Applied Survival Analysis

Regression Modeling of Time to Event Data

DAVID W. HOSMER, Jr.

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

STANLEY LEMESHOW

Department of Biostatistics and Epidemiology
University of Massachusetts
Amherst, Massachusetts

A Wiley-Interscience Publication
JOHN WILEY & SONS, INC.
New York • Chichester • Weinheim • Brisbane • Singapore • Toronto

This text is printed on acid-free paper. ♾

Library of Congress Cataloging in Publication Data:

Hosmer, David W.
Applied survival analysis : regression modeling of time to event data / David W. Hosmer, Jr., Stanley Lemeshow
p. cm. — (Wiley series in probability and statistics)
Includes bibliographical references and indexes.
ISBN 0-471-15410-5 (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. Title. III. Series.
R853.S7H67 1998
610'.7'27—dc21 98-27511

Printed in the United States of America

10 9 8 7 6 5 4 3

Contents

Preface

The study of events involving an element of time has a long and important history in statistical research and practice. Examples chronicling the mortality experience of human populations date from the 1700s [see Hald (1990)]. Recent advances in methods and statistical software have placed a seemingly bewildering array of techniques at the fingertips of the data analyst. It is difficult to find either a subject matter or a statistical journal that does not have at least one paper devoted to use or development of these methods.

In spite of the importance and widespread use of these methods there is a paucity of material providing an introduction to the analysis of time to event data. A course dealing with this subject tends to be more advanced and often is the third or fourth methods course taken by a student. As such, the student typically has a strong background in linear regression methods and usually some experience with logistic regression. Yet most texts fail to capitalize on this statistical and experiential background. The approach is either highly mathematical or does not emphasize regression model building. The goal of this book is to provide a focused text on regression modeling for the time to event data typically encountered in health related studies. For this text we assume 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 (1989). 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.

The increase in the use of statistical methods for time to event data is directly re-

lated to their incorporation into major and minor (specialized) statistical software packages. To a large extent there are no major differences in the capabilities of the various software packages. When a particular approach is available in a limited number of packages it will be noted in this text. In general, analyses have been performed in STATA [Stata Corp. (1997)]. This easy to use package combines reasonably good graphics and excellent analysis routines, is fast, is compatible across Macintosh, Windows and UNIX platforms and interacts well with Microsoft Word 6.0. Other major statistical packages employed at various points during the preparation of this text include BMDP [BMDP Statistical Software (1992)], SAS [SAS Institute Inc. (1989)] and S-PLUS [S-Plus Statistical Sciences (1993)].

This text was prepared in camera ready format using Microsoft Word 6.0.1 on a Power Macintosh platform. Mathematical equations and symbols were built using Math Type 3.5 [Math Type: Mathematical Equation Editor (1997)]. When necessary, graphics were enhanced and modified using MacDraw.

Early on in the preparation of the text we made a decision that data sets used in the text would be made available to readers via the World Wide Web rather than on a diskette distributed with the text. The ftp site at John Wiley & Sons, Inc. for the data in this text is ftp://ftp.wiley.com/public/sci_tech_med/survival. In addition, the data may also be found, by permission of John Wiley & Sons Inc., in the archive of statistical data sets maintained at the University of Massachusetts at Internet address http://www-unix.oit.umass.edu/~statdata in the survival analysis section. Another advantage to having a text web site is that it provides a convenient medium for conveying to readers text changes after publication. In particular, as errata become known to us they will be added to an errata section of the text's web site at John Wiley & Sons, Inc. Another use that we envision for the web is the addition, over time, of new data sets to the statistical data set archive at the University of Massachusetts.

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. One of us (DWH) would like to express special thanks to a friend and colleague, Petter Laake, Head of the Section of Medical Statistics at the University of Oslo, for arranging for a Senior Scientist Visiting Fellowship from the Research Council of Norway that supported a sabbatical leave visit to the Section in Oslo during the winter of 1997. We would like to thank Odd Aalen for reading and commenting on several sections of the text. His advice was most helpful in preparing the material on frailty and additive models in Chapter 9. While in Oslo, and afterwards, Ørnulf Borgan was especially helpful in clarifying some of the details of the counting process approach and graciously shared some, at that time, unpublished research of his and his student, J. K. Grønnesby. Thoughtful and careful commentary by outside reviewers, in particular Daniel Commenges, of the UFR de Santé Publique at the University of Bordeaux II, improved the content and quality of the text.

We are grateful to colleagues in our Department who have contributed to the development of this book. These include Drs. Jane McCusker, Anne Stoddard and Carol Bigelow for the use and insights into the data from the Project IMPACT Study

and Janelle Klar and Elizabeth K. Donohoe for their extraordinarily careful reading of the manuscript and editorial suggestions.

DAVID W. HOSMER, JR.
STANLEY LEMESHOW

Amherst, Massachusetts
August, 1998

CHAPTER 1

Introduction to Regression Modeling of Survival Data

1.1 INTRODUCTION

Regression modeling of the relationship between an outcome variable and independent predictor variable(s) is commonly employed in virtually all fields. The popularity of this approach is due to the fact that biologically 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 between an "average" of the outcome variable and 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 (i.e., the residuals).

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 between a measure of nutritional status (e.g., caloric intake) and various demographic and physical characteristics of the child such as gender, socio-economic status, height and weight, among children between the ages of two and six 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, the usual linear regression model. Suppose instead that the clinician decides to convert the nutrition data into a dichotomous variable that indicated whether the child's diet met specified intake criteria (1 =

yes and 0 = no). If we assume the goal of this analysis is to estimate the "effect" of the various factors via an odds-ratio, then 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. There are many issues involved in the fitting, refinement, evaluation and interpretation of each of these models. However, the clinician would follow the same basic modeling paradigm 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 (1989)]. We use it in this text to motivate our discussion of the similarities and differences between the linear (and the logistic) regression model and regression models appropriate for survival data. In this spirit we begin with an example.

Example

A large HMO wishes to evaluate the survival time of its HIV+ members using a follow-up study. Subjects were enrolled in the study from January 1, 1989 to December 31, 1991. The study ended on December 31, 1995. After a confirmed diagnosis of HIV, members were followed until death due to AIDS or AIDS-related complications, until the end of the study or until the subject was lost to follow-up. We assume that there were no deaths due to other causes (e.g., auto accident). The primary outcome variable of interest is survival time after a confirmed diagnosis of HIV. Since subjects entered the study at different times over a 3-year period, the maximum possible follow-up time is different for each study participant. Possible predictors of survival time were collected at enrollment into the study. Data listed in Table 1.1 for 100 subjects are: TIME: the follow-up time is the number of months between the entry date (ENT DATE) and the end date (END DATE), AGE: the age of the subject at the start of follow-up (in years), DRUG: history of prior IV drug use (1 = Yes, 0 = No), and CENSOR: vital status at the end of the study (1 = Death due to AIDS, 0 = Lost to follow-up or alive).[1] Of many possible covariates, age and prior drug use

[1] Although it may seem odd that if the subject's time to failure is *not* censored the subject receives a "1" for this variable, this is the convention followed in the literature and will be followed throughout this text as well.

were chosen for their potential clinical relevance as well as for statistical purposes to illustrate techniques for continuous and nominal scale predictor variables.

One of the most important differences between the outcome variables modeled via linear and logistic regression analyses and the time variable in the current example is the fact that we may only observe the survival time partially. The variable TIME listed in Table 1.1 actually records two different things. For those subjects who died, it is the outcome variable of interest, the actual survival time. However, for subjects who were alive at the end of the study, or for subjects who were lost, TIME indicates the length of follow-up (which is a partial or incomplete observation of survival time). These incomplete observations are referred to as being *censored.* For example, subject 1 died from AIDS 5 months after being seen in the HMO clinic (CENSOR = 1) while subject 2 was not known to have died from AIDS at the conclusion of the study and had been followed for 6 months (CENSOR = 0). It is possible for a subject to have entered the study 6 months before the end or he/she could have entered the study much earlier, eventually becoming lost to follow-up as a result of moving, failing to return to the clinic or some other reason. For the time being we do not differentiate between these possibilities and consider only the two states: dead (as a result of AIDS) and not known to be dead.

The main goal for a statistical analysis of these data is to fit a model that will yield biologically plausible and interpretable estimates of the effect of age and drug use on survival time, for HIV+ patients. Before beginning any statistical modeling, we should perform a thorough univariate analysis of the data to obtain a clear sense of the distributional characteristics of our outcome variable as well as all possible predictor variables. The fact that some of our observations of the outcome variable, survival time, are incomplete is a problem for conventional univariate statistics such as the mean, standard deviation, median, etc. If we ignore the censoring and treat the censored observations as if they were measurements of survival time, then the resulting sample statistics are not estimators of the respective parameters of the survival time distribution. They are estimators of parameters of a combination of the survival time distribution and a second distribution that depends on survival time as well as statistical assumptions about the censoring mechanism. For example, the average of TIME for subjects 1 and 2 in Table 1.1 is 5.5 months. The number 5.5 months is not an estimate of the mean length of survival. We can say the mean survival is estimated to be *at least* 5.5 months. But how can we appropriately use the fact that the survival time

Table 1.1 Study Entry and Ending Dates, Survival Time (Time), Age, History of IV Drug Use (Drug) and Vital Status (Censor) at Conclusion of Study

ID	Ent Date	End Date	Time	Age	Drug	Censor	ID	Ent Date	End Date	Time	Age	Drug	Censor
1	15May90	14Oct90	5	46	0	1	51	11Nov89	10Feb91	15	33	0	1
2	19Sep89	20Mar90	6	35	1	0	52	1Oct90	31Oct90	1	31	0	1
3	21Apr91	20Dec91	8	30	1	1	53	20Mar90	18Jan91	10	33	0	1
4	3Jan91	4Apr91	3	30	1	1	54	30Jul90	29Aug90	1	50	1	1
5	18Sep89	19Jul91	22	36	0	1	55	17Jul89	14Feb90	7	36	1	1
6	18Mar91	17Apr91	1	32	1	0	56	10Nov90	9Feb91	3	30	1	1
7	11Nov89	11Jun90	7	36	1	1	57	5Mar89	4Jun89	3	42	1	1
8	25Nov89	25Aug90	9	31	1	1	58	2Mar91	1May91	2	32	1	1
9	11Feb91	13May91	3	48	0	1	59	11Sep89	11May92	32	34	0	1
10	11Aug89	11Aug90	12	47	0	1	60	12Sep89	12Dec89	3	38	1	1
11	11Apr90	10Jun90	2	28	1	0	61	8Apr90	6Feb91	10	33	0	0
12	11May91	10May92	12	34	0	1	62	20Apr89	20Mar90	11	39	1	1
13	17Jan89	16Feb89	1	44	1	1	63	31Jan91	2May91	3	39	1	1
14	16Feb91	17May92	15	32	1	1	64	15Sep89	15Apr90	7	33	1	1
15	9Apr91	6Feb94	34	36	0	1	65	7Dec91	7May92	5	34	1	1
16	9Mar91	8Apr91	1	36	0	1	66	4Mar90	1Oct92	31	34	0	1
17	3Aug90	2Dec90	4	54	0	1	67	20Apr89	19Sep89	5	46	1	1
18	10Jun90	8Jan92	19	35	0	0	68	16Jun89	15Apr94	58	22	0	1
19	12Jun91	11Sep91	3	44	1	0	69	1Oct90	31Oct90	1	44	1	1
20	7Jan91	8Mar91	2	38	0	1	70	1Feb91	3May91	3	37	0	0
21	29Aug89	28Oct89	2	40	0	0	71	13May89	10Dec92	43	25	0	1
22	29May89	27Nov89	6	34	1	1	72	9Aug90	8Sep90	1	38	0	1
23	16Nov90	14Nov95	60	25	0	0	73	18Dec91	17Jun92	6	32	0	1
24	9May90	8Apr91	11	32	0	1	74	23Aug90	21Jan95	53	34	0	1
25	10Sep91	9Nov91	2	42	1	0	75	19Jan91	19Mar92	14	29	0	1
26	26Dec91	26May92	5	47	0	1	76	26Aug91	25Dec91	4	36	1	1
27	29May91	27Sep91	4	30	0	0	77	16May91	13Nov95	54	21	0	1
28	1May90	31May90	1	47	1	1	78	20Mar89	19Apr89	1	26	1	1
29	24Mar91	22Apr92	13	41	0	1	79	5Oct91	4Nov91	1	32	1	1
30	18Jul89	17Oct89	3	40	1	1	80	21May91	19Jan92	8	42	0	1
31	16Sep90	15Nov90	2	43	0	1	81	10Jun91	9Nov91	5	40	1	1
32	22Jun89	22Jul89	1	41	0	1	82	31Aug89	30Sep89	1	37	1	1
33	27Apr90	25Oct92	30	30	0	1	83	28Dec91	27Jan92	1	47	0	1
34	16May90	14Dec90	7	37	0	1	84	29Sep90	28Nov90	2	32	1	1
35	19Feb89	20Jun89	4	42	1	1	85	20Nov91	19Jun92	7	41	1	0
36	17Feb90	18Oct90	8	31	1	1	86	2Jul89	1Aug89	1	46	1	0
37	6Aug91	5Jan92	5	39	1	1	87	11Oct91	10Aug92	10	26	1	1
38	10Aug89	10Jun90	10	32	0	1	88	11Oct90	10Oct92	24	30	0	0
39	27Dec90	25Feb91	2	51	0	1	89	5Dec90	5Jul91	7	32	1	1
40	26Apr89	24Jan90	9	36	0	1	90	8Sep89	8Sep90	12	31	1	0
41	4Dec90	3Dec93	36	43	0	1	91	10Apr90	9Aug90	4	35	0	1
42	28Apr91	28Jul91	3	39	0	1	92	11Dec90	9Sep95	57	36	0	1
43	9Jul91	7Apr92	9	33	0	1	93	15Dec90	14Jan91	1	41	1	1
44	31Dec89	1Apr90	3	45	1	1	94	13Jan89	13Jan90	12	36	1	0
45	20Dec89	18Nov92	35	33	0	1	95	22Aug91	21Mar92	7	35	1	1
46	22Jun91	20Feb92	8	28	0	1	96	2Aug91	1Sep91	1	34	1	1
47	11Apr90	11Mar91	11	31	0	1	97	22May91	21Oct91	5	28	0	1
48	22May90	19Jan95	56	20	1	0	98	2Apr90	1Apr95	60	29	0	0
49	11Nov91	10Jan92	2	44	0	0	99	1May91	30Jun91	2	35	1	0
50	18Jan91	19Apr91	3	39	1	1	100	11May89	10Jun89	1	34	1	1

for subject 1 is *exactly* 5 months while that of subject 2 is *at least* 6 months? We return to the univariate descriptive statistics problem shortly.

Suppose for the moment that we have performed the univariate analysis and wish to explore possibilities for an appropriate regression model. In linear regression modeling the first step is usually to examine a scatterplot of the outcome variable versus all continuous variables to see if the "cloud" of data points supports the use of a straight-line model. We also assess if there appears to be anything unusual in the scatter about a potential model. For example, is the linear model plausible except for one or two points? The fact that we have censored data presents a problem for the interpretation of a scatterplot with survival time data. If we were to ignore the censoring in survival time, then we would have an extension of the problem we noted with use of the arithmetic mean as an estimator of the "true" mean. The values obtained from any "line" fit to the cloud of points would not estimate the "mean" at that point. We would only know that the "mean" is *at least* as large as the point on the "line."

Regardless of this "at least" problem, a scatterplot is still a useful and informative descriptive tool with censored survival time data. However, to interpret the plot correctly we must keep track of the different types of observations by using different plotting symbols for the values assigned to the censoring variable. Figure 1.1 presents the scatterplot of TIME versus AGE for the data in Table 1.1, where different plotting symbols are used for the two levels of CENSOR. We formalize the statistical assumptions about the censoring later in Chapter 1, but for the moment we assume that it is independent of the values of survival time and all covariate variables.

Under the independence assumption the censored and non-censored points should be mixed in the plot with the mix dictated by the study design. Any trend in the plot is controlled by the nature and strength of the association between the covariate and survival time. For example, if age has a strong negative association with survival time, then observed survival times should be shorter for older subjects than for younger ones. If all subjects were followed for the *same fixed length of time*, then we would expect to find proportionally more censored observations among younger subjects than older ones. However, if subjects enter the study *uniformly over the study period* and independently of their age, then we would expect an equal proportion of censored observations at all ages. The example data are assumed to be from a study of

this type. We see in Figure 1.1 that the censored and non-censored observations are mixed at about a 4 to 1 ratio at all ages.

In the linear regression model the basic shape of the scatterplot is controlled by the nature and strength of the relationship between the outcome and covariate variables and the fact that the errors follow a normal distribution (a relatively short-tailed symmetric distribution). For example, if the relationship is systematically linear and strongly positive, then the cloud of points should be a tight ellipse oriented from lower left to upper right. If the relationship is weakly linear and positive, then the cloud will be more circular in shape with a left to right orientation. If the relationship is quadratic with a strong association, then the cloud may look like a banana. With survival data the shape of the plot is also controlled by the nature of the systematic relationship between "time" and the covariate, but the distribution of the errors is typically skewed to the right. The shape of the plot in Figure 1.1 is controlled by the strong association in these data between age and survival time, the fact that survival time is skewed to the right and the constraint that subjects can be followed for at most 84 months. The cloud of points in Figure 1.1 is densest for short survival times and slowly

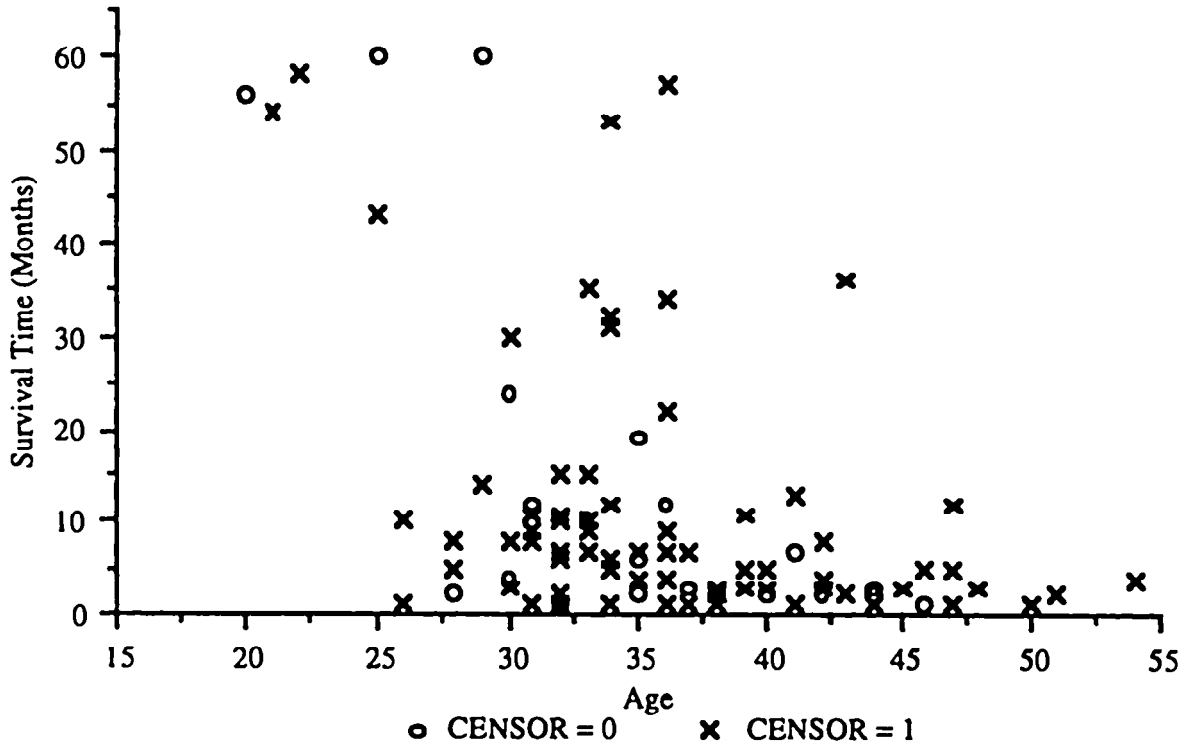


Figure 1.1 Scatterplot of survival time versus age for 100 subjects in the HMO-HIV+ study. The plotting symbols represent values of CENSOR.

trickles out to longer times with the plot truncated at the maximum length of follow-up.

In order to illustrate the shape of the plot when the covariate is strongly positively related to survival time, we reverse the order of age by creating a new variable IAGE = 1000/AGE. The scatterplot of TIME versus the created variable is shown Figure 1.2. In this case we see that the plot has the same shape but in the other direction.

We are still faced with the task of how to use the scatterplot to postulate a model for the systematic component and the issue of identifying an appropriate distribution for the errors. In linear regression when a choice for the parametric model is neither clearly indicated by the scatterplot nor provided by past experience or by some underlying biologic or clinical theory, we can use a technique called "scatterplot smoothing" to yield a non-parametric estimate of the systematic component. Cleveland (1993) discusses scatterplot smoothing and several of the methods are available in the STATA and S-Plus software packages as well as others. A scatterplot smoothing of a plot such as the one in Figure 1.1 could be difficult to interpret since censored and non-censored times have been treated equally. That is, the presence of the censored observations in the smoothing process could, in some examples, make it difficult to visualize the systematic component of the survival times.

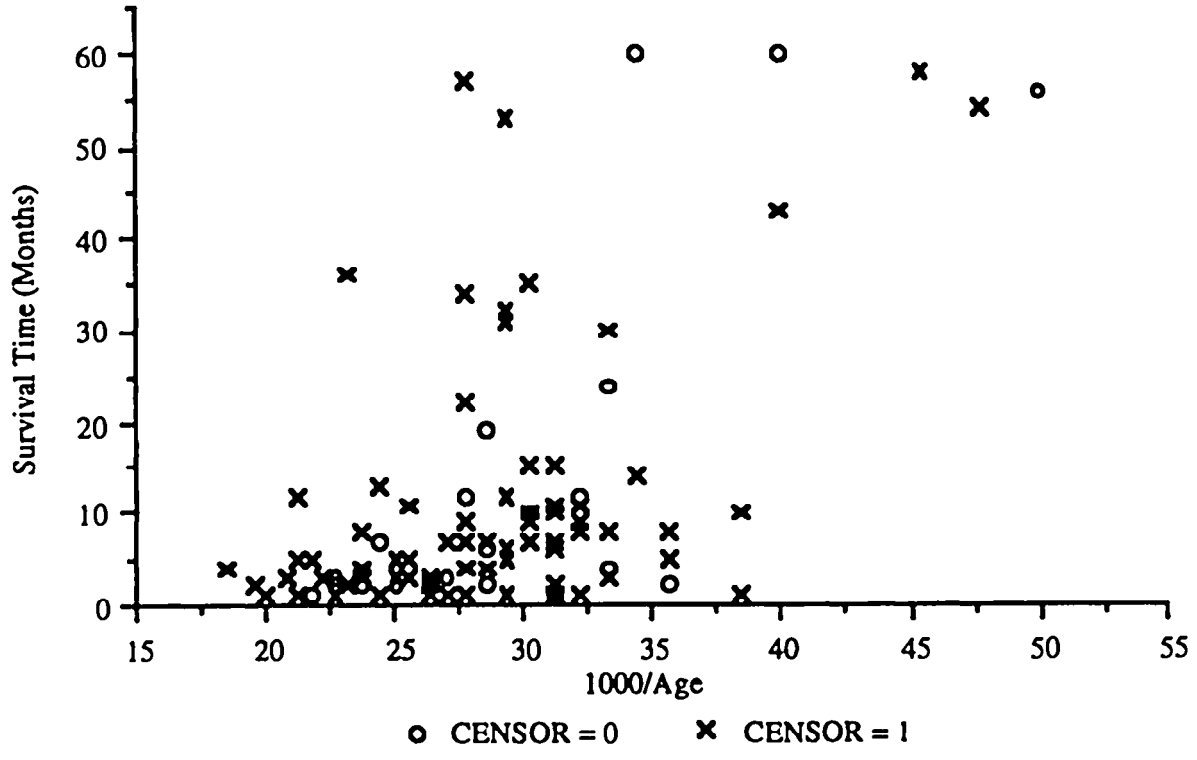


Figure 1.2 Scatterplot of survival time versus 1000/age for 100 subjects in the HMO-HIV+ study. The plotting symbols represent values of CENSOR.

The scatterplot in Figure 1.1 can be used to illustrate other fundamental differences between an analysis of censored survival time and a normal errors linear regression. The dependent variable, TIME, must take on positive values. Thus any model we choose for the systematic component of the model must yield fitted values which are strictly positive. This discourages use of a strictly linear model, as fitted values could be negative, especially for subjects with short survival times. If we look at Figure 1.1 and try to draw a smooth curve (systematic component) which, by eye, best fits the points, it would begin in the top left corner and drop sharply, curving to the lower right. Curves of this basic shape can often be described by a function with the basic form $t = e^{-x}$.

We noted that the distribution of survival times in Figure 1.1 appears to be skewed to the right. The simplest statistical distribution with this characteristic is the exponential distribution. The combination of an exponential systematic component and exponentially distributed errors suggests, as a beginning point, a regression model which is called the *exponential regression model.* If we assume that we have a single independent variable, x, then this model may be expressed as follows:

$$T = e^{\beta_0 + \beta_1 x} \times \varepsilon, \tag{1.1}$$

where T denotes survival time and ε follows the exponential distribution with parameter equal to one and is denoted $E(1)$ in this text.[2] The model in (1.1) has the desired properties of yielding positive values from a "curved" systematic component with a skewed error distribution. Note that this model is not linear in its parameters. However, it may be "linearized" by taking the natural log. (In this text $\log \equiv \log_e \equiv \ln$.) This yields the following model:

$$Y = \beta_0 + \beta_1 x + \theta, \tag{1.2}$$

where $Y = \ln(T)$ and $\theta = \ln(\varepsilon)$. The model in (1.2) looks like the equation for the usual normal errors linear regression model except that the distribution of the errors, θ, is not normal. Instead, the errors follow an "extreme minimum value" distribution. This distribution is not encountered often outside of applications in survival analysis but plays a central role in models of life-length and is often referred to as the Gumbel distribution. The mean of this distribution is 0 and its shape

---

[2] The $E(1)$ density function is $f(\varepsilon) = e^{-\varepsilon}$ and the survivorship function is $S(\varepsilon) = e^{-\varepsilon}$.

parameter is 1 (denoted $G(0,1)$ in this text[3]). The details of this distribution are presented in Lawless (1982). [Other texts such as Evans, Hastings and Peacock (1993) present the distribution of $-\theta = -\ln(\varepsilon)$, the "extreme maximum value" distribution.] The extreme minimum value distribution is derived by considering the statistical distribution of the minimum value from a simple random sample of observations. As the size of the sample increases, the distribution of the minimum value may be shown, after appropriate scaling, to be $G(0,1)$. The notion of a survival time being the minimum of many other times is an appealing, but somewhat simplistic, way to conceptualize survival time. For example, if the survival time of a complex object, such as a computer, depends on the continued survival of each of a large number of components whose failures are independent, then survival of the computer terminates when the first component fails (i.e., the minimum value of many independent, identically distributed, observations of time). The same analogy could be used to characterize the death of a human being.

The use of the distribution $G(0,1)$ in (1.2) is somewhat like using the standard normal distribution in linear regression. The standard normal distribution is denoted $N(0,1)$ in this text. From practical experience we know that, in linear regression, the errors rarely if ever have variance equal to one. The usual assumption is that the variance is neither a function of the outcome variable nor of the independent variables. It is assumed to be constant and equal to the parameter σ^2. This distribution is denoted $N(0,\sigma^2)$. An additional parameter may be introduced into (1.2) by multiplying θ by σ to yield the model

$$y = \beta_0 + \beta_1 x + \sigma \times \theta. \tag{1.3}$$

The distribution of $\sigma \times \theta$ is denoted as $G(0,\sigma)$.

The problem we face now is not only how to fit models like those in (1.1)–(1.3) but how to fit them when some of the observations of the outcome variable are censored. In linear regression with normal errors, *least squares* is the method discussed in regression texts such as Kleinbaum, Kupper, Muller and Nizam (1998) and used by most (probably all) computer software packages. This approach yields estimators with a number of desirable statistical properties. They are normally distributed with variances and covariances whose estimates are available in the output from the regression programs in all software packages. This allows

---

[3] The density function of the $G(0,1)$ is $f(\theta) = e^{[\theta - \exp(\theta)]}$ and the survivorship function is $S(\theta) = e^{-\exp[\theta]}$.

put from the regression programs in all software packages. This allows for the t-distribution, with appropriately chosen degrees-of-freedom, to be used to form confidence intervals and to test hypotheses about individual parameters. The F-distribution, with appropriate degrees-of-freedom, may be used to assess overall model significance. Least squares is an estimation method with its own statistical properties, but it may also be viewed, with normally distributed errors, as a special case of an estimation method called *Maximum Likelihood Estimation* (MLE). We use MLE with an adaptation for censored data to fit the models in (1.1)–(1.3). This allows us to appeal to the well-developed theory for maximum likelihood estimators to test hypotheses and form confidence intervals for individual parameters and to assess overall model significance with the same ease and simplicity of computation as in linear regression.

The simplest way to conceptualize our data is to assume that continued observation of a subject is controlled by two completely independent time processes. The first is the actual survival time associated with the disease of interest. For example, in the HMO-HIV+ study it would be the length of survival after diagnosis as HIV+. The second is the length of time until a subject is lost to follow-up. Again in the HMO-HIV+ study this would be the length of time until the subject moved, died from another cause such as an auto accident, etc. We assume both of these are under observation and that the recorded time represents time to the event that occurred first. Two variables are used to characterize a subject's time, the actual observed time, T, and a censoring indicator variable, C. In this text we use $c=1$ to denote that the observed value of T measures the actual survival time of the subject (i.e., death from the "disease" of interest was the reason follow-up ended on the subject). We use $c=0$ to denote that follow-up ended on the subject for reasons other than death from the disease of interest. Actual observed values of these variables and a covariate for a subject are denoted by lower case letters in the triplet (t,c,x) where x denotes the value of a covariate of interest. For example, the triplet for subject 1 in Table 1.1 with AGE as the covariate is (5, 1, 46), where x = age at the time of enrollment into the study. This triplet states that subject 1 was observed for $t=5$ months when the subject died from AIDS or AIDS-related causes ($c=1$) and was $x=46$ years old at the time of enrollment into the study. The triplet for subject 2 is (6, 0, 35). This triplet states that subject 2 was observed for 6 months before being lost for some reason unrelated to being HIV+ ($c=0$) and was 35 years old at the time of enrollment into the study.

