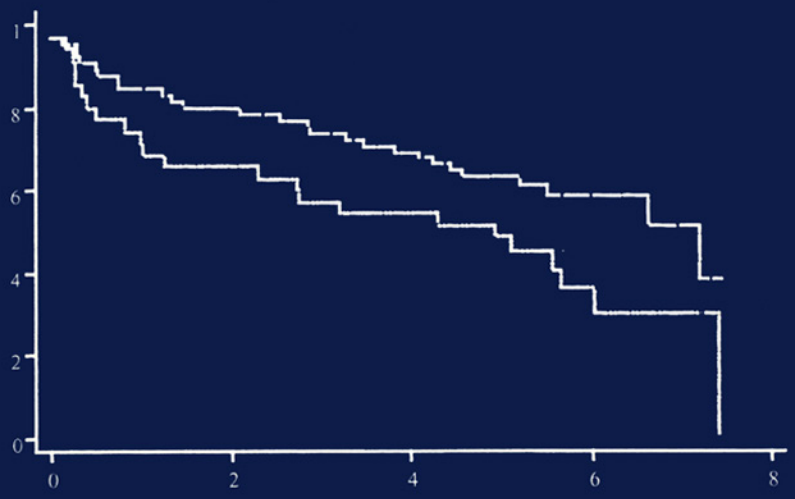
WILEY

Applied Survival Analysis

Regression Modeling of Time-to-Event Data

SECOND EDITION



DAVID W. HOSMER
STANLEY LEMESHOW
SUSANNE MAY

ftp:// SITE AVAILABLE

WILEY SERIES IN PROBABILITY AND STATISTICS

This page intentionally left blank

APPLIED SURVIVAL ANALYSIS

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

Editors: *David J. Balding, Noel A. C. Cressie, Garrett M. Fitzmaurice, Iain M. Johnstone, Geert Molenberghs, David W. Scott, Adrian F. M. Smith, Ruey S. Tsay, Sanford Weisberg*
Editors Emeriti: *Vic Barnett, J. Stuart Hunter, David G. Kendall, Jozef L. Teugels*

A complete list of the titles in this series appears at the end of this volume.

APPLIED SURVIVAL ANALYSIS

Regression Modeling of Time-to-Event Data

Second Edition

DAVID W. HOSMER
University of Massachusetts
School of Public Health and Health Sciences
Department of Public Health
Division of Biostatistics and Epidemiology
Amherst, MA

STANLEY LEMESHOW
The Ohio State University
College of Public Health
Center for Biostatistics
Columbus, OH

SUSANNE MAY
University of California, San Diego
Department of Family & Preventative Medicine
Division of Biostatistics and Bioinformatics
La Jolla, CA

A John Wiley & Sons, Inc., Publication

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data:

Hosmer, David W.
Applied survival analysis : regression modeling of time-to-event data /
David W. Hosmer, Stanley Lemeshow, Susanne May. — 2nd ed.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-471-75499-2 (cloth : alk. paper)
1. Medicine—Research—Statistical methods. 2. Medical sciences—Statistical methods—Computer programs. 3. Regression analysis—Data processing. 4. Prognosis—Statistical methods. 5. Logistic distribution. I. Lemeshow, Stanley. II. May, Susanne. III. Title.
[DNLM: 1. Survival Analysis. 2. Logistic Models. 3. Mathematical Computing. 4. Prognosis. 5. Regression Analysis. WA 950 H827a 2008]
R853.S7H67 2008
610.72'7—dc22 2007035523

Printed in the United States of America.

10 9 8 7 6 5 4

To Trina: Wife, Mother, Athlete, and
Companion in Life's Adventures
D. W. H.

To Elaine: with respect and admiration for her compassion, generosity and appreciation of what is important in our lives and in our world
S. L.

To Bruce: Husband, Friend and Partner
S. M.

This page intentionally left blank

Contents

This page intentionally left blank

Preface to the Second Edition

Since the publication of the first edition nine years ago, analyses using time-to-event methods have increased considerably in all areas of scientific inquiry. We believe that two important reasons for the increase are: (1) the statistical methods for the analysis of time-to-event data are now taught in many intermediate level methods courses and not just advanced courses and (2) the software to perform most of the methods is now available and easy to use in all the major software packages.

The approach taken in the second edition has not changed from the first edition, where the goal was to provide a focused text on regression modeling for the time-to-event data typically encountered in health related studies. As in the first edition, we assume that the reader has had a course in linear regression at the level of Kleinbaum, Kupper, Muller and Nizam (1998) and one in logistic regression at the level of Hosmer and Lemeshow (2000). Emphasis is placed on the modeling of data and the interpretation of the results. Crucial to this is an understanding of the nature of the "incomplete" or "censored" data encountered. Understanding the censoring mechanism is important as it may influence model selection and interpretation. Yet, once understood and accounted for, censoring is often just another technical detail handled by the computer software, allowing emphasis to return to model building, assessment of model fit, and assumptions and interpretation of the results.

In the second edition, we have replaced the HMO–HIV data as the main data set for illustrating methods with a sample of 100 observations from the Worcester Heart Attack Study. We have kept the data from the UMARU Impact Study (UIS), but only use it occasionally. The main modeling data set is a sample of 500 observations from the Worcester Heart Attack Study. Data from the German Breast Cancer Study and the ACTG320 Study are used to demonstrate various modeling and analysis techniques and methods. In short, most of the examples in the text and exercises are new or use new data.

