } \text{RX} = 0) = 0.560 \text{ (P} = 0.038\text{)}$$

## V. The Independence Assumption (pages 403–411)

- A. **Non-informative (independent)** censoring: risk for being censored at time  $t$  does not depend on prognosis for failure at time  $t$ .
1. Typical context: no competing risks; homogeneous sample.
  2. Informative censoring can lead to bias results.
- B. **Non-informative (independent)** censoring with **competing risks**. Any subject in the risk set at time  $t$  with a given set of covariates is just as likely to be censored at time  $t$  as any other subject in the risk set 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. Informative censoring: subject in risk set at time  $t$  is more likely to be censored from a competing risk than other subjects in the risk set at the same time.
  2. Homogeneous: subjects in the risk set having the same values of the predictors of interest.
  3. 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 412–420)

- 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

$KM_{Ca}: S_{Ca}(t = 5) = 0$  and  $Risk_{Ca}(T \leq 5) = 1$

- 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.  $KM_{Ca}: S_{Ca}(t = 5) = 0$  and  $Risk_{Ca}(T \leq 5) = 1$  because KM assumes noninformative 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.  $KM_{Ca}$ :  $S_{Ca}(t = 5) = 0.99$  and  $Risk_{Ca}(T \leq 5) = 0.01$ .
  - 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:

$$CIC(t_j) = \sum_{j'=1}^j \hat{I}_c(t_j) = \sum_{j'=1}^j \hat{S}(t_{j'-1}) \hat{h}_c(t_j)$$

6. Calculation of CIC for another hypothetical dataset.
7. Tests have been developed (Pepe and Mori, 1992) for comparing the equality of **CICs** 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 420–421)

- A.  $CPC_c = \Pr(T_c \leq t | T \geq t)$  where  $T_c =$  time until event **c** occurs,  $T =$  time until any competing risk event occurs.
- B. Formula in terms of CIC:  $CPC_c = CIC_c / (1 - CIC_{c'})$  where  $CIC_{c'} =$  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—Lunn–McNeil (LM) Approach**

(pages 421–427)

A. Allows only one Cox PH model to be fit rather than fitting separate models for each event type (Method 1).

B. **LM** uses an augmented data layout.

1. For  $i$ th subject at time  $t_i$ , layout has  $C$  rows of data, where  $C = \#$  event-types.
2. Dummy variables  $D_1, D_2, \dots, D_C$  are created to distinguish the  $C$  event-types.
3. The Status variable,  $e_c$ ,  $c = 1, \dots, C$ , equals 1 for event-type  $c$  and 0 otherwise.
4. Predictors are denoted by  $X_1, \dots, X_p$ .
5. Example of data layout for Byar dataset.

C. General form of **LM** model (interaction SC model).

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \delta_{21} D_2 X_1 + \delta_{22} D_2 X_2 + \dots + \delta_{2p} D_2 X_p + \delta_{31} D_3 X_1 + \delta_{32} D_3 X_2 + \dots + \delta_{3p} D_3 X_p + \dots + \delta_{C1} D_C X_1 + \delta_{C2} D_C X_2 + \dots + \delta_{Cp} D_C X_p]$$

1. **LM** model for event-type  $g = 1$ :

- a.  $h_1^*(t, \mathbf{X}) = h_{01}^*(t) \times \exp[\beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p]$
- b.  $D_2 = D_3 = \dots = D_C = 0$
- c.  $HR_{g=1}(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\beta_1]$

2. **LM** model for event-type  $g (> 1)$ :

- a.  $h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp[(\beta_1 + \delta_{g1})X_1 + (\beta_2 + \delta_{g2})X_2 + \dots + (\beta_p + \delta_{gp})X_p]$
- b.  $HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[(\beta_1 + \delta_{g1})]$

D. **LM** model for Byar data.

1.  $h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \times \exp[\beta_1 Rx + \beta_2 Age + \dots + \beta_8 SG + \delta_{21} D_2 Rx + \delta_{22} D_2 Age + \dots + \delta_{28} D_2 SG + \delta_{31} D_3 Rx + \delta_{32} D_3 Age + \dots + \delta_{38} D_3 SG]$   
 $g = 1, 2, 3$
2.  $D_2 = \text{CVD}$  and  $D_3 = \text{OTH}$  are (0,1) dummy variables for 3 event-types.

3.  $HR_{Ca}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1]$   
 $HR_{cVd}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_{21}]$   
 $HR_{OTH}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_{31}]$
4. Mathematical formulas for Wald tests and confidence intervals require standard errors for sums ( $\hat{\beta}_1 + \hat{\delta}_{g1}$ ) 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 ( $LM_{alt}$ )

#### Approach (pages 427–430)

- 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  $LM_{alt}$  model.
- C. General form of  $LM_{alt}$  (interaction SC) model:

$$g = 1, \dots, C$$

$$h'_g(t, \mathbf{X}) = h_{0g}'(t) \times \exp[\delta'_{11}D_1X_1 + \delta'_{12}D_1X_2 + \dots + \delta'_{1p}D_1X_p + \delta'_{21}D_2X_1 + \delta'_{22}D_2X_2 + \dots + \delta'_{2p}D_2X_p + \dots + \delta'_{c1}D_cX_1 + \delta'_{c2}D_cX_2 + \dots + \delta'_{cp}D_cX_p]$$

- D. Hazard ratio formula involves only coefficients of product terms for all  $g$ :

$$HR_g(X_1 = 1 \text{ vs. } X_1 = 0) = \exp[\delta'_{g1}], \quad g = 1, 2, 3$$

- a. Statistical inference information directly provided by  $LM_{alt}$  output.
- E. Example of output for  $LM_{alt}$  model using Byar data.

### X. Method 1—Separate Models versus Method 2—LM Approach (pages 431–434)

- A.  $LM$  and  $LM_{alt}$  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  $LM_{alt}$  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 **LM<sub>alt</sub>** models can assess whether an unstratified LM model (called **LM<sub>U</sub>**) is more appropriate than a stratified **LM** model.
  - E. Example of **LM<sub>U</sub>** model involving Byar data.
- XI. Summary** (pages 434–439)
- 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 (noninformative 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 noninformative (i.e., 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  $t$ , from risk  $c$ , 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:

$$h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 Rx + \beta_2 Age + \delta_1(D \times Rx) + \delta_2(D \times Age)],$$

$g = 1, 2$  where  
 $D = 0$  if Ca and  $= 1$  if CVD

**then**

$$HR_{CVD}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\beta_1 + \delta_1]$$

- T F 19. **Given the LM<sub>alt</sub> model** for two cause-specific events, Cancer and CVD:

$$h'_g(t, \mathbf{X}) = h'_{0g}(t) \times \exp[\delta'_{11} D_1 Rx + \delta'_{12} D_1 Age + \delta'_{21} D_2 Rx + \delta'_{22} D_2 Age],$$

$g = 1, 2$  where

$D_1 = 1$  if Ca or 0 if CVD, and  
 $D_2 = 0$  if Ca or 1 if CVD,

**then**

$$HR_{CVD}(Rx = 1 \text{ vs. } Rx = 0) = \exp[\delta'_{21}]$$

- T F 20. The **LM<sub>U</sub>** model that would result if the **LM** model of Question 18 were changed to an unstratified Cox PH model can be written as follows.

$$h^*(t, \mathbf{X}) = h_0^*(t) \exp[\beta_1^* Rx + \beta_2^* Age + \delta_{21}^* (D \times Rx) + \delta_{22}^* (D \times Age)]$$

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 nonrelapse deaths by year 1, 15 relapses during year 2, 40 nonrelapse 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.

<b>Hospital A</b>					<b>Hospital B</b>				
$t_j$	$n_j$	$m_j$	$q_j$	$S(t_j)$	$t_j$	$n_j$	$m_j$	$q_j$	$S(t_j)$
0	100	0	60	1	0	100	0	20	1
2	40	0	0	1	1	80	0	0	1
4	40	20	20	.5	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**

$t_j$	$n_j$	$m_j$	$\hat{h}_{ca}(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_{ca}(t_j)$	CIC( $t_j$ )
0	100	0	0	-	-	-
2	40	0	0	1	0	0
4	40	20	-	-	-	-

**Hospital B**

$t_j$	$n_j$	$m_j$	$\hat{h}_{ca}(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_{ca}(t_j)$	CIC( $t_j$ )
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 ( $\mathbf{G}$ ):  $\mathbf{G} = 0$  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 ( $\mathbf{G}$ ) 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  $i$ th 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<sub>alt</sub> 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<sub>alt</sub> 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)

*(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

$$\text{CIC}_1(t_j) = \sum_{j'=1}^j \hat{I}_1(t_{j'}) = \sum_{j'=1}^j \hat{S}(t_{j'-1}) \hat{h}_1(t_{j'})$$

where  $\hat{h}_1(t_j) = m_{1j}/n_j$ ,  $m_{1j}$  is the number of local recurrent failures at time  $t_j$ , and  $\hat{S}(t_{j-1})$  is the overall (event-free) survival probability for failure from either of the two competing risks at time  $t_{j-1}$ .

**tx = 1 (Treatment A)**

$t_j$	$n_j$	$d_{1j}$	$\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_1(t_j)$	$\text{CIC}_1(t_j)$
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
<b>8</b>	<b>23</b>	<b>1</b>	<b>.0435</b>	<b>.8889</b>	<b>.0387</b>	<b>.0387</b>
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)**

$t_j$	$n_j$	$d_{1j}$	$\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_1(t_j)$	$CIC_1(t_j)$
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
<b>25</b>	<b>6</b>	<b>1</b>	<b>.1667</b>	<b>.3632</b>	<b>.0605</b>	<b>.3554</b>
29	4	0	0	.2421	0	.3554
30	2	1	0	.2421	0	.3554

- Verify the  $CIC_1$  calculation provided at failure time  $t_j = 8$  for persons in the treatment group ( $\mathbf{tx} = 1$ ); that is, use the original data to compute  $\hat{h}_1(t_j)$ ,  $\hat{S}(t_{j-1})$ ,  $\hat{I}_1(t_j)$ , and  $CIC_1(t_j)$ , assuming that the calculations made up to this failure time are correct.
- Verify the  $CIC_1$  calculation provided at failure time  $t_j = 25$  for persons in the placebo group ( $\mathbf{tx} = 0$ ).
- Interpret the  $CIC_1$  values obtained for both the treatment and placebo groups at  $t_j = 30$ .
- How can you calculate the  $CPC_1$  values for both treatment and placebo groups at  $t_j = 30$ ?



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

- State the hazard model formula for the LM model used for the above output.
  - Determine the estimated hazard ratios for the effect of each of the 3 cause-specific events based on the above output.
  - 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  $LM_{alt}$  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

- State the hazard model formula for the  $LM_{alt}$  model used for the above output.
- Determine the estimated hazard ratios for the effect of each of the 3 cause-specific events based on the above output.
- 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 non-informative.
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  $t$ , from risk  $c$ , given that an individual has **not** experienced any of the other competing risks by time  $t$ .

15. F: the correct formula for CPC is:  $\mathbf{CPC}_c = \text{CIC}_c / (1 - \text{CIC}_{c'})$  where  $\text{CIC}_c = .4$  and  $\text{CIC}_{c'} = \text{CIC}$  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  

$$h^*(t, \mathbf{X}) = h_0^*(t) \exp[\gamma_1 D + \beta_1^* \text{Rx} + \beta_2^* \text{Age} + \delta_{21}^* (D \times \text{Rx}) + \delta_{22}^* (D \times \text{Age})]$$
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

$t_j$	$n_j$	$m_j$	$\hat{h}_{ca}(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_{ca}(t_j)$	CIC( $t_j$ )
0	100	0	0	—	—	—
2	40	0	0	1	0	0
4	40	20	0.5	0.4	0.20	0.20

Hospital B

$t_j$	$n_j$	$m_j$	$\hat{h}_{ca}(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_{ca}(t_j)$	CIC( $t_j$ )
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 year 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 <sub>1</sub>	D <sub>2</sub>	G	TMS	AGE
i	t <sub>i</sub>	e <sub>1i</sub>	1	0	G <sub>i</sub>	TMS <sub>i</sub>	AGE <sub>i</sub>
i	t <sub>i</sub>	e <sub>2i</sub>	0	0	G <sub>i</sub>	TMS <sub>i</sub>	AGE <sub>i</sub>

where  $e_{1j} = 1$  if the  $i$ th subject develops a hospital infection, 0 otherwise

$e_{2j} = 1$  if  $i$ th subject dies after surgery, 0 otherwise

$D_1 =$  indicator for hospital infection event

$D_2 =$  indicator for death after surgery event

$$30. h_g^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 G + \beta_2 TMS + \beta_3 AGE + \delta_{21} D_2 G + \delta_{22} D_2 TMS + \delta_{23} D_2 AGE]$$

$g=1,2$

$$31. HR_{HI}(RX = 1 \text{ vs. } RX = 0) = \exp[\beta_1]$$

32. Carry out a likelihood ratio test to compare the following two models. Full (SC Interaction LM) model:

$$h_{g=1,2}^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 G + \beta_2 \text{TMS} + \beta_3 \text{AGE} + \delta_{21} D_2 G + \delta_{22} D_2 \text{TMS} + \delta_{23} D_2 \text{AGE}]$$

Reduced (no-interaction SC LM) model:

$$h_{g=1,2}^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 G + \beta_2 \text{TMS} + \beta_3 \text{AGE}]$$

LR test statistic =  $-2 \ln L_R - (-2 \ln L_F)$  is distributed  $\chi_3^2$  under  $H_0$ : no-interaction model

33.  $h'_{g=1,2}(t, \mathbf{X}) = h'_{0g}(t) \exp[\beta'_1 D_1 G + \beta'_2 D_1 \text{TMS} + \beta'_3 D_1 \text{AGE} + \delta'_{21} D_2 G + \delta'_{22} D_2 \text{TMS} + \delta'_{23} D_2 \text{AGE}]$

34.  $\text{HR}_{\text{HI}}(\text{RX} = 1 \text{ vs. } \text{RX} = 0) = \exp[\beta'_1]$

# **Computer**

---

# **Appendix:**

---

# **Survival**

---

# **Analysis**

---

# **on the**

---

# **Computer**

---

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 three of the most widely used packages. The software packages that we describe are Stata (version 7.0), SAS (version 8.2), and SPSS (version 11.5). 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 Web site at <http://www.sph.emory.edu/~dkleinb/surv2.htm>. On this Web site we also provide many of the other datasets that have been used in the examples and exercises throughout this text. The data on our Web site are provided in several forms: (1) as Stata datasets (with a **.dta** extension), (2) as SAS version 8.2 datasets (with a **.sas7bdat** extension), (3) as SPSS datasets (with a **.sav** extension), and (4) 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 their 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 86 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, 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 tumor(s).