The first step in maximum likelihood estimation is to create the specific likelihood function to be maximized. In simplest terms, the likelihood function is an expression that yields a quantity similar to the probability of the observed data under the model. First, we create a fairly general likelihood function, then we apply the method to the models in (1.1)–(1.3). Suppose that the distribution of survival time for a subject with covariate x and the disease of interest can be described by the cumulative distribution function $F(t,\boldsymbol{\beta},x)$. For example, the value of the function $F(5,\boldsymbol{\beta},46)$ gives the proportion of 46-year-old subjects expected to die from AIDS or AIDS-related causes in less than 5 months. The quantity $\boldsymbol{\beta}$ denotes the parameters of the distribution, which we need to estimate. For example, when we use the models in (1.1)–(1.3) the unknown parameters are $\boldsymbol{\beta}=(\beta_0,\beta_1)$. The *survivorship function* is obtained from the cumulative distribution and is defined as $S(t,\boldsymbol{\beta},x)=1-F(t,\boldsymbol{\beta},x)$. The value of the function $S(5,\boldsymbol{\beta},46)$ gives the proportion of 46 year olds expected to live at least 5 months. To create the likelihood function, we also need a function that we think of, for the moment, as giving the "probability" that the survival time is exactly t. This function is derived mathematically from the distribution function and is called the density function. We denote the density function corresponding to $F(t,\boldsymbol{\beta},x)$ as $f(t,\boldsymbol{\beta},x)$. For example, the value of the function $f(5,\boldsymbol{\beta},46)$ gives the "probability" that a subject 46 years old survives exactly 5 months.[4]

We construct the actual likelihood function by considering the contribution of the triplets $(t,1,x)$ and $(t,0,x)$ separately. In the case of the triplet $(t,1,x)$ we know that the survival time was exactly t. Thus the contribution to the likelihood for this triplet is the "probability" that a subject with covariate value x dies from the disease of interest at time t units. This is given by the value of density function $f(t,\boldsymbol{\beta},x)$. For the triplet $(t,0,x)$ we know that the survival time was at least t. Thus the contribution to the likelihood function of this triplet is the *probability* that a subject with covariate value x survives at least t time units. This probability is given by the survivorship function $S(t,\boldsymbol{\beta},x)$. Under the assumption of independent observations, the full likelihood function is obtained by multiplying the respective contributions of the observed triplets, a value of $f(t,\boldsymbol{\beta},x)$ for a noncensored observation and a value

---

[4] Readers having had some mathematical statistics know that the density function does not yield a probability but a probability per-unit of time over a small interval of time, $f(t,\boldsymbol{\beta},x)=\lim_{\Delta t\to 0}\{F(t+\Delta t,\boldsymbol{\beta},x)-F(t,\boldsymbol{\beta},x)/\Delta t\}$.

of $S(t,\beta,x)$ for censored observations. In general, a concise way to denote the contribution of each triplet to the likelihood is the expression

$$[f(t,\beta,x)]^c \times [S(t,\beta,x)]^{1-c}, \tag{1.4}$$

where $c = 0$ or 1.

We denote the observed data for a sample of n independent observations as (t_i, c_i, x_i) for $i = 1,2,\ldots,n$. Since the observations are assumed to be independent, the likelihood function is the product of the expression in (1.4) over the entire sample and is

$$l(\beta) = \prod_{i=1}^{n} \left\{ [f(t_i,\beta,x_i)]^{c_i} \times [S(t_i,\beta,x_i)]^{1-c_i} \right\}. \tag{1.5}$$

To obtain the maximized likelihood with respect to the parameters of interest, β, we maximize the log-likelihood function,

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i \ln[f(t_i,\beta,x_i)] + (1-c_i) \ln[S(t_i,\beta,x_i)] \right\}. \tag{1.6}$$

Since the log function is monotone, the maximum of (1.5) and (1.6) occur at the same value of β; however, maximizing (1.6) is computationally simpler than maximizing (1.5). The procedure to obtain the values of the MLE involves taking derivatives of $L(\beta)$ with respect to β, the unknown parameters, setting these equations equal to zero, and solving for β.

Before becoming completely involved in maximum likelihood estimation, let us consider the implications and assumptions of our model. There are several key points to be made. We have assumed that we are in "constant contact" with our subjects and thus are able to record the exact time of survival or follow-up. In essence we have treated time as a continuous variable. Scenarios where time is observed less precisely are considered in Chapter 7. We have accounted for the partial information on survival time contained in the censored observations. That is, we have explicitly used the fact that we know survival is at least as large as the recorded follow-up time via the inclusion in the likelihood of the term $S(t,\beta,x)$ for all censored observations. Another key point is that the reasons for observing a censored observation are assumed to be completely unrelated to the disease process of interest. In the example, we assume that being lost to follow-up is unrelated to the progression of

disease in an HIV+ subject. We exclude the possibility that subjects have moved to another location which they perceive to offer better care for an HIV+ individual.

After a careful examination of the scatterplot in Figure 1.1, we arrived at the conclusion that the exponential regression model in (1.1) might be a good starting point to model these data. We also noted that the model in (1.1) could be linearized to the model shown in (1.2) and further generalized by the inclusion of a shape parameter in (1.3). We now apply MLE to each of these models in turn to show that (1.1) and (1.2) are equivalent, with (1.1) yielding fitted values for time and (1.2) for log-time. Comparison of (1.1) and (1.2) to (1.3) requires discussion of the role of the extra shape parameter in the analysis.

Suppose we wish to use a software package to fit the exponential regression model in (1.1) to the data displayed in Figure 1.1. We would find that many packages (e.g., BMDP, EGRET, SAS and STATA) fit, as a default, the model in (1.2). Once this model has been fit, we can convert it by exponentiation to estimate the model in (1.1). The equations to be solved to obtain the MLE of $\boldsymbol{\beta}$ are identical for the models in (1.1) and (1.2). Thus we show in detail the application of MLE to the log-linearized model in (1.2).

The model in (1.2) states that the values of log(survival time) come from a distribution of the form $\beta_0 + \beta_1 x + G(0,1)$. This is the extreme minimum value distribution with mean equal to $\beta_0 + \beta_1 x$ and is denoted $G(\beta_0 + \beta_1 x, 1)$. Another way to describe the model is to subtract the part involving the unknown parameters (the systematic component) from both sides of the equation in (1.2) and note that since this difference, $y - (\beta_0 + \beta_1 x)$, is equal to θ, it is distributed $G(0,1)$. Thus we may obtain the contributions to the likelihood function by substituting the expression $y - (\beta_0 + \beta_1 x)$ into the equations defining the survivorship and density function for $G(0,1)$ as follows:

$$S(y, \boldsymbol{\beta}, x) = e^{-\exp[y-(\beta_0+\beta_1 x)]} \tag{1.7}$$

and

$$f(y, \boldsymbol{\beta}, x) = e^{\{y-(\beta_0+\beta_1 x)-\exp[y-(\beta_0+\beta_1 x)]\}}. \tag{1.8}$$

Substituting the expressions in (1.7) and (1.8) into (1.6) yields the following log-likelihood:

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} c_i \ln\left(e^{\{y_i-(\beta_0+\beta_1 x_i)-\exp[y_i-(\beta_0+\beta_1 x_i)]\}}\right) + (1-c_i)\ln\left(e^{-\exp[y_i-(\beta_0+\beta_1 x_i)]}\right)$$

$$= \sum_{i=1}^{n} c_i\left[y_i-(\beta_0+\beta_1 x_i)\right] - e^{[y_i-(\beta_0+\beta_1 x_i)]}. \qquad (1.9)$$

In order to obtain the MLE of $\boldsymbol{\beta}$, we must take the derivatives of the log-likelihood in (1.9) with respect to β_0 and β_1, set the two resulting expressions equal to zero and solve them for β_0 and β_1. The two equations to be solved are

$$\sum_{i=1}^{n}\left(c_i - e^{[y_i-(\beta_0+\beta_1 x_i)]}\right) = 0 \qquad (1.10)$$

and

$$\sum_{i=1}^{n} x_i\left(c_i - e^{[y_i-(\beta_0+\beta_1 x_i)]}\right) = 0. \qquad (1.11)$$

The equations in (1.10) and (1.11) are nonlinear in β_0 and β_1 and must be solved using an iterative method. It is not important to understand the details of how these equations are solved at this point since any software package we choose to use will have such a method. We used the exponential regression command in STATA, "ereg," to fit this model to the data in Table 1.1 using x = AGE.

Table 1.2 presents the parameter estimates in the column labeled "Coeff." and estimates of the standard error of the estimated parameters in the column labeled "Std. Err." The standard error estimates are obtained from theoretical results of maximum likelihood estimation. The column labeled "z" is the ratio of the estimated coefficient to its estimated standard error and is the Wald statistic for the respective parameter. Under the usual assumptions for maximum likelihood es-

Table 1.2 Estimated Parameters, Standard Errors, z-Scores, Two-Tailed p-Values and 95 Percent Confidence Intervals for the Log-Time Exponential Regression Model Fit to the Data in Table 1.1

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% Conf. Int.
Age	-0.094	0.0158	-5.96	0.00	-0.124, -0.063
Constant	5.859	0.5853	10.01	0.00	4.711, 7.006

timation, the Wald statistic follows the standard normal distribution under the hypothesis that the true parameter value is zero. The last two columns provide a two-tailed p-value and the endpoints of a 95 percent confidence interval computed under these assumptions.

The output in Table 1.2 shows that the maximum likelihood estimates of the two parameters are

$$\hat{\beta}_0 = 5.859 \text{ and } \hat{\beta}_1 = -0.094.$$

In this text the "^" is used to indicate that a particular quantity is the maximum likelihood estimate. We can use the estimates in Table 1.2 in the same manner as is used in linear regression to obtain an equation which provides predicted (i.e., fitted) values of the outcome variable, log-time. The resulting equation is $\hat{y} = 5.859 - 0.094\text{AGE}$. This equation may be converted to one providing fitted values for time by exponentiation, namely

$$\hat{t} = e^{5.859 - 0.094\text{AGE}}.$$

This conversion is similar to that used to convert parameter estimates in logistic regression to estimates of odds ratios. In order to see the results of fitting this model, we add it to the scatterplot that was shown in Figure 1.1. The new scatterplot with the fitted values is presented in Figure 1.3.

Recall that the objective of the analysis was to postulate and then fit a model which would yield positive fitted values and display the curvature observed in Figure 1.1 for the systematic component. Examining the plot in Figure 1.3, we can see that the fitted model has both of these properties. The curve does not go through the middle of the data in the sense that 26 data points lie above the curve and 74 below it. Intuitively, since censored observations represent lower bounds on unobserved survival times, one would expect the curve to be shifted upward. The actual location of the fitted curve on the graph depends on the value of $\hat{\beta}_0$, whose value depends on the percentage of censored observations. It suffices for the moment to note that if 80 percent of the data had been censored, the curve could have fallen above all of the points on the graph. On the whole we find, at least visually, that the model seems to provide an adequate descriptor of the trend in the data.

One possible approach to improving the fitted model would be to see whether the addition of the shape parameter, σ, in (1.3) contributes significantly to the model. The model in (1.3) is a log-Weibull distri-

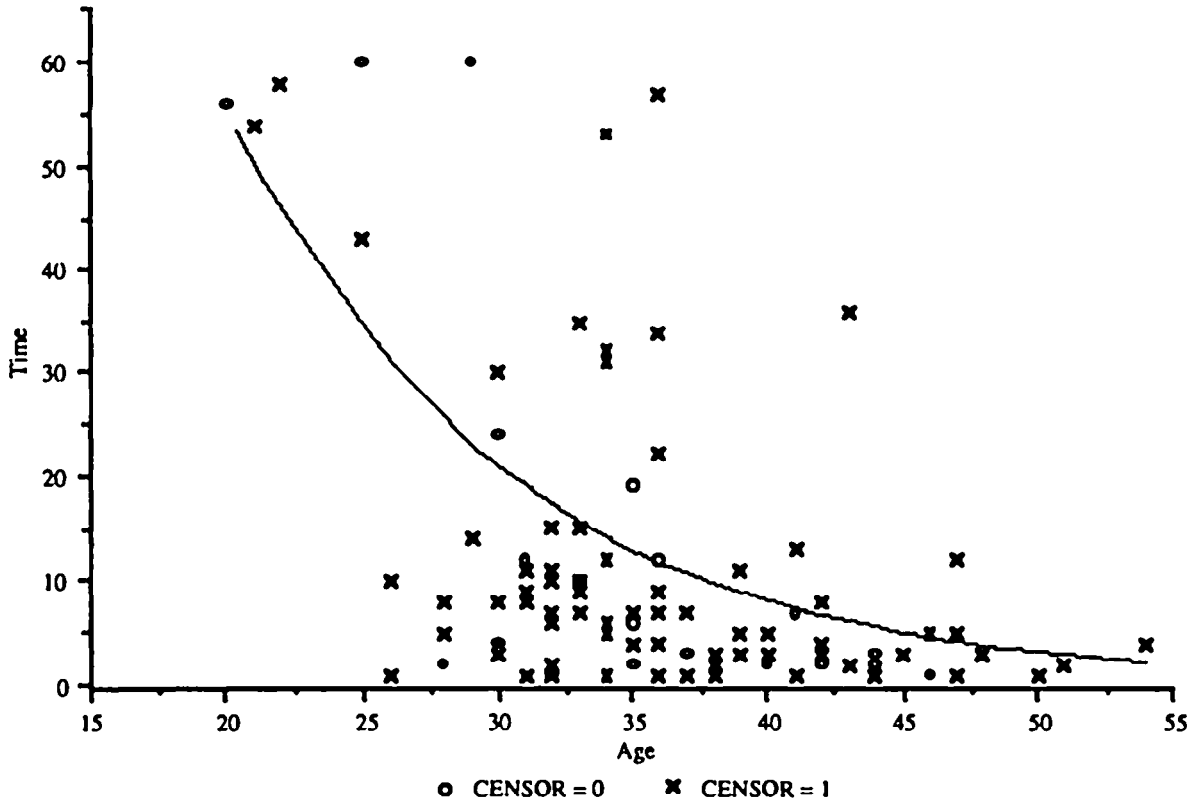


Figure 1.3 Scatterplot of survival time versus age for 100 subjects in the HMO-HIV+ study. The values of censor are the plotting symbol. The smooth curve is the fitted values, $\hat{t} = \exp(5.859 - 0.094\text{AGE})$, from the exponential regression model in Table 1.2.

bution (see Chapter 8). We note, without showing the actual output, that the shape parameter is not significant. (For those interested, this was done by fitting the model using STATA's Weibull Regression command, "weibull.") Thus we conclude that, of the two models considered, model (1.2) describes the data as well as the more complicated model (1.3).

If we were to continue to use the linear regression modeling paradigm to motivate our approach to the analysis of these data, the next step would be to check the scale of age in the systematic component, making sure that the data support a linear model. If not, a suitable transformation must be identified. Once we felt we had done the best possible job of building the systematic component, we would use appropriately formulated regression diagnostics to search for overly influential and/or poorly fit points. This would be followed by an examination of the distribution of the estimated residuals to see if our assumptions about the error component hold. Once convinced that our model was the best fitting model possible, we would provide a clinical

interpretation of the estimated model parameters. This important series of tasks is not addressed at this point, but it provides the approach for much of what follows in this text.

In summary, the HMO-HIV+ example has served to highlight the similarities and, more importantly, the differences we must address when trying to apply the linear regression modeling paradigm to the analysis of survival time data. The fact that we observed "time" places restrictions on the types of models that can be used. Any model must yield positive fitted values and its error component will be more likely to have a skewed distribution (e.g., exponential-like) than a symmetric one such as the normal. In addition, the presence of incompletely observed or censored values of "time" necessitates modifications to the standard maximum likelihood approach to estimation. It is this latter point that tends to make the analysis of survival data more complicated than a typical linear or logistic regression analysis. Thus we present a more detailed discussion of typically encountered censoring mechanisms.

1.2 TYPICAL CENSORING MECHANISMS

It may seem somewhat obvious, but we cannot discuss a censored observation until we have carefully defined an uncensored observation. This point may seem trivial, but in applied settings confusion about censoring may not be due to the incomplete nature of the observations but rather may be the result of an unclear definition of survival time. The observation of survival time, life-length, or whatever other term may be used has two components which must be unambiguously defined: a beginning point where $t = 0$ and a reason or cause for the observation of time to end. For example, in a randomized clinical trial, observation of survival time may begin on the day a subject is randomized to receive one of the treatment protocols. In an occupational exposure study, it may be the day a subject began work at a particular plant. In the HMO-HIV+ study discussed above, it was when a subject met the clinical criteria for being diagnosed as HIV+ and entered the study. In some applications it may not be obvious what the best $t = 0$ point should be. For example, in the HIV+ study, the best $t = 0$ point might be infection date; another choice might be the date of diagnosis; and a third, the criteria used in the example, might be diagnosis *and* enrollment in the study. 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.

In practice, a value of time may be obtained by calculating the number of days (or months, or years, etc.) between two calendar dates. Table 1.1 presents the entry date and end date for the subjects in the HMO-HIV+ study. Most statistical software packages have functions which 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 BMDP, SAS and STATA is January 1, 1960. Subject 1 entered the study on May 15, 1990 which is 11,092 days after the reference date, and died October 14, 1990 which is 11,244 days after the reference date. The interval between these two dates is $11,244 - 11,092 = 152$ days. The number of days is converted into the number of months by dividing by 30.4375 ($= 365.25/12$). Thus the survival time in months for subject 1 is 4.994 ($= 152/30.4375$). It is common, when reporting results in tabular form, to round to the nearest whole number as shown in Table 1.1 (i.e., 5 months). The level of precision used for survival time will depend on the particular application. Clock time may be combined with calendar date to obtain survival time in units of fractions of days.

Two mechanisms that can lead to incomplete observation of time are censoring and truncation. A censored observation is one whose value is incomplete due to random factors for each subject. A truncated observation is one which is incomplete due to a selection process inherent in the study design. The most commonly encountered form of a censored observation is one in which observation begins at the defined time $t = 0$ and terminates before the outcome of interest is observed. Since 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 HMO-HIV+ study, a subject could move out of town, could die in an auto accident or the study could end before death from the disease of interest could be observed. 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 observational study, such as the HMO-HIV+ study, patients may enter the study over a several year enrollment period. As we see from the data reported in Table 1.1, subject 2 entered the study on September 19, 1989 while subject 4 entered on January 3,

1991. In this type of study, each subject's calendar beginning point is assumed to define the $t = 0$ point.

For obvious practical reasons all studies have a point at which observation ends on all subjects; therefore subjects entering at different times will have variable lengths of maximum follow-up time. In the HMO-HIV+ study, subjects were enrolled between January 1, 1989 and December 31, 1991, with follow-up ending December 31, 1995. Thus, the longest any subject could have been followed was 7 years. For example, subject 5 entered the study on September 18, 1989. Thus the longest this subject could have been followed was 6 years and 3.5 months. However, this subject was not followed for the maximum length of time as the subject died of AIDS or AIDS-related causes on July 19, 1991, yielding a survival time of 22 months. Incomplete observation of survival time due to the end of the study is also a right-censored observation.

A typical pattern of entry into a follow-up study is shown in Figure 1.4. This is a hypothetical 2-year-long 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,

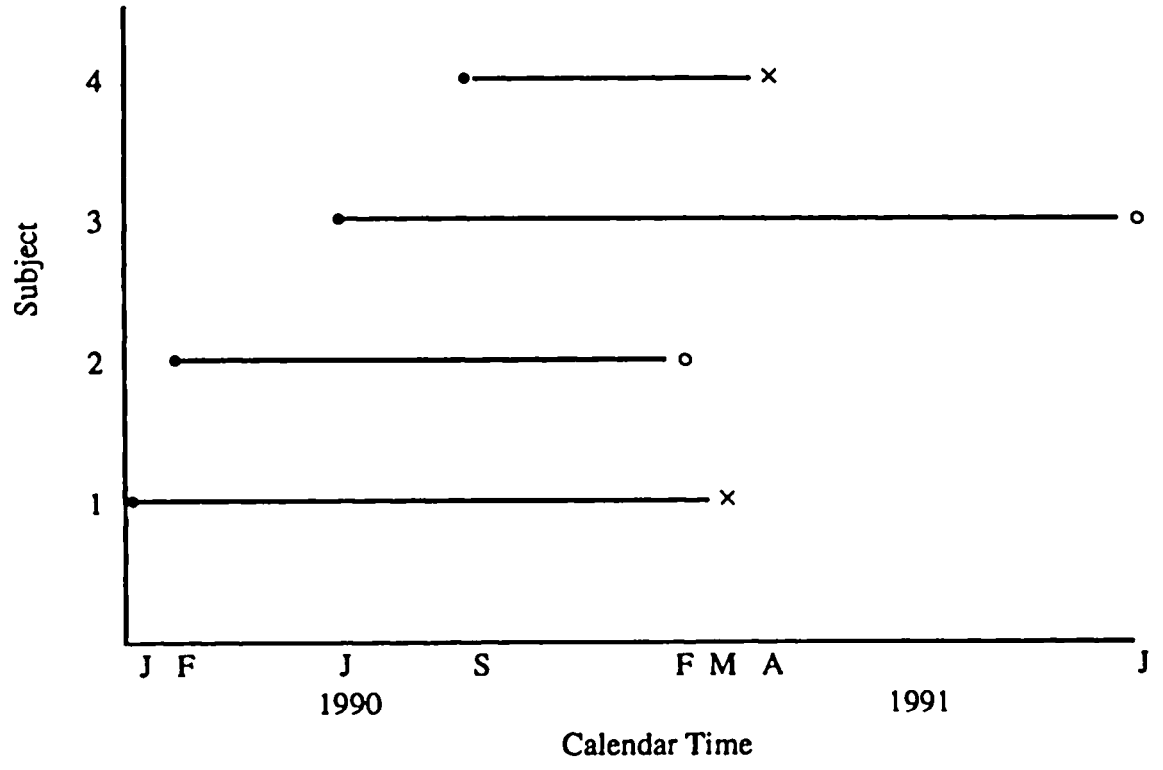


Figure 1.4 Line plot in calendar time for four subjects in a hypothetical follow-up study.

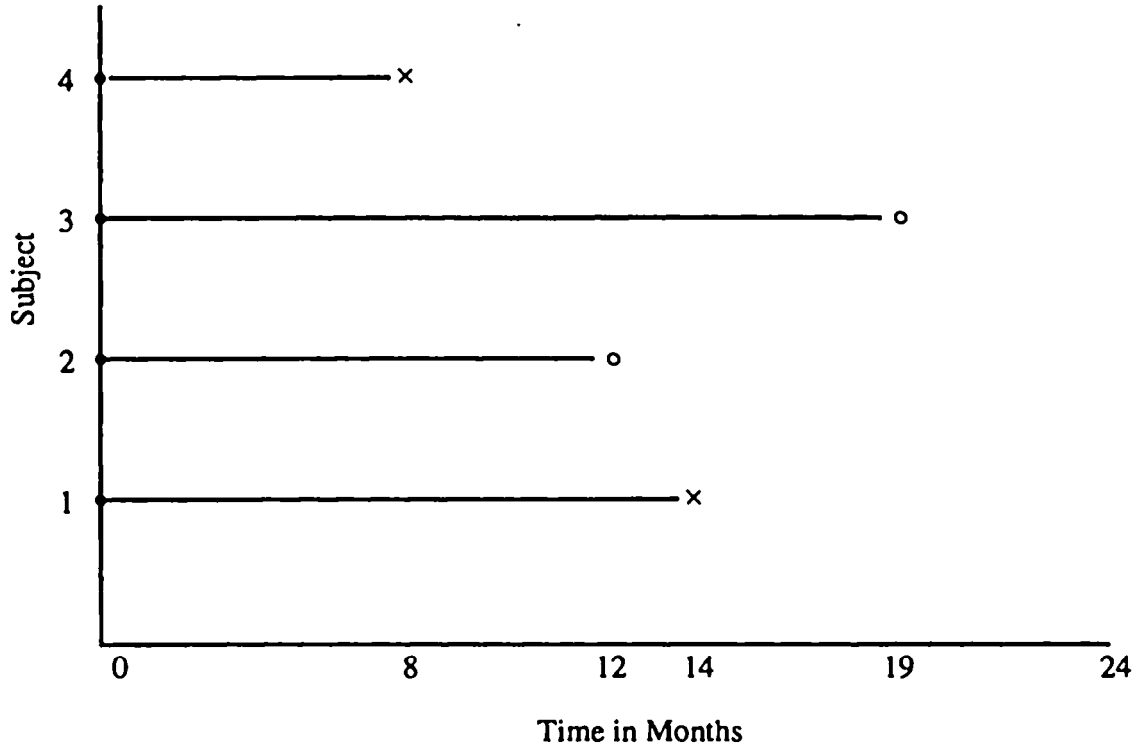


Figure 1.5 Line plot in the time scale for four subjects in a hypothetical follow-up study.

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 right-censored observations of survival time. These data are plotted on the actual time scale in months in Figure 1.5. Note that each subject's time has been plotted as if he or she were enrolled at exactly the same calendar time and were followed until his or her respective end point.

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 was "at risk" for experiencing menarche since 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 censoring mechanism that can occur 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 subect's time is 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. 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 selection process must be taken into account in the analysis. This type of incomplete observation of time is called *right truncation.*

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 to first relapse to IV drug use. Since 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 which assumed that 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 mechanisms and analysis of right-censored data throughout this text since this is the most commonly occurring form of censoring. Modifications of the methods of analysis appropriate for right-censored data to other censoring mechanisms is discussed in Chapter 7.

Prior to the development of a regression model for the relationship between age and survival time among the subjects in the HMO-HIV+ study, we mentioned that the first step in any analysis of survival time, or for that matter any set of data, should be a thorough univariate analysis. In the absence of censoring, this 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 histograms, box and whisker plots, cumulative percent 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 each continuous variable. The presence of censored data in the sample complicates the calculations but not the fundamental goal of univariate analysis. In the next chapter we pre-

sent the methods for univariate analysis in the presence of right-censored data.

1.3 EXAMPLE DATA SETS

In addition to the data from the hypothetical HMO-HIV+ study introduced in this chapter, data from two additional studies will be used throughout the text to illustrate methods and provide data for the end of chapter exercises. The data from all three studies may be obtained from the John Wiley & Sons (ftp://ftp.wiley.com/public/sci_tech_med/survival) web site. They may also be obtained from the web site of statistical data sets at the University of Massachusetts/Amherst in the section on survival data (http://www-unix.oit.umass.edu/~statdata).

Our colleagues, Drs. Jane McCusker, Carol Bigelow, and Anne Stoddard, have provided us with 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.

We refer to the two treatment program sites as A and B in this text. The trial at site A randomized 444 participants and was a comparison of 3- and 6-month modified therapeutic communities which incorporated elements of health education and relapse prevention. Clients in the relapse prevention/health education program (site A) were taught to recognize "high-risk" situations that are triggers to relapse and were taught the skills to enable them to cope with these situations without using drugs. In the trial at site B, 184 clients were randomized to receive either a 6- or 12-month therapeutic community program involving a highly structured life-style in a communal living setting. Our colleagues have published a number of papers reporting the results of this study, see McCusker et. al. (1995, 1997a, 1997b).

As is shown in the coming chapters, the data from the UIS provide a rich setting for illustrating methods for survival time analysis. The small subset of variables from the main study we use in this text is described

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
BECKTOTA	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 (Exit Date – Admission Date)	Days
TIME	Time to Return to Drug Use (Measured from Admission)	Days
CENSOR	Returned to Drug Use	1 = Returned to Drug Use 0 = Otherwise

in Table 1.3. Since the analyses we report in this text are based on this small subset of variables, the results reported here should not be thought of as being in any way comparable to results of the main study. In addition we have taken the liberty in this text 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 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 in such a way as to preserve subject confidentiality.

Another data set has been provided by our colleague Dr. Robert Goldberg of the Department of Cardiology at the University of Massachusetts Medical School. The data come from The Worcester Heart Attack Study (WHAS). The main goal of this study is to describe trends over time in the incidence and survival rates following hospital admission for acute myocardial infarction (AMI). Data have been collected

Table 1.4 Description of the Variables Obtained from the Worcester Heart Attack Study (WHAS), 481 Subjects

Variable	Description	Codes / Values
ID	Identification Code	1–481
AGE	Age at Hospital Admission	Years
SEX	Gender	0 = Male, 1 = Female
CPK	Peak Cardiac Enzymes	International Units (IU/100)
SHO	Cardiogenic Shock Complications	0 = No, 1 = Yes
CHF	Left Heart Failure Complications	0 = No, 1 = Yes
MIORD	MI Order	0 = First, 1 = Recurrent
MITYPE	MI Type	1 = Q-wave, 2 = Not Q-wave 3 = Indeterminate
YEAR	Cohort Year	1 = 1975, 2 = 1978, 3 = 1981, 4 = 1984, 5 = 1986, 6 = 1988
YRGRP	Grouped Cohort Year	1 = 1975 & 1978 2 = 1981 & 1984 3 = 1986 & 1988
LENSTAY	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	Status as of Last Follow-up	0 = Alive 1 = Dead

during ten 1-year periods beginning in 1975 on all AMI patients admitted to hospitals in the Worcester, Massachusetts, metropolitan area. The main data set has information on more than 8,000 admissions. The data in this text were obtained by taking a 10 percent random sample within 6 of the cohort years. In addition only a small subset of variables is included in our data set, and subjects with any missing data were dropped from the sampled data set. Dr. Goldberg and his colleagues have published more than 30 papers reporting the results of various analyses from the WHAS. The reader interested in learning more about the WHAS and its findings should see Goldberg et. al. (1986, 1988, 1989, 1991, 1993) and Chiriboga et al. (1994). A complete list of WHAS papers may be obtained by contacting the authors of this text.

Table 1.4 describes the subset of variables used along 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.

Various survival time variables can be created from the hospital admission date, the hospital discharge date and the date of the last follow-up. Two times have been calculated from these dates and are included in the data set, length of hospital stay (hospital admission to discharge) and total length of follow-up (hospital admission to last follow-up). Each has its own censoring variable denoting whether the subject had died or was alive at hospital discharge or last follow-up, respectively. As noted, the data set we use in this text contains a few key patient demographic characteristics and variables describing the nature of the AMI. One should be aware of the fact that the values of the variable peak cardiac enzymes are unadjusted to the respective hospital norm. The principle rationale for inclusion of this covariate is to provide a continuous covariate that may be predictive of survival and require some sort of non-linear transformation when included in the regression models discussed in this text.

EXERCISES

1. Using the data from the Worcester Heart Attack Study:

(a) Graph length of follow-up versus age using the censoring variable at follow-up as the plotting symbol for each of the pooled cohorts defined by YRGRP. Are the plots basically the same or do they differ in shape in an important way? Is it possible to tell from the shape of the plot if age is a predictor of survival time?

(b) What key characteristics of the data plotted in problem 1(a) should be kept in mind when choosing a possible regression model?

(c) By eye, draw on each of the three scatterplots from problem 1(a) what you feel is the best regression function for a survival time regression model.

(d) Obtain a cross tabulation of YRGRP and the censoring variable FSTAT and compute the percent dead and the percent censored in each of the three groups. What effect do you think the difference in the percent censored should have on the location of the lines drawn in problem 1(e)?

(e) Fit the exponential regression model to the data in each of the three scatterplots and add the fitted values to the plot (e.g., see Figure 1.3). How do the regression fitted values compare to the ones drawn in problem 1(c)? Is the response to problem 1(d) correct?

2. What key characteristics about the observations of total length of follow-up must be kept in mind when considering the computation of simple univariate descriptive statistics?

3. Repeat problems 1 and 2 using time to return to drug use and age in the UIS and grouping by study site.

CHAPTER 2

Descriptive Methods for Survival Data

2.1 INTRODUCTION

In any applied setting, a statistical analysis should begin with a thoughtful and thorough univariate description of the data. The fundamental building block of this analysis is an estimate of the cumulative distribution function. Typically, not much attention is paid to this fact in an introductory course on statistical methods, where directly computed estimators of measures of central tendency and variability are more easily explained and understood. However, routine application of standard formulas for estimators of the sample mean, variance, median, etc., will not yield estimates of the desired parameters when the data include censored observations. In this situation, we must obtain an estimate of the cumulative distribution function in order to obtain values of statistics which do estimate the parameters of interest.