A reading of the Table of Contents for the first four chapters will look as if nothing much has changed. However, the actual text has many changes and additions. For example, the discussions of interactions and the covariate-adjusted survival functions in Chapter Four are greatly expanded. In Chapter Five, we have added variable selection by multivariable fractional polynomials. Changes in Chapter Six follow from a new model based on data from the Worcester Heart

Attack Study studied in Chapter Five. The major change to Chapter Seven is a greatly expanded discussion of time varying covariates with examples. In Chapter Eight we, again, focus on the exponential, Weibull, and log-logistic parametric regression models but have expanded the discussion of each. In Chapter Nine, we have taken advantage of the addition of the capability to fit frailty/random effects models in Stata. Examples are used to compare fitting stratified models to frailty models. The last sections of Chapter Nine contain new material on competing risk models, sample size and power, and using multiple imputation methods to analyze data with missing values.

As we noted, we believe that the increase in the use of statistical methods for time-to-event data is directly related to their incorporation into the major statistical software packages. There are some differences in the capabilities of the various software packages and, when a particular approach is available in a limited number of packages, we note this in the text. Analyses have, for the most part, been performed in Stata Version 9 [Stata Corp. (2005)]. This easy-to-use package combines good graphics and excellent analysis routines, is fast, is compatible across Macintosh, Windows, and UNIX platforms, and interacts well with Microsoft Word 2004 for Mac. Just as we were going to press, Stata Version 10 was released. Among the enhancements in this version is the ability to perform time to event analysis of survey data. Unfortunately we were not able to incorporate that capability into this text. The only other major statistical package employed at various points during the preparation of this text is SAS Version 9.1 [SAS Institute Inc. (2003)].

This text was prepared in camera-ready format using Microsoft Word 2004 for Mac Version 11.3.5 on a PowerBook G4 using Mac OS X Version 10.4.9. Mathematical equations and symbols were built using Math Type Version 5.1 [MathType[5] Mathematical Equation Editor (2004)].

All data may be obtained from the John Wiley & Sons, Inc. ftp site,

ftp://ftp.wiley.com/public/sci_tech_med/survival.

They may also be obtained from the a web site at the University of Massachusetts / Amherst by going the following link and then the section for survival analysis,

http://www.umass.edu/statdata/statdata.

As was the case with the first edition we will have a link at the John Wiley & Sons, Inc. ftp site listed above for errata and corrections.

As in any project with the scope and magnitude of this text, there are many who have contributed directly or indirectly to its content and style and we feel quite fortunate to be able to acknowledge the contributions of others. We thank Rob Goldberg for providing us with a subset of the Worcester Heart Attack Study that we used to create further subsets of 100 and 500 observations. These are used

extensively in the text. We thank Fred Anderson and Gordon FitzGerald for providing a subset of data from the GRACE registry containing time-varying covariates. We thank former faculty colleagues Jane McCusker, Anne Stoddard, and Carol Bigelow for the use and insights into the data from the Project IMPACT Study. We thank the AIDS Clinical Trials Group for making the ACTG 320 data available. We appreciate Ohio State Provost Barbara Snyder's agreeing to allow SL to take a special research assignment (SRA) so that he had the time necessary to work on this book. Not only did Annick Alpérovitch, Carole Dufouil and Christophe Tzourio at INSERM Unit 708 in Paris, France provide an office and an environment conducive for working on this book during the SRA, but they also facilitated obtaining data from the 3C Study Investigators that we were able to use as an exercise in Chapter 7.

We express special thanks to Patrick Royston and Willi Sauerbrei for their helpful suggestions on the text describing fractional polynomials and for comments on numerous other sections of the text. They generously shared with us data from the German Breast Cancer Study that they have analyzed extensively in their publications.

We would like to thank Janice Jones for pointing out the 5731 commas that were missing in the initial draft and for many suggestions that made the text much easier to read. We also thank Charisse Darrell-Fields for inserting Janice's commas into the manuscript. Finally, we thank Tracy McHone for coordinating the printing and organization of the final manuscript.

Over the last nine years we have used the first edition in semester-long course offerings at the University of Massachusetts as well as numerous short courses to audiences around the world. We thank collectively the students in these courses for their comments and insights on how to make things clearer. We hope we have done so in this edition.

DAVID W. HOSMER
STANLEY LEMESHOW
SUSANNE MAY

Stowe, Vermont
Columbus, Ohio
San Diego, California
August, 2007

This page intentionally left blank

CHAPTER 1

Introduction to Regression Modeling of Survival Data

1.1 INTRODUCTION

Regression modeling of the relationship between an outcome variable and one or more independent (predictor) variable(s) is commonly employed in virtually all fields. The popularity of this approach is due to the fact that plausible models may be easily fit, evaluated, and interpreted. Statistically, the specification of a model requires choosing both systematic and error components. The choice of the systematic component involves an assessment of the relationship among the "average" of the outcome variable relative to specific levels of the independent variable(s). This may be guided by an exploratory analysis of the current data and/or past experience. The choice of an error component involves specifying the statistical distribution of what remains to be explained after the model is fit.

In an applied setting, the task of model selection is, to a large extent, based on the goals of the analysis and on the measurement scale of the outcome variable. For example, a clinician may wish to model the relationship among body mass index (BMI, kg/m^2) and caloric intake and gender among teenagers seen in the clinics of a large health maintenance organization (HMO). A good place to start would be to use a model with a linear systematic component and normally distributed errors (i.e., the usual linear regression model). Suppose, instead, that the clinician decides to convert BMI into a 0 – 1 dichotomous variable (taking on the value 1 if BMI > 30) and assess its association with caloric intake and gender. In this case, the logistic regression model would be a good choice. The logistic regression model has a systematic component that is linear in the log-odds and has binomial/Bernoulli distributed errors. While there are many issues involved in the fitting, refinement, evaluation, and interpretation of each of these models, the same basic modeling paradigm would be followed in each scenario.