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 three broad sections, one for each of the following software packages,

- A. **Stata**,
- B. **SAS**, and
- C. **SPSS**.

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 lowercase 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. Commands 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 to that of 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; and
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 all observations had an event (i.e., no censorships). Notice 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 because 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, 0 for a censorship);
- .**t0**—the beginning “time variable”. All observations start at time 0 by default; and
- .**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

```

Category	total	per subject			
		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 and 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

```

clinic	Survivor Function		
	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	.	.

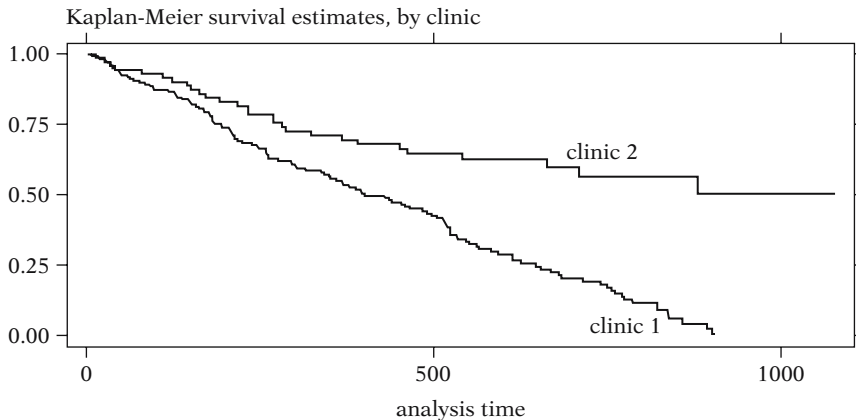
Notice that the survival rate for CLINIC = 2 is higher than CLINIC = 1. Other survival times could have been requested using the **compare( )** option.

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

clinic	Events observed	Events 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 but weight each observation differently. The Wilcoxon test weights the  $j$ th failure time by  $n_j$  (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  $\hat{s}(t_j)$  calculated over all groups combined. This survival estimate  $\hat{s}(t_j)$  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  $j$ 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 is 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 follow provide life table survival estimates, stratified by CLINIC, at the time points (in days) specified by the **interval( )** option.



**Itable 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

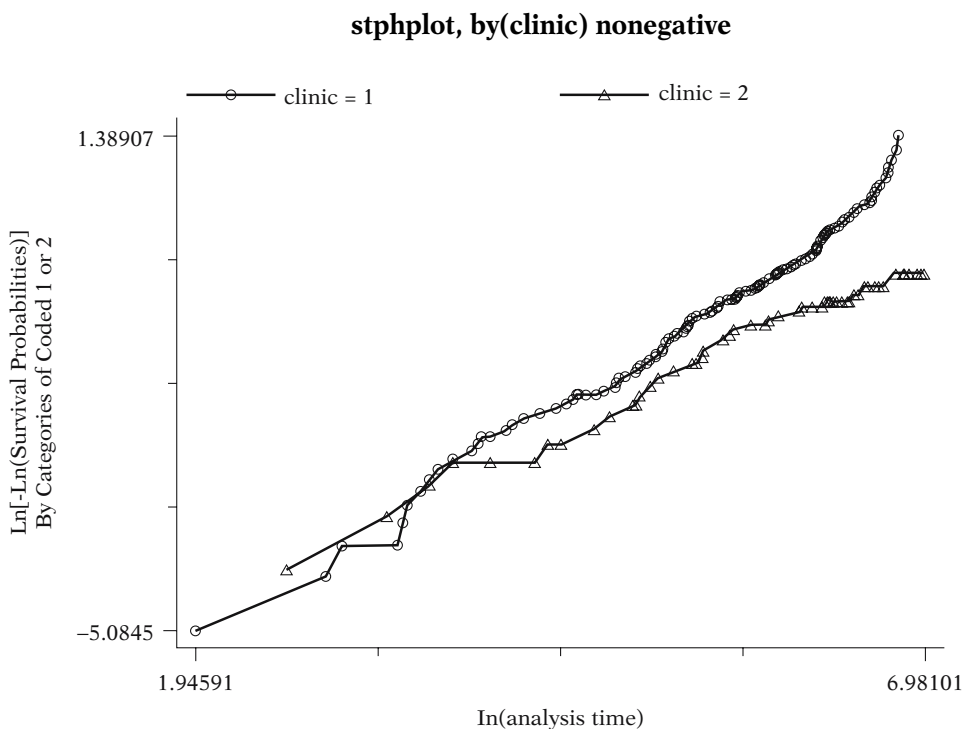
## 2. ASSESSING THE PH ASSUMPTION USING GRAPHICAL APPROACHES

Several 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; and
3. Kaplan-Meier survival estimates and Cox adjusted survival estimates plotted on the same graph.

All three approaches are somewhat subjective yet, it is hoped, 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?

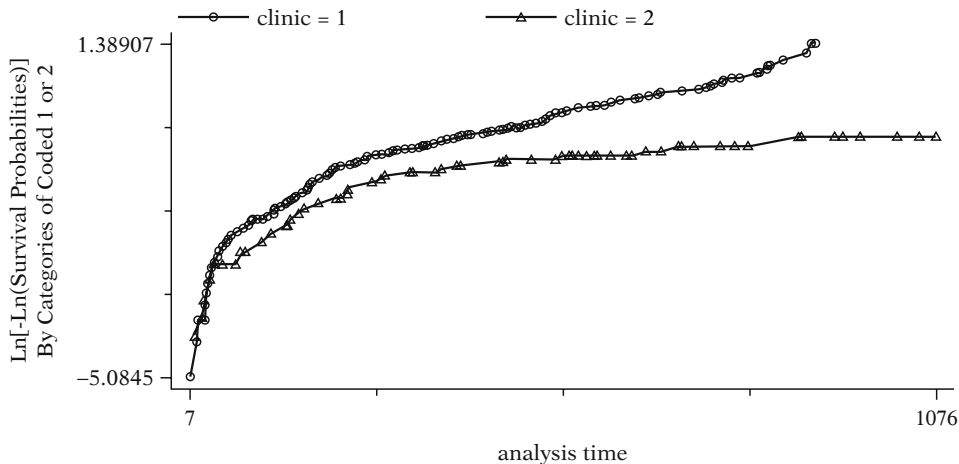
The first two approaches use the **stphplot** command whereas the third approach uses the **stcoxkm** command. The code and output for the log-log Kaplan-Meier survival plots follow.



The left side of the graph seems jumpy for  $CLINIC = 1$  but it only represents a few events. It also looks as if 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) plots  $\log(\text{survival time})$  rather than survival time on the horizontal axis by default. As far as checking the parallel assumption, it does not matter if  $\log(\text{survival time})$  or survival time is on the horizontal axis. However, if the log-log survival curves look like straight lines with  $\log(\text{survival time})$  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.

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 follow.

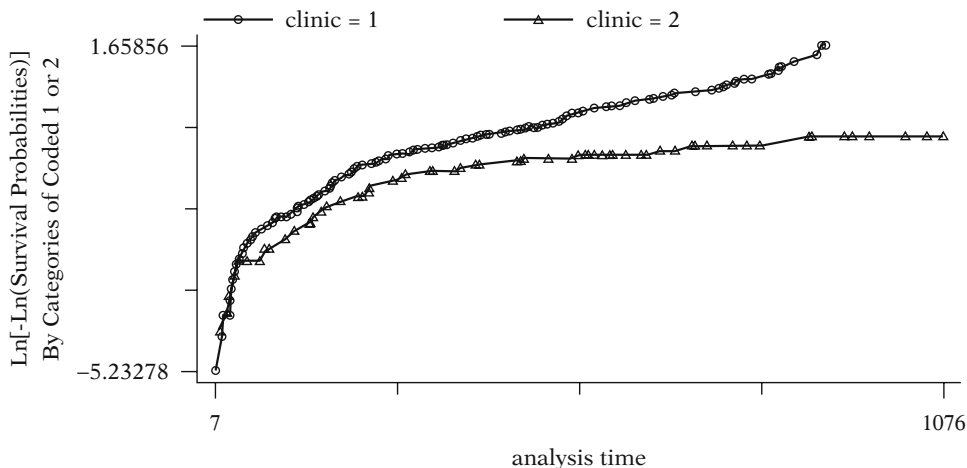
**stphplot, by(clinic) nonegative nolntime**

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**

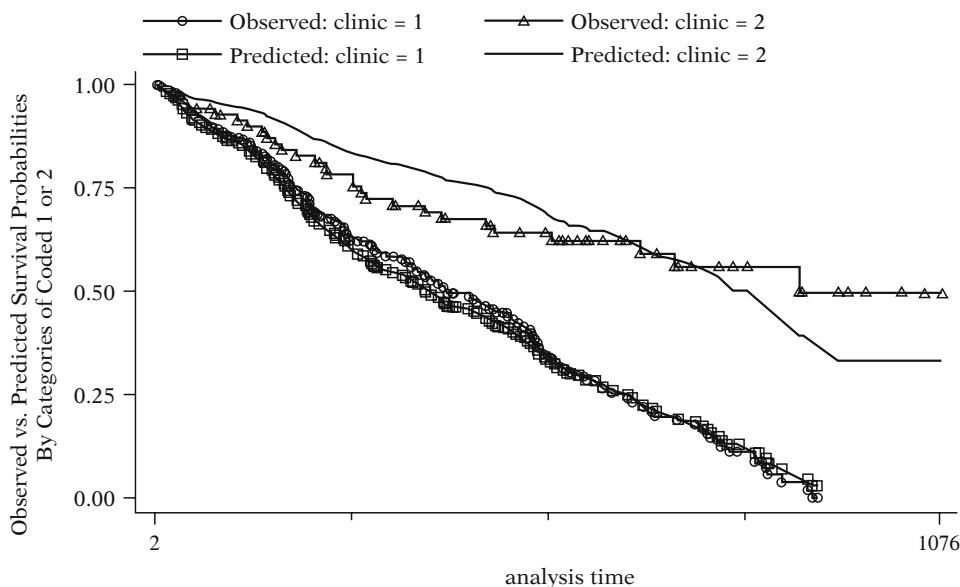
The log-log curves are adjusted for PRISON and DOSE using a stratified COX model on the variable CLINIC. The mean values of PRISON and DOSE are used for the adjustment. The output follows.



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 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
Log likelihood = -673.40242          LR chi2(3) =          64.52
                                      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, and 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
Log likelihood   = -673.40242          LR chi2(3)      =          64.52
                                          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 in the previous 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			
Log likelihood	=	-666.3274	LR chi2(3)	=	64.56
			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

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  $\times$  PRISON and CLINIC  $\times$  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
Log likelihood = -671.59969          LR chi2(5)   =          68.12
                                          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
Log likelihood = -597.714          LR chi2(2)      =     33.94
                               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
Log likelihood  = -596.77891          LR chi2(4)      =          35.81
                                          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).

$$\begin{aligned}
 HR &= \frac{h_0(t) \exp[1\beta_1 + \beta_2 DOSE + (2)(1)\beta_3 + \beta_4 CLIN\_DO]}{h_0(t) \exp[0\beta_1 + \beta_2 DOSE + (2)(0)\beta_3 + \beta_4 CLIN\_DO]} \\
 &= \exp(\beta_1 + 2\beta_3)
 \end{aligned}$$

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 subset 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
Log likelihood     = -104.37135          LR chi2(2)      =           9.70
                                          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 subset data approach (0.9210324).