In the HMO-HIV+ study described in Chapter 1, we assume that the recorded data are continuous and are subject to right censoring only. Remember that time itself is always continuous, but our inability to measure it precisely is an issue that we must deal with. We introduced the cumulative distribution function in Chapter 1 along with its complement, the survivorship function. Simply stated, the cumulative distribution function is the probability that a subject selected at random will have a survival time less than some stated value, t. This is denoted as $F(t) = \Pr(T < t)$. The survivorship function is the probability of observing a survival time greater than or equal to some stated value t, denoted $S(t) = \Pr(T \geq t)$. In most applied settings we are more interested in describing how long the study subjects live, than how quickly they die. Thus estimation (and inference) focuses on the survivorship function.

2.2 ESTIMATION OF THE SURVIVORSHIP FUNCTION

The Kaplan-Meier estimator of the survivorship function [Kaplan and Meier (1958)], also called *the product limit* estimator, is the estimator used by most software packages. This estimator incorporates information from all of the observations available, both uncensored and censored, by considering survival to any point in time as a series of steps defined by the observed survival and censored times. It is analogous to considering a toddler who must take five steps to walk from a chair to a table. This journey of five steps must begin with one successful step. The second step can only be taken if the first was successful. The third step can be taken only if the second (and also the first) was successful. Finally the fifth step is possible only if the previous four were completed successfully. In an analysis of survival time, we estimate the conditional probabilities of "successful steps" and then multiply them together to obtain an estimate of the overall survivorship function.

To illustrate these ideas in the context of survival analysis, we describe estimation of the survivorship function in detail using data for the first five subjects in the HMO-HIV+ study in Table 1.1, as shown in Table 2.1.

The "steps" are intervals defined by a rank ordering of the survival times. Each interval begins at an observed time and ends just before the next ordered time and is indexed by the rank order of the time point defining its beginning. Subject 4's survival time of 3 months is the shortest and is used to define the interval $I_0 = \{t : 0 \le t < 3\} = [0,3)$. The expression in curly brackets, { }, defines a collection or set of values that includes all times beginning with and including 0 and up to, but not including, 3. This is more concisely denoted using the mathematical notation of a square bracket to mean the value is included, a parenthesis to mean the value is not included, and the comma to mean all values in between. We use both notations in this text. The second rank-ordered

Table 2.1 Survival Times and Vital Status (Censor) for Five Subjects from the HMO-HIV+ Study

Subject	Time	Censor
1	5	1
2	6	0
3	8	1
4	3	1
5	22	1

time is subject 1's survival time of 5 months. This survival time, in conjunction with the ordered survival time of subject 4, defines interval $I_1 = \{t : 3 \le t < 5\} = [3,5)$. The next ordered time is subject 2's censored time of 6 months and, in conjunction with subject 1's value of 5 months, defines interval $I_2 = \{t : 5 \le t < 6\} = [5,6)$. The next interval uses subject 3's value of 8 months and the previous value of 6 months and defines $I_3 = \{t : 6 \le t < 8\} = [6,8)$. Subject 5's value of 22 months and subject 3's value of 8 months are used to define the next to last interval $I_4 = \{t : 8 \le t < 22\} = [8,22)$. The last interval is defined as $I_5 = \{t : t \ge 22\} = [22,\infty)$.

All subjects were alive at time $t = 0$ and remained so until subject 4 died at 3 months. Thus, the estimate of the probability of surviving through interval I_0 is 1.0; thus, the estimate of the survivorship function is

$$\hat{S}(t) = 1.0$$

at each t in I_0. Just before time 3 months, five subjects were alive, and at 3 months one subject died. In order to describe the value of the estimator at 3 months, consider a small interval beginning just before 3 months and ending at 3 months. We designate such an interval as $(3-\delta,3]$. The estimated conditional probability of dying in this small interval is 1/5 and the probability of surviving through it is $1-1/5 = 4/5$. At any specified time point, the number of subjects alive is called the number at risk of dying or simply the number at risk. At time 3 months this number is denoted as n_1, the 1 referring to the fact that 3 months is the first observed time. The number of deaths observed at 3 months was 1 but, with a larger sample, more than one could have been observed. To allow for this, we denote the number of deaths observed as d_1. In this more general notation, the estimated probability of dying in the small interval around 3 is d_1/n_1 and the estimated probability of surviving is $(n_1 - d_1)/n_1$. The probability that a subject survives to 3 months is estimated as the probability of surviving through interval I_0 times the conditional probability of surviving through the small interval around 3. Throughout the discussion of the Kaplan-Meier estimator, the word "conditional" refers to the fact that the probability applies to those who survived to the point or interval under consideration. Since we observed the death at exactly 3 months, this estimated probability would be the same no matter how small a value of δ we use to define the interval around 3 months. Thus, we consider the estimate of the survival probability to be at exactly 3 months. The value of this estimate is

$$\hat{S}(3) = 1.0 \times (4/5) = 0.8.$$

We now consider estimation of the survivorship function at each time point in the remainder of interval I_1. No other failure times (deaths) were observed, hence the estimated conditional probability of survival through small intervals about every time point in the interval is 1.0. Cumulative multiplication of these times the estimated survivorship function leaves it unchanged from its value at 3 months.

The next observed failure time is 5 months. The number at risk is $n_2 = 4$ and the number of deaths is $d_2 = 1$. The estimated conditional probability of surviving through a similarly defined small interval at 5 months, $(5-\delta, 5]$, is $(4-1)/4 = 0.75$. By the same argument used at 3 months, the estimate of the survivorship function at 5 months is the product of the respective estimated conditional probabilities,

$$\hat{S}(5) = 1.0 \times (4/5) \times (3/4) = 0.6.$$

No other failure times were observed in I_2, thus the estimate remains at 0.6 through the interval.

The number at risk at the next observed time, 6 months, is $n_3 = 3$ and the number of deaths is zero since subject 2 was lost to follow-up at 6 months. The estimated conditional probability of survival through a small interval at 6 months is $(3-0)/3 = 1.0$. Again, the estimated survivorship function is obtained by successive multiplication of the estimated conditional probabilities and is

$$\hat{S}(6) = 1.0 \times (4/5) \times (3/4) \times (3/3) = 0.6.$$

No failure times were observed in I_3 and the estimate remains the same until the next observed failure time.

The number at risk 8 months after the beginning of the study is $n_4 = 2$ and the number of deaths is $d_4 = 1$. The estimated conditional probability of survival through a small interval at 8 months is $(2-1)/2 = 0.5$. Hence, by the same argument used at 3, 5 and 6 months, the estimated survivorship function at 8 months after the beginning of the study is

$$\hat{S}(8) = 1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) = 0.3.$$

No other failure times were observed in I_4, thus the estimated survivorship function remains constant and equal to 0.3 throughout the interval.

The last observed failure time was 22 months. There was a single subject at risk and this subject died, hence $n_5 = 1$ and $d_5 = 1$. The estimated conditional probability of surviving through a small interval at 22 months is $(1-1)/1 = 0.0$. The estimated survivorship function at 22 months is

$$\hat{S}(22) = 1.0 \times (4/5) \times (3/4) \times (3/3) \times (1/2) \times (0/1) = 0.0.$$

No subjects were alive after 22 months; thus the estimated survivorship function is equal to zero after that point.

Through this example, we have demonstrated the essential features of the Kaplan-Meier estimator of the survivorship function. The estimator at any point in time is obtained by multiplying a sequence of conditional survival probability estimators. Each conditional probability estimator is obtained from the observed number at risk of dying and the observed number of deaths and is equal to " $(n-d)/n$." This estimator allows each subject to contribute information to the calculations as long as they are known to be alive. Subjects who die contribute to the number at risk up until their time of death, at which point they also contribute to the number of deaths. Subjects who are censored contribute to the number at risk until they are lost to follow-up.

The estimate obtained from the data in Table 2.1 is presented in tabular form in Table 2.2. Computer software packages often present an abbreviated version of this table containing only the observed failure times and estimates of the survivorship function at these times with the implicit understanding that it is constant between failure times.

A graph is an effective way to display an estimate of a survivorship function. The graph shown in Figure 2.1 is obtained from the survivorship function in Table 2.2. The graph shows the decreasing step function defined by the estimated survivorship function. It drops at the values of the observed failure times and is constant between observed failure times. An embellishment provided by some software packages, but rarely presented in published articles, is an indicator on the graph where censored observations occurred. The censored time of 6 months appears as a small × in the figure.

Table 2.2 Estimated Survivorship Function Computed from the Survival Times for the Five Subjects from the HMO-HIV+ Study Shown in Table 2.1

Interval	$\hat{S}(t)$
$0 \le t < 3$	1.0
$3 \le t < 5$	0.8
$5 \le t < 6$	0.6
$6 \le t < 8$	0.6
$8 \le t < 22$	0.3
$t \ge 22$	0.0

In our example, no two subjects shared an observation time, and the longest observed time was a failure. Simple modifications to the method described above are required when either of these conditions is not met. Consider a case where a failure and a censored observation have the same recorded value. We assume that, since the censored observation was known to be alive when last seen, its survival time is longer than the recorded time. Thus a censored subject contributes to the number at risk at the recorded time but is not among those at risk immediately after that time. Along the same lines, suppose we have multi-

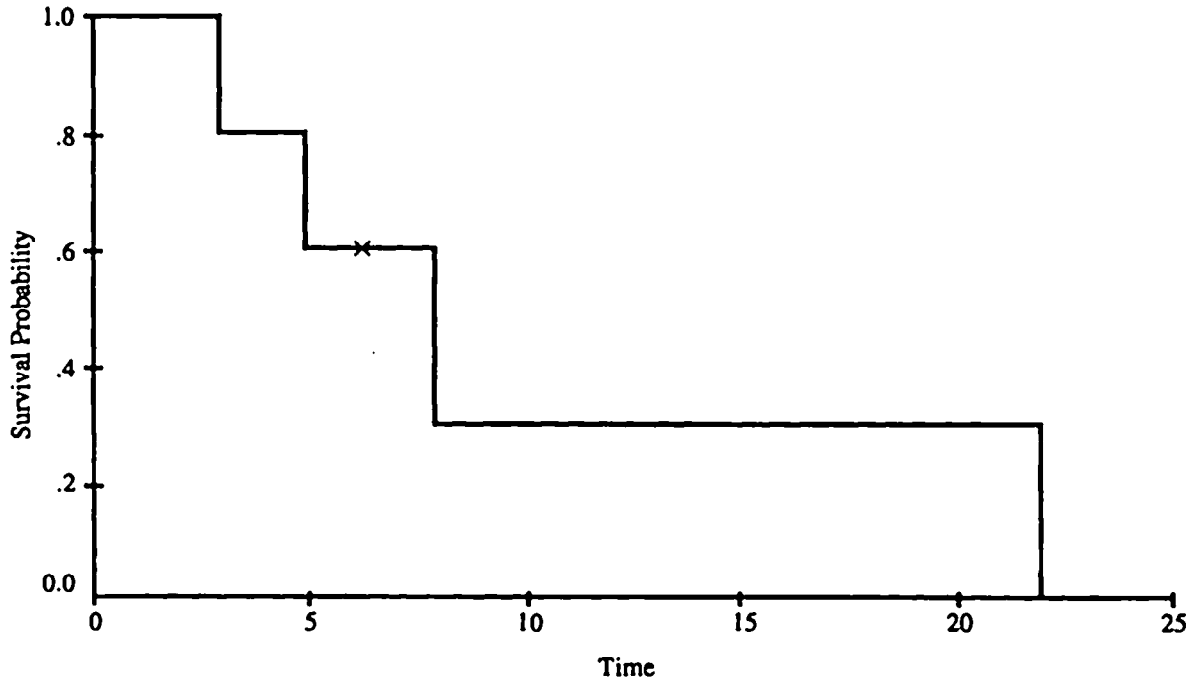


Figure 2.1 Graph of the estimated survivorship function from Table 2.2.

ple failures, $d > 1$, at some time t. It is unlikely that each subject died at the exact same time t; however, we were unable to record the data with any more accuracy. One way to break these ties artificially would be to order the d tied failure times randomly by subtracting a tiny random value from each. For example, if we had observed three values at 8 months we could subtract from each failure time the value of a uniformly distributed random variable on the interval (0, 0.01). This would artificially order the times, yet not change their respective positions relative to the rest of the observed failure times. We would estimate the survivorship function with $d = 1$ at each of the randomly ordered times. The resulting estimate of the survivorship function at the last of the d times turns out to be identical to that obtained using $(n-d)/n$ as the estimate of the conditional probability of survival for all d considered simultaneously. Thus, it is unnecessary to make adjustments for ties when estimating the survivorship function. However, if there are extensive numbers of tied failure times, then a discrete time model may be a more appropriate model choice (see Chapter 7).

If the last observed time corresponds to a censored observation, then the estimate of the survivorship function does not go to zero. Its smallest value is that estimated at the last observed survival time. In this case the estimate is considered to be undefined beyond the last observed time. If both censored and non-censored values occur at the longest observed time, then the protocol of assuming that censoring takes place after failures dictates that $(n-d)/n$ is used to estimate the conditional survival probability at this time. The estimated survivorship function does not go to zero and is undefined after this point. When these types of ties occur, software packages, which provide a tabular listing of the observed survival times and estimated survivorship function, list the censored observations after the survival time, with the value of the estimated survivorship function at the survival time. Simple examples demonstrating each of these situations are obtained by adding additional subjects to the five shown in Table 2.1.

In order to use the Kaplan–Meier estimator in other contexts, we need a more general formulation. Assume we have a sample of n independent observations denoted (t_i, c_i), $i = 1, 2, \ldots, n$ of the underlying survival time variable T and the censoring indicator variable C.[1] Assume that among the n observations there are $m \leq n$ recorded times of failure.

---

[1] Unless stated otherwise we assume recorded values of time are continuous and subject only to right censoring.

We denote the rank-ordered survival times as $t_{(1)} < t_{(2)} < \cdots < t_{(m)}$. In this text, when quantities are placed in rank order we use the same variable notation but place subscripts in parentheses. Let the number at risk of dying at $t_{(i)}$ be denoted n_i and the observed number of deaths be denoted d_i. The Kaplan–Meier estimator of the survivorship function at time t is obtained from the equation

$$\hat{S}(t) = \prod_{t_{(i)} \le t} \frac{n_i - d_i}{n_i} \tag{2.1}$$

with the convention that

$$\hat{S}(t) = 1 \text{ if } t < t_{(1)}.$$

This formulation differs slightly from that described using the data in Table 2.1 in that intervals defined by censored observations are not considered. We saw in the example that conditional survival probabilities are equal to one at censored observations and that the estimate of the survivorship function is unchanged from the value at the previous survival time. Thus the general formula in (2.1) uses only the points at which the value of the estimator changes.

Figure 2.2 presents the graph of the Kaplan–Meier estimate of the

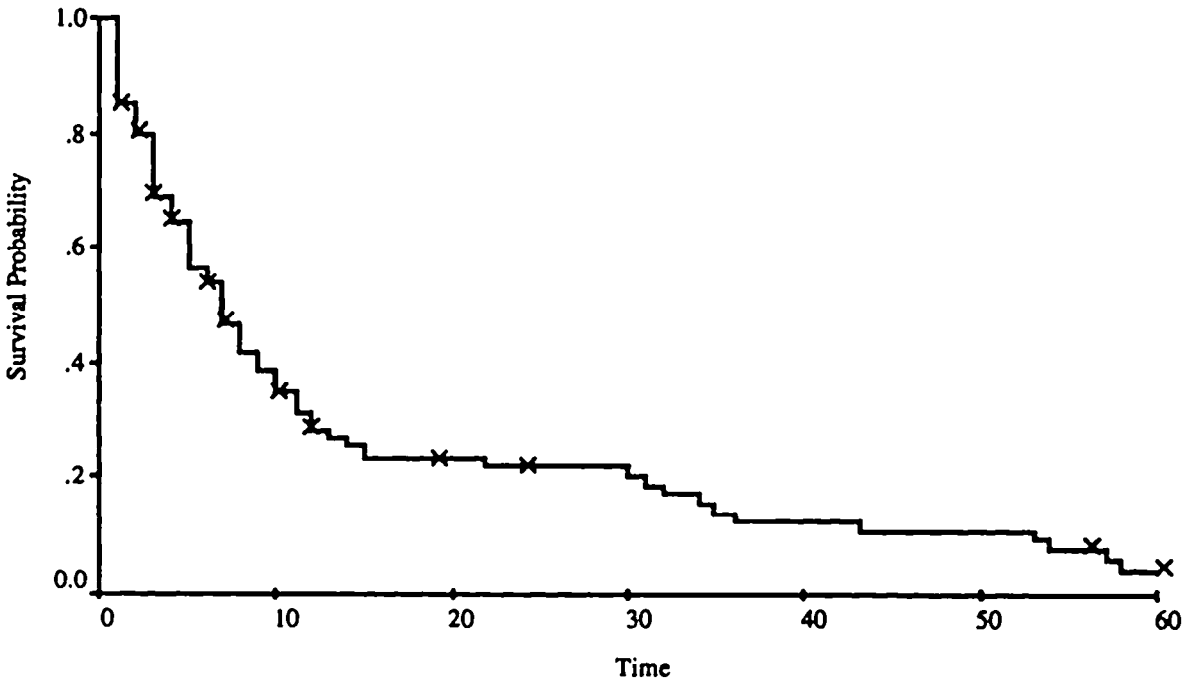


Figure 2.2 Kaplan–Meier estimate of the survivorship function for the HMO-HIV+ study.

survivorship function in (2.1), using all subjects in the HMO-HIV+ study. The construction of the estimate in this case demonstrates conventions for handling tied survival times as well as tied survival and censored times. The data, along with calculations for the beginning and end of the survivorship function, are presented in Table 2.3. The columns in Table 2.3 present the time interval, the number at risk of dying (n), the number of deaths (d), the number of subjects lost to follow-up (c), the estimate of the conditional survival probability $[(n-d)/n]$ and the estimate of the survivorship function $[\hat{S}(t)]$. All quantities are evaluated at the time point defined by the end of the previous interval and the beginning of the current interval.

The first observed survival time is 1 month; thus the value of the estimated survivorship function at each point in the interval [0,1) is 1.0. At 1 month there were 100 subjects at risk. Of these, 15 died and 2 were lost to follow-up (censored), yielding an estimate of the conditional survival probability of $0.85 = (100-15)/100$. The estimate of the survivorship function at 1 month is $0.85 = 1.0 \times 0.85$. The estimate remains at this value at each point in the interval [1,2). At the next observed survival time, 2 months, there were only 83 subjects at risk since 15 died and 2 were lost to follow-up one month before. At 2 months, 5 subjects died and 5 more were lost to follow-up; thus the estimate of the conditional survival probability is $(83-5)/83 = 0.9398$. The estimate of the survivorship function is obtained as the product of the value of the survivorship function just prior to 2 months and the conditional survival probability at 2 months and is $0.85 \times 0.9398 = 0.7988$. The estimate remains at this value throughout the interval [2,4). At the next observed survival time, 4 months, there were 73 subjects at risk, since 5 died and 5 were censored at 2 months. At 4 months, 10 subjects died and 2 were censored. The estimate of the conditional survival probability is

Table 2.3 Partial Calculations of the Kaplan–Meier Estimate Shown in Figure 1.2

Interval	n	d	c	$(n-d)/n$	$\hat{S}$
[0,1)	100	0	0	1.0	1.0
[1,2)	100	15	2	0.85	0.85
[2,4)	83	5	5	0.9398	0.7988
[4,5)	73	10	2	0.8630	0.6894
⋮	⋮	⋮	⋮	⋮	⋮
[58,60)	3	1	0	0.6667	0.0389
[60,60]	2	0	2	1.0	0.0389

$(73-10)/73=0.8630$ and the estimate of the survivorship function is $0.7988\times0.8630=0.6894$. The estimate remains at this value until the next observed survival time, 5 months, at which time 61 subjects are at risk. This process continues, sequentially, considering each observed survival time, until the last observed survival time, which was 58 months. At that time 3 subjects were at risk, 1 died and none were censored. The estimate of the conditional survival probability is $(3-1)/3=0.6667$. The estimate of the survivorship function is $0.0584\times0.6667=0.0389$, where 0.0584 is the value just prior to 58 months. The largest observed time is 60 months, when 2 subjects remained at risk and both were censored. Thus, the estimate of the conditional survival probability is $(2-0)/2=1.0$ and the estimate of the survivorship function remains at the value 0.0389. The function is undefined beyond 60 months, which is denoted in Table 2.3 by recording the last interval as [60,60].

When we have a large study whose mortality experience is presented in calendar time units (such as quarterly, semi-annually, etc.), the life-table estimator of the survivorship function may be used as an alternative to the Kaplan–Meier estimator. The life-table estimator has been used for more than 100 years to describe human mortality experience and is among the earliest examples of the application of statistical methods. It will not play a large role in the analysis of survival data in this text, but we present it because of its historical importance and the fact that it is a grouped-data analog of the Kaplan–Meier estimator. More detail on the various types of life-table estimators may be found in Lee (1992).

In some applied settings the data may be quite extensive with sample sizes in the many hundreds of subjects. In these situations it can be quite cumbersome to tabulate or graph the Kaplan–Meier estimator of the survivorship function. In a sense, the problem faced is similar tc one addressed in a first course on statistical methods: how best to reduce the volume of data but not the statistical information that can be gleaned from it. To this end the histogram is usually introduced as an estimator of the density function and the resulting cumulative percent distribution polygon as an estimator of the cumulative distribution function. This process could be reversed. That is, we might first derive the estimator of the cumulative distribution and, afterwards, compute the histogram as a function of the cumulative distribution. When the data contain censored observations, using the second approach and deriving an estimator of the survivorship function (instead of the cumulative distribution function) is the more feasible tactic. The first step is to define the intervals that will be used to group the data. The goal in the choice of intervals is

the same as for the construction of a histogram—the intervals should be biologically meaningful, yield an adequate description of the data and, if convenient, be of equal width. There are no mechanized rules for construction of the histogram, to guide in the choice of number of intervals. However, the meaningful unit will likely be some multiple of a year.

Once a set of intervals has been chosen, the construction of the estimator follows the basic idea used for the Kaplan–Meier estimator. Suppose we decide to use 6-month intervals. A typical interval will be of the form $[t,t+6)$. As before, let n denote the number of subjects at risk of dying at time t. These subjects are often described as the number who enter the interval alive. As we follow these subjects across the interval, d subjects have survival times and c subjects have censored times in this interval. Thus, not all subjects were at risk of dying for the entire interval. A modification typically employed is to reduce the size of the risk set by one-half of those censored in the interval. The rationale behind this adjustment is that if we assume the censored observations were uniformly distributed over the interval, then the average size of the risk set in the interval is $n-(c/2)$. This average risk set size is used to calculate the estimate of the conditional probability of survival through the interval as $(n-(c/2)-d)/(n-(c/2))$. These estimates of the conditional probabilities are multiplied to obtain the life-table estimator of the survivorship function.

The life-table estimator of the survivorship function for the HMO-HIV+ data using 6-month intervals is shown in Table 2.4. The estimated value of the survivorship function in the first interval is

$$0.5684 = (100-(10/2)-41)/(100-(10/2)).$$

The value in the second interval is computed as

$$0.3171 = 0.5684 \times (49-(3/2)-21)/(49-(3/2)).$$

The remaining values are calculated in a similar fashion.

When we graph the estimate, we have to decide how to represent the actual values. Consider the first interval [0, 6), where the value of the estimated survivorship function is reported in Table 2.4 as 0.5684. If, as in Figure 2.3, we were to represent the graph as a step function, then this interval would be represented by a horizontal straight line of height

Table 2.4 Life-Table Estimator of the Survivorship Function for the HMO-HIV+ Study

Interval	Enter	Die	Censored	$\hat{S}$
[0 , 6)	100	41	10	0.5684
[6 , 12)	49	21	3	0.3171
[12 , 18)	25	6	2	0.2378
[18 , 24)	17	1	1	0.2234
[24 , 30)	15	0	1	0.2234
[30 , 36)	14	5	0	0.1436
[36 , 42)	9	1	0	0.1277
[42 , 48)	8	1	0	0.1117
[48 , 54)	7	1	0	0.0958
[54 , 60)	6	3	1	0.0435
[60 , 66)	2	0	2	0.0435

1 until 6 months when it would drop to 0.5684. Other intervals would be represented in a similar manner. An alternative representation, used by some software packages, is a polygon connecting the value of the estimator drawn at the end of the interval. The first interval would be

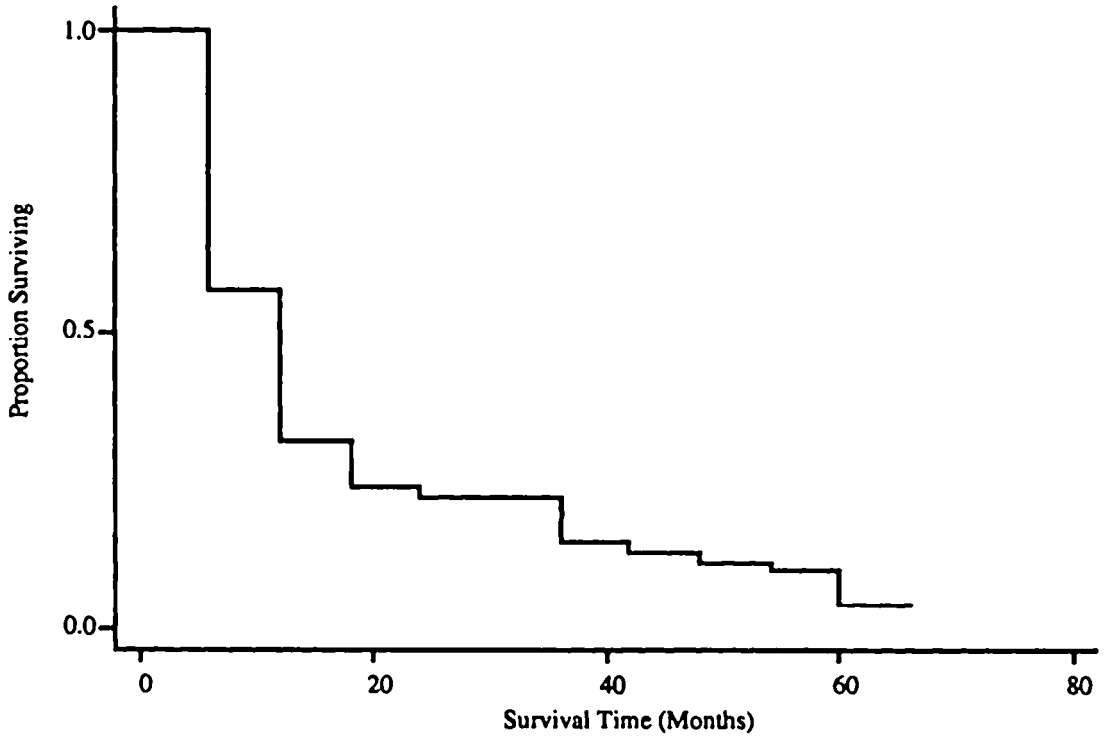


Figure 2.3 Step function representation of life-table estimate of the survivorship function for the HMO-HIV+ study in Table 2.4.

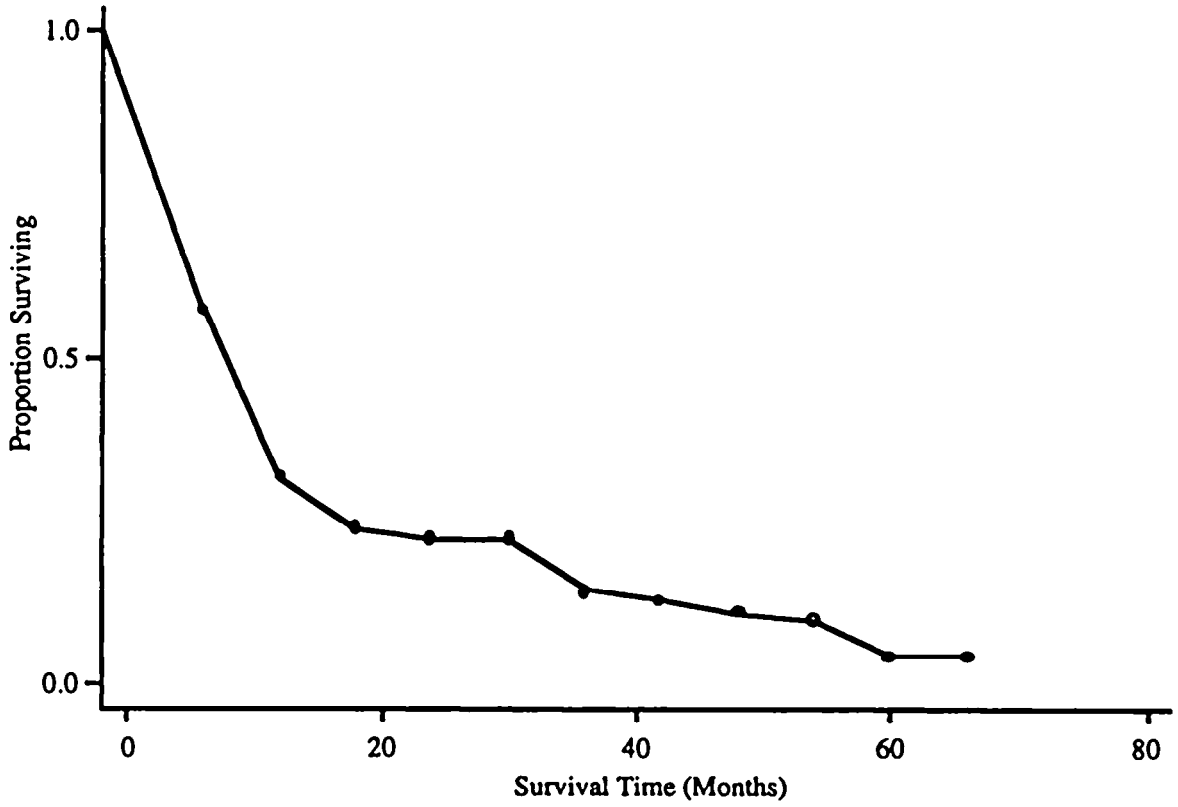


Figure 2.4 Polygon representation of life-table estimate of the survivorship function for the HMO-HIV+ study in Table 2.4.

represented by a point of height 0.5684 plotted at 6 months, the second by a point of height 0.3171 at 12 months, the third by a point of height 0.2378 at 18 months, and so on. These points are then connected by straight lines. The rationale for using the polygon is to better represent the assumed underlying continuous distribution of survival time. Some, but not all, programs will plot a point equal to 1.0 at time zero since, by definition, that is the value of the survivorship function at zero. This point is then connected to the point representing the first interval. The polygonal representation of the life-table estimator from Table 2.4 is shown in Figure 2.4.

Because the graph in Figure 2.4 has been drawn as a polygon, it looks smoother than the step function of the Kaplan–Meier estimator. The life-table estimate in Figure 2.3 in this example does a reasonable job of estimating the survivorship function. Since it is a grouped-data statistic, it is not as precise an estimate as the Kaplan–Meier estimator, which uses the individual values. Later in this chapter we discuss estimation of percentiles of the survival time distribution and these use the Kaplan–Meier estimator.

2.3 USING THE ESTIMATED SURVIVORSHIP FUNCTION

In Section 2.2 we described in detail how to calculate the Kaplan–Meier and life-table estimators of the survivorship function with little if any discussion of how to interpret the resulting estimate or how it may be used to derive point estimates of quantiles of the distribution. One of the biggest challenges in survival analysis is becoming accustomed to using the survivorship function as a descriptive statistic. This function describes the complement of what we typically describe in a set of data. The change from thinking about the percentage of observations less than a value to thinking about the percentage greater than that value, like many things, becomes easier with practice.