This basic modeling paradigm is commonly used in texts taking a data-based approach to either linear or logistic regression [e.g., Kleinbaum, Kupper, Muller and Nizam (1998) and Hosmer and Lemeshow (2000)]. In general we follow this same modeling paradigm in this text to motivate our study of regression models where the dependent variable measures the time to the occurrence of an event of

interest. However, as we will see shortly, the fact that *time* to an event is the outcome of interest requires us to think carefully about what actually has been measured. Also the fact that time is a dynamic process provides challenges in formulating a model that are not present in settings where a typical linear or logistic regression model might be applied. In this spirit, we begin with an example.

Example

Throughout this book, we use a number of different data sets to illustrate the methods and provide grist for the exercises at the end of each chapter. Some, but not all, of these are described in Section 1.3. One is a subset of the data from the Worcester Heart Attack Study (WHAS) provided to us by its principal investigator, Dr. Robert J. Goldberg. Briefly, the goal of the WHAS is to study factors and time trends associated with long-term survival following acute myocardial infarction (MI) among residents of the Worcester, Massachusetts, Standard Metropolitan Statistical Area (SMSA). The study began in 1975 and has collected data approximately every other year, with the most recent cohort being subjects who experienced an MI in 2001. The main study has data on over 11,000 subjects, and we will focus our analyses on two samples from the main study. We present one such sample of 100 subjects in Table 1.1. These data are referred to as the WHAS100 data in this text. Suppose our goal for the data in Table 1.1 is to study the effects of gender, age, and body mass index (kg/m^2) at time of hospitalization for the MI on length of survival. Typical regression modeling questions might include: (1) Do women have a more favorable survival experience over time than men? (2) In what way do the age and BMI at admission affect survival over time? (3) Are the effects of age and BMI the same for men and women? Before we can discuss a regression model to address these questions, we need to consider what outcome variable we are going to model. If the outcome is time to an event, then what is the event and how do we define time to it? Suppose we consider the event of interest to be death from any cause following hospitalization for an MI and we define the time to it as the number of days from admission to the hospital until death. The next step in the regression modeling paradigm is to specify the systematic component. Because we have followed subjects over time, it seems logical that the systematic component should be the "mean" of this dynamic process and how it changes as a function of covariates. Prior experience in linear and logistic regression provides little guidance on how to do this. The first few chapters of this book are devoted to providing the necessary background and methods to begin to address this question as well as specification of the error component. The remainder of the text considers application of the methods to different time-to-event scenarios.

Returning to our outcome variable, each subject in Table 1.1 has a date recorded for when the last follow up occurred. Vital status reports whether the subject was dead or alive on that date. For those subjects who died, the reported

date of death and the value presented for follow-up time is the actual value of the outcome of interest: survival time following hospitalization for an MI. For example, subject 5 in Table 1.1 was admitted to the hospital on February 9, 1995, and, 1205 days later, died on May 29, 1998. Subject 10 was admitted to the hospital on July 22, 1995, and was still alive at the time of his last follow up, December 31, 2001. For this subject, all we know is that his survival time exceeds the follow up time of 2719 days. Hence the observation of survival time is incomplete. The statistical term used to describe the process producing this type of incomplete observation is called "censoring" and the observation is referred to as being "censored." In general, incomplete observation of time to an event can occur in several ways and we provide an overview of them in the next section. Methods for handling incompletely observed time-to-event data in regression models is a central theme in this text.

1.2 TYPICAL CENSORING MECHANISMS

We cannot discuss a censored observation until we have carefully defined an uncensored observation. This point may seem rather obvious, but in applied settings confusion, about censoring may not be due to the fact that some observations are incomplete but may instead be the result of an unclear definition of survival time.[1] The observation of survival time has two components that must be unambiguously defined: a beginning point (i.e., when the "clock starts") and an endpoint that is reached when the event of interest occurs (i.e., when the "clock stops"). The point where analysis time, t, is zero is denoted $t = 0$. In the WHAS example, observation began on the day a subject was admitted to the hospital following an MI. In a randomized clinical trial, observation of survival time usually begins on the day a subject is randomized to receive one of the treatment protocols. In an occupational exposure study, $t = 0$ may be the day a subject began work at a particular plant. In some applications, the best $t = 0$ point may not be obvious. For example, in the WHAS study, other beginning points might be the date of discharge from the hospital or the actual moment that the MI occurred. Observation may end at the time when a subject literally "dies" from the disease of interest, or it may end upon the occurrence of some other non-fatal, well-defined, condition such as meeting clinical criteria for remission of a cancer. The survival time is the distance on the time scale between these two points.

---

[1] In this text, we use interchangeably the terms time to event, survival time, and life length to describe the outcome variable. In any example, we choose the one that seems most appropriate but we have a preference for survival time.

Table 1.1 Study ID, Admission Date, Follow Up Date, Length of Hospital Stay, Follow Up Time (Days), Vital Status at Follow Up, Age at Admission (Years), Gender, and Body Mass Index (kg/m^2) (BMI) for 100 Subjects in the Worcester Heart Attack Study