## 5. ASSESSING THE PH ASSUMPTION USING A STATISTICAL TEST

The **stphptest** command can be used to perform a statistical test of the PH assumption. 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 **stphptest** 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 Schoenfeld scaled 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). In the parentheses are the names of newly defined variables; **schoen\*** creates SCHOEN1, SCHOEN2, and SCHOEN3 whereas **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 and 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

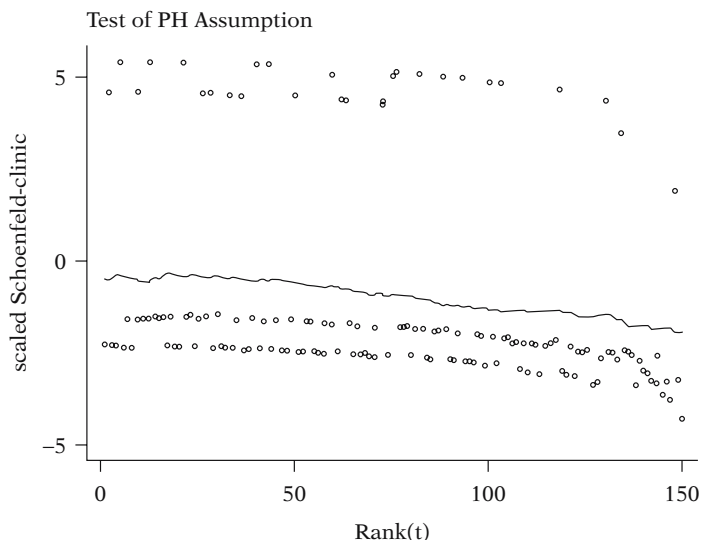
Time: Rank(t)

	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 because 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 horizontally).

## 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).

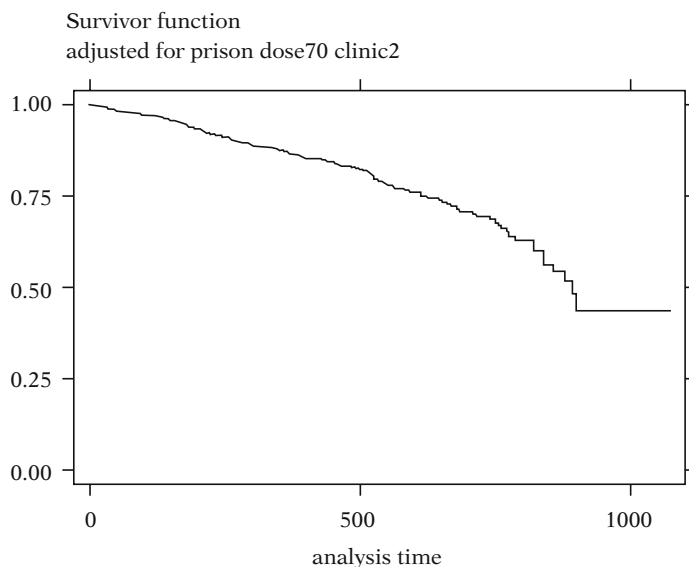
```
sts graph, adjustfor(prison dose clinic)
```

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
```

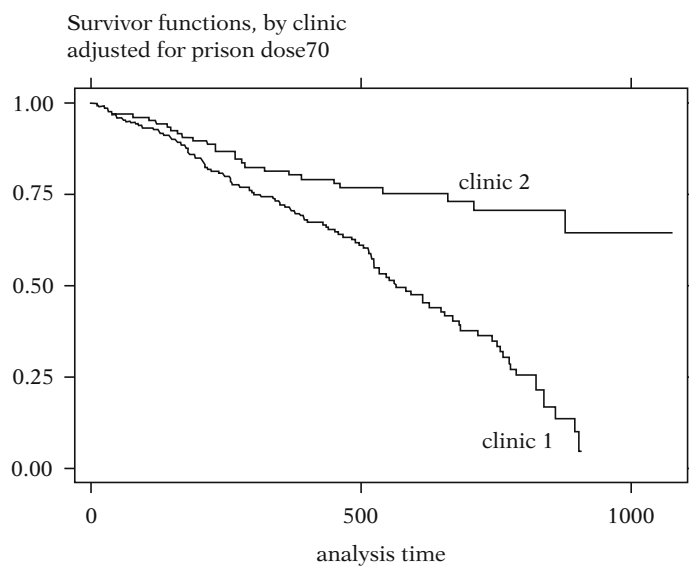
These variables (`PRISON`, `CLINIC2`, and `DOSE70`) produce the desired pattern of covariate when each is set to zero. The following code produces the desired results.

```
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)**



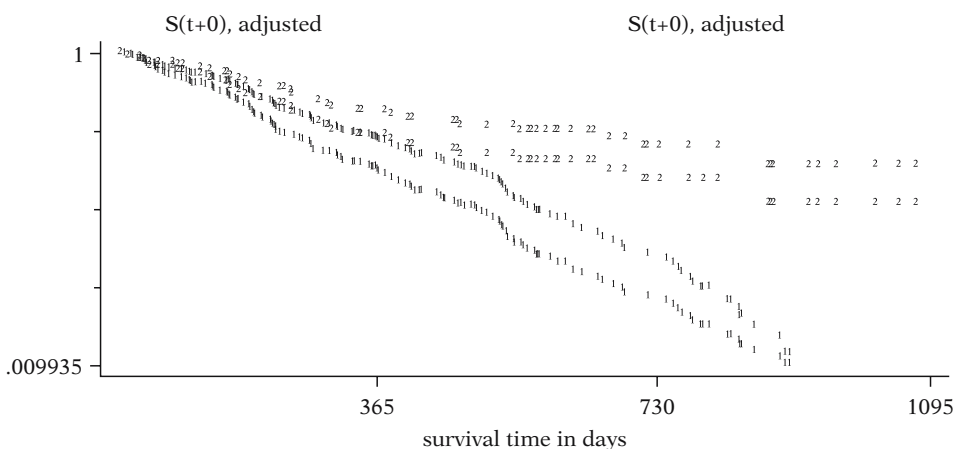
The adjusted curves suggest there is a strong effect from CLINIC on survival.

Suppose the interest is in comparing adjusted survival plots of  $\text{PRISON} = 1$  to  $\text{PRISON} = 0$  stratified by  $\text{CLINIC}$ . In this setting, the **sts graph** command cannot be used directly because we cannot simultaneously define both levels of prison ( $\text{PRISON} = 1$  and  $\text{PRISON} = 0$ ) 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  $\text{PRISON} = 0$  is defined as baseline and once where  $\text{PRISON} = 1$  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)
```

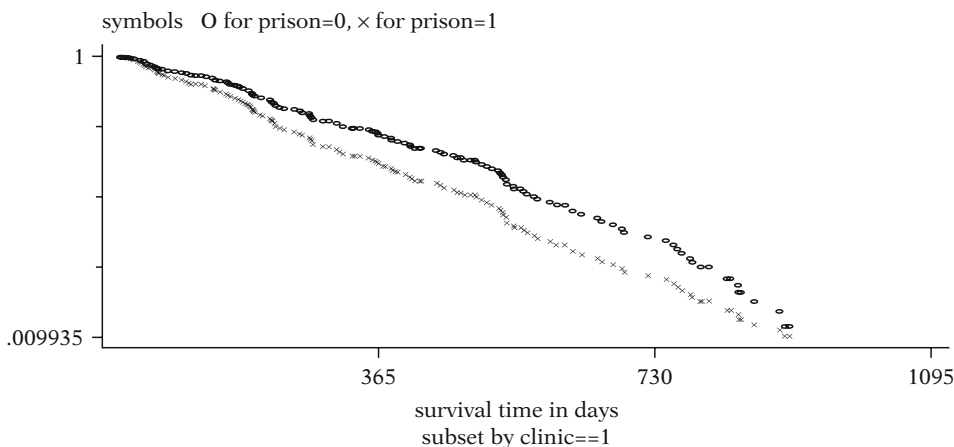
The variables  $\text{SCOX1}$  and  $\text{SCOX0}$  contain the survival estimates for  $\text{PRISON} = 1$  and  $\text{PRISON} = 0$ , respectively, adjusting for dose and stratifying by clinic. The **graph** command is used to plot these estimates. If you are using a higher version of Stata than Stata 7.0 (e.g., Stata 8.0), then you should replace the **graph** command with the **graph7** command. The code and output follow.

```
graph scox0 scox1 survt, twoway symbol([clinic] [clinic]) xlabel(365,730,1095)
```



We can also graph  $\text{PRISON} = 1$  and  $\text{PRISON} = 0$  with the data subset where  $\text{CLINIC} = 1$ . The option **twoway** requests a two-way scatterplot. 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, × for prison=1") title("subsetting 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 **tv** 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,  $\text{DOSE} \times \ln(\_t)$ , 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
Log likelihood  = -672.51694          LR chi2(4)      =          66.29
                                          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  $\text{DOSE} \times \ln(.t)$  is .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  $(.t \geq 365)$  as taking the value 1 if survival time is  $\geq 365$  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
Log likelihood = -668.57443          LR chi2(4)   =          74.17
                                          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 whereas 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 10 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 clinic dose hv1 hv2, nohr
No. of subjects   =           238                Number of obs   =           360
No. of failures   =           150
Time at risk     =           95812
Log likelihood    =   -668.57443                LR chi2(4)         =           74.17
                                                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 their 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 **tv** 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 are used for illustration. First a parametric model using the exponential distribution is demonstrated. The code and output follow.

**streg prison dose clinic, dist(exponential) nohr**

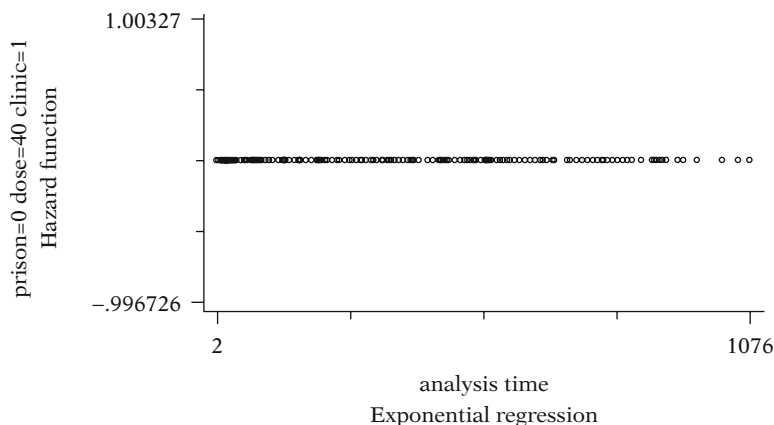
Exponential regression -- log relative-hazard form

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			
Log likelihood	=	-260.98467	LR chi2(3)	=	60.89
			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  $\lambda p t^{p-1}$ . If  $p = 1$  then the Weibull distribution is also an exponential distribution ( $h(t) = \lambda$ ). 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

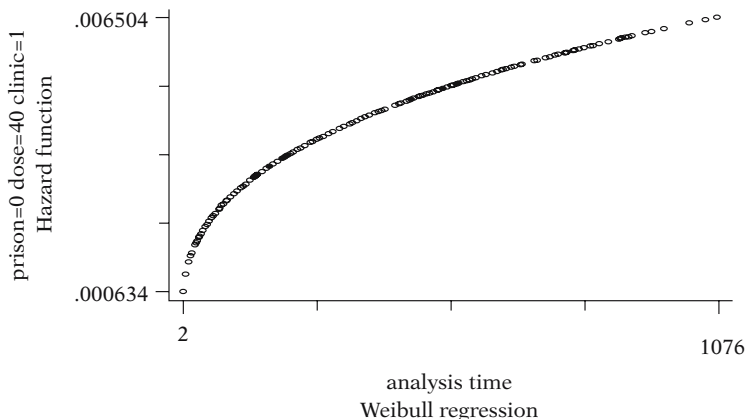
No. of subjects	=	238	Number of obs	=	238
No. of failures	=	150			
Time at risk	=	95812			
Log likelihood	=	-260.98467	LR chi2(3)	=	60.89
			Prob > chi2	=	0.0000

<i>_t</i>	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  $\beta_j$  and the AFT parameter  $\alpha_j$  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
```

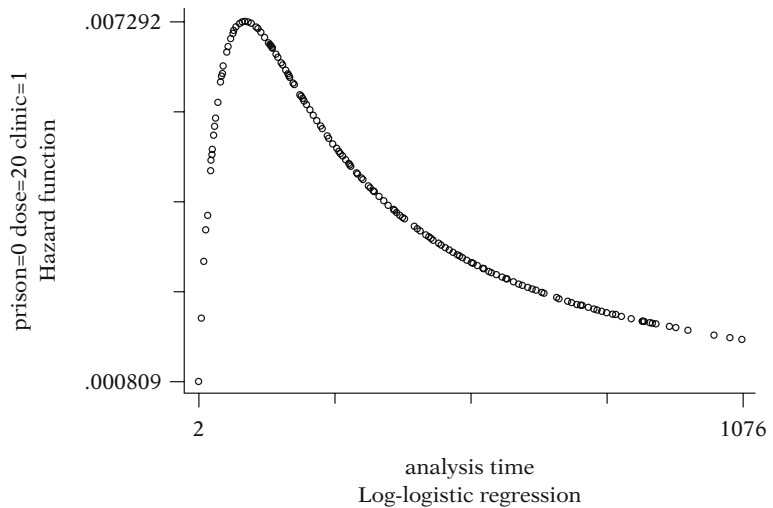
```
No. of subjects   =          238           Number of obs   =          238
No. of failures   =           150
Time at risk      =          95812
Log likelihood    = -270.42329           LR chi2(3)      =          52.18
                                           Prob > chi2     =          0.0000
```

	_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)
```

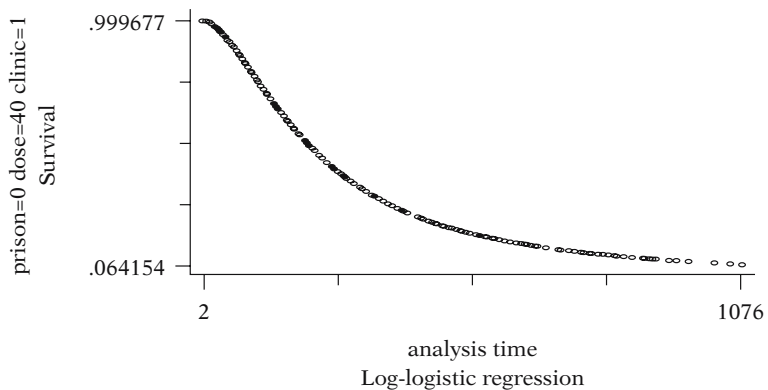




The hazard function (in contrast to the Weibull hazard function) first increases and then decreases.

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 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 three to five 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 and  $\lambda$  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
Log likelihood    = -272.65273                LR chi2(2)      =          38.87
                                                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  $\alpha$  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 code follows.