The survivorship function estimate shown in Figure 2.2 descends sharply at first and then tails off gradually, reaching its minimum value of 0.04 at 60 months. The initial steep descent shows that there were many subjects who died shortly after enrollment in the study. The relatively long right tail is a result of the few subjects who had long survival times. The minimum value of the survivorship function is not zero since the largest observed time was a censored observation. The shape of the curve depends on the observed survival times and the proportion of observations that are censored. If many subjects in the HMO-HIV+ study had long survival times with the same pattern of censored observations, then the curve would descend slowly at first and then more rapidly until the minimum is reached. If the survival times were more evenly distributed over the 60 months, then the curve would descend gradually to its minimum value. The pattern of enrollment in a follow-up study can influence the shape of the curve. A study with a 2-year enrollment period and 5 years overall length with many late entries is likely to have more censored observations and thus a different looking estimated survivorship function than the same study with many early entries. Many factors influence the shape of the survivorship function, and thus it is difficult to make accurate statements about what a "typical" survivorship function will look like.

In most, if not all, applied settings we will need a confidence interval estimate for the survivorship function as well as point and confidence interval estimates of various quantiles of the survival time distribution. We begin by discussing confidence interval estimation of the survivorship function.

Several different approaches may be taken when deriving an estimator for the variance of the Kaplan–Meier estimator. We derive it

from a technique which is referred to as the *delta method* and is based on a first-order Taylor series expansion. This method is presented in general terms in Appendix 1. The Kaplan–Meier estimator at any time t may be viewed as a product of proportions. Rather than derive a variance estimator of this product, we derive one for its log since the variance of a sum is simpler to calculate than variance of a product. The log of the Kaplan–Meier estimator is

$$\ln\left(\hat{S}(t)\right) = \sum_{t_{(i)} \le t} \ln\left(\frac{n_i - d_i}{n_i}\right) = \sum_{t_{(i)} \le t} \ln\left(\hat{p}_i\right),$$

where

$$\hat{p}_i = (n_i - d_i)/n_i .$$

If we consider the observations in the risk set at time $t_{(i)}$ to be independent Bernoulli observations with constant probability, then $\hat{p}_i$ is an estimator of this probability and an estimator of its variance is $\left(\hat{p}_i(1-\hat{p}_i)\right)/n_i$. As shown in Appendix 1, the variance of the log of variable X is approximately:

$$\mathrm{Var}[\ln(X)] \cong \frac{1}{\mu_X^2}\sigma_X^2, \tag{2.2}$$

where the mean and variance of X are denoted μ_X and σ_X^2, respectively. An estimator for the variance is obtained by replacing μ_X and σ_X^2 in (2.2) with estimators of their respective values. Applying this result to $\ln\left(\hat{p}_i\right)$ yields the estimator

$$\hat{\mathrm{Var}}[\ln(\hat{p}_i)] \cong \frac{1}{\hat{p}_i^2}\frac{\hat{p}_i(1-\hat{p}_i)}{n_i} \cong \frac{d_i}{n_i(n_i - d_i)} .$$

If we assume that observations at each time are independent, then the estimator of the variance of the log of the survivorship function is

$$\hat{\mathrm{Var}}\left[\ln\left(\hat{S}(t)\right)\right] = \sum_{t_{(i)} \le t} \hat{\mathrm{Var}}\left[\ln(\hat{p}_i)\right]$$
$$= \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)} . \qquad (2.3)$$

An estimator of the variance of the survivorship function is obtained by another application of the delta method shown in Appendix 1. This time an approximation is applied to find the variance of an exponentiated variable and is

$$\mathrm{Var}(e^X) \cong \left(e^{\mu_X}\right)^2 \sigma_X^2. \qquad (2.4)$$

Using the fact that $\hat{S}(t) = e^{\ln\left(\hat{S}(t)\right)}$, we let X stand for $\ln\left(\hat{S}(t)\right)$, σ_X^2 stand for the variance estimator in (2.3) and approximate μ_X by $\ln\left(\hat{S}(t)\right)$ in expression (2.4). Then we obtain Greenwood's formula [Greenwood (1926)] for the variance of the survivorship function:

$$\hat{\mathrm{Var}}\left(\hat{S}(t)\right) = \left(\hat{S}(t)\right)^2 \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)}. \qquad (2.5)$$

The method shown to derive the estimator in (2.5) is, in some sense, the "traditional" approach in that it may be found in most texts on survival analysis published prior to 1990. In contrast, the texts by Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993) consolidate a large number of results derived from applications of theory based on counting processes and martingales. This theory is well beyond the scope of this text, but we mention it here as it has allowed development of many useful tools and techniques for the study of survival time. The current thrust in the development of software is based on the counting process paradigm as its methods and tools may be used to analyze, in a relatively uncomplicated manner, some rather complex problems. The estimator in (2.5) may also be obtained from the counting process approach.

The counting process approach to the analysis of survival time plays a central role in many of the methods discussed in this text. A brief presentation of the central ideas behind the counting process formulation of survival analysis is given in Appendix 2. We will use results

from this theory to provide justification for estimators, confidence interval estimators and hypothesis testing methods.

After obtaining the estimated survivorship function, we may wish to obtain pointwise confidence interval estimates. The counting process theory has been used to prove that the Kaplan–Meier estimator and functions of it are asymptotically normally distributed [Andersen, Borgan, Gill and Keiding (1993, Chapter IV) or Fleming and Harrington (1991, Chapter 6)]. Thus, we may obtain pointwise confidence interval estimates for functions of the survivorship function by adding and subtracting the product of the estimated standard error times a quantile of the standard normal distribution. We could apply this theory directly to the Kaplan–Meier estimator using the variance estimator in (2.5). However, this approach could easily lead to confidence interval endpoints that are less than zero or greater than one. In addition, the assumption of normality implicit in the use of the procedure may not hold for the small to moderate sample sizes often seen in typical problems. To address these problems, Kalbfleisch and Prentice (1980, page 15) suggest that confidence interval estimation should be based on the function

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right],$$

called the *log-log survivorship function*. One advantage of this function over the survivorship function is that its possible range is from minus to plus infinity. The expression for the variance of the log-log survivorship function is obtained from a second application of the delta method for a log transformed variable shown in (2.2). The estimator of the variance of the log-log survivorship function is

$$\hat{\mathrm{Var}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\} = \frac{1}{\left[\ln\left(\hat{S}(t)\right)\right]^2}\sum_{t_{(i)}\le t}\frac{d_i}{n_i(n_i-d_i)}. \tag{2.6}$$

The endpoints of a $100(1-\alpha)$ percent confidence interval for the log-log survivorship function are given by the expression

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm z_{1-\alpha/2}\hat{\mathrm{SE}}\left\{\ln\left[-\ln\left(\hat{S}(t)\right)\right]\right\}, \tag{2.7}$$

where $z_{1-\alpha/2}$ is the upper $\alpha/2$ percentile of the standard normal distribution and $\hat{\mathrm{SE}}(\cdot)$ represents the estimated standard error of the argu-

ment, which in this case is the positive square root of (2.6). If we denote the lower and upper endpoints of this confidence interval as $\hat{c}_l$ and $\hat{c}_u$, it follows that the lower and upper endpoints of the confidence interval for the survivorship function are

$$\exp\left[-\exp(\hat{c}_u)\right] \text{ and } \exp\left[-\exp(\hat{c}_l)\right], \tag{2.8}$$

respectively. That is, the lower endpoint from (2.7) yields the upper endpoint in (2.8). These are the endpoints reported by most, if not all, software packages for each observed value of survival time. The confidence interval is valid only for values of time over which the Kaplan–Meier estimator is defined, which is basically the observed range of survival times. Borgan and Leistøl (1990) studied this confidence interval and found that it performed well for sample sizes as small as 25 with up to 50 percent right-censored observations.

Figure 2.5 presents the Kaplan–Meier estimator of the survivorship function for the HMO-HIV+ study and the upper and lower pointwise 95 percent confidence bands computed using (2.8). The endpoints of the pointwise confidence intervals are connected to form a "confidence band." (Recall that any time one has a collection of individual 95 percent confidence interval estimates, the probability that they all contain their respective parameters is much less than 95 percent.) An alternative presentation used by some software packages connects the endpoints of the confidence intervals with vertical lines. This is useful for small data sets, but for large data sets the resulting graph becomes cluttered with too many lines, and we lose the visual conciseness seen in Figure 2.5. This figure demonstrates some of the properties of the log-log-based confidence interval estimator. The intervals are skewed for large and, though harder to see in Figure 2.5, small values of the estimated survivorship function and are fairly symmetric around 0.5. The direction of skewness is opposite for the two tails, toward zero for values of the estimated survivorship function near one and toward one for values near zero. In all cases, the endpoints lie between zero and one. In Figure 2.5, the confidence intervals further support the observation of a survivorship function describing many early deaths with a few deaths near the maximum of 5 years of follow-up.

Simultaneous confidence bands for the entire survivorship function are not as readily available as the pointwise estimates, since they require percentiles for statistical distributions not typically computed by software packages. The band proposed by Hall and Wellner (1980) is discussed in some detail in Andersen, Borgan, Gill and Keiding (1993) and

Fleming and Harrington (1991). It is also discussed in Marubini and Valsecchi (1995). A table of percentiles obtained from Hall and Wellner (1980) is provided in Appendix 3. Given the tabled percentiles, confidence bands based on the estimated survivorship function itself, or its log-log transformation, are not difficult to calculate. Borgan and Leistøl (1990) show that the performance of the Hall and Wellner confidence bands is comparable for both functions and is adequate for samples as small as 25 with up to 50 percent censoring. To maintain consistency with the pointwise intervals calculated in (2.8), which are based on the log-log transformation, we present the Hall and Wellner bands for the transformed function. Hall and Wellner, as well as Borgan and Leistøl, recommend that these confidence bands be restricted to values of time smaller than or equal to the largest observed survival time, e.g., the largest non-censored value of time denoted $t_{(m)}$. The endpoints of the $100(1-\alpha)$ percent confidence bands in the interval $[0, t_{(m)}]$ for the log-log transformation are

$$\ln\left[-\ln\left(\hat{S}(t)\right)\right] \pm H_{\hat{a},\alpha} \frac{\left(1+n\hat{\sigma}^2(t)\right)}{\sqrt{n}\left|\ln\left(\hat{S}(t)\right)\right|}, \tag{2.9}$$

where

$$\hat{\sigma}^2(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i(n_i - d_i)},$$

the estimator of the variance of the log of the Kaplan–Meier estimator from (2.3), and $H_{\hat{a},\alpha}$ is a percentile from Appendix 3, where

$$\hat{a} = n\hat{\sigma}^2(t_{(m)}) \Big/ \left[1 + n\hat{\sigma}^2(t_{(m)})\right].$$

If we denote the lower and upper endpoints of this confidence band as $\hat{b}_l$ and $\hat{b}_u$, then the lower and upper endpoints of the confidence band for the survivorship function are

$$\exp\left[-\exp\left(\hat{b}_u\right)\right] \text{ and } \exp\left[-\exp\left(\hat{b}_l\right)\right]. \tag{2.10}$$

To obtain the bands for the survivorship function from the HMO-HIV+ study, we note that the largest observed survival time is 58 months and $\hat{\sigma}^2(58) = 0.423$. Most software packages will provide either the values of the estimated variance of the log of the Kaplan–Meier estimator

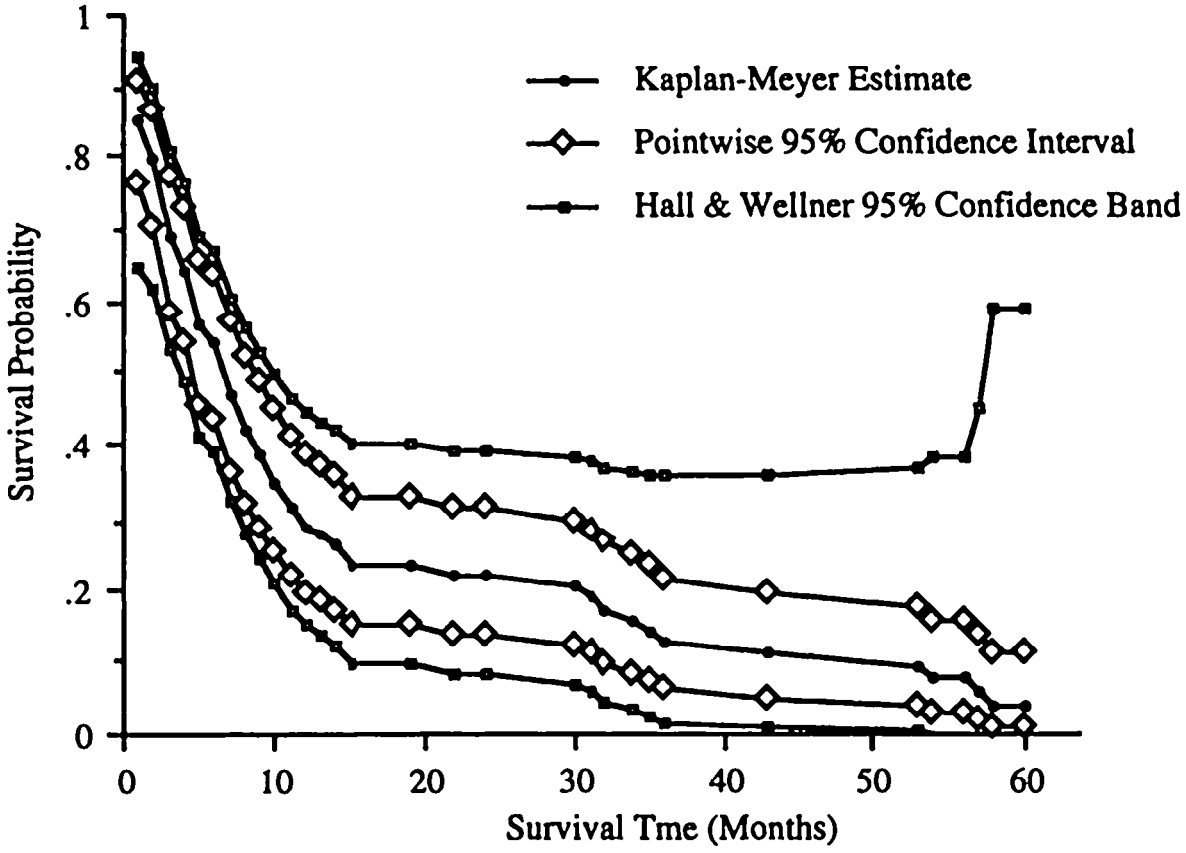


Figure 2.5 Kaplan–Meier estimate, pointwise 95% confidence intervals, and Hall and Wellner 95% confidence bands for the survivorship function for the HMO–HIV+ study.

or those of the Greenwood estimator of the variance of the survivorship function. The values of $\hat{\sigma}^2(t)$ are easily obtained by dividing the Greenwood estimator by the square of the Kaplan–Meier estimator. To obtain the percentile from Appendix 3 we compute

$$\hat{a} = (100 \times 0.423)/(1 + 100 \times 0.423) = 0.98$$

and note that, since both $H_{0.9,0.95}$ and $H_{1.0,0.95}$ equal 1.358, linear interpolation of tabled values is not necessary and we use 1.358. In cases when $\hat{a} < 0.9$, linear interpolation between two tabled values may be required to obtain the most accurate value. To obtain the confidence bands, we compute the endpoints in (2.9) and (2.10) for each observed value of time. We can ignore the censoring since the estimated survivorship function and its variance are constant between observed failure times. These endpoints may be plotted, along with the estimated survivorship function, restricting the plot to the interval [0,58]. This plot is

also shown in Figure 2.5. The increased width of the confidence bands relative to the pointwise confidence intervals is seen in this figure. The increased width is needed to assure that the probability is 95 percent that each of the individual 95 percent confidence interval estimates simultaneously covers its respective parameter. In particular, we note the lack of precision in the band for times near the maximum of 58 months. The bands do support the observation of many early deaths and a few at or near the maximum follow-up time of 60 months.

The estimated survivorship function and its confidence intervals and/or bands provide a useful descriptive measure of the overall pattern of survival times. However, it is often useful to supplement the presentation with point and interval estimates of key quantiles. The estimated survivorship function may be used to estimate quantiles of the survival time distribution in the same way that the estimated cumulative distribution of, say, height or weight may be used to estimate quantiles of its distribution. This may be done graphically and the graphical procedure can be codified into a formula for analytic calculations based on the tabular form of the estimate.

The quantiles most frequently reported by software packages are the three quartile boundaries of the survival time distribution. To obtain graphical estimates, begin on the percent survival (y) axis at the quartile of interest and draw a horizontal line until it first touches the estimated survivorship function. A vertical line is drawn down to the time axis to obtain the estimated quartile. In order for the estimate to be finite, the horizontal line must hit the survivorship function. Thus, the minimum possible estimated quantile which has a finite value is the observed minimum of the survivorship function, and only quantiles within the observed range of the estimated survivorship function may be estimated. For example, if the range was from 1.0 to 0.38 then we could estimate the 75th and 50th percentiles but not the 25th percentile. Graphically determined estimates of the three quartile boundaries, denoted $\hat{t}_{75}$, $\hat{t}_{50}$ and $\hat{t}_{25}$, based on the Kaplan–Meier estimate of the survivorship function for the data in Table 2.1 are shown in Figure 2.6.

The graphical method is easy to use, but it is not especially precise. The method may be described in a formula, from which a more accurate numerical value may be determined from a tabular presentation of the estimated survivorship function. We illustrate the method by estimating the median or second quartile, $\hat{t}_{50}$, and we then generalize it into a formula that may be used for any quantile. By referring to Table 2.2, and Figure 2.6 we see that the horizontal line hits the survivorship funct-

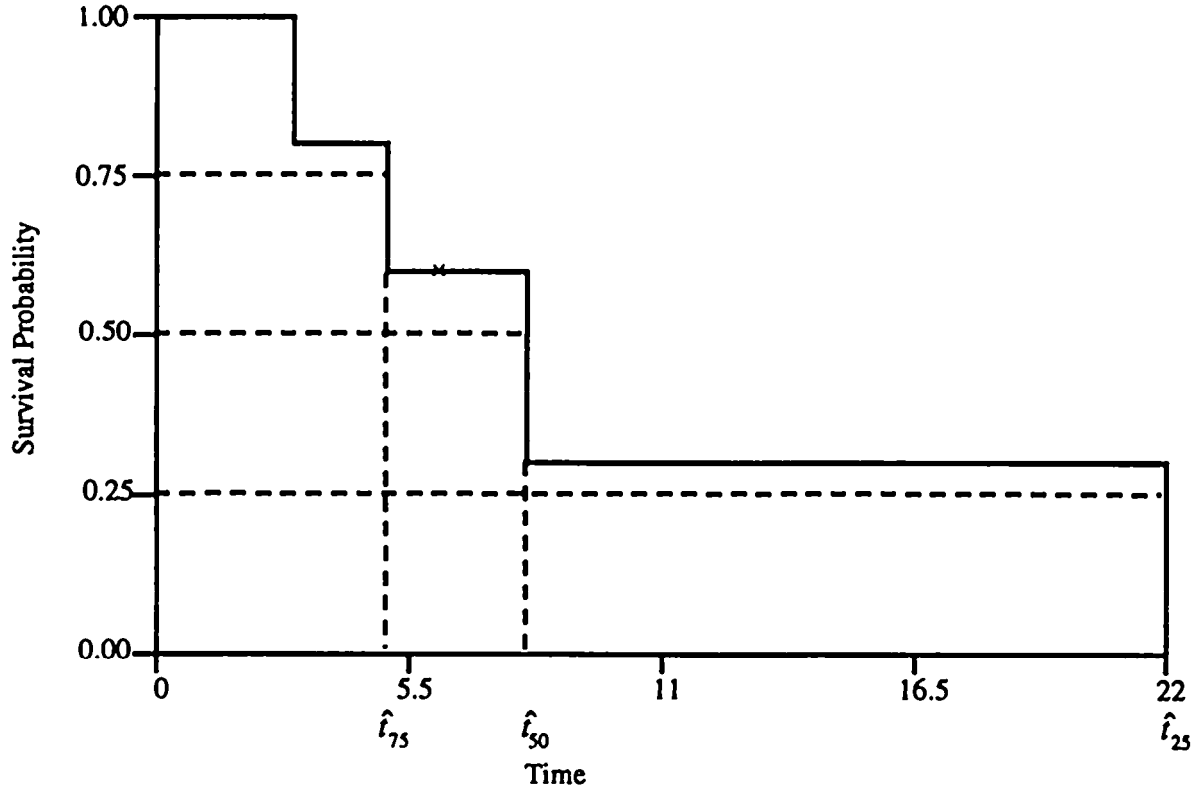


Figure 2.6 Kaplan–Meier estimate of the survivorship function for the data in Table 2.1 and graphically determined estimates of the quartiles.

ion at the riser connecting steps ending, respectively (looking right to left), at 8 and 5 months. The vertical line hits at exactly 8 months. Thus the estimated median survival time in this example is $\hat{t}_{50} = 8$. A formula to describe this estimator is

$$\hat{t}_{50} = \min\{t : \hat{S}(t) \leq 0.50\}.$$

The formula says to proceed as if you are walking up a set of stairs from the right to the riser where the horizontal line hits. The estimate is the time value defining the left-most point of the step you're standing on. If we assume that the riser is attached at the top and bottom, then the description also works when the horizontal line hits one of the steps. The estimate is, again, the value of time defining the left-most point of the step. In general, the estimate of the pth percentile is

$$\hat{t}_p = \min\{t : \hat{S}(t) \leq (p/100)\}.$$

The estimates of the other quartiles from Table 2.2 are $\hat{t}_{75} = 5$ and $\hat{t}_{25} = 22$.

For the full data set for the HMO-HIV+ study, the estimates of the three quartiles are $\hat{t}_{75}=3$, $\hat{t}_{50}=7$ and $\hat{t}_{25}=15$. The interpretation of these values is that we estimate that 75 percent will live at least three months, half are estimated to live at least 7 months, and only 25 percent are estimated to live at least 15 months.

A confidence interval estimate for the quantiles can add further understanding about possible values for the parameter being estimated. Approximate confidence intervals may be obtained by appealing to the theory that, for large samples, the quantile estimator is normally distributed with mean equal to the quantile being estimated. An estimator of the variance of this distribution may be obtained from an application of the delta method, as outlined in Collet (1994) and discussed in greater detail from the counting process approach in Andersen, Borgan, Gill and Keiding (1993). The suggested estimator for the variance of the estimator of the pth percentile is

$$\hat{\mathrm{Var}}\left(\hat{t}_p\right)=\frac{\hat{\mathrm{Var}}\left(\hat{S}(\hat{t}_p)\right)}{\left[\hat{f}(\hat{t}_p)\right]^2}. \tag{2.11}$$

The numerator of (2.11) is Greenwood's estimator and the denominator is an estimator of the density function of the distribution of survival time. The estimator of the density function used by many software packages is

$$\hat{f}(\hat{t}_p)=\frac{\hat{S}(\hat{u}_p)-\hat{S}(\hat{l}_p)}{\hat{l}_p-\hat{u}_p}. \tag{2.12}$$

The values $\hat{u}_p$ and $\hat{l}_p$ are chosen such that $\hat{u}_p<\hat{t}_p<\hat{l}_p$ and most often are obtained from the equations shown below:

$$\hat{u}_p=\max\left\{t:\hat{S}(t)\geq(p/100)+0.05\right\} \text{ and } \hat{l}_p=\min\left\{t:\hat{S}(t)\leq(p/100)-0.05\right\}. \tag{2.13}$$

While values other than 0.05 could have been used in (2.13), 0.05 seems to work well in practice and is used by a number of statistical packages. The endpoints of a $100(1-\alpha)$ percent confidence interval are

$$\hat{t}_p \pm z_{1-\alpha/2}\hat{\mathrm{SE}}(\hat{t}_p), \tag{2.14}$$

where $\hat{\mathrm{SE}}(\hat{t}_p) = \sqrt{\hat{\mathrm{Var}}(\hat{t}_p)}$.

Evaluation of (2.11) through (2.14) is most easily illustrated with an example. In the HMO-HIV+ study, the estimated median survival time is $\hat{t}_{50} = 7$ months. The value of $\hat{u}_{50}$ is the largest value of time, t, such that $\hat{S}(t) \geq 0.55$. After sorting on survival time and listing the values of the Kaplan–Meier estimator, we find that $\hat{S}(5) = 0.56$ and $\hat{S}(6) = 0.54$, hence $\hat{u}_{50} = 5$. The value of $\hat{l}_{50}$ is the smallest value of t, time, such that $\hat{S}(t) \leq 0.45$. From the same listing we find that $\hat{S}(7) = 0.47$ and $\hat{S}(8) = 0.42$, hence $\hat{l}_{50} = 8$. Thus the estimate of the density function in (2.12) is

$$\hat{f}(\hat{t}_{50}) = \frac{\hat{S}(5) - \hat{S}(8)}{8-5} = \frac{0.56 - 0.42}{8-5} = 0.0467.$$

The value of Greenwood's estimator at $t = 7$ months is

$$\hat{\mathrm{Var}}\left(\hat{S}(7)\right) = 0.002672$$

and evaluation of (2.11) yields

$$\hat{\mathrm{Var}}(\hat{t}_{50}) = \frac{0.00267}{[0.0467]^2} = 1.224.$$

The end points of the 95 percent confidence interval for median survival time are

$$7 \pm 1.96 \times \sqrt{1.224} = (4.8,\ 9.2).$$

Table 2.5 Estimated Quartiles, Estimated Standard Errors and 95% Confidence Intervals for Survival Time in the HMO-HIV+ Study

Quantile	Estimate	Std. Err.	95% CIE
75	3	0.59	1.8, 4.2
50	7	1.11	4.8, 9.2
25	15	7.45	1.4, 29.6

Table 2.5 presents the estimated survival times for the quartiles, their estimated standard errors, and 95 percent confidence intervals. The results in Table 2.5 further quantify our previous observation of many early deaths with a few at nearly the maximum of follow-up. We note that the confidence interval is quite wide for the 25th percentile. After 15 months only 17 subjects remained at risk. The lack of precision in the confidence interval estimate for this percentile is due to the smaller number of subjects at risk. In general, the right tail of the survivorship function is estimated with considerably less precision than the left tail.

The confidence interval estimator in (2.14) requires that we compute an estimator of the density function at the estimator of the quantile, and the endpoints depend on the assumption that the distribution of the estimated quantile is normal. The sensitivity of the confidence interval to the choice of estimator of the density and the assumption of normality has not been studied. Brookmeyer and Crowley (1982) proposed an alternative method which does not require estimation of the density function [this is discussed in general terms in Andersen, Borgan, Gill and Keiding (1993)]. In this method, the confidence interval for a quantile consists of the values t such that

$$\frac{\left|\hat{S}(t) - p/100\right|}{\hat{\mathrm{SE}}\left(\hat{S}(t)\right)} \leq z_{1-\alpha/2}.$$

The expression on the left side is a test statistic for the hypothesis $H_o: S(t) = p/100$. The confidence interval is the set of values of t for which we would fail to reject the hypothesis. In other words, it is the set of observed survival times for which the confidence interval estimates for the survivorship function contain the quantile. This interval may be determined graphically in a manner similar to Figure 2.6 by drawing a horizontal line from $p/100$ to where it intersects the step functions defining the upper and lower pointwise confidence intervals. The endpoints of the confidence interval are found by drawing vertical lines down to the time axis. If the software package provides the capability to list the endpoints of the confidence intervals for the estimated survivorship function, then the upper and lower endpoints can be precisely determined. Alternative test statistics based on transformations of the estimated survivorship function, such as the log-log transformation, could be used equally well. Brookmeyer and Crowley recommend that this interval be used when there are no tied survival times. The data from the HMO-HIV+ study contain many tied survival times and thus it

would be inappropriate to use the Brookmeyer–Crowley limits in a definitive analysis. However, these data may be used to illustrate the calculations for the median survival time.

Table 2.6 lists the values of the estimated survivorship function and the endpoints of 95 percent confidence intervals determined by the log-log transformation for survival times around the median value of 7 months.

In Table 2.6, we see that the confidence interval estimate at 4 months does not contain 0.5, while at 5 months it does contain 0.5. Thus the lower endpoint of the Brookmeyer–Crowley interval is 5 months. We see that the confidence interval at 9 months does not contain 0.5, while the interval at 8 months does contain it. Hence, the upper limit is 8 months. Brookmeyer–Crowley limits could be determined in a similar manner for other quantiles, though those for the median are most often calculated and reported by software packages. The Brookmeyer–Crowley confidence interval for the median of (5, 8) is comparable to the interval (4.8, 9.2) from Table 2.5, which was based on the large sample distribution of the estimator of the median.

In the analysis of survival time, the sample mean is not as important a measure of central tendency as it is in other settings. (The exception is in fully parametric modeling of survival times when the estimator of the mean, or a function of it, provides an estimator of a parameter vital to the analysis and interpretation of the data. We discuss parametric modeling in Chapter 8.) This is due to the fact that censored survival time data are most often skewed to the right and, in these situations, the median usually provides a more intuitive measure of central tendency. For the sake of completeness, we describe how the estimator of the mean

Table 2.6 Listing of Observed Survival Times, the Estimated Survivorship Function and Individual 95% Confidence Limits for Values of Time near the Estimated Median Survival Time of 7 months for the HMO-HIV+ Data

Time	Estimate	95% CIE
4	0.64	0.54, 0.66
5	0.56	0.46, 0.66
6	0.54	0.43, 0.64
7	0.47	0.36, 0.57
8	0.42	0.32, 0.52
9	0.39	0.28, 0.49

and the estimator of its variance are calculated and illustrate their use with examples from the HMO-HIV+ study.

Computational questions arise if the largest observation is censored, in which case one has two choices: (1) Use only the observed survival times (in which case the estimator is biased downwards) or (2) use all observations (in which case one "pretends" that the largest observation was actually a survival time, but the estimator is interpreted conditionally on the observed range). There is no uniform agreement on which is the best approach. For example, SAS (PROC LIFETEST) uses the former approach while BMDP (1L) uses the latter approach. In the absence of censoring, both approaches yield the usual arithmetic mean.

The estimator used for the mean is obtained from a mathematical result which states that, for a positive continuous random variable, the mean is equal to the area under the survivorship function. From mathematical methods of calculus this may be represented as the integral of the survivorship function over the range, that is,

$$\mu = \int_0^\infty S(u)\, du.$$

If we restrict the variable to the interval $[0,t^*]$, then the mean of the variable in this interval is

$$\mu(t^*) = \int_0^{t^*} S(u) du.$$

The estimator is obtained by using the Kaplan–Meier estimator of the survivorship function. The reason for restricting the range over which the mean is calculated is that the Kaplan–Meier estimator is undefined beyond the largest value of time. The value of t^* used depends on which of the two previously described approaches is chosen. Recall that the observed ordered survival times are denoted $t_{(i)}$, $i = 1,\ldots,m$. We denote the largest observed value of time in the sample as $t_{(n)}$. The two approaches to calculating the estimator of the mean correspond to defining $t^* = t_{(m)}$, that is, using the interval $[0,t_{(m)}]$, or defining $t^* = t_{(n)}$, i.e., using the interval $[0,t_{(n)}]$. In situations where the largest observed value of time is an observed failure time, the two approaches yield identical estimators.

The value of the estimator is the area under the step function defined by the Kaplan–Meier estimator and the particular interval chosen.

To illustrate the calculation, consider the data in Table 2.1 for which the estimated survivorship function is presented in Table 2.2 and is graphed in Figure 2.1. In this example, the largest observed value of time is 22 months and it represents a survival time. Thus, the value of the estimated mean is the area under the step function shown in Figure 2.1. This area is the sum of the areas of four rectangles defined by the heights of the four steps and the four observed survival times. The actual calculation is performed as follows (refer to Table 2.2):

$$\begin{aligned}\hat{\mu}(22) &= 1.0\times[3-0]+0.8\times[5-3]+0.6\times[8-5]+0.3\times[22-8]\\ &= 10.6.\end{aligned}$$

This is the value which would be reported by both BMDP and SAS.