ID	Admission Date	Follow Up Date	Length of Stay	Follow Up Time	Vital Status	Age at Admission	Gender	BMI
1	3/13/95	3/19/95	4	6	Dead	65	Male	31.4
2	1/14/95	1/23/96	5	374	Dead	88	Female	22.7
3	2/17/95	10/4/01	5	2421	Dead	77	Male	27.9
4	4/7/95	7/14/95	9	98	Dead	81	Female	21.5
5	2/9/95	5/29/98	4	1205	Dead	78	Male	30.7
6	1/16/95	9/11/00	7	2065	Dead	82	Female	26.5
7	1/17/95	10/15/97	3	1002	Dead	66	Female	35.7
8	11/15/94	11/24/00	56	2201	Dead	81	Female	28.3
9	8/18/95	2/23/96	5	189	Dead	76	Male	27.1
10	7/22/95	12/31/02	9	2719	Alive	40	Male	21.8
11	10/11/95	12/31/02	6	2638	Alive	73	Female	28.4
12	5/26/95	9/29/96	11	492	Dead	83	Male	24.7
13	5/21/95	3/18/96	6	302	Dead	64	Female	27.5
14	12/14/95	12/31/02	10	2574	Alive	58	Male	29.8
15	11/8/95	12/31/02	7	2610	Alive	43	Male	23.0
16	10/8/95	12/31/02	5	2641	Alive	39	Male	30.1
17	10/17/95	5/12/00	6	1669	Dead	66	Male	32.0
18	10/30/95	1/5/03	9	2624	Dead	61	Male	30.7
19	12/10/95	12/31/02	6	2578	Alive	49	Male	25.7
20	11/23/95	12/31/02	5	2595	Alive	53	Female	30.1
21	10/5/95	2/5/96	6	123	Dead	85	Male	18.4
22	11/5/95	12/31/02	8	2613	Alive	69	Female	37.6
23	9/9/95	10/22/97	4	774	Dead	54	Male	29.0
24	9/9/95	3/13/01	14	2012	Dead	82	Male	19.9
25	12/15/95	12/31/02	4	2573	Alive	67	Female	28.3
26	12/3/95	1/19/01	11	1874	Dead	89	Female	23.4
27	10/18/95	12/31/02	2	2631	Alive	68	Male	26.4
28	3/16/95	6/4/00	7	1907	Dead	78	Male	28.2
29	10/25/95	4/15/97	5	538	Dead	56	Male	24.1
30	10/6/95	1/18/96	4	104	Dead	85	Female	36.7
31	9/3/95	9/9/95	4	6	Dead	72	Male	28.0
32	6/30/95	5/1/99	5	1401	Dead	50	Male	20.4
33	7/22/95	12/22/02	8	2710	Dead	81	Female	28.6
34	9/17/95	1/5/98	4	841	Dead	85	Female	20.2
35	3/21/97	8/16/97	6	148	Dead	84	Female	23.6
36	2/23/97	12/31/02	12	2137	Alive	75	Male	23.7
37	1/1/97	12/31/02	16	2190	Alive	61	Male	23.4
38	1/18/97	12/31/02	5	2173	Alive	48	Male	33.5
39	1/19/97	4/25/98	8	461	Dead	83	Female	19.6
40	3/18/97	12/31/02	10	2114	Alive	82	Male	25.8
41	2/3/97	12/31/02	4	2157	Alive	62	Male	30.9
42	5/17/97	12/31/02	5	2054	Alive	39	Male	24.2
43	3/8/97	12/31/02	5	2124	Alive	45	Male	31.7
44	2/23/97	12/31/02	4	2137	Alive	65	Male	26.2
45	6/14/97	1/5/03	18	2031	Dead	76	Female	32.4
46	7/7/97	12/31/02	9	2003	Alive	77	Female	24.6
47	4/27/97	12/31/02	9	2074	Alive	68	Male	21.3
48	5/15/97	2/13/98	7	274	Dead	73	Male	26.5
49	7/26/97	12/31/02	4	1984	Alive	64	Male	28.0
50	7/17/97	12/31/02	6	1993	Alive	80	Male	36.0
51	9/9/97	12/31/02	7	1939	Alive	84	Female	22.3

Table 1.1 Continued

ID	Admission Date	Follow Up Date	Length of Stay	Follow Up Time	Vital Status	Age at Admission	Gender	BMI
52	6/19/97	9/3/00	4	1172	Dead	43	Female	25.3
53	8/20/97	11/17/97	3	89	Dead	87	Female	18.8
54	8/28/97	1/3/98	7	128	Dead	70	Female	18.6
55	9/9/97	12/31/02	17	1939	Alive	80	Male	25.5
56	9/1/97	9/15/97	11	14	Dead	64	Female	24.4
57	9/3/97	6/10/00	5	1011	Dead	59	Female	29.9
58	9/24/97	10/30/01	6	1497	Dead	92	Male	24.4
59	9/19/97	12/31/02	3	1929	Alive	51	Male	34.8
60	4/17/97	12/31/02	1	2084	Alive	41	Male	27.3
61	10/21/97	2/5/98	6	107	Dead	90	Male	24.8
62	10/2/97	12/27/98	4	451	Dead	83	Male	21.8
63	1/8/97	12/31/02	3	2183	Alive	61	Male	27.4
64	11/11/97	12/31/02	7	1876	Alive	64	Male	26.2
65	11/7/97	5/31/00	3	936	Dead	82	Male	26.9
66	4/20/97	4/18/98	5	363	Dead	91	Female	27.6
67	6/18/97	5/1/00	5	1048	Dead	48	Male	31.6
68	10/29/97	12/31/02	12	1889	Alive	63	Male	23.3
69	4/29/97	12/31/02	5	2072	Alive	81	Male	28.4
70	11/8/97	12/31/02	7	1879	Alive	52	Male	32.6
71	11/17/97	12/31/02	4	1870	Alive	65	Male	32.0
72	11/28/97	12/31/02	5	1859	Alive	74	Male	25.0
73	5/19/97	12/31/02	5	2052	Alive	62	Male	30.2
74	12/11/97	12/31/02	4	1846	Alive	60	Female	29.3
75	5/10/97	12/31/02	7	2061	Alive	71	Male	32.3
76	10/6/97	12/31/02	3	1912	Alive	73	Male	31.5
77	12/21/97	12/31/02	5	1836	Alive	43	Male	28.6
78	11/22/97	3/16/98	7	114	Dead	80	Male	33.4
79	10/31/97	2/4/02	7	1557	Dead	72	Male	21.8
80	6/28/97	12/27/00	5	1278	Dead	57	Male	23.6
81	12/21/97	12/31/02	3	1836	Alive	80	Female	28.4
82	10/2/97	12/31/02	6	1916	Alive	76	Male	28.0
83	9/14/97	12/31/02	3	1934	Alive	53	Male	24.2
84	9/25/97	12/31/02	10	1923	Alive	44	Male	32.6
85	12/2/97	1/15/98	3	44	Dead	71	Male	23.1
86	9/26/97	12/31/02	6	1922	Alive	64	Male	31.8
87	10/24/97	7/25/98	5	274	Dead	86	Male	21.1
88	11/27/97	12/31/02	7	1860	Alive	72	Female	25.2
89	4/12/97	3/23/02	4	1806	Dead	73	Female	22.9
90	2/15/97	12/31/02	6	2145	Alive	85	Female	26.1
91	10/22/97	4/22/98	5	182	Dead	60	Male	23.2
92	6/27/97	12/31/02	4	2013	Alive	63	Male	35.5
93	1/17/97	12/31/02	5	2174	Alive	80	Female	20.6
94	12/12/97	5/24/02	4	1624	Dead	74	Male	30.1
95	11/4/97	5/10/98	10	187	Dead	79	Female	16.8
96	11/4/97	12/31/02	4	1883	Alive	48	Female	32.1
97	12/24/97	4/19/02	3	1577	Dead	32	Female	39.9
98	11/26/97	1/27/98	8	62	Dead	86	Female	14.9
99	8/10/97	12/31/02	16	1969	Alive	56	Male	29.1
100	3/26/97	2/13/00	7	1054	Dead	74	Male	32.9