```
streg dose prison clinic, dist(weibull) frailty(gamma) nohr
```

The **frailty()** option requests that a frailty model be run. The output follows.







**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 12th through 20th 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			
Log likelihood	=	-460.07958	Wald chi2(3)	=	11.25
			Prob > chi2	=	0.0105

(standard errors adjusted for clustering on id)

	_t	_d	Coef.	Robust 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 **conditional 1** in Chapter 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
Log likelihood   = -319.85912          Wald chi2(3) =          7.11
                                          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 **conditional 2**) is a slight variation of the conditional 1 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 conditional 2 approach, the starting time at risk gets reset to zero for each subsequent event. The following code creates data suitable for using the conditional 2 approach.

```
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 conditional 2 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 conditional 2 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
Log likelihood  = -363.16022          Wald chi2(3) = 11.99
                                     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 conditional 1 and conditional 2 approaches vary slightly.

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

Log likelihood    = -184.73658          LR chi2(3)       =          8.04
                                                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 analysis 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 **stsplot** 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 as follows.

**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 are 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 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 is written in capital letters for readability. However, SAS is *not* case sensitive. If a program is written with lowercase letters, SAS reads them as uppercase. 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 A drive (i.e., on a disk). 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 'A:\';
```

The user is free to define his or 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.ADDICTS. Alternatively, a temporary SAS dataset could be created and used for these procedures.

Text that you do not wish SAS to process should 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:

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; and
8. Modeling recurrent events.

#### 1. DEMONSTRATING PROC LIFETEST TO OBTAIN KAPLAN–MEIER 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 proportional hazard 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	Number Failed	Number Left
	Survival	Failure	Standard Error		
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

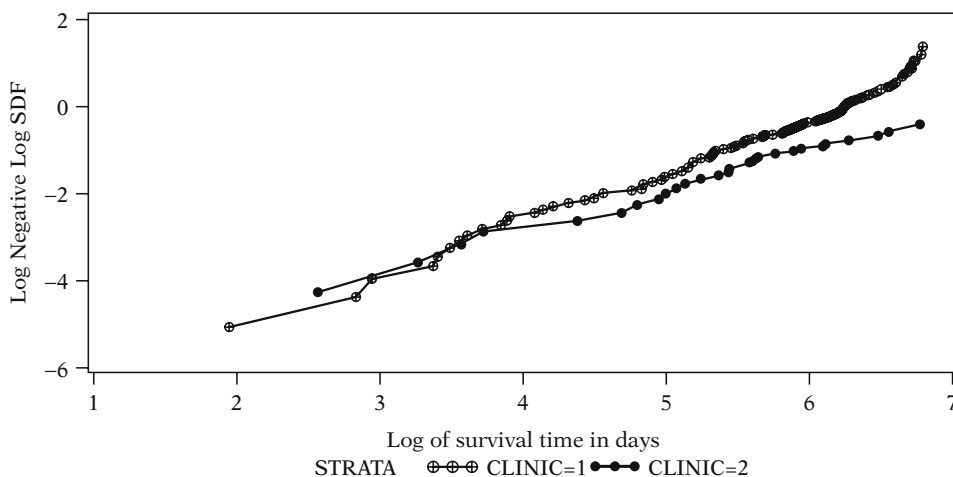
SURVT			Survival	Number Failed	Number Left
	Survival	Failure	Standard Error		
0.00	1.0000	0	0	0	75
2.00*	.	.	.	0	74
13.00	0.9865	0.0135	0.0134	1	73
26.00	0.9730	0.0270	0.0189	2	72
.	.	.	.	.	.
.	.	.	.	.	.

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) plots  $\log(\text{survival time})$  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  $\log(\text{survival time})$  or survival time is on the horizontal axis. However, if the log–log survival curves look like straight lines with  $\log(\text{survival time})$  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 what and 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;
```

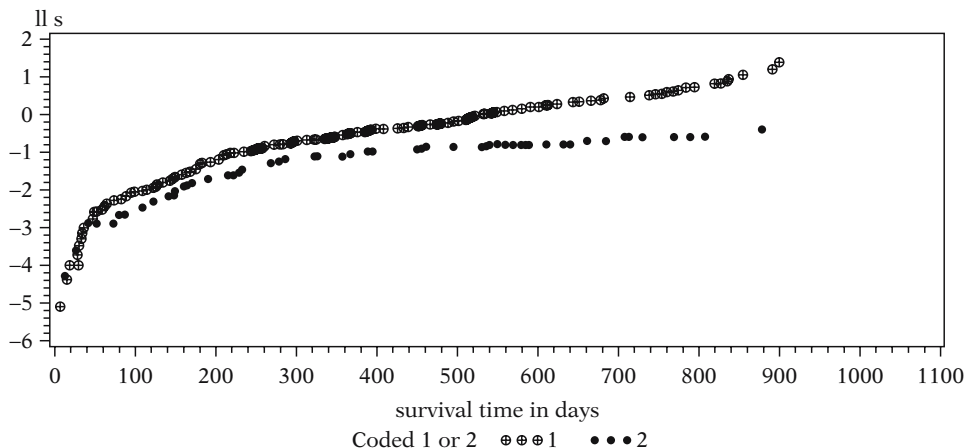
The first 10 observations from PROC PRINT are listed below.

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 plots 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 GLOT DATA=CAT;
PLOT LLS*SURVT=CLINIC;
RUN;
```



The plot has survival time (in days) rather than the default  $\log(\text{survival time})$ . 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 to 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 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

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
PRISON	1	0.32650	0.16722	3.8123	0.0509	1.386	0=none, 1=prison record
DOSE	1	-0.03540	0.00638	30.7844	<.0001	0.965	methadone dose (mg/day)
CLINIC	1	-1.00876	0.21486	22.0419	<.0001	0.365	Coded 1 or 2

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.

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 option `RL` in the `MODEL` statement of `PROC PHREG` provides 95% confidence intervals for the hazard ratio estimates.

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE CLINIC/TIES=EXACT RL;
RUN;
```

The output is shown below.

```

                                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 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 added to the model.

```

PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)=PRISON DOSE CLINIC CLIN_PR CLIN_DO;
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

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 vs. PRISON = 0 where DOSE = 0 and CLINIC = 0. This is a meaningless hazard ratio because CLINIC is coded 1 or 2 and DOSE is always greater than zero (all patients are on methadone).

The Wald chi-square p-values for the two product terms are 0.0872 for CLIN\_PR and 0.3333 for CLIN\_DO. 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 1343.199 for the full model and 1346.805 for the reduced model. The test is a 2-degree of freedom test because 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.

### 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.

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=PRISON DOSE;
STRATA CLINIC;
RUN;
```

The output of the parameter estimates follows.

The PHREG Procedure						
Analysis of Maximum Likelihood Estimates						
Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	0.38877	0.16892	5.2974	0.0214	1.475
DOSE	1	-0.03514	0.00647	29.5471	<.0001	0.965

Notice there is no parameter estimate for CLINIC because 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 because 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. 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 and CLINIC is put in the strata statement.

```
DATA ADDICTS2;
SET REF.ADDICTS;
CLIN_PR=CLINIC*PRISON;
CLIN_DO=CLINIC*DOSE;
RUN;
```

```
PROC PHREG DATA=ADDICTS2;
MODEL SURVT*STATUS(0)=PRISON DOSE CLIN_PR CLIN_DO;
STRATA CLINIC;
RUN;
```

The output of the parameter estimates 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

Note with the interaction model that the hazard ratio for PRISON = 1 vs. PRISON = 0 for CLINIC = 1 controlling for DOSE is  $\exp(\beta_1 + \beta_3)$ , and the hazard ratio for PRISON = 1 vs. 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).

$$\begin{aligned} HR &= \frac{h_0(t) \exp[1\beta_1 + \beta_2 DOSE + (2)(1)\beta_3 + \beta_4 CLIN\_DO]}{h_0(t) \exp[0\beta_1 + \beta_2 DOSE + (2)(0)\beta_3 + \beta_4 CLIN\_DO]} \\ &= \exp(\beta_1 + 2\beta_3) \end{aligned}$$

By substituting in the parameter estimates we obtain an estimated hazard ratio for PRISON at  $\exp(1.08716 + 2(-0.58467)) = 0.921$  among those at CLINIC = 2.

An alternative approach allowing for interaction with CLINIC and the other covariates is obtained by running two models: one subset on the observations where CLINIC = 1 and the other subset 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 limits the number of observations for analyses. A TITLE statement can also be added to the procedure. The output containing the parameter estimates subset 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 follow containing the parameter estimates subset 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 vs. 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 proportional hazard 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 proportional hazard 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 run the full model. 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.

```
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 shown 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 data set using the OUT= option. The variable to be ranked is SURVT (the survival time variable). 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 proportional hazard 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 p-value for testing the proportional hazard 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 shown in the column on the right and matches the p-value that was obtained using PROC CORR.

## 5. OBTAINING COX ADJUSTED SURVIVAL CURVES

We can 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 is 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 these 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.

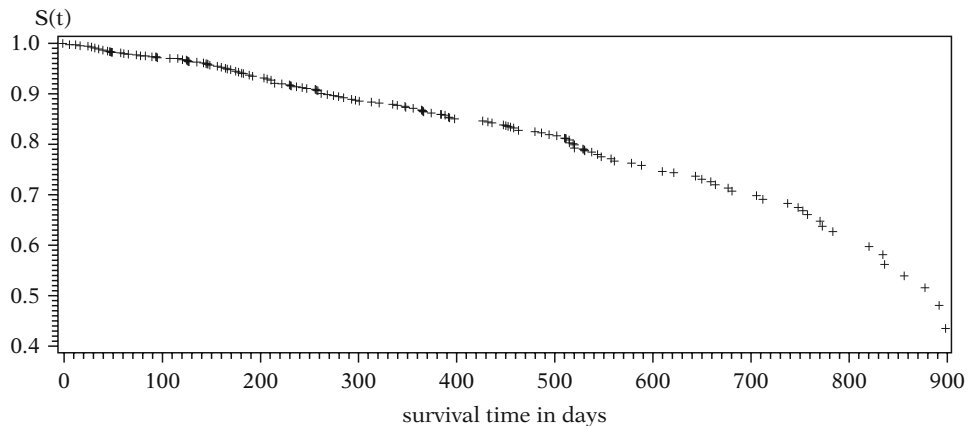
```
DATA IN1;
INPUT PRISON DOSE CLINIC;
CARDS;
0 70 2
;

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) does not use the NOMEAN option.

The output for PROC Gplot follows.



Adjusted survival for prison = 0, dose = 70, clinic = 2

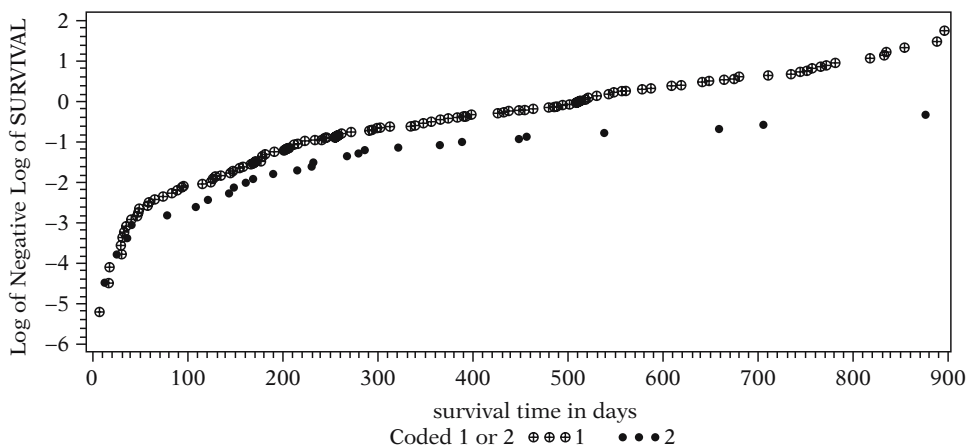
For Ex2 we wish to create and output a dataset (called OUT2) that contains the adjusted survival estimates from a Cox model stratified by CLINIC using the mean values of PRISON and DOSE. An input dataset need not be specified because by default the mean values of PRISON and DOSE will be used if the NOMEAN 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 second `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.



Log–log curves stratified by clinic, adjusted for prison, dose

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.