For sake of illustration, suppose that the value recorded at 22 months was a censored observation. If we use the interval [0,22] (BMDP's method), we would report the estimated mean as $\hat{\mu}(22)=10.6$. If we use the interval [0,8] (SAS's method), then we would report the estimated mean as $\hat{\mu}(8)=6.4$. This is the area of the first three rectangles in Figure 2.1. In this example, the two estimates of the mean are quite different since the largest observation, 22 months, is much larger than the largest observed survival time, 8 months.

The equation defining the estimator based on the observed range of survival times only is

$$\hat{\mu}(t_{(m)}) = \sum_{i=1}^{m} \hat{S}(t_{(i-1)})\left(t_{(i)} - t_{(i-1)}\right), \tag{2.15}$$

where $\hat{S}(t_{(0)})=1.0$ and $t_{(0)}=0.0$. The equation defining the estimator for the entire observed range of data is

$$\hat{\mu}(t_{(n)}) = \hat{\mu}(t_{(m)}) + (1-c_{(n)})\hat{S}(t_{(m)})(t_{(n)} - t_{(m)}), \tag{2.16}$$

where $c_{(n)}$ denotes the censoring status, (0, 1), of this observation. Each term in the summation in (2.15) denotes the calculation of the area of one of the rectangles defined by the Kaplan–Meier estimator and two observed times. Note that the estimators in (2.15) and (2.16) are identical when the largest observation and the largest observed survival time are equal.

We recommend that the estimator based on the entire observed range of the data (2.16) be used since the one based on the observed

range of survival times (2.15) does not use the information on survival available in times larger than the largest survival time. We note that if those observations that are long and censored had actually been observed survival times, then the estimated mean survival time would have been increased substantially. However, there may be situations (e.g., when there is considerable uncertainty in measuring the longest censored time [$t_{(n)}$ in (2.16)], when the estimator based on survival times only is preferred.

The estimator of the variance of the sample mean is neither particularly intuitive nor easy to motivate, so we just provide it and demonstrate the calculation. In the case of no censored data, it reduces to the usual "sample variance divided by the sample size" estimator. Andersen, Borgan, Gill and Keiding (1993) present a mathematical derivation of the estimator of the mean and its variance, as well as results which show that the standard normal distribution may be used to form a confidence interval estimator. The equation defining the estimator of the variance of the sample mean computed using (2.15) is as follows:

$$\hat{\mathrm{Var}}(\hat{\mu}(t_{(m)})) = \frac{n_d}{n_d - 1} \sum_{i=1}^{m-1} \frac{A_i^2 d_i}{n_i(n_i - d_i)}, \tag{2.17}$$

where $n_d = \sum_{i=1}^{m} d_i$ denotes the total number of subjects with an observed survival time and

$$A_i = \sum_{j=i}^{m-1} \hat{S}(t_{(j)})(t_{(j+1)} - t_{(j)}).$$

The estimator of the variance using (2.16) is obtained by "pretending" that the largest observed time is an observed survival time for purposes of the summation in (2.17), but n_d is not changed. An example will help distinguish between the two cases. The data in Table 2.1 yielded an estimated mean $\hat{\mu}(22) = 10.6$. Evaluation of the estimator in (2.17) yields

$$\begin{aligned} \hat{\mathrm{Var}}[\hat{\mu}(22)] &= \frac{4}{4-1}\left[\frac{7.6^2}{5(5-1)} + \frac{6.0^2}{4(4-1)} + \frac{4.2^2}{2(2-1)}\right] \\ &= 19.61 \end{aligned}$$

where

$$7.6 = A_1 = 0.8(5-3) + 0.6(8-5) + 0.3(22-8),$$
$$6.0 = A_2 = 0.6(8-5) + 0.3(22-8)$$

and

$$4.2 = A_3 = 0.3(22-8).$$

Assume for the moment that the largest value, 22 months, is a censored observation and that we use (2.16) to estimate the mean. Then the estimate of the variance is

$$\hat{\text{Var}}[\hat{\mu}(22)] = \frac{3}{3-1}\left[\frac{7.6^2}{5(5-1)} + \frac{6.0^2}{4(4-1)} + \frac{4.2^2}{2(2-1)}\right]$$
$$= 22.06.$$

If we restrict estimation of the mean to observed survival times and estimate the mean using (2.15), then the estimate of the variance obtained by evaluating (2.17) is

$$\hat{\text{Var}}[\hat{\mu}(8)] = \frac{3}{3-1}\left[\frac{3.4^2}{5(5-1)} + \frac{1.8^2}{4(4-1)}\right]$$
$$= 1.27,$$

where

$$3.4 = A_1 = 0.8(5-3) + 0.6(8-5)$$

and

$$1.8 = A_2 = 0.6(8-5).$$

Approximate confidence intervals are obtained using percentiles from the standard normal distribution. Using the data in Table 2.1, the endpoints of a 95 percent confidence interval are $10.6 \pm 1.96\sqrt{19.61}$. This is shown only for purposes of illustration since the sample size is only five with four survival times and any asymptotic theory will not hold. In practice, the estimated mean and its estimated standard error would typically be included in the table containing the estimates of the key quantiles and their estimated standard errors.

For the whole HMO-HIV+ study the estimate of the mean using all of the observed times is $\hat{\mu}(60) = 14.67$ and the estimated variance from (2.17) is 3.93, yielding a 95 percent confidence interval of (10.78,

18.56). We note that, in this example, the largest survival time was 58 months and $\hat{\mu}(58) = 14.59$. Thus, the means from the two approaches are not too different. The right skewness evident in the plot of the survivorship function shown in Figure 2.2 is further quantified by the difference between the estimate of the median (7 months) and the estimate of the mean (approximately 15 months). In these data, as is the case with most analyses of survival time, the median is the better measure of central tendency.

2.4 COMPARISON OF SURVIVORSHIP FUNCTIONS

After providing a description of the overall survival experience in the study, we usually turn our attention to a comparison of the survivorship experience in key subgroups in the data. These groups might be defined by treatment arms in a clinical trial or by other key factors thought to be related to survival. The goals in this analysis are identical to those of the two sample *t*-test, the nonparametric rank sum test and the one-way analysis of variance. Namely, we wish to quantify differences between groups through point and interval estimates of key measures. Standard statistical procedures, such as those named above, may be used without modification when there are no censored observations.

Since survival data are typically right skewed, we would likely use rank-based non-parametric tests followed by estimates and confidence intervals of medians (and possibly other quantiles) within groups. Modifications of these procedures are required when censored observations are present in the data. These tests are described and illustrated with the HMO-HIV+ study data beginning with methods for comparing two groups.

When comparing groups of subjects, it is always a good idea to begin with a graphical display of the data in each group. In studies of survival time, we should graph the Kaplan–Meier estimator of the survivorship function for each of the groups. In the HMO-HIV+ study, a variable thought to be related to the survival experience of the subjects was a history of IV drug use, coded 0 = No and 1 = Yes. Figure 2.7 presents the graphs of the estimated survivorship functions for these two groups of subjects.

Both groups show a similar pattern of survival: a rapidly descending survivorship function with a long right tail. This is the result of a number of early deaths and a few subjects with survival near the maximum follow-up time. Since the estimated survivorship functions do not go to

zero, we know that the largest observation in each group was a censored value. The figure also shows a separation of the functions for the two groups. The estimated survivorship function for the non-IV drug users lies completely above that for the IV drug users. In general, the pattern of one survivorship function lying above another means the group defined by the upper curve lived longer, or had a more favorable survival experience, than the group defined by the lower curve. In other words, at any point in time the proportion of subjects estimated to be alive is greater for one group (represented by the upper curve) than the other (represented by the lower curve). Estimates of the within-group statistics such as the median are computed using the methods described in Section 2.3. The statistical question is whether the observed difference seen in Figure 2.7 is significant.

A number of statistical tests have been proposed to answer this question, and most software packages provide results from at least two of these tests. However, comparison of the results obtained by different packages can become confusing due to small but annoying differences in terminology and methods used to calculate the tests. The original developers [Mantel (1966), Peto and Peto (1972), Gehan (1965), Breslow (1970), Prentice (1978)] of these tests sought ways to extend tests used with non-censored data to the censored data setting.

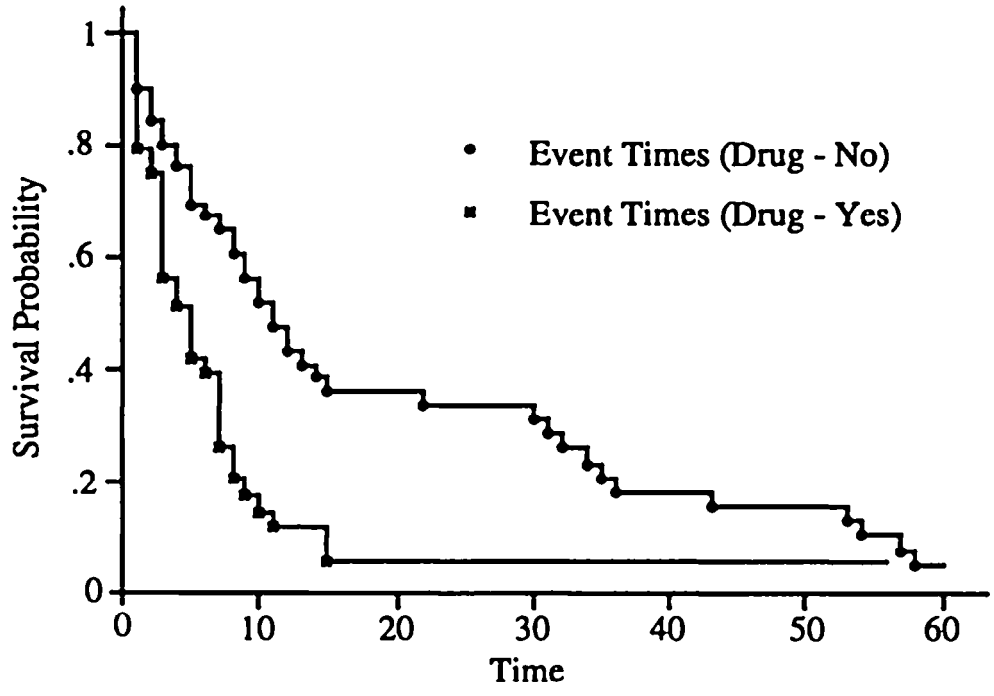


Figure 2.7 Estimated survivorship functions for subjects with and without a history of IV drug use.

The derivation and algebraic representation of the tests can, at times, seem complex and confusing. Lawless (1982) presents a concise summary of the traditional approach to the development of these tests, based on the theory of nonparametric tests, using exponentially ordered scores. However, in recent years, these tests have been reexamined from the counting process point of view and have been shown to be special cases of a more general class of counting process based tests. These results are summarized in Andersen, Borgan, Gill and Keiding (1993).

The calculation of each test is based on a contingency table of group by status at each observed survival time, as shown in Table 2.7. In this table, the number at risk at observed survival time $t_{(i)}$ is denoted by n_{0i} in Group 0 and by n_{1i} in group 1; the number of observed deaths in each of the these two groups is denoted by d_{0i} and d_{1i}, respectively; the total number at risk is denoted by n_i; and the total number of deaths is denoted by d_i. The contribution to the test statistic at each time is obtained by calculating the expected number of deaths in group 1 or 0, assuming that the survivorship function is the same in each of the two groups. This yields the usual *row total times column total divided by grand total* estimator. For example, using group 1, the estimator is

$$\hat{e}_{1i} = \frac{n_{1i} d_i}{n_i}. \tag{2.18}$$

Most software packages base their estimator of the variance of d_{1i} on the hypergeometric distribution, defined as follows:

$$\hat{v}_{1i} = \frac{n_{1i} n_{0i} d_i (n_i - d_i)}{n_i^2 (n_i - 1)}. \tag{2.19}$$

Table 2.7 Table Used for Test of Equality of the Survivorship Function in Two Groups at Observed Survival Time $t_{(i)}$

Event/Group	1	0	Total
Die	d_{1i}	d_{0i}	d_i
Not Die	$n_{1i} - d_{1i}$	$n_{0i} - d_{0i}$	$n_i - d_i$
At Risk	n_{1i}	n_{0i}	n_i

The contribution to the test statistic depends on which of the various tests is used, but each may be expressed in the form of a ratio of weighted sums over the observed survival times. These tests may be defined in general as follows:

$$Q = \frac{\left[\sum_{i=1}^{m} w_i(d_{1i} - \hat{e}_{1i})\right]^2}{\sum_{i=1}^{m} w_i^2 \hat{v}_{1i}}. \tag{2.20}$$

Under the null hypothesis that the two survivorship functions are the same, and assuming that the censoring experience is independent of group, and that the total number of observed events and the sum of the expected number of events is large, then the significance level for Q may be obtained using the chi-square distribution with one degree-of-freedom [i.e., $p = \Pr(\chi^2(1) \geq Q)$]. Exact methods of inference for use with small samples have been implemented in the software package StatXact 3 (1995) but will not be discussed in this text.

The most frequently used test is based on weights equal to one, $w_i = 1$. In this case, the test mimics the well-known Mantel–Haenszel test of the hypothesis that the stratum specific odds-ratio is equal to one [see Mantel (1966) for further details]. However, this test is most often called the log-rank test, due to Peto and Peto (1972). The test is related to a test proposed by Savage (1956) for noncensored data, and BMDP calls it the generalized Savage test.

Gehan (1965) and Breslow (1970) generalized the Wilcoxon rank-sum test to allow for censored data. This test uses weights equal to the number of subjects at risk at each survival time, $w_i = n_i$, and is called the Wilcoxon or generalized Wilcoxon test by most software packages.

SAS's lifetest procedure provides two ways of obtaining the same test, but different variance estimators are used. In SAS, if we define the grouping variable to be a stratification variable, the variance estimator $\hat{v}_{1i}$ is used. If we use SAS's test option, then the variance estimator

$$\hat{v}_{1i}^{*} = \frac{n_{1i} n_{0i}}{n_i^2}$$

is used, which assumes that $d_i = 1$; there are no tied failure times. Thus, in any one example, we may obtain test statistics of similar magnitude

but with slightly different values. Because survival time is often recorded in discrete units that may lead to ties, we recommend that the variance estimator $\hat{v}_{1i}$ be used.

The choice of weight influences the type of differences in the survivorship function the test is most apt to detect. The generalized Wilcoxon test, since it uses weights equal to the number at risk, will put relatively more weight on differences between the survivorship functions at smaller values of time. The log-rank test, since it uses weights equal to one, will place more emphasis than does the generalized Wilcoxon test on differences between the functions at larger values of time. Other tests have been proposed that use weight functions intermediate between these, for example, Tarone and Ware (1977) suggested using $w_i = \sqrt{n_i}$.

Peto and Peto (1972) and Prentice (1978) suggested using a weight function that depends more explicitly on the observed survival experience of the combined sample. The weight function is a modification of the Kaplan-Meier estimator and is defined in such a way that its value is known just prior to the observed failure. The value of any estimated survivorship function at a particular observed failure time is known only after the observation is made. The property of having the value known in advance of the actual observed failure is referred to as *predictable* in counting process terminology. This theory is needed to prove results concerning the distribution of the test statistics. The modified estimator of the survivorship function is

$$\tilde{S}(t) = \prod_{t_{(j)} \le t} \left(\frac{n_j + 1 - d_j}{n_j + 1} \right) \tag{2.21}$$

and the weight used is

$$w_i = \tilde{S}(t_{(i-1)}) \times \frac{n_i}{n_i + 1}. \tag{2.22}$$

Note that when $d_i = 1$ the weight is equal to the modified estimator, that is, $w_i = \tilde{S}(t_{(i)})$, which is an assumption made in the implementation of this test in BMDP. In the example demonstrating the calculations, we will use both the correct version of the weight given in (2.22) as well as BMDP's implementation. In subsequent examples, only the BMDP version of the Peto–Prentice test will be discussed, as it is the only software package providing this test.

Harrington and Fleming (1982) suggested a class of tests that incorporates features of both the log-rank and the Peto and Prentice tests. They suggest using the Kaplan–Meier estimator raised to a power, as the weight, namely

$$w_i = \left[\hat{S}(t_{(i-1)})\right]^{\rho}.$$

If the power is $\rho = 0$ then $w_i = 1$ and the test is the log-rank test. However, if $\rho = 1$ then the weight is the Kaplan–Meier estimator at the previous survival time, a weight similar to that of the Peto and Prentice test. This test has been implemented in the S-PLUS software package.

The principle advantage of the Peto–Prentice and Harrington-Fleming tests over the generalized Wilcoxon test is that they weight relative to the overall survival experience. The generalized Wilcoxon test uses the size of the risk set and hence weights depend both on the censoring as well as the survival experience. If the pattern of censoring is markedly different in each of the groups, then this test may either reject or fail to reject, not on the basis of similarity or differences in the survivorship functions, but on the pattern of censoring. For this reason most software packages will provide information as to the pattern of censoring in each of the two groups. This information should be checked for comparability—especially when the results of several of these tests are provided and yield markedly different significance levels.

A problem can occur if the estimated survivorship functions cross one another. This means that in some time intervals one group will have a more favorable survival experience, while in other time intervals the other group will have the more favorable experience. This situation is analogous to having interaction present when applying Mantel–Haenszel methods to a stratified contingency table. Unfortunately, tests for the homogeneity across strata may not be used in most survival time applications, because data in tables like Table 2.7 will be too thin to satisfy the necessary large sample criteria. Fleming, Harrington and O'Sullivan (1987) proposed a test that addresses the problem by using, as a test statistic, the maximum observed difference between the two survivorship functions. This test has not been implemented in any software package. We consider methods based on regression modeling to address this issue in Chapter 7. For the time being, our only check is via a visual examination of the plot of the Kaplan–Meier estimator for the two groups being compared. If one or more of the various tests fails to reject a difference, and if we see that the curves cross, then this "interaction" may be present.

It is not possible to provide a categorical rank ordering of the values of the test statistics. The actual calculated values will depend on the observed survival and censoring times.

In order to illustrate the computation of each of the tests, we have chosen a small subset of subjects in each of the two drug use groups in the HMO-HIV+ study. These data are listed in Table 2.8. Column 1 of Table 2.9 lists the eight distinct survival times. Columns 2 through 5 present the quantities defined by the notation shown in Table 2.7, and columns 6 and 7 present quantities defined in equations (2.18) and (2.19). Columns 8 through 11 present values for the weight functions for the four tests, where "LR" stands for log-rank test weights, "WL" stands for generalized Wilcoxon test weights, "TW" stands for Tarone–Ware weights and "PP" stands for Peto–Prentice weights. The calculated values of the test statistics and their respective p-values are shown in Table 2.10. The difference between the values of the log-rank and generalized Wilcoxon tests in Table 2.10 reflects the fact that the two groups differed most at the later observed survival times. The significance levels in Table 2.10 are provided only for the purpose of illustrating the calculations since, with only 4 events in each group and an expected number of events in group 1 of 5.45, the assumption that the sample sizes are large is a bit tenuous.

Recall the Kaplan–Meier estimates of the survivorship functions for the two drug groups in the whole HMO-HIV+ study, shown in Figure 2.7. Note that the two curves do not cross at any point, indicating that the previously described problem of "interaction" may not be present. An inspection of the proportion of values that are censored and the pattern of censoring (not shown) indicates that the censoring experience of the two groups is similar. Thus it would appear that the assumptions necessary for using the tests for equality of the survivorship functions seem to hold. Table 2.11 presents the values of the test statistics.

In Table 2.11, all tests are highly significant and support the impression from Figure 2.7 that those with a prior history of drug use tended

Table 2.8 Listing of Data from the Two Drug Use Groups in the HMO-HIV+ Study Used to Illustrate the Tests for the Comparison of Two Survivorship Functions

Drug Use Group	Ordered Observed Survival Times
No	3, 4*, 5, 22, 34
Yes	2, 3, 4, 7*, 11

* Denotes a censored observation.

Table 2.9 Listing of Quantities Needed to Calculate the Tests for the Equality of Two Survivorship Functions

							Weights			
Time	d_{1i}	n_{1i}	d_i	n_i	$\hat{e}_{1i}$	$\hat{v}_{1i}$	LR	WL	TW	PP
2	0	5	1	10	0.500	0.250	1	10	3.16	0.909
3	1	5	2	9	1.110	0.432	1	9	3.00	0.818
4	0	4	1	7	0.571	0.245	1	7	2.64	0.636
5	1	3	1	5	0.600	0.240	1	5	2.23	0.530
11	0	2	1	3	0.667	0.222	1	3	1.73	0.398
22	1	2	1	2	1.000	0	1	2	1.41	0.265
34	1	1	1	1	1.000	0	1	1	1.00	0.133

to die sooner than those who did not have a history of drug use. In practice, one could provide additional support for this conclusion by presenting the estimates of the within-group median survival times along with confidence interval estimates.

Each of the tests used to compare the survivorship experience in two groups may be extended to compare more than two groups. For example, the survivorship experience of three or four racial groups could be compared. In the HMO-HIV+ study, it was hypothesized that age might be related to survival. Since age is a continuous variable, one approach to assessing a potential relationship is to use regression modeling. This is discussed in detail in Chapter 3. An approach used in practice, for preliminary analyses that can yield easily understood summary measures, is to break a continuous variable into several groups of interest and use methods for grouped data on the categorized variable. We use this approach with groups based on the following intervals for age: {[20 – 29], [30 – 34], [35 – 39], [40 – 54]}. Table 2.12 presents the number of subjects, the number of deaths, the median survival time and

Table 2.10 Listing of the Test Statistics and *p*-Values for the Equality of Two Survivorship Functions Computed from Table 2.9

Statistic	Value	*p*-Value
Log-rank	1.512	0.219
Generalized Wilcoxon	1.250	0.264
Tarone–Ware	1.363	0.243
Peto-Prentice (Correct wt.)	1.327	0.249
Peto-Prentice (BMDP)	1.423	0.233

Table 2.11 Test Statistics and *p*-Values for the Equality of the Survivorship Functions for the Two Drug Use Groups in the HMO-HIV+ Study

Statistic	Value	*p*-Value
Log-rank	11.856	<0.001
Generalized Wilcoxon	10.910	<0.001
Tarone–Ware	12.336	<0.001
Peto–Prentice (BMDP)	11.497	<0.001

associated 95 percent confidence interval for each age group.

The estimated median survival time is 43 months for the youngest age group in Table 2.12, which is considerably larger than the estimated median in each of the other three groups. This suggests that these young subjects may have a more favorable survival experience than older subjects. However, the estimated standard error of the estimated median is 32.8 and the symmetric normal theory confidence interval covers the entire observed range of time. This problem arises because there are only 12 subjects in this age group, the minimum value of the estimated survivorship function is 0.24 at 58 months and the largest observations are two censored values at 60 months. The medians and confidence intervals for the other three groups suggest that survival experience worsens with age. The goal in the four-group comparison will be to evaluate whether trends seen in the medians persist when the entire survival experience of the groups is compared. Before presenting the graphs of the Kaplan–Meier estimates of the survivorship functions for the four age groups, we present the details of the extension of the two-group tests to the multiple-group situation.

If we assume that there are K groups, then the calculations of the test statistics are based on a two by K table for each observed survival time. The general form of this table is presented in Table 2.13. In a manner

Table 2.12 Number of Subjects, Events and Estimated Median Survival Time in Four Age Groups in the HMO-HIV+ Study

Age Group	Freq	Deaths	Median	95% CIE
20–29	12	8	43	*
30–34	34	29	9	6.3, 11.7
35–39	25	20	7	4.5, 9.5
40–54	29	23	4	2.5, 5.5

* Estimated standard error too large to compute a CIE.

similar to the two-group case, we estimate the expected number of events for each group under an assumption of equal survivorship functions as

$$\hat{e}_{ki} = \frac{d_i n_{ki}}{n_i}, \quad k = 1,2,\ldots,K. \tag{2.23}$$

We compare the observed and expected numbers of events for $K-1$ of the K groups. The reason for this will be explained shortly. The easiest way to denote the $K-1$ comparisons is to use vector notation to represent both observed and estimated expected number of events as follows:

$$\mathbf{d}_i' = (d_{1i}, d_{2i}, \ldots, d_{K-1i}),$$

and

$$\hat{\mathbf{e}}_i' = (\hat{e}_{1i}, \hat{e}_{2i}, \ldots, \hat{e}_{K-1i}).$$

The difference between these two vectors is

$$(\mathbf{d}_i - \hat{\mathbf{e}}_i)' = (d_{1i} - \hat{e}_{1i}, d_{2i} - \hat{e}_{2i}, \ldots, d_{K-1i} - \hat{e}_{K-1i}). \tag{2.24}$$

For convenience, we have used the first $K-1$ of the K groups, but any collection of $K-1$ groups could equally well be used.

To obtain a test statistic, we need an estimator of the covariance matrix of $\mathbf{d}_i$. The elements of this matrix are obtained assuming that the observed number of events follows a multivariate central hypergeometric distribution [see Johnson and Kotz (1997)]. The diagonal elements of the $(K-1)\times(K-1)$ matrix, denoted $\hat{\mathbf{V}}_i$, are

$$\hat{v}_{kki} = \frac{n_{ki}(n_i - n_{ki})d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k = 1,2,\ldots,K-1, \tag{2.25}$$

and the off-diagonal elements are

$$\hat{v}_{kli} = -\frac{n_{ki}n_{li}d_i(n_i - d_i)}{n_i^2(n_i - 1)}, \quad k,l = 1,2,\ldots,K-1, k \neq l. \tag{2.26}$$

The various multiple-group versions of the two-group test statistics are obtained by computing a weighted difference between the observed and expected number of events. The weights used at each distinct survival time can be any of the weights used in the two-group test, denoted in general at time $t_{(i)}$ by w_i. To obtain a formula for the test statistic, we

Table 2.13 Table Used for the Test for the Equality of the Survivorship Function in K Groups at Observed Survival Time $t_{(i)}$

Event/Group	1	2	...	k	...	K	Total
Die	d_{1i}	d_{2i}	...	d_{ki}	...	d_{Ki}	d_i
Not Die	$n_{1i}-d_{1i}$	$n_{2i}-d_{2i}$	...	$n_{ki}-d_{ki}$	...	$n_{Ki}-d_{Ki}$	n_i-d_i
At Risk	n_{1i}	n_{2i}	...	n_{ki}	...	n_{Ki}	n_i

define a $K-1$ by $K-1$ diagonal matrix denoted $\mathbf{W}_i = \text{diag}(w_i)$. This matrix has the value of the weight, w_i, at time $t_{(i)}$ in all $K-1$ positions along the diagonal of the matrix. The test statistic to compare the survivorship experience of the K groups is

$$Q = \left[\sum_{i=1}^{m} \mathbf{W}_i(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]' \left[\sum_{i=1}^{m} \mathbf{W}_i \hat{\mathbf{V}}_i \mathbf{W}_i\right]^{-1} \left[\sum_{i=1}^{m} \mathbf{W}_i(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]. \qquad (2.27)$$

The reason we use only $K-1$ of the K possible observed to expected comparisons is to prevent the matrix in the center of the right-hand side of (2.27) from being singular. The value of the test statistic in (2.27) is the same, regardless of which collection of $K-1$ groups are used.

The expression on the right-hand side of (2.27) may look intimidating to those not familiar with matrix algebra calculations, but when $K=2$ it simplifies to the more easily understood statistic defined in (2.20). Most software packages providing statistics for several definitions of the weight use (2.27). These packages typically provide only the test statistic and a p-value. One exception is SAS's lifetest procedure, which provides the individual elements in (2.24)–(2.25) for the log-rank and generalized Wilcoxon tests when the group variable is defined as a stratum variable. Under the hypothesis of equal survival functions, and if the summed estimated expected number of events is large, then Q will be approximately distributed as chi-square with $K-1$ degrees-of-freedom, and the p-value is $p = \Pr(\chi^2(K-1) \geq Q)$. The remarks made earlier about how the choice of weights in the two-group case can affect the ability of the test to detect differences apply to the multiple-group case as well.

The log-rank test, $w_i = 1$, has the following easily computed, conservative, approximation:

$$Q_c = \sum_{k=1}^{K} \frac{(d_{k+} - \hat{e}_{k+})^2}{\hat{e}_{k+}} < Q,$$

where

$$d_{k+} = \sum_{i=1}^{m} d_{ki},$$

and $\hat{e}_{k+}$ is defined similarly. If we calculate Q_c and reject the hypothesis of equal survival experience, then we would reject using Q.

The estimated survivorship functions for the four age groups are shown in Figure 2.8. The figure confirms our preliminary observations based on estimates of median survival times. We see that the survivorship function for the youngest group lies completely above those of the other three groups. It has a long right tail and does not go to zero since two observations are censored at 60 months. For the first 15 months, the estimated survivorship functions for the youngest three age groups follow the trend observed in the medians. In this interval, the three functions are, for the most part, inversely ordered by age. The functions for the middle two age groups cross four times between 15 and 45 months, suggesting that the survival experience for these two age groups may be similar in this range. The estimated survivorship function for the oldest age group lies completely below that of the other three groups for 34 months. This suggests that we should begin our analysis with a test for the overall equality of the survivorship experience. If we find that the experience of at least one group is different from the others, we should construct single degree-of-freedom contrasts to examine between-group differences, as is typically done in analysis of variance methods.