In practice, a value of time is obtained by calculating the number of days (or months, or years, etc.) between two calendar dates. Table 1.1 shows the admission date and the follow up date for the subjects in this sample from the WHAS study. Most statistical software packages have functions that allow the user to manipulate calendar dates in a manner similar to other numeric variables. They do this by creating a numeric value for each calendar date, which is defined as the number of days from some predetermined reference date. For example, the reference date used by most, if not all, packages is January 1, 1960. Subject 5 entered the study on February 9, 1995, which is 12,823 days after the reference date, and died May 29, 1998, which is 14,028 days after the reference date. The interval between these two dates is $14,028 - 12,823 = 1,205$ days. The number of days can be converted into the number of months by dividing by $30.4375 = (365.25/12)$. Thus, the survival time in months for subject 5 is $39.589 = (1,205/30.4375)$. It is common, when reporting results in tabular form, to round months to the nearest whole number, e.g., 40 months. The level of precision used in reporting and analyzing survival time should depend on the particular application.

Two mechanisms can lead to incomplete observation of time: censoring and truncation. A censored observation is one whose value is incomplete due to factors that are random for each subject. A truncated observation is incomplete due to a selection process inherent in the study design. The most commonly encountered form of a censored observation is one where observation begins at the defined time $t = 0$ and terminates before the outcome of interest is observed. Because the incomplete nature of the observation occurs in the right tail of the time axis, such observations are said to be *right censored.* For example, in the WHAS study, a subject could move out of town or still be alive at the last follow up. In a study where right censoring is the only type of censoring possible, observation on subjects may begin at the same time or at varying times. For example, in a test of computer life length, we may begin with all computers started at exactly the same time. In a randomized clinical trial or in an observational study, such as the WHAS study, patients may enter the study over several years. As we see in Table 1.1, subject 2 entered the study on January 14, 1995, while subject 50 entered on July 17, 1997. In this type of study, regardless of calendar time, each subject's time of enrollment is assumed to define the $t = 0$ point.

For obvious practical reasons, all studies have a point when observation ends on all subjects; therefore subjects entering at different times will have variable lengths of maximum follow-up time. In the WHAS study, the last follow up date is December 31, 2002. Subject 13 entered the study on May 21, 1995. Thus the longest this subject could have been followed is 7 years, 7 months, and 10 days. However, this subject was not followed for the maximum length of time because the subject died on March 18, 1996, yielding a survival time of 302 days. Incomplete observation of a survival time due to the end of the study or follow-up is considered a right censored observation because the process by which subjects entered the study is random at the subject level.

A typical pattern of entry into a follow-up study is shown in Figure 1.1. This is a hypothetical 2-year study in which patients are enrolled during the first year. We see that subject 1 entered the study on January 1, 1990, and died on March 1, 1991. Subject 2 entered the study on February 1, 1990, and was lost to follow-up on February 1, 1991. Subject 3 entered the study on June 1, 1990, and was still alive on December 31, 1991, the end of the study. Subject 4 entered the study on September 1, 1990, and died on April 1, 1991. Subjects 2 and 3 have survival times that are right-censored. These data are plotted on the analysis time scale, in months, in Figure 1.2. Note that each subject's time is plotted as if he or she were enrolled at exactly the same calendar time and were followed until his or her respective end point. The two figures illustrate the difference between collecting data in calendar time and then converting it to analysis time.

In some studies, there may be a clear definition of the beginning time point; but subjects may not come under actual observation until after this point has passed. For example, in modeling age at menarche, suppose we define the zero value of time as 8 years. Suppose a subject enters the study at age 10, still not having experienced menarche. We know that this subject could have experienced menarche after age 8 but, due to the study design, was not enrolled in the study until age 10. This subject would not enter the analysis until time 10. This type of incomplete observation of time is called *left truncation* or *delayed entry*. Another example would be to study survival time in the WHAS among those discharged from the hospital alive. Here subjects stay in the hospital for varying lengths of time but we do not begin to study them until they "leave the front door."