```

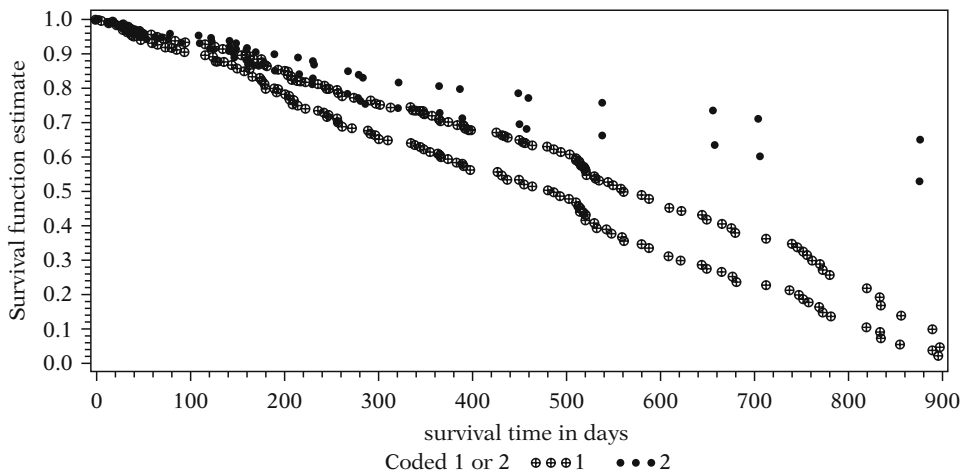
DATA IN3;
INPUT PRISON DOSE;
CARDS;
1 70
0 70
;

PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)= PRISON DOSE;
STRATA CLINIC;
BASELINE COVARIATES=IN3 OUT=OUT4 SURVIVAL=S3/NOMEAN;
RUN;

PROC GLOT DATA=OUT3;
PLOT S3*SURVT=CLINIC;
TITLE adjusted survival stratified by clinic for both levels
of prison;
RUN;

```

The PROC GLOT output follows.



Adjusted survival stratified by clinic for both levels of prison

For the above graph, the PH assumption is not assumed for CLINIC because 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 proportional hazard 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.,  $DOSE \times (TIME - 100)$ ). 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 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 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;
RUN;
```

```
PROC PHREG DATA=REF.ADDICTS;
MODEL SURVT*STATUS(0)=CLINIC PRISON DOSE HV;
IF SURVT >= 365 THEN HV = CLINIC; ELSE HV = 0;
RUN;
```

The output for the model with two Heaviside functions follows.

**The PHREG Procedure**  
**Analysis of Maximum Likelihood Estimates**

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
PRISON	1	0.37770	0.16840	5.0304	0.0249	1.459
DOSE	1	-0.03551	0.00644	30.4503	<.0001	0.965
HV1	1	-0.45956	0.25529	3.2405	0.0718	0.632
HV2	1	-1.82823	0.38595	22.4392	<.0001	0.161

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  $\exp(-0.45956) = 0.632$  and the estimated hazard ratio for CLINIC at 400 days is  $\exp(-1.82823) = 0.161$ .

The output for the model with one Heaviside function follows.

**The PHREG Procedure**  
**Analysis of Maximum Likelihood Estimates**

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio
CLINIC	1	-0.45956	0.25529	3.2405	0.0718	0.632
PRISON	1	0.37770	0.16840	5.0304	0.0249	1.459
DOSE	1	-0.03551	0.00644	30.4503	<.0001	0.965
HV	1	-1.36866	0.46139	8.7993	0.0030	0.254

Notice 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  $\exp(-0.45956) = 0.632$  and the estimated hazard ratio for CLINIC at 400 days is  $\exp((-0.45956) + (-1.36866)) = 0.161$ . 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.

Suppose it is believed that the hazard ratio for CLINIC = 2 vs. 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. The code follows (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;
```

## 7. RUNNING PARAMETRIC MODELS WITH PROC LIFEREG

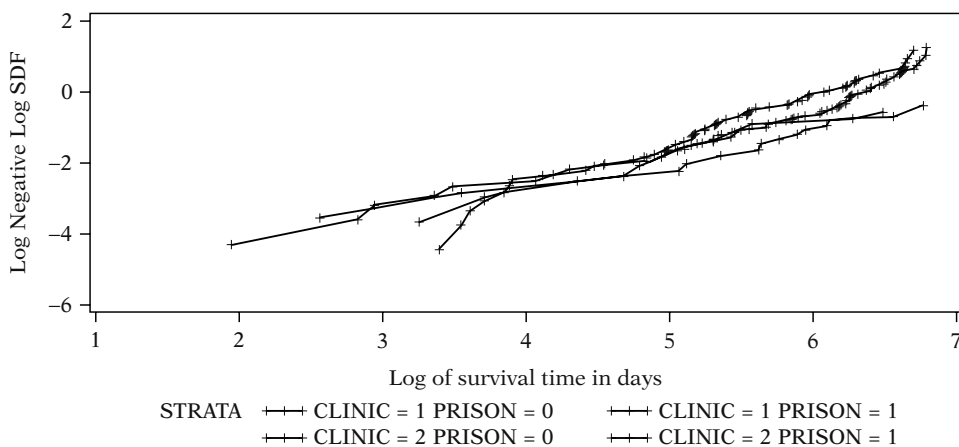
PROC LIFEREG runs parametric accelerated failure time (AFT) models rather than proportional hazards (PH) models. Whereas the key assumption of a PH model is that hazard ratios are constant over time, the key assumption for an AFT model is that survival time accelerates (or decelerates) by a constant 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 survival plots produced by PROC LIFETEST follow.



The log-log curves do not look straight but for illustration we proceed as if the Weibull assumption were appropriate. First an exponential model is 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 exponential 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  $\exp(-(3.6843 + 50(0.0289) + 2(0.8806))) = .001$ . 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 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 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 were extended from three to five 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  $\log(\text{time})$ , 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 follow 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 whereas 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. SAS version 8.2 does not support frailty survival models.

## 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 12th through 20th 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 processes data layout. The following code runs a model with the 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 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 nonrobust 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 differently compared to the corresponding model in Stata.

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 a 1st, 2nd, 3rd, or 4th event. This approach is called **conditional 1** 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;
STRATA INTERVAL;
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 **conditional 2**) is a slight variation of the conditional 1 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 conditional 2 approach, the starting time at risk gets reset to zero for each subsequent event. The following code creates data suitable for using the conditional 2 approach.

```
DATA BLADDER2 ;
  SET REF.BLADDER ;
  START2=0 ;
  STOP2=STOP-START ;
  RUN ;
```

The new dataset (BLADDER2) copies the data from REF.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 the conditional 2 approach 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 conditional 1 and conditional 2 approaches vary slightly.

The counting process data layout with multiple observations per subject need not only apply to recurrent event data but can also be used for 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 that 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 are used to illustrate these procedures. The addicts dataset was obtained from a 1991 Australian study by Coplehorn 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 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 survival time variable (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  
  /INTERVAL 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:

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; and
6. Running an extended Cox model.

SPSS version 11.5 does not provide commands to run parametric survival models, frailty models, or models using a counting processes 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 survival time variable (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 because 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 and does not all initially appear in the output window (called the SPSS viewer). In order to make it easier to view the output, try right-clicking inside the output and then select SPSS Rtf Document Object → Open to open up a text output window. The output is easier to view and can also be edited from this window.



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 STATUS 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. Click 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. To edit (and better view) the output, right-click on the output and then select SPSS Rtf Document Object → Open in order to open up a text output window.

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 four event-times from CLINIC = 1 and CLINIC = 2 as well for the log rank, Breslow, and Tarone–Ware tests follows.

#### Survival Analysis for SURVT Survival time (days)

Factor CLINIC = 1.00

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
2.00	censored			0	162
7.00	endpoint	.9938	.0062	1	161
17.00	endpoint	.9877	.0087	2	160
19.00	endpoint	.9815	.0106	3	159
28.00	censored			3	158
28.00	censored			3	157
29.00	endpoint	.9752	.0122	4	156
.	.				

Factor CLINIC = 2.00

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
2.00	censored			0	74
13.00	endpoint	.9865	.0134	1	73
26.00	endpoint	.9730	.0189	2	72
35.00	endpoint	.9595	.0229	3	71
41.00	endpoint	.9459	.0263	4	70
.	.				

Test Statistics for Equality of Survival Distributions for CLINIC

	Statistic	df	Significance
Log Rank	27.89	1	.0000
Breslow	11.63	1	.0007
Tarone-Ware	17.60	1	.0000

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 to those described above for Kaplan-Meier 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 1000 by 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 is created containing the KM survival estimates. Then another new variable is created containing the log–log of the survival estimates. Finally the log–log survival estimates are 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, STATUS as the status variable, and CLINIC as the factor variable as described above. Then click the Save button. This opens a dialog 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 and defining the new variable in the dialog 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 → Scatter and then clicking on Simple and then Define in the Scatterplot dialog 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 survival time variable (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 because 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  $\exp(\beta)$ ). 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 proportional hazards assumption is assumed to hold for all three covariates using this Cox model (the output follows).

#### Omnibus Tests of Model Coefficients<sup>a,b</sup>

-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

<sup>a</sup> Beginning Block Number 0, initial Log Likelihood function: -2 Log likelihood: 1411.324

<sup>b</sup> 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 survival time variable (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.

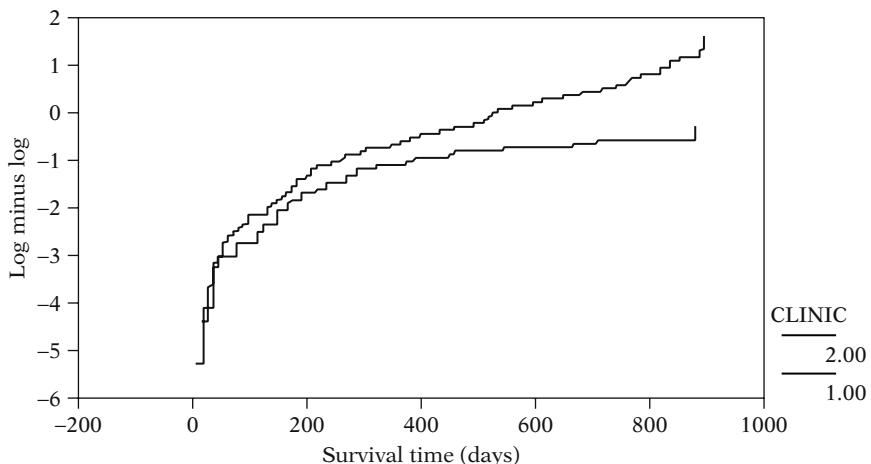
##### COXREG

```
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



LML Functions at mean of covariates

Notice that there are parameter estimates for PRISON and DOSE but not CLINIC because 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. Now click on DOSE(Mean) in the window called Covariate Values Plotted at. Type in the value 70 in the box underneath called Change Value and 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 proportional hazards 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.
4. Run correlations on the survival rankings with the Schoenfeld residuals.
5. The p-value for testing whether the correlation is zero between the ranked survival time and the covariate's Schoenfeld residuals is the p-value for the statistical test that the PH assumption is violated. 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 dialog box appears that is called Cox Regression: Save New Variables. Check Partial Residuals (under Diagnostics) 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 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 dialog box and click on `Continue`. Check `Deleted` in the box called `Unselected Cases Are`. Click `OK` and only observations with events will be kept in the dataset.