The values of the four test statistics using their respective weights in (2.27) are given in Table 2.14. Since each statistic is significant at beyond the 1 percent level, we reject the hypothesis that the survivorship functions for the four age groups are the same. We follow the test for overall group differences in survival experience with contrasts to try and describe more precisely the source(s) of the significance of the overall test. The BMDP package, program 1L, offers this option by allowing the user to specify a trend test and to input a set of coefficients to test for trend when the groups are not equally spaced. The SAS package lifetest procedure has a test option that provides a trend test for a numeric covariate. The test does not yield the same numeric value as the trend test in BMDP. We describe the test used in BMDP as it follows directly from the multiple group test in (2.27). The null hypothesis is that the survivorship functions are equal and the alternative is that they are rank-ordered and follow the trend specified by the coefficients denoted by the vector $\mathbf{c}' = (c_1, c_2, \ldots, c_{K-1})$. If the groups are equally

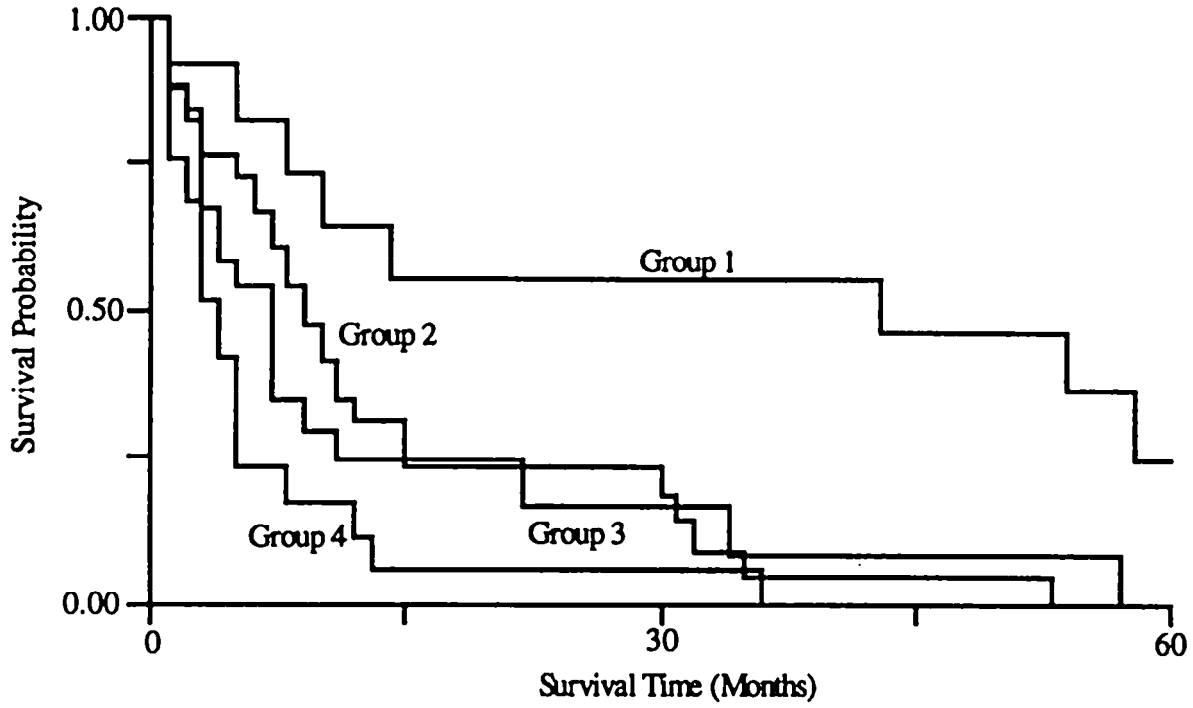


Figure 2.8 Estimated survivorship functions for the four age groups in the HMO-HIV+ study.

spaced, we may use $c_k = k$. The age groups we used in the HMO-HIV+ study are not equally spaced so we will use a vector of coefficients whose values are the midpoints of the four groups, i.e.,

$$\mathbf{c}' = (25, 32.5, 37.5, 47.5).$$

Any linear transformation of these coefficients would yield the same value of the test statistic. The statistic to test for trend, with one degree-of-freedom, is

$$Q_{\text{trend}} = \frac{\left[\mathbf{c}'\sum_{i=1}^{m}\mathbf{W}_i(\mathbf{d}_i - \hat{\mathbf{e}}_i)\right]^2}{\mathbf{c}'\left[\sum_{i=1}^{m}\mathbf{W}_i\hat{\mathbf{V}}_i\mathbf{W}_i\right]\mathbf{c}}. \tag{2.28}$$

The p-value is computed using the chi-square distribution with one degree-of-freedom, i.e., $p = \Pr(\chi^2(1) \geq Q_{\text{trend}})$. Table 2.15 presents the statistics and their p-values for the test of trend among the four age groups in the HMO-HIV+ study. These values are each just slightly

Table 2.14 Test Statistics, Degrees-of-Freedom and *p*-Values for the Equality of the Survivorship Functions for the Four Age Groups in the HMO-HIV+ Study

Statistic	Value	df	*p*-Value
Log-rank	19.906	3	<0.01
Generalized Wilcoxon	14.143	3	<0.01
Tarone–Ware	16.956	3	<0.01
Peto–Prentice (BMDP)	15.665	3	<0.01

smaller than the values in Table 2.14, providing strong evidence for a trend in survival experience that is inversely related to age. We explore this relationship in more detail when we consider regression modeling in the next chapter.

In the examples we have used from the HMO-HIV+ study to illustrate the comparison of the survivorship functions over groups, the magnitude of the test statistics has not varied too dramatically with the choice of weight, and the significance or non-significance of all test statistics has been consistent. However, this is not always the case and to illustrate this we use some data provided to us by our colleagues Drs. Carol Bigelow and Penny Pekow (at the University of Massachusetts) and Dr. Kathy Meyer (at Baystate Medical Center in Springfield, Massachusetts). These data were used as part of Ms. Shiaw-Shyuan Yuan's Masters degree project [Yuan (1993)]. The purpose of the study was to determine factors which predict the length of time low birth weight infants (<1500 grams) with bronchopulmonary dysplasia (BPD) were treated with oxygen. The data were collected retrospectively for the period December 1987 to March 1991. Beginning in August 1989, the treatment of BPD changed to include the use of surfactant replacement therapy. This was done with parental permission since, at the time, this therapy was considered experimental. A total of 78 infants met the study criteria, with 35 receiving surfactant replacement therapy and 43

Table 2.15 Trend Test Statistics, Degrees-of-Freedom and *p*-Values for the Equality of the Survivorship Functions among the Four Age Groups in the HMO-HIV+ Study

Statistic	Value	df	*p*-Value
Log-rank	19.066	1	<0.01
Generalized Wilcoxon	14.080	1	<0.01
Tarone–Ware	16.673	1	<0.01
Peto–Prentice (BMDP)	15.536	1	<0.01

not receiving this therapy. Five babies were still on oxygen at their last follow-up visit and represent censored observations. We refer to this study as the BPD study.

The outcome variable is the total number of days the baby required supplemental oxygen therapy. Figure 2.9 presents the Kaplan–Meier estimates of the survivorship functions for two groups defined by use of surfactant replacement therapy. The estimated median number of days of therapy for those babies who did not have surfactant replacement therapy (group 0) is 107 {95 percent CIE: (55.3, 158.7)}, and the estimated median number of days for those who had the therapy (group 1) is 71 {95 percent CIE: (33.3, 108.7)}. The median number of days of therapy for the babies not on surfactant is about 1.5 times longer than those using the therapy, but there is considerable overlap in the confidence intervals. The plots of the survivorship functions in Figure 2.9 indicate a progressively larger difference in the survivorship experience between the two groups over time. Table 2.16 presents test statistics and associated p-values for the equality of the survivorship functions. The Wilcoxon test is not significant at the 5 percent level, but the log-rank test is significant. The difference in the magnitude of the test statistics is due to the difference in the weights used. The Wilcoxon test uses a weight equal to the size of the risk set and thus is more likely to detect early differences. The log-rank test uses a weight equal to one and is more likely to detect later differences in the survivorship functions.

In any statistical analysis in which more than one test can be used, we need to make a decision about which results we will report. The log-rank test is the most frequently used and reported test for the comparison of survivorship functions. For most analyses, at least when each test has roughly the same level of significance, reporting only the results of the log-rank test is appropriate. When the tests give different results, then more than one result should be reported. This will provide the reader with a clearer picture as to where the survivorship functions are different. The current example demonstrates the importance of computing several of the tests. Most packages have both the log-rank and generalized Wilcoxon tests, and we recommend that both be computed. To our knowledge, only BMDP computes the Tarone–Ware and Peto–Prentice tests. The pattern of censoring can influence the magnitude of the tests, but the values of the Tarone–Ware and Peto–Prentice tests tend to be intermediate between the log-rank and Wilcoxon tests.

We conclude our presentation of the tests for comparison of survivorship functions with a brief discussion of the assumptions underlying the tests and the types of alternative hypotheses the tests have the power

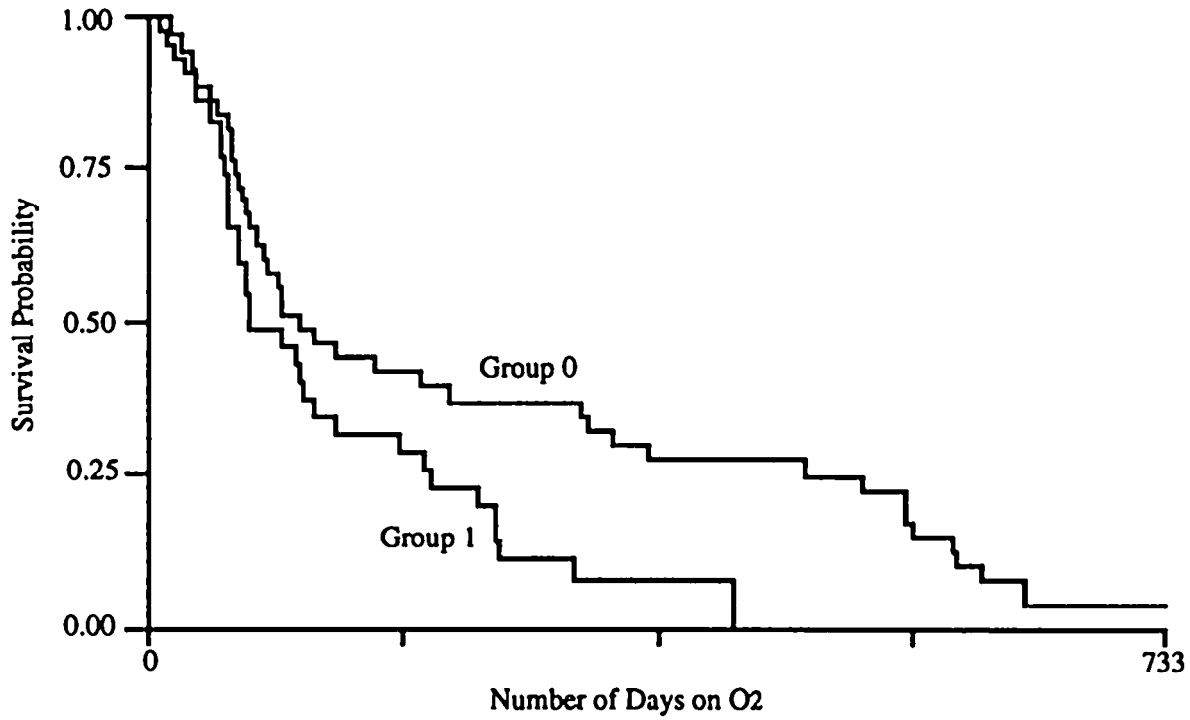


Figure 2.9 Estimated survivorship functions defined by surfactant use in the BPD study (0 = No surfactant, 1 = Surfactant).

to detect. Recall that the Kaplan–Meier estimator assumes that censoring is independent of survival time. In addition, the tests assume that the censoring is independent of the group. Problems in study design and data collection can lead to differential effects due to censoring, and the best protection is a carefully designed study. However, it is good practice to examine the censoring pattern in the data.

In general, we cannot over-emphasize the importance of a careful study of the plot of the Kaplan–Meier estimates of the survivorship functions. Any tests comparing these functions, and within-group point estimates of quantiles, should support what is seen in the plot. The plot is also the basic diagnostic tool to determine whether the tests described

Table 2.16 Test Statistics and p-Values for the Equality of the Survivorship Functions for Two Groups Defined by Surfactant Use in the BPD Study

Statistic	Value	df	p-Value
Log-rank	5.618	1	0.018
Generalized Wilcoxon	2.490	1	0.115
Tarone–Ware	3.698	1	0.055
Peto–Prentice (BMDP)	2.534	1	0.111

previously should be used or, if used, have any chance of detecting a difference. The alternative hypothesis that the tests are most likely to detect is a monotonic ordering of the survivorship functions (e.g., they lie one above another). The tests have little to no power to detect differences when the survivorship functions cross one another. An example of a worst-case scenario is when the survivorship functions for two groups have the same median and cross each other once at that value. For the early times one group has the more favorable survival experience, but for later times the other group does. None of the tests described in this section are able to detect this kind of difference. This is a situation analogous to the presence of interaction in a Mantel–Haenszel analysis of stratified contingency tables. Unfortunately, tests for interaction used with a Mantel–Haenszel analysis, such as the Breslow–Day test [Breslow and Day (1980)], can't be used, due to small cell frequencies in tables such as Table 2.13. In this case, one approach that can be used is to subdivide the sample on the basis of the stratification variable and then test for group differences within the strata. This approach is limited by the study size, as we can spread the data over only so many strata. Eventually there are too few subjects per stratum to reliably estimate the survivorship function. However, in practice, there may be one or two clinically plausible variables to use for stratification purposes. These types of differences, or interactions, between survivorship functions are much more clearly addressed using the regression modeling approach to be discussed in Chapter 3.

2.5 OTHER FUNCTIONS OF SURVIVAL TIME AND THEIR ESTIMATORS

The Kaplan–Meier estimator of the survivorship function has been, and continues to be, the most frequently used estimator, largely due to the fact that it is routinely calculated by most software packages. To motivate the discussion of another estimator, we begin by presenting a different representation of the survivorship function. If we assume that the underlying time random variable is absolutely continuous, then we may express the survivorship function as

$$S(t) = e^{-H(t)}, \tag{2.29}$$

where $H(t) = -\ln(S(t))$. The expression in (2.29) suggests that estimators of the survivorship function could be based on an estimator of $S(t)$

(e.g., the Kaplan–Meier estimator) or via an estimator of $H(t)$. Aalen (1975, 1978), Nelson (1969, 1972) and Altshuler (1970) have proposed an easily computed estimator of $H(t)$, which we refer to as the Nelson-Aalen estimator.

The work by Aalen is considered to be one of the landmark contributions to the field, as virtually all recent statistical developments for the analysis of survival time have been based on the counting process approach he used to derive his version of the estimator of $H(t)$. The statistical theory and use of this estimator in various applied settings are discussed in detail in Andersen, Borgan, Gill and Keiding (1993) and in Fleming and Harrington (1984, 1991). We will use results derived from the counting process theory to justify various techniques discussed in this text. We will not present the counting process approach in any detail since fully appreciating and understanding it requires having had calculus-based courses in mathematical statistics and probability theory.

Without providing any details as to its derivation (a heuristic argument is given later in this section), the Nelson–Aalen estimator of $H(t)$ is

$$\tilde{H}(t) = \sum_{t_{(i)} \le t} \frac{d_i}{n_i}. \tag{2.30}$$

An estimator of the survivorship function, based on (2.30), is

$$\tilde{S}(t) = e^{-\tilde{H}(t)}. \tag{2.31}$$

One theoretical problem is that the expression in (2.29) is valid for continuous time, but the estimator in (2.31) is discrete. However, the estimator in (2.31) provides the basis for the estimator of the survivorship function used with the proportional hazards regression model discussed in Chapter 3. For this reason, we consider it in some detail.

Even though packages may not provide the Nelson–Aalen estimator of the survivorship function, it is remarkably easy to compute. In the absence of ties, one merely sorts the data into ascending order on the time variable. The size of the risk set at $t_{(i)}$ is $n-i+1$ and the estimator, $\tilde{H}(t)$ in (2.30), is obtained as the cumulative sum of the zero-one censoring indicator variable divided by the size of the risk set. The Nelson-Aalen estimator of the survivorship function is obtained by evaluating the expression in (2.31). When ties are present, one sorts the data into ascending order on time and into descending order on the censoring

variable within values of time. Sorting in this way places the censored observations after the events when ties occur. One then calculates a variable equal to $n-i+1$, and uses a procedure such as STATA's collapse command, or the means procedure in SAS, to provide summary statistics at each value of time observed. One needs to obtain the maximum value of $n-i+1$ among the tied time values and the total number of events and/or censored observations. This reduced data set is used to calculate the Nelson-Aalen estimator using the cumulative sum described for the case where there are no ties.

Peterson (1977) proposed another estimator, which is based on the Kaplan–Meier estimator of the cumulative hazard function, as follows:

$$\hat{H}(t) = -\ln\left(\hat{S}(t)\right) = -\ln\left(\prod_{t_{(i)}\le t}\left(\frac{n_i - d_i}{n_i}\right)\right) = -\ln\left(\sum_{t_{(i)}\le t}\left(\frac{n_i - d_i}{n_i}\right)\right)$$

$$= \sum_{t_{(i)}\le t} -\ln\left(1 - \frac{d_i}{n_i}\right).$$

One may show, by using a Taylor series expansion (see Appendix 1), that $d_i/n_i \le -\ln(1-d_i/n_i)$ for each survival time. Thus, the Nelson–Aalen estimator of the survivorship function will always be greater than or equal to the Kaplan–Meier estimator. If the size of the risk sets relative to the number of events is large, then $d_i/n_i \cong -\ln(1-d_i/n_i)$ and there will be little practical difference between the Nelson-Aalen and the Kaplan-Meier estimators of the survivorship function.

The HMO-HIV+ study provides a good illustration of a situation in which there is little practical difference between the two estimators. Table 2.17 presents the results of collapsing the sample of 100 observations to obtain the necessary within-time summary statistics at each observed value of time: the frequency of occurrence (freq), the number of events (d), the size of the risk set (n), the Nelson–Aalen estimator, $\tilde{H}(t)$, the Nelson–Aalen estimator of the survivorship function, $\tilde{S}(t)$ and, for comparison, the Kaplan–Meier estimator, $\hat{S}(t)$. For example, at 3 months the values of the estimators are

$$\tilde{H}(3) = \frac{15}{100} + \frac{5}{83} + \frac{10}{73} = 0.347,$$

$$\tilde{S}(3) = e^{-0.347} = 0.707,$$

and

$$\hat{S}(3) = \left(1 - \frac{15}{100}\right) \times \left(1 - \frac{5}{83}\right) \times \left(1 - \frac{10}{73}\right) = 0.689.$$

Table 2.17 Summary Table Used to Calculate the Nelson-Aalen Estimator of the Survivorship Function for the HMO-HIV+ Study

Time	freq	d	n	$\tilde{H}(t)$	$\tilde{S}(t)$	$\hat{S}(t)$
1	17	15	100	0.150	0.861	0.850
2	10	5	83	0.210	0.810	0.799
3	12	10	73	0.347	0.707	0.689
4	5	4	61	0.413	0.662	0.644
5	7	7	56	0.538	0.584	0.564
6	3	2	49	0.579	0.561	0.541
7	7	6	46	0.709	0.492	0.470
8	4	4	39	0.812	0.444	0.422
9	3	3	35	0.897	0.408	0.386
10	4	3	32	0.991	0.371	0.350
11	3	3	28	1.098	0.333	0.312
12	4	2	25	1.178	0.308	0.287
13	1	1	21	1.226	0.294	0.273
14	1	1	20	1.276	0.279	0.260
15	2	2	19	1.381	0.251	0.232
19	1	0	17	1.381	0.251	0.232
22	1	1	16	1.444	0.236	0.218
24	1	0	15	1.444	0.236	0.218
30	1	1	14	1.515	0.220	0.202
31	1	1	13	1.592	0.204	0.187
32	1	1	12	1.675	0.187	0.171
34	1	1	11	1.766	0.171	0.156
35	1	1	10	1.866	0.155	0.140
36	1	1	9	1.977	0.138	0.125
43	1	1	8	2.102	0.122	0.109
53	1	1	7	2.245	0.106	0.093
54	1	1	6	2.412	0.090	0.078
56	1	0	5	2.412	0.090	0.078
57	1	1	4	2.662	0.070	0.058
58	1	1	3	2.995	0.050	0.039
60	2	0	2	2.995	0.050	0.039

The values at other times are obtained in a similar manner. Figure 2.10 presents graphs of the the Nelson–Aalen and Kaplan–Meier estimators. We see little practical difference between the two estimators, even though $\tilde{S}(t) \geq \hat{S}(t)$ at every observed value of time.

The function $H(t)$ is an important analytic tool for the analysis of survival time data. In much of the survival analysis literature it is called the *cumulative hazard function*, but in the counting process literature it is related to a function called the *cumulative* or *integrated intensity process*. The term "hazard" is used to describe the concept of the risk of "failure" in an interval after time t, conditional on the subject having survived to time t. The word "cumulative" is used to describe the fact that its value is the "sum total" of the hazard up to time t. At this point we focus on the hazard function itself, as it plays a central role in regression modeling of survival data.

Consider a subject in the HMO-HIV+ study who has a survival time of 7 months. For this subject to have died at 7 months, he/she had to be alive at 6 months. The hazard at 7 months is the failure rate "per month," conditional on the fact that the subject has lived 6 months. This is not the same as the unconditional failure rate "per month" at 7

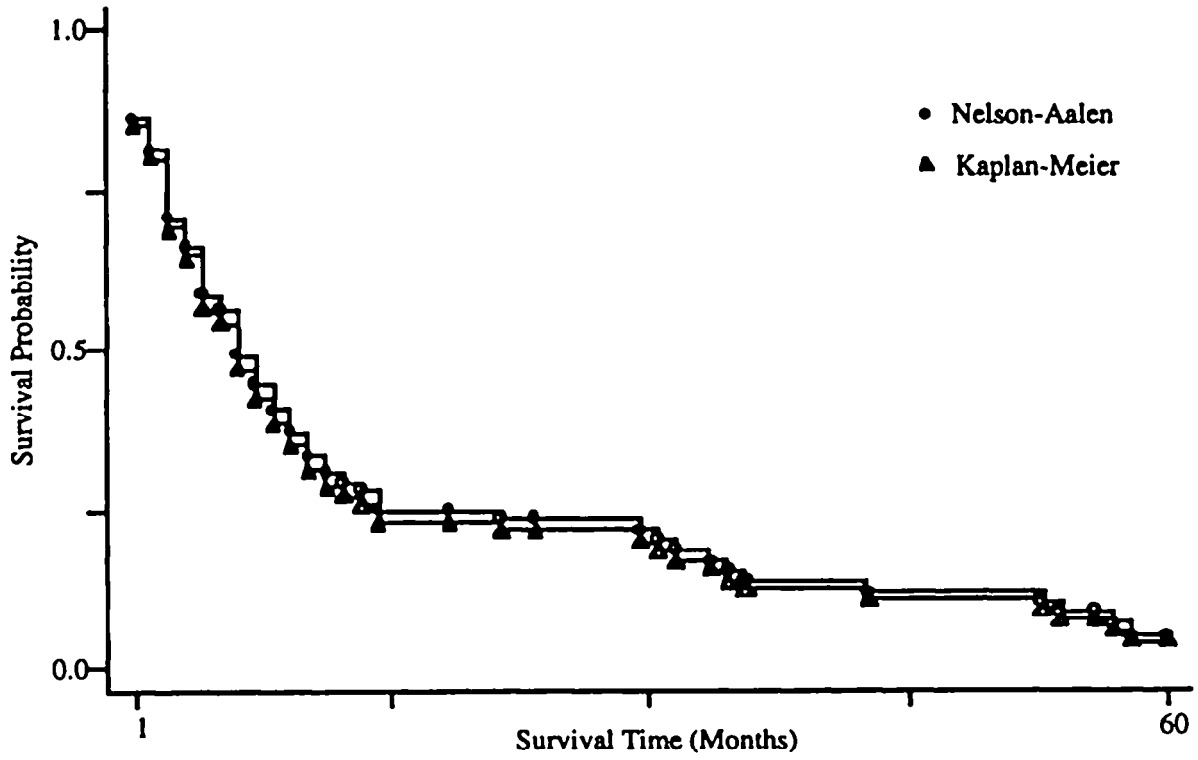


Figure 2.10 Graphs of the Nelson-Aalen and Kaplan-Meier estimators of the survivorship function from the HMO-HIV+ study.

months. The unconditional rate applies to subjects at time zero and, as such, does not use the information available as the study progresses about the survival experience in the sample. This accumulation of knowledge, over time, is generally referred to as *aging*. For example, of 100 subjects who enroll in a study, what fraction is expected to die at 7 months? The conditional failure rate applies only to that subset of the sample that has survived to a particular time, thus it accounts for the aging that has taken place in the sample.

The data from the HMO-HIV+ study can be used to demonstrate the difference between the conditional and the unconditional failure rate. If we assume that there were no censored observations in the study, the "freq" column in Table 2.17 gives the number of deaths. The first two columns of Table 2.17 are a typical presentation of grouped data. A histogram based on these data provides a graphical estimator of the unconditional failure rate.

To construct the histogram, we divide the follow-up time into 10 intervals, each of width 6 months. Each interval is represented graphically by a rectangle with height equaling the frequency drawn over the interval. To construct a relative histogram we divide each frequency by the total sample size. At this point we must decide what we wish to use as the appropriate unit of time. If we do nothing, we implicitly let 6 months denote "one unit" of time. If we wish to have "one unit" equal "one month" then we must further divide by 6. For other intervals of time, we would divide by the correct multiple of interval width and unit. If we divide by 6, the heights of the rectangles give us the relative proportions of the *total* number of subjects beginning at time "zero" who had a survival time in each interval, and the area of each rectangle is the observed unconditional failure rate per month in that interval.

For each time, t, the histogram estimator, $\hat{f}(t)$, is

$$\hat{f}(t) = \frac{(\text{freq})/(\text{width})}{n}, \tag{2.32}$$

where "freq" denotes the number of survival times in the interval, "width" denotes the width of the interval relative to the definition of "one unit" and n is the total sample size. The fact that the numerator of the estimator is expressed relative to the total sample size makes it an unconditional estimator. This is further reflected by the fact that the total area of the histogram rectangles is one, meaning that each subject has been counted once and only once in the presentation of the data.

The interval grouped-data estimator of the hazard function is, for all values of time, t, in an interval,

$$\hat{h}(t) = \frac{(\text{freq})/(\text{width})}{n(t)}, \tag{2.33}$$

where the quantity $n(t)$ is used somewhat imprecisely to denote the number of subjects still alive (at risk) at the beginning of the current interval. The area of the rectangle formed by graphing $\hat{h}(t)$ versus t estimates the conditional, on $n(t)$, per-month failure rate in the interval. The sum of the areas of the rectangles up to and including an interval is an estimate of the cumulative hazard. Since subjects are at risk until they actually die or are censored, they may be counted more than once and the sum of the areas of the rectangles may be greater than one.

Figure 2.11 presents the graphs of the histogram and hazard function estimators of the unconditional and conditional failure rates, computed from the data in Table 2.17, using 6-month intervals (e.g., (0,6], (6,12],...,(54,60]). The shaded rectangles of the histogram, which estimate the overall, unconditional per-month failure rate, are initially high and then drop rapidly, staying consistently low to 60 months. This pattern reflects the many early deaths; relatively few subjects had survival times throughout the period of follow-up. This was described by the Kaplan–Meier estimator in Figure 2.2. On the other hand, the open rectangles of the hazard function estimate the failure rate in the current interval, given that a subject is alive at the beginning of the interval. This pattern is not as consistent as that seen in the shaded histogram due to the fact that each rectangle is based on fewer subjects than the previous one. In other words, the variability is greater in the estimator of the hazard than the histogram. The graph indicates a relatively high initial failure rate which drops and then rises again.

The histogram estimator in (2.32) is useful for providing an estimate of the unconditional rate only when there are no censored observations. It may be modified to handle censored observations by using the difference between the values of the Kaplan–Meier (or Nelson–Aalen) estimator of the survivorship function at the two endpoints of the interval. The hazard function estimator in (2.33) may be modified to accommodate censored observations by having censored values of time contribute to the count in the denominator but not in the numerator. To provide a better approximation of the number at risk over the whole interval in settings in which there are large numbers of subjects and/or the inherently continuous time variable has been recorded at a few dis-

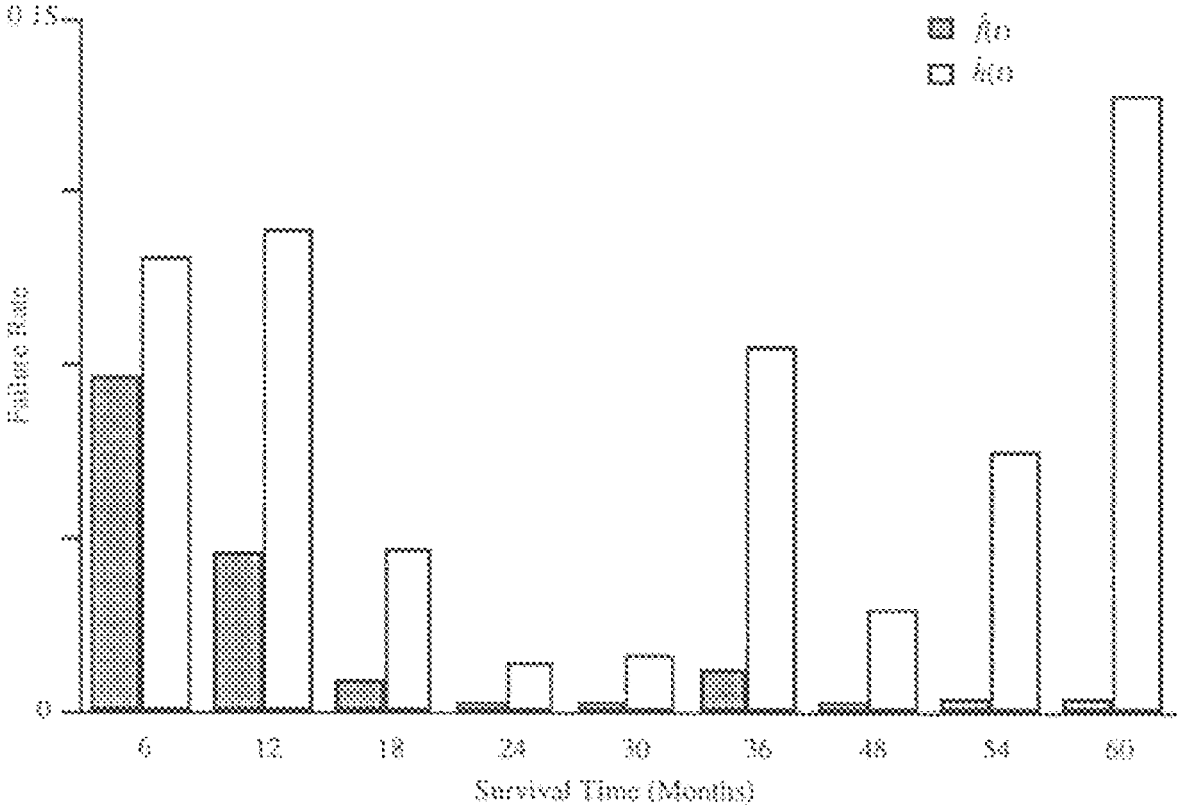


Figure 2.11 Graphs of the histogram estimator (shaded) of the unconditional failure rate and the hazard function estimator (open) of the conditional failure rate from the HMO-HIV+ study.

crete time points, the estimator of the hazard may use a denominator in which the number at risk at the beginning of the interval is reduced by one-half the number of subjects who failed, were censored or were lost for other reasons [see Lee (1992)].

Considering Figure 2.11, it is logical to postulate a function of time that describes, in a concise fashion, the form of either the unconditional or conditional failure rate, which may then be used to express the survivorship function as a function of time. If we can answer this question, then we have taken an important first step toward a more comprehensive analysis that will enable us to study which factors affect survival, namely parametrizing this function with a regression-like model.

As we think about the problem of trying to develop a function to describe survival time in the presence of censored data, we focus attention on the hazard function since it incorporates any aging that might take place. Figure 2.11 may be useful for general descriptive purposes but it is, in a sense, too discrete to be of use in developing a more precise function of time to describe the hazard function. What we would like is a more "continuous" time analysis. If we let the interval width

shrink to the point where it is one measurement unit wide (i.e., one month in the HMO-HIV+ study), then the right-hand side of the estimator of the hazard function in (2.33) is d_i/n_i at observed survival times and is zero elsewhere.

Figure 2.12 presents a scatterplot of the pairs $\left(t_{(i)}, d_i/n_i\right)$, $i = 1,2,\ldots,31$ and a lowess smooth[2] of the plot [see StataCorp (1997), ksm command]. The smoothing done here is for illustrative purposes [see Andersen, Borgan, Gill and Keiding (1993) for a more complete discussion of smoothed estimators of the hazard function]. One difficulty with the plot in Figure 2.12 is that the hazard function should be estimated to be 0 at times when no deaths occurred. The smoothed curve in Figure 2.12 does not incorporate these 0 values. However, the goal in this section is to begin to make the transition from fully nonparametric to regression models discussed in subsequent chapters. Figure 2.12, while not totally correct, does serve to guide the reader in the direction of these regression models.