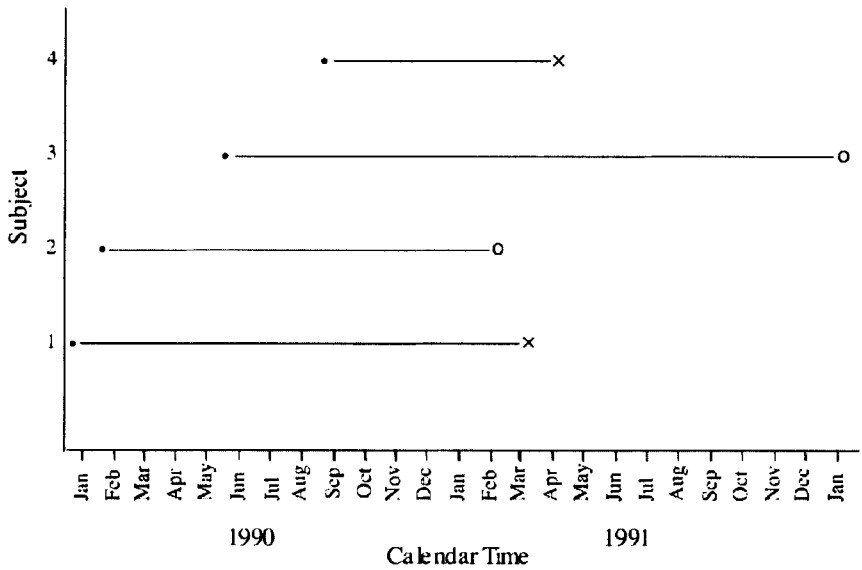


Figure 1.1 Line plot in calendar time for four subjects in a hypothetical follow-up study.

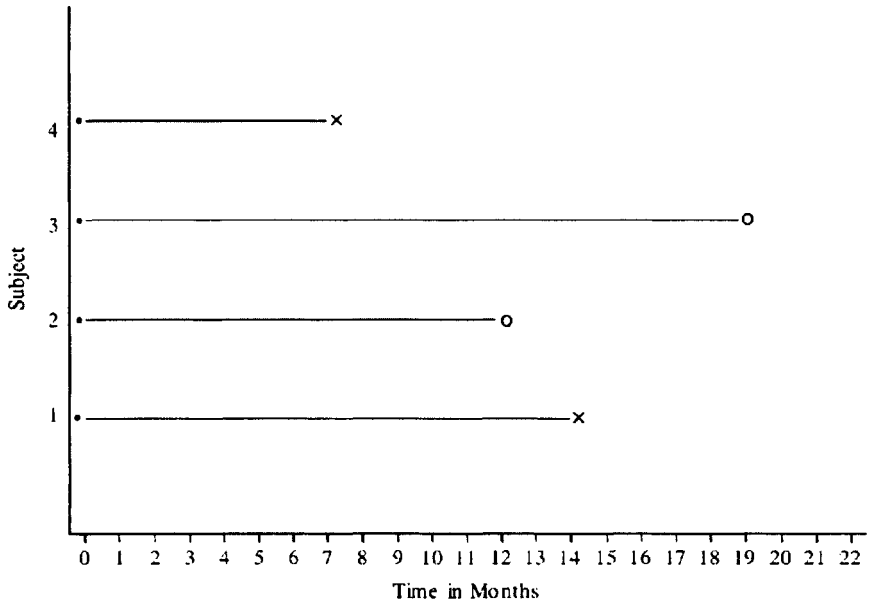


Figure 1.2 Line plot in the time scale for four subjects in a hypothetical follow-up study.

Another censoring mechanism that sometimes occurs in practice is *left censoring*. An observation is left censored if the event of interest has already occurred when observation begins. For example, in the study of age at menarche, if a subject enrolls in the study at age 10 and has already experienced menarche, this subject's time is left censored. In the WHAS study, if we begin observation at seven days post admission then subjects who die in the first week are left censored.

A less common form of incomplete observation occurs when the entire study population has experienced the event of interest before the study begins (i.e., subjects have been selected because they have experienced the event of interest). This is sometimes referred to as *length biased sampling* and it must be accounted for in the analysis. An example would be a study of risk factors for time to diagnosis of colorectal cancer among subjects in a cancer registry with this diagnosis. In this study, being in the cancer registry represents a selection process assuring that time to the event is known for each subject. This type of incomplete observation of time is called *right truncation*. Because this type of data occurs relatively infrequently in practice, we do not consider it further in this text. Readers interested in learning more about the analysis of right truncated data are referred to Klein and Moeschberger (2003).

In some practical settings, one may not be able to observe time continuously. For example, in a study of educational interventions to prevent IV drug use, the protocol may specify that subjects, after completion of their "treatment," will be

contacted every 3 months for a period of 2 years. In this study, the outcome might be time of first relapse to IV drug use. Because subjects are contacted every 3 months, time is only accurately measured to multiples of 3 months. Given the discrete nature of the observed time variable, it would be inappropriate to use a statistical model that assumed the observed values of time were continuous. Thus, if a subject reports at the 12-month follow-up that she has returned to drug use, we know only that her time is between 9 and 12 months. Data of this type are said to be *interval censored.*

We consider methods for the analysis of right censored data throughout this text because this is the most commonly occurring type of censoring. The next most common forms of incomplete observation are left truncation and interval censoring. Modifications of the methods to handle these mechanisms are discussed in Chapter 7.