Create the variable that contains the ranking of survival times by selecting `Transform → Ranked Cases`. Select the survival time variable (`SURVT`) in 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 DO	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 dialog box 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\_ times DOSE. Enter the expression LN(T\_)\*DOSE into the dialog 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  $(T_ \geq 365) * CLINIC$ . 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 dose clinic T.COV_
/CRITERIA=PIN(.05) POUT(.10) ITERATE(20)
```

Note that SPSS recognizes the expression  $(T_ \geq 365)$  as taking the value 1 if survival time is  $\geq 365$  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  $\exp(-0.460) = 0.632$  and 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  $\geq 365$  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 hv2= (T.>= 365)* clinic
COMPUTE hv1= (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  $\exp(-1.828) = 0.161$ . These results are consistent with the estimates obtained from the previous model with one Heaviside function.

# **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  $t$ ;  $h(t)$  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.

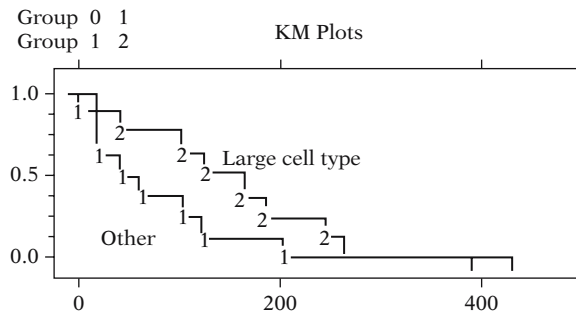
19.

$t_{(j)}$	$m_j$	$q_j$	$R(t_{(j)})$
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 survive $\geq 3.0$ 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.



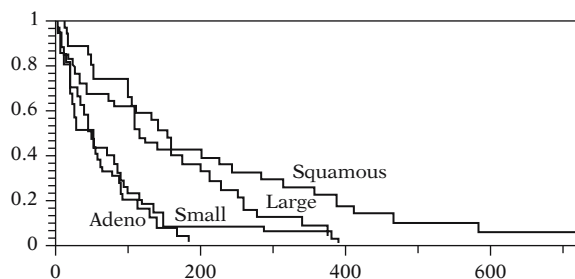
Group	Events observed	Events expected
1	102	93.45
2	26	34.55
Total	128	128.00

Log rank =  $\chi^2(2) = 3.02$

p-value =  $\Pr > \chi^2 = 0.0822$

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.

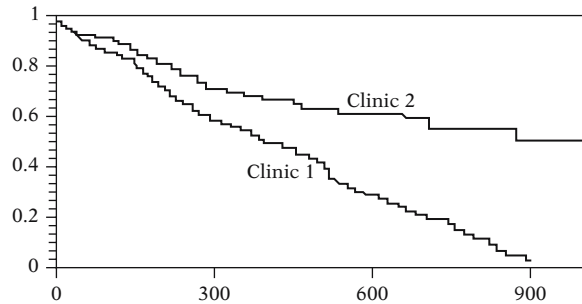
Group	Events observed	Events expected
1	26	34.55
2	26	15.69
3	45	30.10
4	31	47.65
Total	128	128.00

Log rank =  $\chi^2(2) = 25.40$

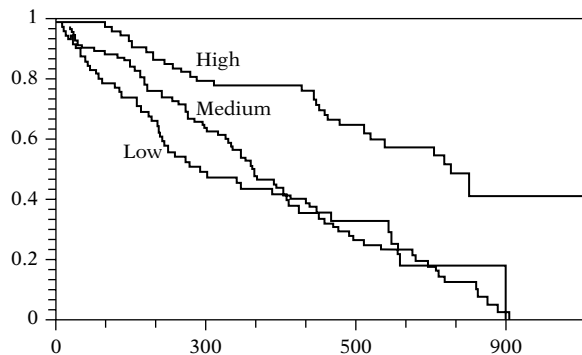
P-value =  $\Pr > \chi^2 = 0.0000$

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 =  $\chi^2(2) = 33.02$

P-value =  $\Pr > \chi^2 = 0.0000$

The test statistic is highly significant, indicating that these three curves are not equivalent.

## Chapter 3

1. a. 
$$h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 T1 + \beta_2 T2 + \beta_3 PS + \beta_4 DC + \beta_5 BF + \beta_6 (T1 \times PS) + \beta_7 (T2 \times PS) + \beta_8 (T1 \times DC) + \beta_9 (T2 \times DC) + \beta_{10} (T1 \times BF) + \beta_{11} (T2 \times BF)]$$
- b. **Intervention A:**  $\mathbf{X}^* = (1, 0, PS, DC, BF, PS, 0, DC, 0, BF, 0)$   
**Intervention C:**  $\mathbf{X} = (-1, -1, PS, DC, BF, -PS, -PS, -DC, -DC, -BF, -BF)$   

$$HR = \frac{h(t, \mathbf{X}^*)}{h(t, \mathbf{X})} = \exp[2\beta_1 + \beta_2 + 2\beta_6 PS + \beta_7 PS + 2\beta_8 DC + \beta_9 DC + 2\beta_{10} BF + \beta_{11} BF]$$
- c.  $H_0: \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$  in the full model.  
 Likelihood ratio test statistic:  $-2 \ln \hat{L}_R - (-2 \ln \hat{L}_F)$ , which is approximately  $\chi_6^2$  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 1a above)
- d. The two models being compared are:  
 Full model ( $F$ ):  $h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 T1 + \beta_2 T2 + \beta_3 PS + \beta_4 DC + \beta_5 BF]$

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  $\chi^2_2$  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 =  $264.90 - 264.70 = 0.21$ ;  $\chi^2$  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  $\exp[.8051 \pm 1.96(.3252)] = (1.183, 4.231)$  whereas the corresponding 95% confidence interval in the model without AGE (Model 1) is given by  $\exp[.8595 \pm 1.96(.3116)] = (1.282, 4.350)$ . 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  $CHR = 1$ :  $\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[0.8051 + 0.0856(\overline{AGE})]}$   
 For  $CHR = 0$ :  $\hat{S}(t, \mathbf{X}) = [\hat{S}_0(t)]^{\exp[0.0856(\overline{AGE})]}$
- 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 test for significance of this variable is significant below the .01 level.
3. a. Full model ( $F = \text{Model 1}$ ):  $h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 Rx + \beta_2 \text{Sex} + \beta_3 \log \text{WBC} + \beta_4 (Rx \times \text{Sex}) + \beta_5 (Rx \times \log \text{WBC})]$   
 Reduced model ( $R = \text{model 4}$ ):  
 $h(t, \mathbf{X}) = h_0(t) \exp[\beta_1 Rx + \beta_2 \text{Sex} + \beta_3 \log \text{WBC}]$   
 $H_0: \beta_4 = \beta_5 = 0$   
 LR statistic =  $144.218 - 139.030 = 5.19$ ;  $\chi^2$  with 2 d.f. under  $H_0$ ; not significant at 0.05, though significant 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  $Rx \times \log \text{WBC}$  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 ( $\text{Sex} = 0$ ):  $\widehat{HR} = \exp[0.405] = 1.499$   
 Females ( $\text{Sex} = 1$ ):  $\widehat{HR} = \exp[0.405 + 2.013(1)] = 11.223$
- e. Model 2 is preferred to Model 3 if one decides that the coefficients for the variables  $Rx$  and  $Rx \times \text{Sex}$  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 for 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  $\log \text{WBC}$  is a confounder; otherwise  $\log \text{WBC}$  is not a confounder. Note that the  $\log \text{WBC}$  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  $\log \text{WBC}$ .
- f. Model 2 appears to be best, because there is significant interaction of  $Rx \times \text{Sex}$  ( $P = .023$ ) and because  $\log \text{WBC}$  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 ( $P \ll .01$ ), 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 are 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  $LR = -2 \ln L_R - (-2 \ln L_F)$  where  $R$  denotes the reduced (PH) model obtained when all  $\delta$ s 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 non-significant 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  $LR = -2 \ln L_R - (-2 \ln L_F)$ , 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  $\delta_1$  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  $\delta_1$  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

- 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.
- The adjusted survival curves 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 time.
- $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \text{prison} + \beta_2 \text{dose}]$ ,  $g = 1, 2$   
 This is a no-interaction model because the regression coefficients for prison and dose are the same for each stratum.
- 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.
- Version 1:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_{1g} \text{prison} + \beta_{2g} \text{dose}]$ ,  $g = 1, 2$   
 Version 2:  $h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_1 \text{prison} + \beta_2 \text{dose} + \beta_3 (\text{clinic} \times \text{prison}) + \beta_4 (\text{clinic} \times \text{dose})]$ ,  $g = 1, 2$

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).  $LR = -2 \ln L_R - (-2 \ln L_F)$  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 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

- 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.
- The  $P(PH)$  value for the  $Tx$  variable is 0, indicating that the PH assumption is not satisfied for the treatment variable based on this goodness-of-fit test.
- $h(t, \mathbf{X}) = h_0(t)\exp[\beta_1(Tx)g_1(t) + \beta_2(Tx)g_2(t) + \beta_3(Tx)g_3(t)]$

where

$$g_1(t) = \begin{cases} 1 & \text{if } 0 \leq t < 250 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_2(t) = \begin{cases} 1 & \text{if } 250 \leq t < 500 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$

$$g_3(t) = \begin{cases} 1 & \text{if } t \geq 500 \text{ days} \\ 0 & \text{if otherwise} \end{cases}$$

- 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 intervals 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 chemotherapy 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.

$$\gamma = \frac{\exp[\alpha_0 + \alpha_1(2) + \alpha_2PRISON + \alpha_3DOSE + \alpha_4PRISDOSE]}{\exp[\alpha_0 + \alpha_1(1) + \alpha_2PRISON + \alpha_3DOSE + \alpha_4PRISDOSE]}$$

$$= \exp(\alpha_1)$$

$$\hat{\gamma} = \exp(0.698) = 2.01$$

$$95\% \text{ CI} = \exp[0.698 \pm 1.96(0.158)] = (1.47, 2.74)$$

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.

$$HR = \frac{\exp[\beta_0 + \beta_1(2) + \beta_2PRISON + \beta_3DOSE + \beta_4PRISDOSE]}{\exp[\beta_0 + \beta_1(1) + \beta_2PRISON + \beta_3DOSE + \beta_4PRISDOSE]}$$

$$= \exp(\beta_1)$$

$$\hat{HR} = \exp(-0.957) = 0.38$$

$$95\% \text{ CI} = \exp[-0.957 \pm 1.96(0.213)] = (0.25, 0.58)$$

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  $\hat{\beta}_1 = -(0.698)(1.370467) = -0.957$ , 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:

$$t = [-\ln S(t)]^{1/p} \times \frac{1}{\lambda^{1/p}} = [-\ln(0.5)]^{1/p} \times \exp[\beta_0 + 2\beta_1 + \beta_2 + 50\beta_3 + 100\beta_4]$$

$\hat{t}$  (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:

$$t = [-\ln(0.5)]^{1/p} \times \exp[\beta_0 + 1\beta_1 + \beta_2 + 50\beta_3 + 100\beta_4]$$

$\hat{t}$  (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  $S(t) = 0.5$  (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 ( $\alpha$ ) is constant at 1. Frailty is defined as a multiplicative random effect on the hazard  $h(t|\alpha) = \alpha h(t)$ . If  $\alpha = 1$  then  $h(t|\alpha) = h(t)$ , and there is no frailty.

## Chapter 8

- Cox PH Model for **CP** approach to Defibrillator Study:

$$h(t, \mathbf{X}) = h_0(t) \exp[\beta \mathbf{tx} + \gamma \text{smoking}]$$

where  $\mathbf{tx} = 1$  if treatment A, 0 if treatment B,  
**smoking status** = 1 if ever smoked, 0 if never smoked
- 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).
- No-interaction SC model for marginal approach:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta \mathbf{tx} + \gamma \text{smoking}], \quad g = 1, 2, 3$$

Interaction SC model for marginal approach:

$$h_g(t, \mathbf{X}) = h_{0g}(t) \exp[\beta_g \mathbf{tx} + \gamma_g \text{smoking}], \quad g = 1, 2, 3$$
- $\mathbf{LR} = -2 \ln L_R - (-2 \ln L_F)$  is approximately  $\chi^2$  with 4 df under

$H_0$ : no-interaction SC model is appropriate, where R denotes the reduced (no interaction SC) model and F denotes the full (interaction SC) model.
- 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.
- 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.
- 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.
- The number in the risk set ( $\mathbf{n}_j$ ) remains unchanged through day 68 because every subject who failed by this time was still at risk for a later event.
- Subjects 3, 6, 10, 26, and 31 all fail for the third time at day 98 and are not followed afterwards.
- Subjects 9, 15, and 28 fail for the second time at 79 days, whereas subject #16 is censored at 79 days.
- Subjects 4, 14, 15, 24, and 29 were censored between days 111 and 112.
- 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.

13. i. None of the above.  
 ii. 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.”
14. Use the information provided in Table T.2 to complete the data layouts for plotting the following survival curves.
- a.  $S_1(t) = \Pr(T_1 > t)$  where  $T_1 =$  time to first event from study entry

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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
<b>51</b>	<b>7</b>	<b>2</b>	<b>0</b>	<b><math>0.19 \times 5/7 = 0.14</math></b>
<b>57</b>	<b>5</b>	<b>2</b>	<b>0</b>	<b><math>0.14 \times 3/5 = 0.08</math></b>
<b>58</b>	<b>3</b>	<b>2</b>	<b>0</b>	<b><math>0.08 \times 1/3 = 0.03</math></b>
<b>61</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b><math>0.03 \times 0/1 = 0.00</math></b>

- b. **Conditional**  $S_{2c}(t) = \Pr(T_{2c} > t)$  where  $T_{2c} =$  time to second event from first event.