The smooth of the pointwise estimates of the hazard agrees with our original impression drawn from Figure 2.11 that the conditional risk is relatively high, drops and then rises. On the basis of this observation, we might postulate that the hazard function is a quadratic function of time,

$$h(t) = \theta_0 + \theta_1 t + \theta_2 t^2.$$

Suppose for the moment that we have a parametric form for the hazard function. We need to link the hazard function in a more direct way to the survivorship function. Since we assume the time variable is absolutely continuous, the cumulative hazard is, by methods of calculus,

$$H(t) = \int_0^t h(u)\, du, \tag{2.34}$$

and by (2.29)

$$S(t) = e^{-\int_0^t h(u)\, du}. \tag{2.35}$$

Those readers familiar with calculus will recognize the right-hand side of (2.34) as the integral of the hazard function over the time interval $[0,t]$. For readers not familiar with calculus, the estimator in (2.30) can

---

[2] For those unfamiliar with scatterplot smoothing methods, the purpose is to remove some of the "noise" in the plot by computing, for each y in the plot, a weighted average of the other y's near it.

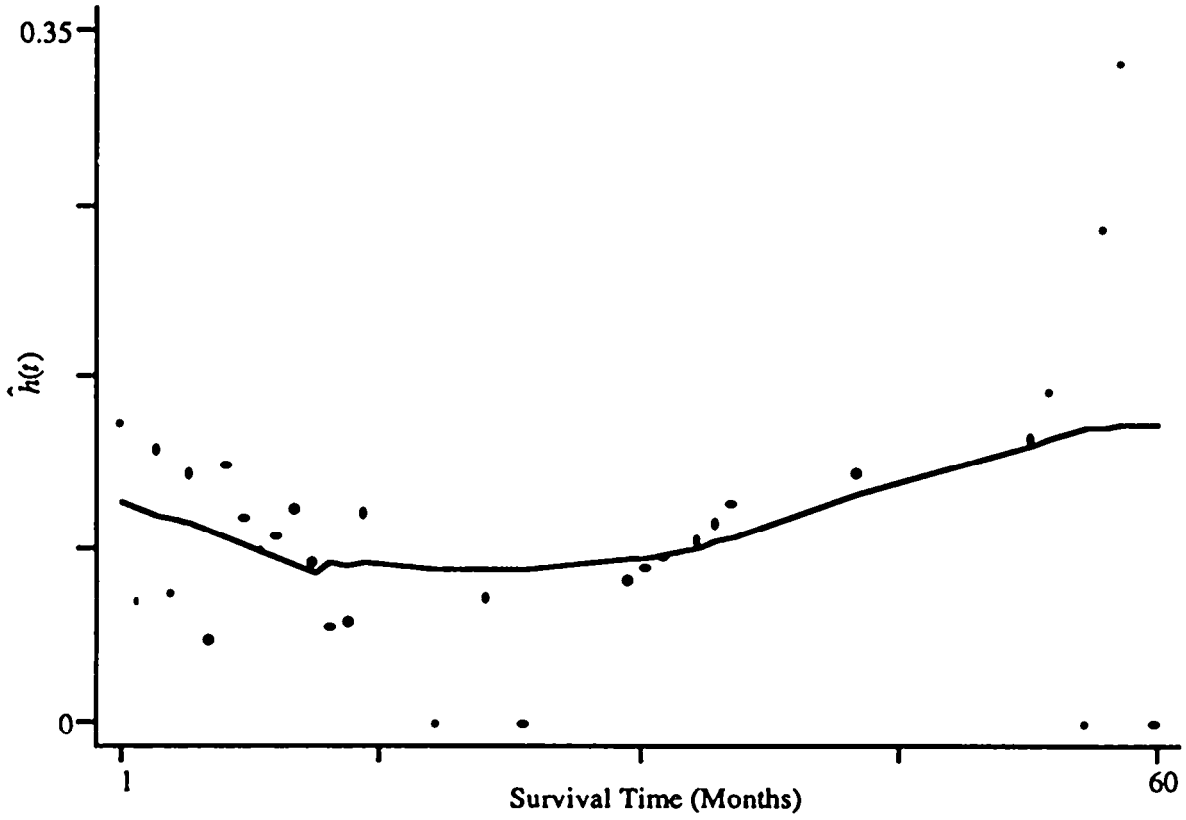


Figure 2.12 Scatterplot of the pointwise estimator of the hazard function, d_i/n, and its lowess smooth from the HMO-HIV+ study.

serve as a convenient mental model of what is being computed in (2.34). Another representation of the hazard function may be obtained by taking the log of (2.35) and then differentiating with respect to t yielding

$$h(t) = \frac{f(t)}{S(t)}, \tag{2.36}$$

where $f(t)$ denotes the probability density function for the time random variable. Those not familiar with methods of calculus may think of the function $f(t)$ as what the histogram estimator in (2.32) becomes if we use larger and larger sample sizes and the width of each interval used in its construction becomes quite small. A similar intuitive argument may be applied to the hazard function estimator in (2.33) to motivate the expression in (2.36).

As noted above, one way to envision the hazard function is to think of it as a limiting, $n \to \infty$, version of the estimator in (2.33). In this argument, we let the width of each interval become quite small and, in the

end, we have a function which describes the failure rate in the next instant following t. The expressions in (2.35)–(2.36) show that if we can specify the hazard function, then it is, in principle, relatively easy to obtain an expression for any of the other functions of survival time. The advantage of using the hazard function is that it characterizes the aging process as a function of time.

To obtain a better understanding of the hazard function and how it specifies the survivorship function, we consider various possible parametric models. A discussion of parametric survival time models is presented in Chapter 8. The goal here is see how this function describes the aging process.

The simplest possible model is for the hazard function to be constant, not depending on time [i.e., $h(t) = \theta$]. This hazard function states that at any particular time the chance that a subject "dies" in the next instant does not depend on how long the subject has survived. For example, in Figure 2.12 the average value of the plotted pointwise estimates of the hazard function is about 0.1. Thus, the constant hazard model is $\hat{h}(t) = 0.1$. The interpretation of this hazard function is that there is about a 10 percent chance that a subject will die in the next month, regardless of how long he/she has already survived. This model for the hazard may be clinically plausible in some studies of human populations when the follow-up time is relatively short. For example, the chance that a "healthy" 35-year-old person dies in the next year is about the same as that of a healthy 36- or 37- or 38- or 39-year-old subject.

The next simplest model is for the hazard to be a linear function of time, $h(t) = \theta_0 + \theta_1 t$. For example, an approximate straight-line fit to the plotted points in Figure 2.12 yields the model $\hat{h}(t) = 0.07 + 0.001t$. The interpretation is that at the beginning of the study subjects had about a 7 percent chance of dying in the next month, and this increases at about 0.1 percent per month. Since the hazard function must be greater than zero, the values of the parameters are constrained. For example, the model $h(t) = 0.12 - 0.004t$ describes the hazard in the first 30 months in Figure 2.12, but yields negative values after 30 months. This leads to the clinically implausible situation of positive probability of infinite physical life. Therefore, we have to use special methods when fitting hazard functions to observed data, since simple least squares regression methods will not be appropriate. We discuss these methods in detail in the next chapter.

On the basis of the lowess smooth in Figure 2.12, we postulated a quadratic function for the hazard function for the HMO-HIV+ study.

This is a more complicated function than the linear or constant model, but a life process of decreasing risk followed by increasing risk is clinically plausible. If one conceptualizes the risk of death in the next "instant" from birth to age 80, the function decreases for the first 5 or so years, remains fairly constant for 40 or so years and then begins to rise rapidly. This is more of a "bathtub" shape and requires a more complex function to describe it than a simple quadratic [see Lawless (1982)].

The major point is that the hazard function itself says a great deal about the fundamental underlying life-length process being studied. Specifying a fully parametric model leads to a specific life-length process. In some settings we may need this level of specificity, but in others it may not be necessary or flexible enough. This point will be dealt with directly in the next chapter.

The univariate descriptive methods discussed in this chapter, computed for the whole study or within a few subgroups, are an important first step in any analysis of survival time; however, these methods cannot be used to address the more sophisticated questions that can typically be addressed through regression modeling techniques. In Chapter 1 we discussed the general similarities and differences between regressions using dependent variables such as weight or disease status and regressions using survival time (with and without censoring) as the dependent variable. At this point, we are in a position to consider the regression methods for survival data in more detail.

Other texts presenting descriptive as well as other methods for survival data include: Collett (1994), Cox and Oakes (1984), Klein and Moeschberger (1997), Kleinbaum (1996), Le (1997), Lee (1992), Miller (1981), Marubini and Valsecchi (1995) and Parmar and Machin (1995).

EXERCISES

1. Listed below are values of survival time (length of follow-up) for 6 males and 6 females from the WHAS. Right-censored times are denoted by a "+" as a superscript.

Males: 1, 3, 4^+, 10, 12, 18
Females: 1, 3^+, 6, 10, 11, 12^+

Using these data, compute by hand (and verify hand calculations when possible with a software package) the following:

(a) The Kaplan-Meier estimate of the survivorship function for each gender.

(b) Pointwise 95 percent confidence intervals for the survivorship functions estimated in problem 1(a).

(c) The Hall and Wellner 95 percent confidence bands for the survivorship functions estimated in problem 1(b).

(d) Point and 95 percent confidence interval estimates of the 25th, 50th and 75th percentiles of survival time distribution for each gender.

(e) The mean survival time for each gender using all available times.

(f) A graph of the estimated survivorship functions for each gender computed in problem 1(a) along with the pointwise and overall 95 percent limit computed in problems 1(b) and 1(c).

2. Repeat problem 1 using data from grouped cohort 1 (1975 and 1978) from the Worcester Heart Attack Study. All calculations for this problem should be done using a software package.

3. Repeat problem 1 using grouped cohort 1 (1975 and 1978) from the WHAS with four groups defined by the age intervals: [24, 60], [61, 65], [66, 75] and [76, 99]. In this subgroup of the data, 60, 65 and 75 are approximately the three quartiles of the age distribution.

4. Compute by hand, and verify hand calculations with a software package, the log-rank, generalized Wilcoxon, and Peto–Prentice tests for the equality of two survivorship functions estimated in problem 1(a).

5. Repeat problem 4 using data from grouped cohort 1 (1975 and 1978) of the WHAS. Do the results of the test support what is seen in the graphs of the estimated survivorship functions?

6. Repeat problem 4 using data from grouped cohort 1 (1975 and 1978) of the WHAS with four groups defined by the age intervals: [24, 60], [61, 65], [66, 75] and [76, 93]. Using the midpoints of the four age intervals, test for trend using the test statistic defined in (2.28). In addition, test whether the survivorship experience for the middle two age groups is the same or different from the youngest and oldest age groups.

7. For the purposes of this problem restrict analyses to WHAS data from grouped cohort 1 (1975 and 1978). Prepare a table of descriptive statistics for survival time (length of follow-up) for each of the patient characteristic variables in Table 1.4. For age use the four groups in problem 6 above, and for CPK use two groups defined by the median.

8. Expand the analyses in problem 7 to include estimates from all 3 cohort groups combined. Note that in this problem the final age interval should be [76, 99].

CHAPTER 3

Regression Models for Survival Data

3.1 INTRODUCTION

In considering regression modeling of survival data, the first question we have to answer is: What are we going to model? Specifically, what will play the role of the systematic component in a regression model? The inherent aging process that is present when subjects are followed over time is what distinguishes survival time from other dependent variables. The presence of censoring in the data makes the study of survival time more interesting from a statistical research perspective, but from a practical point of view, it is an annoying technical detail that must be dealt with when we fit models. Of the functions describing the distribution of survival time discussed in Chapter 2, the hazard function best and most directly captures the essence of the aging process. Thus, a natural place to begin is to explore how to incorporate the hazard function into the heuristic approach to regression modeling presented in Chapter 1.

In Chapter 1 we used a scatterplot of data to motivate a regression model in which the log of survival time had a linear systematic component and an extreme minimum value error component. Assuming that the value of the covariate, x, is fixed and does not change over time, the model, as shown in (1.3), is

$$y = \beta_0 + \beta_1 x + \sigma \times \varepsilon^*, \tag{3.1}$$

where $y = \ln(t)$ and $\varepsilon^* = \ln(\varepsilon)$. Expressed on the time scale, the model is multiplicative and of the form

$$t = \left(e^{\beta_0+\beta_1 x}\right) \times \varepsilon^{\sigma}. \tag{3.2}$$

As expressed in (3.1) and (3.2), survival time is determined by a systematic component (the $\beta_0 + \beta_1 x$ part) and by an error component (the ε part). When we choose a particular parametric distribution for the error component in (3.1) or (3.2), we have also chosen a specific parametric structure for the hazard function. For example, if we assume that the value of the shape parameter in (3.2) is $\sigma = 1$, then the distribution of the error component in (3.2) is exponential with parameter equal to one, and the hazard function for a subject with covariate equal to x is

$$h(t, x, \boldsymbol{\beta}) = e^{-(\beta_0+\beta_1 x)}. \tag{3.3}$$

Two points should be noted: (1) the hazard function does not depend on time; its value is determined by the covariate x and the unknown parameters β_0 and β_1, and (2) the hazard function and systematic component in the regression model are inversely related.

The fact that the hazard does not depend on time means that the risk of "failure" is the same no matter how long the subject has been followed. In Chapter 1, we considered the age of the subject in the HMO-HIV+ study as the covariate. The hazard function in (3.3) states that the risk of dying is determined solely by the age of the subject at the time of HIV+ diagnosis, and not by the time that has elapsed since enrollment in the study, $t = 0$. This assumption of a constant hazard may be unrealistic in many applied settings and should be examined carefully. We discuss this and other methods for model checking in Chapter 6.

One simple way to provide for a nonconstant hazard function is to assume that the shape parameter, σ, in (3.2) is not equal to 1. In this case, the error component has a Weibull distribution with parameters 1 and σ. Survival time has a Weibull distribution with one parameter equal to the systematic component in (3.1) and the second parameter equal to σ. The equation for the hazard function for (3.2) is

$$h(t, x, \boldsymbol{\beta}, \lambda) = \frac{\lambda t^{\lambda-1}}{\left(e^{\beta_0+\beta_1 x}\right)^{\lambda}}, \tag{3.4}$$

where we have set $\lambda = 1/\sigma$ to obtain a more concise expression. Considered as a function of survival time, the hazard function in (3.4) increases over time if $\lambda > 1$ and decreases if $\lambda < 1$. Because it can increase or decrease, the hazard function in (3.4) is more flexible than the

constant hazard in (3.3). However, the change in the hazard function must be monotonic. For example, it would not be a good model if the hazard function first decreases and then increases (as is the case for human life over a many year period). Therefore, it still may not be suitable in certain applied settings.

The inverse relationship between the parameterization of the hazard and the systematic component is a result of the assumption that the distribution of the error component is exponential or Weibull. For example, if the value of the hazard function is 0.10, then the mean survival time is 10. Most software packages fit exponential regression models using the parameterization in (3.1).

In essence, the models described by (3.1)–(3.4) indicate that we are trying to accomplish two goals simultaneously. The model must describe the basic underlying distribution of survival time (error component), but it must also characterize how that distribution changes as a function of the covariates (systematic component). In some applied settings it is important to use a model that accomplishes both goals, but in other settings a model that addresses only the latter one is sufficient.

If we want a model to predict the life-length of a particular brand of computer hard disk as a function of temperature and relative humidity, we need it to address both goals. The desired end product of the statistical modeling is an equation that may be used to predict survival time of the hard disk for specific operating conditions. Fully parametric models such as those in (3.1)–(3.4) may be required, and a comprehensive study of such models is provided in the texts by Lawless (1982) and Nelson (1982). We consider several of these in Chapter 8.

On the other hand, we are often in a setting where we may wish to see if a combination of drug therapies improves survival of HIV+ patients when compared to a single drug therapy. In this case, a complete description of survival time is of secondary importance to a description of how the new therapy modifies the survival experience relative to the old one. In this example, we need to estimate parameters that can be used to compare the survival experience of the two treatment groups, and this comparison may need to be adjusted for other patient characteristics such as age or IV drug use. The regression models in (3.1)–(3.4) could be used to accomplish this goal. However, the assumptions required for their error components may be unnecessarily stringent, given that the desired inferences will be based solely on the parameters in the systematic portion of the model. Models used to describe survival time in a comparative sense are often called *semiparametric regression models* and are the major focus of this text.

3.2 SEMIPARAMETRIC REGRESSION MODELS

We noted in the previous chapter that we can describe the distribution of survival time in one of two equivalent ways. We can specify the density function of a parametric distribution or we can specify the hazard function. The advantage of the latter approach is that we directly address the aging process; but, as shown previously, it does not easily lend itself to the use of scatterplots to motivate regression models. The latter approach may also be preferred in a setting where the end products of the statistical analysis are estimated parameters that compare the survival experience of selected subgroups. By specifying a model through the hazard function, we may address specific questions such as how survival is related to the treatments under study and other subject characteristics.

Suppose we wish to compare the survival experience of cancer patients on two different therapies adjusting for age and gender, patient characteristics known to be associated with survival time. A natural place to begin is to put a regression model type structure on the hazard function. In general we specify the hazard function as a function of time and the covariates. In the hypothetical example there are three covariates: treatment, age and gender. For ease of notation assume for the remainder of this section and the next that there is one covariate denoted x. A regression model for the hazard function that addresses the study goal is

$$h(t,x,\beta) = h_0(t)r(x,\beta). \tag{3.5}$$

The hazard function, as expressed in (3.5), is the product of two functions. The function, $h_0(t)$, characterizes how the hazard function changes as a function of survival time. The other function, $r(x,\beta)$, characterizes how the hazard function changes as a function of subject covariates. The functions must be chosen such that $h(t,x,\beta) > 0$. Note that $h_0(t)$ is the hazard function when $r(x,\beta) = 1$. When the function $r(x,\beta)$ is such that $r(x=0,\beta) = 1$, $h_0(t)$ is frequently referred to as the *baseline hazard function*. Under the model in (3.5) the ratio of the hazard functions for two subjects with covariate values denoted x_1 and x_0 is

$$\mathrm{HR}(t,x_1,x_0) = \frac{h(t,x_1,\beta)}{h(t,x_0,\beta)},$$

$$\begin{aligned} \mathrm{HR}(t, x_1, x_0) &= \frac{h_0(t) r(x_1, \beta)}{h_0(t) r(x_0, \beta)} \\ &= \frac{r(x_1, \beta)}{r(x_0, \beta)} \quad . \end{aligned} \tag{3.6}$$

The hazard ratio (HR) depends only on the function $r(x, \beta)$. If the ratio function $\mathrm{HR}(t, x_1, x_0)$ is easily interpreted, then the actual form of the baseline hazard function is of little importance.

Cox (1972) was the first to propose the model in (3.5) when he suggested using $r(x, \beta) = \exp(x\beta)$. With this parameterization the hazard function is

$$h(t, x, \beta) = h_0(t) e^{x\beta} \tag{3.7}$$

and the hazard ratio is

$$\mathrm{HR}(t, x_1, x_0) = e^{\beta(x_1 - x_0)}. \tag{3.8}$$

This model is referred to in the literature by a variety of terms, such as the *Cox model*, the *Cox proportional hazards model* or simply the *proportional hazards model*. Part of the appeal of the Cox model is the interpretation of (3.8) as a "relative risk"-type ratio. For example, when a covariate is dichotomous, such as gender, with a value of $x_1 = 1$ for males and $x_0 = 0$ for females, the hazard ratio in (3.8) becomes

$$\mathrm{HR}(t, x_1, x_0) = e^{\beta}.$$

If the value of the coefficient is $\beta = \ln(2)$, then the interpretation is that males are "dying" at twice the rate of females. We defer further discussion of the interpretation of the ratio in (3.8) as a function of the coefficients to Chapter 4.

The Cox model in (3.7) is the most frequently used form of the hazard function in (3.5). The term *proportional hazards* refers to the fact that in (3.7) the hazard functions are multiplicatively related, that is, their ratio is constant over survival time. This is an important assumption and methods for assessing its validity are presented in Chapter 6. Other parametrizations have been considered, most notably additive models. One example of an additive model is the *additive relative hazard model* whose hazard function is

$$h(t,x,\beta) = h_0(t)(1 + x\beta). \tag{3.9}$$

Software packages, such as BMDP and EGRET, offer the user the choice of using (3.7) or (3.9) or a mix of the two. We discuss these and other additive models in Chapter 9. Other more generally parametrized positive functions have been suggested [see Andersen, Borgan, Gill and Keiding (1993, Chapter VII)], but none are in wide practical use. We focus primarily on (3.7), the proportional hazards model, as it is the most frequently used model in applied settings.

The hazard functions in (3.5), (3.7) and (3.9) are called semiparametric functions since they do not explicitly describe the baseline hazard function, $h_0(t)$. It was noted at the beginning of this chapter that one way to specify the distribution of survival time is through the hazard function. Thus, a natural question is: What is the survivorship function for a model with hazard function (3.5)? If we use the relationship shown in (2.29), then the survivorship function is

$$S(t,x,\beta) = e^{-H(t,x,\beta)} \tag{3.10}$$

where $H(t,x,\beta)$ is the cumulative hazard function at time t for a subject with covariate x. We have assumed that survival time is absolutely continuous, in which case the value of the cumulative hazard function may be expressed, using methods of calculus, as

$$\begin{aligned} H(t,x,\beta) &= \int_0^t h(u,x,\beta)\, du \\ &= r(x,\beta)\int_0^t h_0(u)\, du \\ &= r(x,\beta)H_0(t). \end{aligned} \tag{3.11}$$

For those not comfortable with the methods of calculus, the expression in (3.11) may be thought of as a measure of the cumulative baseline risk, $H_0(t)$, which is modified by the function, $r(x,\beta)$, for a subject with covariate x. Substituting the result (3.11) into (3.10), the survivorship function for the general semiparametric hazard function is

$$S(t,x,\beta) = e^{-r(x,\beta)H_0(t)}.$$

Thus it follows that

$$S(t,x,\beta) = \left[e^{-H_0(t)}\right]^{r(x,\beta)} = \left[S_0(t)\right]^{r(x,\beta)}, \tag{3.12}$$

where $S_0(t) = e^{-H_0(t)}$ is the baseline survivorship function.

Under the Cox model, the survivorship function is

$$S(t,x,\beta) = \left[S_0(t)\right]^{\exp(x\beta)}. \tag{3.13}$$

The form of the expression for the survivorship function in (3.13) is a consequence of the multiplicative relationship between the baseline hazard function and the exponential function that describes the effect of the covariates. The value of the baseline survivorship function is always between zero and one (true of any survivorship function). Suppose the covariate is age, denoted a, which we model using $x = a - \bar{a}$. The baseline survivorship function corresponds to a subject whose age is equal to the mean age, $\bar{a}$, of the data. Assuming that the risk associated with age is positive (as is usually the case), then $\beta > 0$, and for $a > \bar{a}$ it follows that $x > 0$, $\exp(x\beta) > 1$ and $S(t,x,\beta) < S_0(t)$. The interpretation is that the survivorship experience is less favorable for age a than at the mean age. In other words, at any point in time, the proportion of subjects alive at age a is smaller than the proportion alive at age $\bar{a}$. Similarly, if age is $a < \bar{a}$, then $x < 0$, $\exp(x\beta) < 1$ and $S(t,x,\beta) > S_0(t)$, implying that the survivorship experience is more favorable at age a than at the mean age.

In the next section, we consider estimation of the parameters in the proportional hazards model.

3.3 FITTING THE PROPORTIONAL HAZARDS REGRESSION MODEL

A brief introduction to the use of maximum likelihood to fit regression models to survival time data was provided in Chapter 1. The models fit in Chapter 1 correspond to those given in (3.1) (3.4), where both the systematic and, more importantly, the error components are fully specified. This complete specification allowed for an explicit expression for the likelihood function. We noted in Chapter 1 that the maximum likelihood approach described was used by most software packages to fit

these models. The natural place to begin is with an exploration of whether the likelihood equation given in (1.5) can be used with the proportional hazards model in (3.7).

Assume we have n independent observations each containing information on the length of time a subject was observed, a single covariate whose value is determined at the time observation begins and remains at that value throughout the follow-up of the subject, and whether the observation was a survival time or was right censored. The data are denoted by the triplet (t_i, x_i, c_i), $i = 1,2,...,n$. In order to apply the likelihood function given in (1.5) to the survivorship function in (3.13), we need to obtain an expression for the density function. An application of methods from calculus shows that the density function is the ratio of the hazard function to the survivorship function [see (2.36)], yielding the expression

$$f(t, x, \beta) = h(t, x, \beta) \times S(t, x, \beta). \tag{3.14}$$

Substituting (3.14) into the likelihood equation in (1.5) yields

$$l(\beta) = \prod_{i=1}^{n} \left\{ \left[h(t_i, x_i, \beta) \times S(t_i, x_i, \beta) \right]^{c_i} \times \left[S(t_i, x_i, \beta) \right]^{1-c_i} \right\},$$

and further algebraic simplification yields

$$l(\beta) = \prod_{i=1}^{n} \left\{ \left[h(t_i, x_i, \beta) \right]^{c_i} \times \left[S(t_i, x_i, \beta) \right] \right\}. \tag{3.15}$$

As noted in Chapter 1, the estimate of the parameter, β, is the value that maximizes the log-likelihood function. The log-likelihood function, obtained by taking the log of the likelihood (3.15) and substituting expressions for the hazard function in (3.7) and the survivorship function in (3.13), is

$$L(\beta) = \sum_{i=1}^{n} \left\{ c_i \ln\left[h_0(t_i) \right] + c_i x_i \beta + e^{x_i \beta} \ln\left[S_0(t_i) \right] \right\}. \tag{3.16}$$

Full maximum likelihood requires that we maximize (3.16) with respect to the unknown parameter of interest, β, and the unspecified baseline

hazard and survivorship functions. The proportional hazards model in (3.7) is chosen in order to avoid having to explicitly specify the error component of the model; therefore, it is not possible to use the log-likelihood function in (3.16). This problem is discussed in some detail in Kalbfleisch and Prentice (1980).

Cox (1972) proposed using an expression he called a "partial likelihood function" that depends only on the parameter of interest. He speculated that the resulting parameter estimators from the partial likelihood function would have the same distributional properties as full maximum likelihood estimators. Rigorous mathematical proofs of this conjecture came later, and the counting process approach based on martingales, as detailed in Andersen, Borgan, Gill and Keiding (1993, Chapter VII) and Fleming and Harrington (1991, Chapter 4), simplified earlier work. At this point, it is not vital that one understand the mathematics of these details. An intermediate level of presentation of the construction of the partial likelihood is provided in Collett (1994). The essential idea is similar to the one used to generate the conditional logistic regression model for matched case-control studies or other stratified designs that introduce a large number of nuisance parameters into the model [see Hosmer and Lemeshow (1989, Chapter 7)]. In the present setting, the partial likelihood is given by the expression

$$l_p(\beta) = \prod_{i=1}^{n} \left[\frac{e^{x_i\beta}}{\sum_{j \in R(t_i)} e^{x_j\beta}} \right]^{c_i}, \tag{3.17}$$

where the summation in the denominator is over all subjects in the risk set at time t_i, denoted by $R(t_i)$. Recall that the risk set consists of all subjects with survival or censored times greater than or equal to the specified time.

The expression in (3.17) assumes that there are no tied times, and it is often modified to exclude terms when $c_i = 0$, yielding

$$l_p(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)}\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j\beta}}, \tag{3.18}$$

where the product is over the m distinct ordered survival times and $x_{(i)}$ denotes the value of the covariate for the subject with ordered survival time $t_{(i)}$. The log partial likelihood function is

$$L_p(\beta) = \sum_{i=1}^{m} \left\{ x_{(i)}\beta - \ln\left[\sum_{j \in R(t_{(i)})} e^{x_j\beta} \right] \right\}. \tag{3.19}$$

We obtain the maximum partial likelihood estimator by differentiating the right hand side of (3.19) with respect to β, setting the derivative equal to zero and solving for the unknown parameter. The derivative of (3.19) with respect to β is

$$\begin{aligned} \frac{\partial L_p(\beta)}{\partial \beta} &= \sum_{i=1}^{m} \left\{ x_{(i)} - \frac{\sum_{j \in R(t_{(i)})} x_j e^{x_j\beta}}{\sum_{j \in R(t_{(i)})} e^{x_j\beta}} \right\} \\ &= \sum_{i=1}^{m} \left\{ x_{(i)} - \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j \right\} \\ &= \sum_{i=1}^{m} \left\{ x_{(i)} - \bar{x}_{w_i} \right\}, \end{aligned} \tag{3.20}$$

where

$$w_{ij}(\beta) = \frac{e^{x_j\beta}}{\sum_{l \in R(t_{(i)})} e^{x_l\beta}}$$

and

$$\bar{x}_{w_i} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_j .$$

We note that equation (3.20) looks different from the corresponding equation for the exponential regression model (1.11). The main difference is that equation (3.20) does not incorporate the actual values of survival time. In fact, the estimator obtained when setting the derivative in (3.20) equal to zero and solving for β yields the value such that the sum of the risk-set-weighted means of the covariate is equal to the sum of the covariate over the non-censored subjects.

Another expression for the derivative in (3.20) is obtained by taking the log of (3.17) and differentiating with respect to β, yielding

$$\frac{\partial L_p(\beta)}{\partial \beta} = \sum_{i=1}^{n} c_i \left\{ x_i - \frac{\sum_{j \in R(t_i)} x_j e^{x_j \beta}}{\sum_{j \in R(t_i)} e^{x_j \beta}} \right\}. \tag{3.21}$$

Most software packages provide the maximum partial likelihood estimator. We denote the solution to (3.20) and (3.21) as $\hat{\beta}$.

The estimator of the variance of the estimator of the coefficient is obtained in the same manner as variance estimators are obtained in most maximum likelihood estimation applications. The estimator is the inverse of the negative of the second derivative of the log partial likelihood at the value of the estimator. In particular, taking the derivative of (3.20) we obtain the following expression:

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \left\{ \frac{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} \right] \left[\sum_{j \in R(t_{(i)})} x_j^2 e^{x_j \beta} \right] - \left[\sum_{j \in R(t_{(i)})} x_j e^{x_j \beta} \right]^2}{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta} \right]^2} \right\}. \tag{3.22}$$

The form of this expression may be simplified by using the definition of $w_{ij}(\beta)$ following (3.20). The simplified expression is

$$\frac{\partial^2 L_p(\beta)}{\partial \beta^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij} \left(x_j - \bar{x}_{w_i} \right)^2. \tag{3.23}$$

The negative of the second derivative of the log partial likelihood in (3.22) or (3.23) is called the *observed information*, and we will denote it as

$$\mathbf{I}(\beta) = -\frac{\partial^2 L_p(\beta)}{\partial \beta^2}. \tag{3.24}$$

Later in this chapter we will consider models containing more than one covariate and the result in (3.24) will be called the *observed information*

Table 3.1 Estimated Coefficient, Standard Error, z-Score, Two-Tailed *p*-Value and 95% Confidence Interval for the Proportional Hazards Model Containing Age

Variable	Coeff.	Std. Err.	z	P>|z|	95% CIE
AGE	0.0814	0.0174	4.67	<0.001	0.047, 0.116

matrix. The estimator of the variance of the estimated coefficient is the inverse of (3.24) evaluated at $\hat{\beta}$ and is

$$\hat{\text{Var}}(\hat{\beta}) = \mathbf{I}(\hat{\beta})^{-1}. \tag{3.25}$$

The estimator of the standard error, denoted $\hat{\text{SE}}(\hat{\beta})$, is the positive square root of the variance estimator in (3.25).