Prior to considering any regression modeling, the first step in the analysis of survival time, or for that matter any set of data, should be a thorough univariate analysis. In the absence of censoring and truncation, this analysis would use the techniques covered in an introductory course on statistical methods. The exact combination of statistics used would depend on the application. It might include graphical descriptors such as histograms, box and whisker plots, cumulative percentage distribution polygons or other methods. It would also include a table of descriptive statistics containing point estimates and confidence intervals for the mean, median, standard deviation, and various percentiles of the distribution of survival time. The presence of censored data in the sample complicates the calculations but not the fundamental goal of univariate analysis. In the next chapter we present methods for univariate analysis of right censored survival time.

1.3 EXAMPLE DATA SETS

In addition to the data from the WHAS study presented in Table 1.1, data are available from a larger sample from the entire WHAS study. These data are new to this revision and not the same data used from the WHAS in the first edition. Three additional studies are used throughout the text to illustrate methods and provide data for exercises presented at the end of each chapter. All data may be obtained from the John Wiley & Sons web site,

ftp://ftp.wiley.com/public/sci_tech_med/survival.

They may also be obtained from the web site for statistical services at the University of Massachusetts at Amherst by going to the datasets link and then the section on survival data,

http://www.umass.edu/statdata/statdata.

As noted previously, the data from the WHAS study have been provided to us by Dr. Robert J. Goldberg of the Department of Cardiology at the University of Massachusetts Medical School. The main goal of this study is to describe factors associated with trends over time in the incidence and survival rates following hospital admission for acute myocardial infarction (MI). Data have been collected during 13 one-year periods beginning in 1975 and extending through 2001 on all MI patients admitted to hospitals in the Worcester, Massachusetts Standard Metropolitan Statistical Area. The main data set has information on more than 11,000 admissions. Several variables that provide us the opportunity to demonstrate and discuss various aspects of modeling time-to-event data were added to the data collection in the later three cohorts. The data in this text were obtained by taking an approximately 23 percent random sample from the cohort years 1997, 1999, and 2001, yielding 500 subjects. This data set is called the WHAS500 study in this text. In addition, only a small subset of the variables from the main study is included in our data set. Dr. Goldberg and his colleagues have published more than 30 papers reporting the results of various analyses from the WHAS. For an example of a recent publication from the study see Goldberg et al. (2005) as well as Goldberg et. al. (1986, 1988, 1989, 1991, 1993) and Chiriboga et al. (1994).

Table 1.2 describes the subset of variables used, with their codes and values. One should not infer that results reported and/or obtained in exercises in this text are comparable in any way to analyses of the complete data from the WHAS.

Our colleagues, Drs. Jane McCusker, Carol Bigelow and Anne Stoddard, provided a data set used extensively in the first edition of this text. It is a subset of data from the University of Massachusetts AIDS Research Unit (UMARU) IMPACT Study (UIS). This was a 5-year (1989–1994) collaborative research project (Benjamin F. Lewis, P.I., National Institute on Drug Abuse Grant #R18-DA06151) comprised of two concurrent randomized trials of residential treatment for drug abuse. The purpose of the study was to compare treatment programs of different planned durations designed to reduce drug abuse and to prevent high-risk HIV behavior. The UIS sought to determine whether alternative residential treatment approaches are variable in effectiveness and whether efficacy depends on planned program duration. These data were used to illustrate model building in the first edition of this book and are being retained for use in the second edition primarily for end of chapter exercises. The small subset of variables from the main study we use in this text is described in Table 1.3.

Because the analyses we report in this text are based on this small subset of variables, the results reported here should not be considered as being in any way comparable to results from the main study. In addition, we have taken the liberty of simplifying the study design by representing the planned duration as short versus long. Thus, short versus long represents 3 months versus 6 months planned duration at site A, and 6 months versus 12 months planned duration at site B. The time variable considered in this text is defined as the number of days from admission to one of the two sites to self-reported return to drug use. The censoring variable is coded 1 for return to drug or lost to follow-up and 0 otherwise. The

Table 1.2 Description of the Variables Obtained from the Worcester Heart Attack Study (WHAS), 500 Subjects

Variable	Description	Codes / Values
id	Identification Code	1 – 500
age	Age at Hospital Admission	Years
gender	Gender	0 = Male, 1 = Female
hr	Initial Heart Rate	Beats per minute
sysbp	Initial Systolic Blood Pressure	mmHg
diasbp	Initial Diastolic Blood Pressure	mmHg
bmi	Body Mass Index	kg/m^2
cvd	History of Cardiovascular Disease	0 = No, 1 = Yes
afb	Atrial Fibrillation	0 = No, 1 = Yes
sho	Cardiogenic Shock	0 = No, 1 = Yes
chf	Congestive Heart Complications	0 = No, 1 = Yes
av3	Complete Heart Block	0 = No, 1 = Yes
miord	MI Order	0 = First, 1 = Recurrent
mitype	MI Type	0 = non Q-wave, 1 = Q-wave
year	Cohort Year	1 = 1997, 2 = 1999, 3 = 2001
admitdate	Hospital Admission Date	mm/dd/yy
disdate	Hospital Discharge Date	mm/dd/yy
fdate	Date of last Follow Up	mm/dd/yy
los	Length of Hospital Stay	Days between Hospital Discharge and Hospital Admission
dstat	Discharge Status from Hospital	0 = Alive, 1 = Dead
lenfol	Total Length of Follow-up	Days between Date of Last Follow-up and Hospital Admission Date
fstat	Vital Status at Last Follow-up	0 = Alive 1 = Dead

study team felt that a subject who was lost to follow-up was likely to have returned to drug use. The original data have been modified to preserve subject confidentiality.