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S(t_{(j)}) = S(t_{(j-1)}) \times \Pr(T_1 > t   T_1 \geq t)$
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

(Continued)

*(Continued)*

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S_{(t_{(j)})} = S_{(t_{(j-1)})} \times \Pr(T_1 > t \mid T_1 \geq t)$
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
<b>40</b>	<b>6</b>	<b>2</b>	<b>0</b>	<b><math>0.17 \times 4/6 = 0.12</math></b>
<b>41</b>	<b>4</b>	<b>1</b>	<b>0</b>	<b><math>0.12 \times 3/4 = 0.09</math></b>
<b>42</b>	<b>3</b>	<b>1</b>	<b>0</b>	<b><math>0.09 \times 2/3 = 0.06</math></b>
<b>46</b>	<b>2</b>	<b>1</b>	<b>0</b>	<b><math>0.06 \times 1/2 = 0.03</math></b>
<b>47</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b><math>0.03 \times 0/1 = 0.00</math></b>

- c. **Marginal**  $S_{2m}(t) = \Pr(T_{2m} > t)$  where  $T_{2m}$  = time to second event from study entry.

$t_{(j)}$	$n_j$	$m_j$	$q_j$	$S_{(t_{(j)})} = S_{(t_{(j-1)})} \times \Pr(T_1 > t \mid T_1 \geq t)$
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
<b>79</b>	<b>9</b>	<b>3</b>	<b>1</b>	<b><math>0.25 \times 6/9 = 0.17</math></b>
<b>80</b>	<b>5</b>	<b>2</b>	<b>0</b>	<b><math>0.17 \times 3/5 = 0.10</math></b>
<b>81</b>	<b>3</b>	<b>2</b>	<b>0</b>	<b><math>0.10 \times 1/3 = 0.03</math></b>
<b>97</b>	<b>1</b>	<b>1</b>	<b>0</b>	<b><math>0.03 \times 0/1 = 0.00</math></b>

15. 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):

$$h_1(t, \mathbf{X}) = h_{01}(t) \exp[\beta_{11}tx + \beta_{21}\text{num} + \beta_{31}\text{size}]$$

2. Censored subjects have bladder metastasis (event = 2) or other metastasis (event = 3).  
 3. Cause-specific no-interaction model for bladder metastasis (event = 1):

$$h_2(t, \mathbf{X}) = h_{02}(t) \exp[\beta_{12}tx + \beta_{22}\text{num} + \beta_{32}\text{size}]$$

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  $t_j = 8$  for persons in the treatment group ( $\mathbf{tx} = 1$ ):

$$\hat{h}_1(8) = 1/23 = 0.0435$$

$$\begin{aligned} \hat{S}(4) &= \hat{S}(3) \Pr(T > 4 | T \geq 4) = 0.9630(1 - 2/26) \\ &= 0.9630(0.9231) = 0.8889 \end{aligned}$$

$$\hat{I}_1(8) = \hat{h}_1(8) \hat{S}(4) = 0.0435(.8889) = 0.0387$$

$$CIC_1(8) = CIC_1(4) + 0.0387 = 0 + 0.0387 = 0.0387$$

- b. Verify the  $CIC_1$  calculation provided at failure time  $t_j = 25$  for persons in the placebo group ( $\mathbf{tx} = 0$ ):

$$\hat{h}_1(25) = 1/6 = 0.1667$$

$$\begin{aligned} \hat{S}(23) &= \hat{S}(21) \Pr(T > 23 | T \geq 23) = 0.4150(1 - 1/8) \\ &= 0.4150(0.875) = 0.3631 \end{aligned}$$

$$\hat{I}_1(25) = \hat{h}_1(25) \hat{S}(23) = 0.1667(.3631) = 0.0605$$

$$\begin{aligned} CIC_1(25) &= CIC_1(23) + 0.0605 = 0.2949 + 0.0605 \\ &= 0.3554 \end{aligned}$$

- c. Interpret the  $CIC_1$  values obtained for both the treatment and placebo groups at  $t_j = 30$ .

For  $\mathbf{tx} = 1$ ,  $CIC_1(t_j = 30) = 0.3087$  and for  $\mathbf{tx} = 0$ ,  $CIC_1(t_j = 30) = 0.3554$ .

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 ( $tx = 0$ ), 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  $CPC_1$  values for both treatment and placebo groups at  $t_i = 30$ :

The formula relating CPC to CIC is given by

$CPC_c = CIC_c / (1 - CIC_{c'})$  where  $CIC_c = CIC$  for cause-specific risk event = 1 and  $CIC_{c'} = CIC$  from risks for events = 2 or 3 combined

For  $tx = 1$ ,  $CIC_1(t_i = 30) = 0.3087$  and for  $tx = 0$ ,  $CIC_1(t_i = 30) = 0.3554$ .

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  $CIC_{c'}$  involves the following calculations.

$tx = 1$  (Treatment A)

$t_j$	$n_j$	$d_{1j}$	$\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_1(t_j)$	$CIC_{1'}(t_j)$
0	27	0	0	—	—	—
2	27	1	<b>.0370</b>	<b>1</b>	<b>.0370</b>	<b>.0370</b>
3	26	2	<b>.0769</b>	<b>.9630</b>	<b>.0741</b>	<b>.1111</b>
4	24	0	0	.8889	0	<b>.1111</b>
<b>8</b>	<b>23</b>	<b>1</b>	<b>.0435</b>	<b>.8889</b>	<b>.0387</b>	<b>.1498</b>
9	21	1	<b>.0476</b>	<b>.8116</b>	<b>.0386</b>	<b>.1884</b>
10	20	1	<b>.0500</b>	<b>.7729</b>	<b>.0386</b>	<b>.2270</b>
15	17	1	<b>.0588</b>	<b>.7343</b>	<b>.0432</b>	<b>.2702</b>
16	15	1	<b>.0667</b>	<b>.6479</b>	<b>.0432</b>	<b>.3134</b>
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	<b>.2500</b>	<b>.3779</b>	<b>.0945</b>	<b>.4079</b>
29	2	0	0	.2835	0	.4079
30	1	0	0	.2835	0	.4079

tx = 0 (Placebo)

$t_j$	$n_j$	$d_{1j}$	$\hat{h}_1(t_j)$	$\hat{S}(t_{j-1})$	$\hat{I}_1(t_j)$	$CIC_1(t_j)$
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	<b>.0476</b>	<b>.8413</b>	<b>.0400</b>	<b>.0400</b>
6	20	2	<b>.1000</b>	<b>.8013</b>	<b>.0801</b>	<b>.1201</b>
7	18	1	<b>.0556</b>	<b>.7212</b>	<b>.0401</b>	<b>.1602</b>
10	16	1	<b>.0625</b>	<b>.6811</b>	<b>.0426</b>	<b>.2028</b>
12	15	1	<b>.0667</b>	<b>.6385</b>	<b>.0426</b>	<b>.2454</b>
14	13	0	0	.6835	0	.2454
16	12	1	<b>.0833</b>	<b>.5534</b>	<b>.0461</b>	<b>.2915</b>
17	10	0	0	.4612	0	.2915
18	9	0	0	.4150	0	.2915
21	8	1	<b>.1250</b>	<b>.4150</b>	<b>.0519</b>	<b>.3434</b>
23	7	0	0	.3632	0	.3434
<b>25</b>	<b>6</b>	<b>1</b>	<b>.1667</b>	<b>.3632</b>	<b>.0605</b>	<b>.4039</b>
29	4	0	0	.2421	0	.4039
30	2	0	0	.2421	0	.4039

From these tables, we find that for tx = 1,  $CIC_{1'}(t_i = 30) = 0.4079$  and for tx = 0,  $CIC_{1'}(t_j = 30) = 0.4039$ .

Thus, for tx = 1,  $CPC_1(t_j = 30) = CIC_1/(1 - CIC_{1'}) = 0.3087/(1 - 0.4079) = 0.5213$  and for tx = 0,  $CPC_1(t_j = 30) = CIC_1/(1 - CIC_{1'}) = 0.3554/(1 - 0.4039) = 0.5962$ .

6. a. What is the effect of treatment on survival from having a local recurrence of bladder cancer and is it significant?

$$HR_1(\text{tx} = 1 \text{ vs. } \text{tx} = 0) = 0.535 (= 1/1.87),$$

$$\text{p-value} = 0.250, \text{ N.S.}$$

- b. What is the effect of treatment on survival from developing metastatic bladder cancer and is it significant?

$$HR_2(\text{tx} = 1 \text{ vs. } \text{tx} = 0) = 0.987,$$

$$\text{p-value} = .985, \text{ N.S.}$$

- c. What is the effect of treatment on survival from other metastatic cancer and is it significant?

$$HR_3(\text{tx} = 1 \text{ vs. } \text{tx} = 0) = 0.684 (= 1/1.46),$$

$$\text{p-value} = .575, \text{ N.S.}$$



7. a. State the hazard model formula for the LM model used for the above output.

$$h_{g=1,2,3}^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 tx + \beta_2 \text{num} + \beta_3 \text{size} \\ + \delta_1(\text{txd2}) + \delta_2(\text{numd2}) \\ + \delta_3(\text{sized2}) + \delta_4(\text{txd3}) \\ + \delta_5(\text{numd3}) + \delta_6(\text{sized3})]$$

where d2 = 1 if bladder metastasis and 0 otherwise,  
and  
d3 = 1 if or other metastasis and 0 otherwise

- b. Determine the hazard ratios for the effect of each of the 3 cause-specific events based on the above output.

$$\begin{aligned} \text{HR}_1(\text{tx} = 1 \text{ vs. } \text{tx} = 0) &= \exp(-0.6258) \\ &= 0.535 (= 1/1.87) \end{aligned}$$

$$\begin{aligned} \text{HR}_2(\text{tx} = 1 \text{ vs. } \text{tx} = 0) &= \exp(-0.6258 + .6132) \\ &= 0.987 \end{aligned}$$

$$\begin{aligned} \text{HR}_3(\text{tx} = 1 \text{ vs. } \text{tx} = 0) &= \exp(-0.6258 + .2463) \\ &= 0.684 (= 1/1.46) \end{aligned}$$

- c. Verify that the hazard ratios computed in Part b are identical to the hazard ratios computed in Question 6. The HRs are identical.

8. a. State the hazard model formula for the LM<sub>alt</sub> model used for the above output.

$$h_{g=1,2,3}'(t, \mathbf{X}) = h_{0g}'(t) \exp[\delta'_{11} \text{txd1} + \delta'_{12} \text{numd1} + \delta'_{13} \text{sized1} \\ + \delta'_{21} \text{txd2} + \delta'_{22} \text{numd2} \\ + \delta'_{23} \text{sized2} + \delta'_{31} \text{txd3} \\ + \delta'_{32} \text{numd3} + \delta'_{33} \text{sized3}]$$

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. Determine the hazard ratios for the effect of each of the three cause-specific events based on the above output.

$$\begin{aligned} \text{HR}_1(\text{tx} = 1 \text{ vs. tx} = 0) &= \exp(-0.6258) \\ &= 0.535 (= 1/1.87) \end{aligned}$$

$$\text{HR}_2(\text{tx} = 1 \text{ vs. tx} = 0) = \exp(-0.0127) = 0.987$$

$$\begin{aligned} \text{HR}_3(\text{tx} = 1 \text{ vs. tx} = 0) &= \exp(-0.3796) \\ &= 0.684 (= 1/1.46) \end{aligned}$$

- c. Verify that the hazard ratios computed in Part b are identical to the hazard ratios computed in Questions 6 and 7.

Corresponding hazard ratios are identical.

9. State the formula for a no-interaction SC LM model for these data.

$$h_{g}^*(t, \mathbf{X}) = h_{0g}^*(t) \exp[\beta_1 \text{tx} + \beta_2 \text{num} + \beta_3 \text{size}]$$

$g=1,2,3$

Assumes  $\text{HR}_1(X) = \text{HR}_2(X) = \text{HR}_3(X)$  for any X variable  
e.g.,  $R_x = 0$  vs.  $R_x = 1$ :

$$\text{HR}_1(\text{tx}) = \text{HR}_2(\text{tx}) = \text{HR}_3(\text{tx}) = \exp[\beta_1]$$

10. Describe how you would test whether a no-interaction SC LM model would be more appropriate than an interaction SC LM model.

Carry out the following likelihood ratio test:

$$H_0: \delta_{gj} = 0, \quad g = 2, 3; \quad j = 1, 2, 3$$

where  $\delta_{gj}$  is coefficient of  $D_g X_j$  in the interaction SC LM model

Likelihood Ratio test:  $LR = -2 \log L_R - (-2 \log L_F)$   
approx  $\chi_G^2$  under  $H_0$

R = no-interaction SC (reduced) model

F = interaction SC (full) model

# References

---

Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. 1993. *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

AREDS Research Group. 2003. Potential public health impact of age-related eye disease study results. *Arch. Ophthalmol.* 121: 1621–1624.

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*. New York: Plenum, 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.

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. *Stat. Meth. Med. Res.* 3: 244–262.

Cox, D.R. and Oakes, D. 1984. *Analysis of Survival Data*. London: Chapman & Hall.

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.

Freireich, E.O. et al. 1963. The effect of 6-mercaptopymine on the duration of steroid induced remission in acute leukemia. *Blood*, 21: 699–716.

Gebski, V. 1997. *Analysis of Censored and Correlated Data (ACCORD)*. Eastwood, NSW, Australia: Data Analysis and Research Technologies.

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. *Ann. Stat.* 16: 1141–1154.

Gutierrez, R.G. 2002. Parametric frailty and shared frailty survival models. *Stata J.* 2: 22–44.

Harrel, F. and Lee, K. 1986. *Proceedings of the Eleventh Annual SAS User's Group International*. 823–828.

Harris, E. and Albert, A. 1991. *Survivorship Analysis for Clinical Studies*. New York: Marcel Dekker.

Hosmer, D.W. and Lemeshow, S. 1989. *Applied Survival Analysis*. New York: Wiley.

Kalbfleisch, J.D. and Prentice, R.L. 1980. *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Kay, R. 1986. Treatment effects in competing-risk analysis of prostate cancer data. *Biometrics* 42: 203–211.

Kleinbaum, D.G. and Klein, M. 2002. *Logistic Regression—A Self Learning Text*, Second Edition. New York: Springer-Verlag, Chapter 11.

Kleinbaum, D.G., Kupper, L.L., and Morgenstern, H. 1982. *Epidemiologic Research: Principles and Quantitative Methods*. New York: Van Nostrand Reinhold.

Kleinbaum, D.G., Kupper, L.L., and Muller, K.A. 1987. *Applied Regression Analysis and Other Multivariable Methods*, Second Edition. Belmont, CA: Wadsworth.

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.

Lee, E.T. 1980. *Statistical Methods for Survival Data Analysis*. Belmont, CA: Wadsworth.

Lin, D.Y. and Wei, L.J. 1989. The robust inference for the Cox proportional hazards model. *J. Amer. Stat. 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 Res. Hum. Retroviruses* 4: 305–316.

Pepe, M.S. and Mori, M. 1993. Kaplan–Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Stat. Med.* 12: 737–751.

Prentice, R.L. and Marek. 1979. A qualitative discrepancy between censored data rank tests. *Biometrics* 34.

- 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.
- 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. *Control. Clin. 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.