Cancer clinical trials are a rich source for examples of applications of methods for the analysis of time to event. Willi Sauerbrei and Patrick Royston have graciously provided us with data obtained from the German Breast Cancer Study Group, which they used to illustrate methods for building prognostic models (Sauerbrei and Royston, 1999). In the main study, a total of 720 patients with primary node positive breast cancer were recruited between July 1984, and December 1989, (see Schmoor, Olschweski and Schumacher M. 1996 and Schumacher et al. (1994)). Data used in this text are for 686 subjects with complete data on the covariates in Table 1.4.

Table 1.3 Description of Variables in the UMARU IMPACT Study (UIS), 628 Subjects

Variable	Description	Codes/Values
id	Identification Code	1–628
age	Age at Enrollment	Years
beck	Beck Depression Score at Admission	0.000–54.000
hercoc	Heroin/Cocaine Use During 3 Months Prior to Admission	1 = Heroin & Cocaine 2 = Heroin Only 3 = Cocaine Only 4 = Neither Heroin nor Cocaine
ivhx	IV Drug Use History at Admission	1 = Never 2 = Previous 3 = Recent
ndrugtx	Number of Prior Drug Treatments	0 – 40
race	Subject's Race	0 = White 1 = Other
treat	Treatment Randomization Assignment	0 = Short 1 = Long
site	Treatment Site	0 = A 1 = B
lot	Length of Treatment (Measured from Admission)	Days
time	Time to Return to Drug Use (Measured from Admission)	Days
censor	Returned to Drug Use	1 = Returned to Drug Use 0 = Otherwise

Another clinical trial data set used in this text was provided by the AIDS Clinical Trials Group (ACTG 320). The data come from a double-blind, placebo-controlled trial that compared the three-drug regimen of indinavir (IDV), open label zidovudine (ZDV) or stavudine (d4T), and lamivudine (3TC) with the two-drug regimen of zidovudine or stavudine and lamivudine in HIV-infected patients (Hammer et al., 1997). Patients were eligible for the trial if they had no more than 200 CD4 cells per cubic millimeter and at least three months of prior zidovudine therapy. Randomization was stratified by CD4 cell count at the time of screening. The primary outc ome measure was time to AIDS defining event or death. Because efficacy results met a pre-specified level of significance at an interim analysis, the trial was stopped early. Variables and codes for these data are provided in Table 1.5.

Table 1.4 Description of Variables in the German Breast Cancer Study (GBCS), 686 Subjects

Variable	Description	Codes/Values/ Range
id	Study ID	1 – 686
diagdate	Date of Diagnosis	ddMonthyyyy
recdate	Date of Recurrence Free Survival	ddMonthyyyy
deathdate	Date of Death	ddMonthyyyy
age	Age at Diagnosis	Years
menopause	Menopausal Status	0 = No, 1 = Yes
hormone	Hormone Therapy	0 = No, 1 = Yes
size	Tumor Size	mm
grade	Tumor Grade	1 – 3
nodes	Number of Nodes involved	1 – 51
prog_recp	Number of Progesterone Receptors	1 – 2380
estrg_recp	Number of Estrogen Receptors	1 – 1144
rectime	Time to Recurrence	Days
censrec	Recurrence Censoring	0 = Censored 1 = Recurrence
survtime	Time to Death	Days
censdead	Death Censoring	0 = Censored 1 = Death

EXERCISES

One of the most effective graphical tools that can be employed in regression modeling is a scatter plot of the outcome versus continuous covariates. For example, in linear regression, such a plot can provide guidance as to the plausibility of a linear relationship between the mean of the outcome and the covariate as well as the distribution about the line (i.e., the error component).

1. Using the data from the Worcester Heart Attack Study in Table 1.1, obtain a scatter plot of follow up time versus age. If possible, use the value of the vital status variable as the plotting symbol.
   (a) In what ways is the visual appearance of this plot different from a scatter plot in a typical linear regression setting?
   (b) By eye, draw on the scatter plot from problem 1(a) what you feel is the best regression function for a survival time regression model.

Table 1.5 Description of Variables in the AIDS Clinical Trials Group Study (ACTG 320), 1151 Subjects

Variable	Description	Codes/Values
id	Identification Code	1–1156
time	Time to AIDS diagnosis or death	Days
censor	Event indicator for AIDS defining diagnosis or death	1 = AIDS defining diagnosis or death 0 = Otherwise
time_d	Time to death	Days
censor_d	Event indicator for death (only)	1 = Death 0 = Otherwise
tx	Treatment indicator	1 = Treatment includes IDV 0 = Control group (treatment regimen without IDV)
txgrp	Treatment group indicator	1 = ZDV + 3TC 2 = ZDV + 3TC + IDV 3 = d4T + 3TC 4 = d4T + 3TC + IDV
strat2	CD4 stratum at screening	0 = CD4 ≤ 50 1 = CD4 > 50
sex	Sex	1 = Male 2 = Female
raceth	Race/Ethnicity	1 = White Non-Hispanic 2 = Black Non-Hispanic 3 = Hispanic (regardless of race) 4 = Asian, Pacific Islander 5 = American Indian, Alaskan Native 6 = Other/unknown
ivdrug	IV drug use history	1 = Never 2 = Currently 3 = Previously
hemophil	Hemophiliac	1 = Yes 0 = No
karnof	Karnofsky Performance Scale	100 = Normal; no complaint; no evidence of disease 90 = Normal activity possible; minor signs/symptoms of disease 80 = Normal activity with effort; some signs/symptoms of disease 70 = Cares for self; normal activity/ active work not possible
cd4	Baseline CD4 count (derived from multiple measurements)	Cells/milliliter
priorzdv	Months of prior ZDV use	Months
age	Age at Enrollment	Years