- Wei, L.J., Lin, D.Y., and Weissfeld, L. 1989. Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *J. Amer. Stat. Assoc.* 84: 408.
- Zeger, S.L. and Liang, C.Y. 1986. Longitudinal data analysis for discrete and continuous outcomes. *Biometrics* 42: 121–130.

# Index

---

- Accelerated failure time (AFT), *see* AFT entries
- Acceleration factor, 266–267
  - exponential, 269–271
  - with frailty, 303, 309
  - log-logistic, 281
  - Weibull, 276
- Addicts dataset, 464
  - data analysis, 230–234, 248
  - with SAS programming, 508–538
  - with SPSS programming, 542–555
  - with STATA programming, 467–502
- Additive failure time model, 285
- Adjusted survival curves, 104
  - log-log plots, 159
  - observed versus expected plots, 150
  - stratified Cox procedure, 180
  - using Cox PH model, 103–107, 116, 118
- AFT (accelerated failure time) assumption, 266–268
- AFT models, 265–266, 282–284, 309, 313–317
- Age-Related Eye Disease Study (AREDS), 359–364
- AIC (Akaike’s information criterion), 286

- Average hazard rate, 24
  - in Cox PH model, 94–95, 134
  - in extended Cox model, 219
  - in stratified Cox model, 181
- Baseline hazard function, 94, 134
- Biased results, 405
- Binary regression, 290
- Bladder cancer dataset, 464–465
- Bladder cancer patients
  - comparison of results for, 355–357
  - counting process for first 26 subjects, 339
  - hypothetical subjects, 336–337
  - interaction model results for, 354
  - no-interaction model results for, 354
- Byar data, 399–400
  - cause-specific competing risk analysis, 401–403
  - Lunn-McNeil models, 422, 427
- Cause-specific hazard function, 400
- Censored data, 5–8
  - interval-censored, 286, 289
  - left-censored, 7, 286
  - right-censored, 7, 286
- Censoring, 5
  - non-informative (independent), 403–407
  - informative (dependent), 405–406
- CIC (cumulative incidence curves), 393, 412–420
- Competing risks, 4, 392, 396
  - examples of data, 396–398
  - CIC, 412
  - CPC (Conditional Probability Curves), 420
  - independence assumption, 403
  - Lunn-McNeil models, 421, 427
  - separate models for different event types, 400–403
- Complementary log-log binary model, 293
- Complementary log-log link function, 292
- Computer
  - augmented (Lunn-McNeil approach) data layout for, 422
  - counting process data layout for, 337–338
  - general data layout for, 15–19
  - marginal approach data layout for, 348–349
- Conditional failure rate, 11
- Conditional probability curves (CPC), 420–421
- Conditional survival function, 295
- Confounding effect, 26–27
- Counting process (CP) approach, 334–336
  - example, 336–337
  - general data layout, 338–344
  - results for example, 346–347
- Cox adjusted survival curves
  - using SAS, 525–529
  - using SPSS, 549–550
  - using Stata, 485–488
- Cox likelihood, 111–115
  - extended for time dependent variables, 239–242
- Cox PH cause-specific model, 400
- Cox proportional hazards (PH) model
  - adjusted survival curves using, 103–107
  - computer example using, 86–94
  - extension of, *see* Extended Cox model
  - formula for, 94–96
  - maximum likelihood estimation of, 98–100
  - popularity of, 96–98
  - review of, 214–216
  - using SAS, 514–518
  - using SPSS, 548
  - using Stata, 476–481
- CP, *see* Counting process approach
- CPC (conditional probability curves), 420–421
- Cumulative incidence, viii
- Cumulative incidence curves (CIC), 393, 412–420
- Data layout for computer, 15–19
- Datasets, 464–465
- Decreasing Weibull model, 13
- Discrete survival analysis, 293



- Empirical estimation, 344
- Estimated  $-\ln(-\ln)$  survivor curves, 136
- Estimated survivor curves, 25
- Evans County Study,
  - Cox proportional hazards (PH) model application to, 124–126
  - Kaplan-Meier survival curves for, 73–75
  - multivariable example using, 29–31
  - ordered failure times for, 41–42
  - survival data from, 119–122
- Event, 4
- Event types, different, separate models for, 400–403
- Expected versus observed plots, 145–150
- Exponential regression, 12
  - accelerated failure-time form, 319
  - log relative-hazard form, 319
- Extended Cox likelihood, 239–242
- Extended Cox model, 110, 219
  - application to Stanford heart transplant data, 235–239
  - application to treatment of heroin addiction, 230–234
  - hazard ratio formula for, 221–223
  - using SAS, 530–533
  - using SPSS, 552–555
  - using Stata, 488–492
  - time-dependent variables, 219–221
- Failure, 4
- Failure rate, conditional, 11
- Flemington-Harrington test, 65
- Frailty component, 295
- Frailty effect, 300
- Frailty models, 294–308
  - using Stata, 499–502
- Gamma distribution, 296
- Gamma frailty, 301
- Gastric carcinoma data, 253–254
- General stratified Cox (SC) model, 180–181
- Generalized gamma model, 284
- GOF, *see* Goodness-of-fit entries
- Gompertz model, 285
- Goodness-of-fit (GOF) testing approach, 151–153
- Goodness-of-fit (GOF) tests, 136
- Hazard function, 8, 9–10
  - cause-specific, 400
  - probability density function and, 262–263
- Hazard ratio, 32
  - formula for Cox PH model, 100–103
  - formula for extended Cox model, 221–223
- Heaviside function, 227
- Increasing Weibull model, 13
- Independence assumption, 403–411
- Information matrix, 346
- Informative censoring, in competing risks survival analysis, 405
- Instantaneous potential, 49
- Interactions, 27
- Interval-censored data, 286, 289–294
- Inverse-Gaussian distribution, 296
- Kaplan-Meier (KM) curves, 46
  - example of, 51–55
  - general features of, 56–57
  - log-log survival curves, 141
- Left-censored data, 7–8, 286
- Leukemia remission-time data, 17, 26
  - Cox proportional hazards (PH) model application to, 86–94
  - increasing Weibull for, 13
  - Kaplan-Meier survival curves for, 51–55, 75–77
  - log-log KM curves for, 141–145
  - recurrent event data for, 335
  - exponential survival, 263–265
  - stratified Cox (SC) model application to, 176–188
- Likelihood function, 111–115
  - for Cox PH model,
  - for extended Cox model, 239–242
  - for parametric models, 286–289

- Likelihood ratio (LR) statistic, 89
- LM, *see* Lunn-McNeil approach
- Log-log plots, 137–145
- Log–log survival curves, 137–145
- Log-logistic regression, 277–282
  - accelerated failure-time form, 320
- Log-rank test, 46
  - alternatives to, 63–68
  - for several groups, 61–63
  - for two groups, 57–61
- Logit link function, 292
- Lognormal survival models, 13, 284
- Lunn-McNeil (LM) approach, 393, 421–427
  - alternative, 427–434
- Macular degeneration data set, 359–364
  - results for, 361–362
  - Marginal probability, 414
- Maximum likelihood (ML) estimation of
  - Cox PH model, 98–100
- Multiplicative model, 285
- No-interaction assumption in stratified
  - Cox model, 182–188
- Noninformative censoring, 403–407
- Observed versus expected plots, 145–150
- Parametric approach using shared frailty, 357–359
- Parametric survival models,
  - defined, 260
  - examples
    - exponential model, 263–265, 268–277
    - log-logistic model, 277–282
    - Weibull model, 272–277
  - likelihood function, 286–289
  - other models, 284–286
  - SAS use, 533–538
  - Stata use, 492–499
- Pepe-Mori test, 421, 437
- Peto test, 63
- PH assumption
  - assessment using time-dependent covariates, 153–157, 224–229
  - evaluating (Chapter 4), 132–171
    - meaning of, 107–111
    - assessment using goodness of fit test with Schonfeld residuals, 151–153
    - assessment using Kaplan-Meier log-log survival curves, 546–547
    - assessment using observed versus expected plots, 145–150
    - assessment using SAS, 522–525
    - assessment using SPSS, 550–552
    - assessment using Stata, 473–476, 483–485
- PH model, Cox, *see* Cox proportional hazards (PH) model
- PO (proportional odds) assumption, 292
- Precision, 92
- Probability, 11
- Probability density function, 262–263
- PROC LIFEREG (SAS), 533–538
- PROC LIFETEST (SAS), 510–514
- PROC PHREG, 514–518
- Product limit formula, 46
- Proportional hazards, *see* PH entries
- Proportional odds (PO) assumption, 292
- Recurrent event survival analysis (Chapter 8), 332–390
  - Counting process approach, 336–344
  - definition of recurrent events, 4, 332
  - examples of recurrent event data, 334–336
  - other approaches for analysis 347–353
  - parametric approach using shared frailty, 357–359
  - SAS modeling, 538–542
  - Stata modeling, 502–508
  - survival curves with, 364–367
- Right-censored data, 7, 286
- Risk set, 22
- Robust estimation, 344–346
- Robust standard error, 347
- Robust variance, 345
- SAS, 508–542
  - assessing PH assumption, with statistical tests, 522–525

- demonstrating PROC LIFETEST, 510–514
- modeling recurrent events, 538–542
- obtaining Cox adjusted survival curves, 525–529
- running Cox proportional hazards (PH) model, with PROC PHREG, 514–518
- running extended Cox model, 530–533
- running parametric models, with PROC LIFEREG, 533–538
- running stratified Cox (SC) model, 519–521
- SC, *see* Stratified Cox model
- Schonfield residuals, 151–152, 522–523
- Score residuals, 346
- Semi-parametric model, 96, 261
- Sensitivity analysis (with competing risks), 408–411
- Shape parameter, 272
- Shared frailty, 306–308, 358
  - recurrent events analysis using, 357–359
- Shared frailty model, 306
- SPSS, 542–555
  - assessing PH assumption with statistical tests, 550–552
  - using Kaplan-Meier log-log survival curves, 546–547
  - estimating survival functions, 544–546
  - running Cox proportional hazards (PH) model, 548
  - running extended Cox model, 552–555
  - running stratified Cox (SC) model, 549–550
- Stanford Heart Transplant Study
  - extended Cox model application to, 235–239
  - transplants versus nontransplants, 244–245
- Stata, 60, 465–508
  - assessing PH assumption using graphical approaches, 473–476
  - using statistical tests, 483–485
  - estimating survival functions, 469–473
  - modeling recurrent events, 502–508
  - obtaining Cox adjusted survival curves with, 485–488
  - running Cox proportional hazards (PH) model, 476–481
  - running extended Cox model, 488–492
  - running frailty models, 499–502
  - running parametric models, 492–499
  - running stratified Cox (SC) model, 481–483
- Step functions, 9
- Strata variable, 347
- Stratification variables, several, 188–193
- Stratified Cox (SC) model, 176–180, 364–367
  - for analyzing recurrent event data, 347–353
  - conditional approaches, 347–351
  - general, 180–181
  - graphical view of, 193–194
  - marginal approach, 347–351
  - using SAS, 519–521
  - using SPSS, 549–550
  - using Stata, 481–483
  - Subdistribution function, 419
- Survival curves
  - adjusted, 104
  - using Cox PH model, 103–107
  - Cox adjusted, *see* Cox adjusted survival curves
  - with recurrent events, 364–367
- Survival functions
  - conditional, 295
  - probability density function and, 262–263
  - SAS estimation, 525–529
  - SPSS estimation, 544–546
  - Stata estimation, 469–473
  - unconditional, 295
- Survival time, 4
- Survival time variable, 14
- Survivor function, 8–9
- Tarone-Ware test statistic, 64
- Time-dependent covariates, assessing PH assumption using, 153–157

- Time-dependent variables, 134
  - definition and examples of, 216–219
  - extended Cox model for, 219–221
- Time-independent variables, PH
  - assumption and, 224–229
- Unconditional survival function, 295
- Unshared frailty, 306
- Veterans Administration Lung Cancer
  - Data
    - Kaplan–Meier survival curves for, 62–63
    - model with no frailty, 296
    - proportional hazards assumption
      - evaluation for, 161–164
      - with several stratification variables, 188–193
      - stratified Cox (SC) model application to, 201–204
- Wald statistic, 89
- Weibull model, 272–277
- Weibull regression
  - accelerated failure-time form, 322, 325
  - log relative-hazard form, 323, 325
- Wilcoxon test, 64