As an example, we can use the data from the HMO-HIV+ study to fit a model containing age of the subject as the covariate. The results are shown in Table 3.1. The value of the estimated coefficient is $\hat{\beta} = 0.0814$, and the estimated standard error of the estimated coefficient is $\hat{\text{SE}}(\hat{\beta}) = 0.0174$.

Typically, the first steps following the fit of a regression model are the assessment of the significance of the coefficient and the formation of a confidence interval. We discuss methods that can be used for each of these tasks.

We begin by presenting three different tests to assess the significance of the coefficient: the partial likelihood ratio test, the Wald test and the score test.

The partial likelihood ratio test, denoted G, is calculated as twice the difference between the log partial likelihood of the model containing the covariate and the log partial likelihood for the model not containing the covariate. Specifically,

$$G = 2\left\{L_p(\hat{\beta}) - L_p(0)\right\}, \tag{3.26}$$

where

$$L_p(0) = -\sum_{i=1}^{m} \ln(n_i), \tag{3.27}$$

and n_i denotes the number of subjects in the risk set at observed survival time $t_{(i)}$.

Under the null hypothesis that the coefficient is equal to zero (along with other mathematical conditions), this statistic will follow a chi-square distribution with 1 degree-of-freedom. This distribution can be used to obtain *p*-values to test the significance of the coefficient. The mathematical details using a counting process approach to the partial likelihood may be found in Andersen, Borgan, Gill and Keiding (1993) and Fleming and Harrington (1991). In practice, the "sufficiently" large sample size cited for likelihood ratio tests translates in this case to having the number of observed noncensored survival times be large.

Software packages fitting the proportional hazards model typically provide the value of the log partial likelihood for the fitted model and the value of G. For the example in Table 3.1, these values are $L_p(\hat{\beta}) = -288.518$ and $G = 21.350$. We can use (3.26) to obtain the log partial likelihood of model zero[1] as

$$L_p(0) = L_p(\hat{\beta}) - G/2 = (-288.518) - (21.35/2) = -299.195.$$

The significance level for the test is $\Pr(\chi^2(1) \geq 21.35) < 0.001$, so we reject the null hypothesis and conclude that age is significantly related to survival time. We defer discussion of the interpretation of the coefficient until the next chapter.

Another test for significance of the coefficient can be computed from the ratio of the estimated coefficient to its estimated standard error. This ratio is commonly referred to as a Wald statistic. Under the same mathematical assumptions required for the log partial likelihood ratio test, the Wald statistic will follow a standard normal distribution. The Wald statistic and its *p*-value are typically reported by software packages. Some statistical packages report the square of the Wald statistic, which follows a chi-square distribution with one degree-of-freedom. Unlike normal errors linear regression where the square of the *t*-statistic for the coefficient in a univariable model is equal to the *F*-test for significance, the Wald and log partial likelihood ratio test are not numerically related. The equation for the Wald statistic is

$$z = \frac{\hat{\beta}}{\hat{\mathrm{SE}}(\hat{\beta})} \tag{3.28}$$

---

[1] This will be useful later when we extend the partial likelihood ratio test to the mutivariable regression setting.

and the value shown in Table 3.1 is

$$z = (0.0814/0.0174) = 4.67.$$

The two-tailed p-value is $\Pr(|z| > 4.67) < 0.001$.

The third test one is likely to encounter is the score test. The test statistic is the ratio of the derivative of the log partial likelihood, equation (3.20), to the square root of the observed information, equation (3.24), all evaluated at $\beta = 0$. The equation for the score test is

$$z^* = \left. \frac{\partial L_p / \partial \beta}{\sqrt{\mathbf{I}(\beta)}} \right|_{\beta=0}. \tag{3.29}$$

Under the hypothesis that the coefficient is equal to zero and the same mathematical conditions required for the Wald and partial likelihood ratio tests, this statistic follows a standard normal distribution. The value of the score test for the example in Table 3.1 is $z^* = 4.69$ and the two-tailed p-value is $\Pr(|z^*| > 4.69) < 0.001$. The score test, when computed by a software package such as SAS, may be reported as the square of the value of (3.29), which will follow a chi-square distribution with one degree-of-freedom under the null hypothesis.

In practice, the numeric values of the three tests $\left(\sqrt{G},\ z \text{ and } z^*\right)$ should be quite similar and thus lead one to draw the same conclusion about the significance of the coefficient. In situations where there is disagreement, making it necessary to choose one test, the partial likelihood ratio test is the preferred choice.

A clear advantage of the score test is that it may be computed without evaluating the maximum partial likelihood estimator of the coefficient. For this reason, the score test has gained some favor as a test to use in model building applications in which evaluation of the estimator is computationally intensive. We return to consider this point further when we discuss variable selection in Chapter 5.

The confidence interval for the coefficient shown in Table 3.1 is called the Wald-statistic-based interval. Its endpoints are based on the same assumptions as the Wald test for significance, i.e., that the estimator is distributed normally with standard error estimated by the square root of (3.25). The endpoints of a $100(1-\alpha)$ percent confidence interval for the coefficient are

$$\hat{\beta} \pm z_{1-\alpha/2}\hat{\mathrm{SE}}(\hat{\beta}).$$

The endpoints of the 95 percent confidence interval shown in Table 3.1 are computed as

$$0.0814 \pm 1.96 \times 0.0174,$$

yielding the interval $0.047 \leq \beta \leq 0.116$. The interval does not include zero and is consistent with the results of all three tests of significance. We conclude that age is associated with survival time.

Up to this point we have considered models in which only one covariate is of interest. One advantage of using regression in any statistical analysis is the ability to include multiple covariates in the model simultaneously. The proportional hazards model may be formulated to include a variety of covariates. We now focus on the extension of the model to include a collection of p covariates whose values are measured on each individual at the time follow-up begins and remain fixed over time. Covariates whose values change over time, often referred to as *time-dependent* or *time-varying* covariates, as well as other covariate scenarios are discussed in Chapter 7.

Let the p covariates for subject i be denoted by the vector $\mathbf{x}_i' = (x_{i1}, x_{i2}, \ldots, x_{ip})$. This vector may be any collection of covariates: continuous covariates, design variables for nominal scale covariates, products of covariates (interactions) and other higher order terms. Denote the triplet of observed time, covariates and censoring variable as $(t_i, \mathbf{x}_i, c_i)$, $i = 1, 2, \ldots, n$. The partial likelihood for the multivariable model is obtained by replacing the single covariate, x, in (3.18) with the vector of covariates, $\mathbf{x}$. Its expression is so similar to (3.18) that it will not be repeated.

There are p equations, one for each covariate, similar to (3.20) which, when set equal to zero and solved, yield the maximum partial likelihood estimators. We denote the vector of coefficients as $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$. The equation for the kth covariate is

$$\frac{\partial L_p(\boldsymbol{\beta})}{\partial \beta_k} = \sum_{i=1}^{m} \left\{ x_{(ik)} - \frac{\sum_{j \in R(t_{(i)})} x_{jk} e^{\mathbf{x}_j'\boldsymbol{\beta}}}{\sum_{j \in R(t_{(i)})} e^{\mathbf{x}_j'\boldsymbol{\beta}}} \right\}$$

$$= \sum_{i=1}^{m} \left\{ x_{(ik)} - \bar{x}_{w,k} \right\}, \tag{3.30}$$

where

$$\bar{x}_{w_i k} = \sum_{j \in R(t_{(i)})} w_{ij}(\beta) x_{jk}$$

and

$$w_{ij}(\beta) = \frac{e^{x_j'\beta}}{\sum_{l \in R(t_{(i)})} e^{x_l'\beta}}.$$

We use $x_{(ik)}$ to denote the value of covariate x_k for the subject with observed ordered survival time $t_{(i)}$. We denote the maximum partial likelihood estimator as $\hat{\beta}' = (\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p)$.

The elements of the p by p information matrix are obtained by extending the definition in (3.24) to include all second-order partial derivatives, namely

$$\mathbf{I}(\beta) = -\frac{\partial^2 L(\beta)}{\partial \beta^2}.$$

The general form of the elements in this matrix is obtained from (3.23). The diagonal elements are

$$\frac{\partial^2 L_p(\beta)}{\partial \beta_k^2} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij} \left(x_{jk} - \bar{x}_{w_i k} \right)^2 \tag{3.31}$$

and the off-diagonal elements are

$$\frac{\partial^2 L_p(\beta)}{\partial \beta_k \partial \beta_l} = -\sum_{i=1}^{m} \sum_{j \in R(t_{(i)})} w_{ij} \left(x_{jk} - \bar{x}_{w_i k} \right) \left(x_{jl} - \bar{x}_{w_i l} \right). \tag{3.32}$$

The estimator of the covariance matrix of the maximum partial likelihood estimator is obtained by extending (3.25) and is the inverse of the observed information matrix evaluated at the maximum partial likelihood estimator,

$$\hat{\mathrm{Var}}(\hat{\beta}) = \mathbf{I}(\hat{\beta})^{-1}. \tag{3.33}$$

Software packages typically provide the value of the estimated standard error for all estimated coefficients in the model. Most packages provide the user with the option of obtaining the full estimated covariance matrix for the estimated parameters.

Consider a model for the HMO-HIV+ study that contains age, IV drug use and their product (interaction). This model may be used to determine whether the association of age with survival time is different for subjects with and without a history of IV drug use. The model is used here to present the results of fitting a multivariable model and to demonstrate how the partial likelihood ratio test may be used to assess the significance of subsets of parameters. We present the results of fitting the model in Table 3.2.

The log partial likelihood ratio test is not only the easiest test to compute, but is also the best of the three tests for assessing the significance of the fitted model. Its value is obtained from (3.26). The log partial likelihood for model 0 is the same for this example as in the univariable model in Table 3.1, $L_p(0) = -299.193$. The log partial likelihood for the fitted model is $L_p(\hat{\beta}) = -281.684$ and the value of the log partial likelihood ratio test is

$$G = 2[(-281.684) - (-299.193)] = 35.02.$$

Under the null hypothesis that all three coefficients are simultaneously equal to zero and, under the mathematical regularity and large sample conditions referred to above, G will follow a chi-square distribution with three degrees-of-freedom (one for each coefficient). The significance level for the test in this example is $\Pr(\chi^2(3) \geq 35.02) < 0.001$, providing evidence that at least one of the coefficients in the model is significantly associated with survival time.

The computation of both the score and Wald tests for the multiple

Table 3.2 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Proportional Hazards Model Containing Age, History of IV Drug Use and Their Interaction

Variable	Coeff.	Std. Err.	z	P>|z|	95% CIE
AGE	0.094	0.0229	4.11	<0.001	0.049, 0.139
DRUG	1.186	1.2565	0.94	0.345	-1.277, 3.649
AGExDRUG	-0.007	0.0337	-0.20	0.841	-0.073, 0.059

proportional hazards regression model requires matrix calculations. Specifically, we denote the vector of first partial derivatives whose elements are given in (3.29) as $\mathbf{u}(\beta)$. Under the hypothesis that all coefficients are equal to zero, and under the mathematical conditions needed for the partial likelihood ratio test, the vector of scores $\mathbf{u}(0) = \mathbf{u}(\beta)|_{\beta=0}$ will be distributed as multivariate normal with mean vector equal to zero and covariance matrix given by the information matrix evaluated at the coefficient vector equal to zero, $\mathbf{I}(0) = \mathbf{I}(\beta)|_{\beta=0}$. The elements in this matrix are obtained by evaluating the expressions in (3.31) and (3.32) with the coefficient vector equal to zero. The score test statistic is

$$\mathbf{u}'(0)[\mathbf{I}(0)]^{-1}\mathbf{u}(0),$$

which is distributed asymptotically as chi-square with p degrees-of-freedom. The Wald test is obtained from equivalent theory which states that, under the null hypothesis, the estimator of the coefficient, $\hat{\beta}$, will be asymptotically normally distributed with mean vector equal to zero and a covariance matrix that is estimated by the expression in (3.33). The multiple variable Wald test statistic is

$$\hat{\beta}'\mathbf{I}(\hat{\beta})\hat{\beta},$$

which is also distributed asymptotically as chi-square with p degrees-of-freedom. Both the score and Wald test require matrix calculations that, while not difficult from a purely technical perspective, are inconvenient to perform in most packages. This is in contrast to the partial likelihood ratio test which is easily performed from readily available output. For this reason we will not make extensive use of the multiple variable score and Wald tests in this text. The values for the multiple variable score and Wald tests for the model in Table 3.2 are 35.146 and 32.167, respectively, each with p-value < 0.001.

In contrast to the multiple variable Wald test, the univariate Wald tests based on individual estimated coefficients can provide guidance, during the model building process, as to possible variables that might be eliminated from the model without compromising model performance. The individual significance levels in Table 3.2 suggest that age may be significant, but the picture is not as clear with respect to IV drug use and its interaction with age. To explore this further we fit a reduced model that excludes the interaction term. The results are shown in Table 3.3.

Table 3.3 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Proportional Hazards Model Containing Age and History of IV Drug Use

Variable	Coeff.	Std. Err.	z	P>|z|	95% CIE
AGE	0.092	0.0185	4.97	0.001	0.056, 0.128
DRUG	0.941	0.2555	3.68	0.001	0.440, 1.442

The Wald tests for both remaining coefficients are significant. The partial likelihood ratio test for the excluded interaction term, keeping age and IV drug use in the model, is obtained by comparing the values of the log partial likelihood function for the models in Tables 3.2 and 3.3. This test is analogous to the partial F-test in linear regression, in that two models that have a common set of covariates are being compared. As in any multivariable analysis, we must make sure that both models have been fit to the same set of data. Since the HMO-HIV+ study does not have any missing data, this is not an issue in this example. The value of the log partial likelihood function for the reduced model is –281.704 which, when compared to that of the larger model, yields a test statistic whose value is

$$G = 2[(-281.684) - (-281.704)] = 0.04.$$

Under the null hypothesis that the interaction variable has a coefficient equal to zero, given that age and history of IV drug use are in the model, this statistic will follow a chi-square distribution with one degree-of-freedom. The significance level for the test in this case is $p = 0.841$, indicating that the interaction term does not contribute to the model.

In summary, the basic techniques for fitting the proportional hazards model are identical to those used in other modeling scenarios, such as the linear, logistic and Poisson regression models. Maximum likelihood methods are used to obtain estimators of the coefficients and their standard errors. We use log-likelihood functions in a standard manner to obtain test statistics that are used with the chi-square distribution to assess the overall significance of the model and to compare nested models. The only difference between the analysis of the proportional hazards model and other models is that the likelihood function is a partial, rather than a full, likelihood function.

3.4 FITTING THE PROPORTIONAL HAZARDS MODEL WITH TIED SURVIVAL TIMES

The partial likelihood function methods described in the previous section are based on the assumption that there were no tied values among the observed survival times. Since most, if not all, applied settings are likely to have some tied observations, modifications are needed. A number of approaches to handle tied data have been suggested and, of these, three are used by software packages: an exact expression that is derived in Kalbfleisch and Prentice (1980) and approximations due to Breslow (1974) and Efron (1977). The analyses presented in the previous section were all based on the Breslow approximation described below. An alternative to an approximate partial likelihood is to use one of the discrete time models discussed in Chapter 7.

We will not present the expression for the exact partial likelihood. The basis for its construction is to assume that the d ties at a particular survival time are due to lack of precision in measuring survival time. Thus the tied values could actually have been observed in any one of the $d!$ possible arrangements of their values. The exact partial likelihood is obtained by modifying the denominator of (3.18) to include each of these arrangements. The SAS software package includes the option of using the exact partial likelihood.

The approximations derived by Breslow (1974) and Efron (1977) are designed to provide expressions that are more easily computed than the exact partial likelihood, yet that still account for the fact that ties are present among the observed values of survival time. For ease of notation, we present the approximations to the exact partial likelihood for the case when the model contains a single covariate. The Breslow approximation uses as the partial likelihood

$$l_{pl}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)+}\beta}}{\left[\sum_{j \in R(t_{(i)})} e^{x_j \beta}\right]^{d_i}}, \tag{3.34}$$

where d_i denotes the number of subjects with survival time $t_{(i)}$ and $x_{(i)+}$ is equal to the sum of the covariate over the d_i subjects, that is, $x_{(i)+} = \sum_{j \in D(t_{(i)})} x_j$, where $D(t_{(i)})$ represents the subjects with survival times equal to $t_{(i)}$. The Efron approximation is a bit more complicated and

yields a slightly better approximation to the exact partial likelihood than the Breslow approximation. It uses as the partial likelihood

$$l_{p2}(\beta) = \prod_{i=1}^{m} \frac{e^{x_{(i)*}\beta}}{\prod_{k=1}^{d_i} \left[\sum_{j \in R(t_{(i)})} e^{x_j\beta} - \frac{k-1}{d_i} \sum_{j \in D(t_{(i)})} e^{x_j\beta} \right]}. \tag{3.35}$$

Note that when $d_i = 1$, the terms in the numerators and denominators of (3.18), (3.34) and (3.35) are identical.

The maximum partial likelihood estimator for β in the presence of ties is obtained in the same manner as in the non-tied data case, with the exception that derivatives are taken with respect to the unknown parameter in the log of either the Breslow (1974) or Efron (1977) approximation to the partial likelihood. These equations are similar in form to (3.20)–(3.21). The estimator of the variance of the estimated coefficient is obtained from the second partial derivative evaluated at the value of the estimator, and results are similar to (3.23)–(3.25).

The HMO-HIV+ study provides a good setting for a comparison of the estimators obtained from the three forms of the partial likelihood in the presence of tied survival times. In this study, there are 31 distinct survival times among the 100 subjects, with the number of deaths at a particular time ranging from 1 to 17. If there are major differences in the estimators obtained from the three versions of the partial likelihood with ties, it should be apparent in this example because there are many tied survival times. The values of the estimator using each of the three methods are shown in Table 3.4 for the model containing age and IV drug use.

The results shown in Table 3.4 support the fact that the Efron (1977) method of correcting for tied survival times yields estimates closer to those obtained from the exact partial likelihood than estimates obtained from the Breslow (1974) approximation. While this is true in a strict numeric sense, all three point estimates are close to one another. The Breslow estimates differ from the exact estimates by 6–8 percent and the Efron estimates differ by 0.5 percent. The estimated standard errors are nearly identical. Hence, we would reach the same scientific conclusion using the estimates from the Breslow partial likelihood as we would using the estimates from the other two partial likelihoods. Thus, given a choice, one would prefer to use the Efron approximation, but in this example, the Breslow approximation yields acceptably close estimates.

Table 3.4 Estimated Coefficients and Standard Errors for Age and IV Drug Use Obtained from the Exact Partial Likelihood, Breslow and Efron Approximations

	AGE		DRUG	
Method	Coeff.	Std. Err.	Coeff.	Std. Err.
Exact	0.0977	0.0187	1.0226	0.2572
Breslow	0.0915	0.0185	0.9414	0.2555
Efron	0.0971	0.0186	1.0167	0.2562

The Breslow (1974) approximation is available in many software packages. The Efron (1977) approximation is available in the SAS and S-Plus packages. In many applied settings there will be little or no practical difference between the estimators obtained from the two approximations. Because of this, and since the Breslow approximation is more commonly available, unless stated otherwise, analyses presented in this text will be based on it.

3.5 ESTIMATING THE SURVIVORSHIP FUNCTION OF THE PROPORTIONAL HAZARDS REGRESSION MODEL

An estimator of the survivorship function of the proportional hazards model is available as an option in most software packages. This estimator may be used to describe the survival experience of subgroups of subjects of particular interest, adjusted for other covariates. This particular application is discussed in detail in Chapter 4. In this section we present how the estimator itself is obtained.

The expression for the survivorship function can be found in (3.13) and is repeated here for convenience:

$$S(t,\mathbf{x},\boldsymbol{\beta}) = \left[S_0(t)\right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}. \tag{3.36}$$

This indicates that once we have an estimator of the regression coefficients, all we need is an estimator of the baseline survivorship function. A likelihood-based approach, which assumes that the hazard is constant between observed survival times, is the foundation of the method. The details may be found in Lawless (1982) and are sketched here. A derivation of the estimator from the counting process approach is discussed

by both Fleming and Harrington (1991) and Andersen, Borgan, Gill and Keiding (1993).

The essential idea of the likelihood approach is to mimic the arguments that lead to the Kaplan–Meier estimator of the survivorship function described in Chapter 2, equation (2.1). The key point in that development is the use of the quantity $\hat{\alpha}_i = 1 - d_i/n_i$ as an estimator of the conditional survival probability at observed ordered survival time $t_{(i)}$. The Kaplan–Meier estimator of the survivorship function is the product of estimators of the individual conditional survival probabilities. The expression for the conditional survival probability that leads to this estimator is $\alpha_i = S(t_{(i)})/S(t_{(i-1)})$. To extend this argument to the proportional hazards model, we define the conditional baseline survival probability as $\alpha_i = S_0(t_{(i)})/S_0(t_{(i-1)})$, and the conditional survival probability is

$$\frac{S(t_{(i)}, \mathbf{x}, \boldsymbol{\beta})}{S(t_{(i-1)}, \mathbf{x}, \boldsymbol{\beta})} = \left\{ \frac{\left[S_0(t_{(i)})\right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}}{\left[S_0(t_{(i-1)})\right]^{\exp(\mathbf{x}'\boldsymbol{\beta})}} \right\} = \left\{ \frac{S_0(t_{(i)})}{S_0(t_{(i-1)})} \right\}^{\exp(\mathbf{x}'\boldsymbol{\beta})} = \alpha_i^{\exp(\mathbf{x}'\boldsymbol{\beta})}.$$

Maximum likelihood methods are employed conditional on the partial likelihood estimator of the regression coefficients in the model, $\hat{\boldsymbol{\beta}}$. In order to simplify the notation, we let $\hat{\theta}_l = \exp\left(\mathbf{x}'\hat{\boldsymbol{\beta}}\right)$, and the estimator of the conditional baseline survival probability is obtained by solving the equation

$$\sum_{l \in D_i} \frac{\hat{\theta}_l}{1 - \alpha_i^{\hat{\theta}_l}} = \sum_{l \in R_i} \hat{\theta}_l, \tag{3.37}$$

where R_i denotes the subjects in the risk set at ordered observed survival time $t_{(i)}$ and D_i denotes the subjects in the risk set with survival times equal to $t_{(i)}$.

If there are no tied survival times, D_i contains one subject and the solution to (3.37) is

$$\hat{\alpha}_i = \left[1 - \frac{\hat{\theta}_i}{\sum_{l \in R_i} \hat{\theta}_l} \right]^{\hat{\theta}_i^{-1}}. \tag{3.38}$$

If there are tied survival times, the solution to (3.37) is obtained using iterative methods. The estimator of the baseline survivorship function is the product of the individual estimators of the conditional baseline survival probabilities

$$\hat{S}_0(t) = \prod_{t_{(i)} \le t} \hat{\alpha}_i, \tag{3.39}$$

where $\hat{\alpha}_i$ is the solution to (3.37). This estimator is used in some software packages, for example, SAS and STATA. Other packages may use an approximation to the solution for (3.36) due to Breslow (1974). To obtain this solution, one replaces $\alpha_i^{\hat{\theta}_l}$ on the left-hand side of (3.37) with the approximation $\alpha_i^{\hat{\theta}_l} \approx 1 + \hat{\theta}_l \ln(\alpha_i)$. The solution to (3.36) is then

$$\tilde{\alpha}_i = \exp\left[-d_i \Big/ \sum_{l \in R_i} \hat{\theta}_l\right], \tag{3.40}$$

and the estimator of the baseline survivorship function is again the product of the individual conditional survival probabilities. One uses (3.39) with the estimator in (3.40).

The estimator of the survivorship function in (3.36) is obtained by substituting the estimators of the baseline survivorship function and the estimator of the coefficients using covariate values of interest. Software packages typically provide the value of the estimator of the survivorship function using the observed time and covariates for all (noncensored as well as censored) subjects.

Some software packages provide an estimator of the baseline hazard function, which is a simple function of the estimator of the conditional survival probabilities, namely

$$\hat{h}_0(t_{(i)}) = 1 - \hat{\alpha}_i.$$

The individual pointwise estimators of the baseline hazard function will typically be too "noisy" or unstable (see Figure 2.12) to use themselves. However, by using smoothing methods referred to in Chapter 1, one may get a sense of the shape of the underlying baseline hazard function.

The estimator of the cumulative baseline hazard function is more practical to use since it is less noisy than the estimator of the baseline

hazard function. Its estimator is obtained using the expression for the survivorship function shown in (2.29), namely

$$\hat{S}_0(t) = e^{-\hat{H}_0(t)},$$

thus the estimator of the cumulative baseline hazard function is

$$\hat{H}_0(t) = -\ln\left[\hat{S}_0(t)\right].$$

The estimator of the cumulative hazard function for a specific value of the covariates is

$$\begin{aligned}\hat{H}(t,\mathbf{x},\hat{\boldsymbol{\beta}}) &= -\ln\left[\hat{S}(t,\mathbf{x},\hat{\boldsymbol{\beta}})\right] \\ &= -e^{\mathbf{x}'\hat{\boldsymbol{\beta}}}\ln\left[\hat{S}_0(t)\right], \end{aligned} \tag{3.41}$$

which, when graphed as a function of time, may provide a useful graphical descriptor of the "risk" experience.

We do not present an application of the estimators of the cumulative hazard function or survivorship function in this chapter. We defer it to Chapter 4, where we discuss the interpretation of the coefficients from a fitted proportional hazards model, the assumption of proportional hazards and graphical presentation of fitted models.

EXERCISES

1. Using the data from the WHAS for grouped cohort 1 (1975 and 1978), with length of follow-up as the survival time variable and status at last follow-up as the censoring variable, do the following:

(a) Fit the proportional hazards model containing age, sex, peak cardiac enzymes, left heart failure complications and MI order.

(b) Assess the significance of the model using the partial log likelihood ratio test. If it is possible in the software package, assess for the significance of the model using the score and Wald tests. Is the statistical decision the same for the three tests?

(c) Using the univariate Wald tests, which variables appear not to contribute to the model? Fit a reduced model and test for the significance of the variables removed using the partial log likelihood ratio test.

(d) Fit the reduced model in problem 1(c) using the Breslow, Efron and exact methods for tied survival times. Compare the estimates of the

coefficients and standard errors obtained from the three methods for handling tied survival times. Are the results similar or different?

(e) Estimate the baseline survivorship function for the model fit in problem 1(c). Graph the estimated baseline survivorship function versus survival time. What covariate pattern is the "baseline" subject for the fitted model?

(f) Repeat problem 1(e) using age centered at the median age of 65 years. Explain why the range of the estimated survivorship functions in problems 1(e) and 1(f) are different.

(g) Using the model fit in problem 1(f) estimate the value of the survivorship function for each subject at his or her respective observed value of time. Graph the values of the estimated survivorship function versus survival time. Why is there scatter in this plot that was not present in the graphs in problems 1(e) and 1(f)?

2. Repeat problem 1 for each of the other grouped cohorts.

3. Repeat problem 1 using all the data from the WHAS (i.e., ignore cohort).

CHAPTER 4

Interpretation of a Fitted Proportional Hazards Regression Model

4.1 INTRODUCTION

The interpretation of a fitted proportional hazards model requires that we be able to draw practical inferences from the estimated coefficients in the model. We begin by discussing the interpretation of the coefficients for nominal (Section 4.2) and continuous (Section 4.3) scale covariates. In Section 4.4 we discuss the issues of statistical adjustment and the interpretation of estimated coefficients in the presence of statistical interaction. The chapter concludes with a discussion of the interpretation of fitted values from the model and covariate adjusted survivorship functions.

In any regression model, the estimated coefficient for a covariate represents the rate of change of a function of the dependent variable per-unit change in the covariate. Thus, to provide a correct interpretation of the coefficients, we must determine the functional relationship between the independent and dependent variables, and we must define the unit change in the covariate that is likely to be of interest.

In Chapter 3 we recommended that the hazard function be used in regression analysis to study the effect of one or more covariates on survival time. The first step in the process of interpreting the coefficients is to determine what transformation of the hazard function is linear in the coefficients. In the family of generalized linear models (i.e., linear, logistic, Poisson and other regression models) this linearizing transformation is known as the *link function* [see McCullagh and Nelder (1989)]. This same terminology can be applied to proportional hazards regression models.

The proportional hazards model can be used when the primary goal of the analysis is to estimate the effect of study variables on survival time. Suppose, for the moment, that we have a regression model containing a single covariate. Since the hazard function for the proportional hazards regression model is

$$h(t,x,\beta)=h_0(t)e^{x\beta},$$

it follows that the link function is the natural log transformation. We denote the log of a hazard function as $g(t,x,\beta)=\ln[h(t,x,\beta)]$. Thus, in the case of the proportional hazards regression model, the log-hazard function is

$$g(t,x,\beta)=\ln[h_0(t)]+x\beta. \tag{4.1}$$

The difference in the log-hazard function for a change from $x=a$ to $x=b$ is

$$\begin{aligned}[g(t,x=a,\beta)-g(t,x=b,\beta)]&=\{\ln[h_0(t)]+a\beta\}-\{\ln[h_0(t)]+b\beta\}\\&=a\beta-b\beta\\&=(a-b)\beta.\end{aligned} \tag{4.2}$$

Note that, since the baseline hazard function, $h_0(t)$, appears in both log hazards, it subtracts itself out. Thus, the difference in the log hazards does not depend on time. This critical *proportional hazards* assumption is examined in detail in Chapter 6, when we discuss methods for assessing model adequacy and assumptions.

The log hazard is the correct function to use to assess the effect of change in a covariate. However, it is not as easily interpreted as the expression we obtain when we exponentiate (4.2), namely

$$\begin{aligned}\mathrm{HR}(t,a,b,\beta)&=\exp[g(t,x=a,\beta)-g(t,x=b,\beta)]\\&=\frac{h(t,a,\beta)}{h(t,b,\beta)}\\&=e^{(a-b)\beta}.\end{aligned} \tag{4.3}$$

The quantity defined in (4.3) is the hazard ratio, and it plays the same role in interpreting and explaining the results of a survival analysis that

the odds ratio plays in a logistic regression.[1] We return to this point in the next section.

The results in (4.2) and (4.3) are important as they provide the method that must be followed to interpret the coefficients in any proportional hazards regression model correctly. The presence of censored observations of survival time in the data does not alter the interpretation of the coefficients. Censoring is an estimation issue that was dealt with when we constructed the partial likelihood function, see (3.17). Once we have accounted for the censoring, we can ignore it.

4.2 NOMINAL SCALE COVARIATE

We begin by considering the interpretation of the coefficient for a dichotomous covariate. Dichotomous or binary covariates occur regularly in applied settings. They may be truly dichotomous (e.g., gender) or they may be derived from continuous covariates (e.g., age greater than 40 years).

Assume for the moment that we have a model containing a single dichotomous covariate, denoted X, coded 0 or 1. Following the procedure described in (4.2), the first step in interpreting the coefficient for X is to calculate the difference in the log hazard for a one unit change in the covariate. This yields

$$g(t,1,\beta)-g(t,0,\beta)=(1-0)\beta=\beta.$$

Thus, in the special case when the dichotomous covariate is coded zero and one, the coefficient is equal to the change of interest in the log hazard. We can exponentiate, following (4.3), the value of the difference in log hazards to obtain the hazard ratio

$$\mathrm{HR}(t,1,0,\beta)=e^{\beta}. \tag{4.4}$$

The form of the hazard ratio in (4.4) is identical to the form of the odds ratio from a logistic regression model for a dichotomous covariate. The difference is that, in the current context, it is a ratio of rates rather than of odds. In order to expand on this difference, suppose that we followed a large cohort of males and females for 5 years and noted

---

[1] See Hosmer and Lemeshow (1989) Chapter 3 for a detailed discussion of the interpretation of the coefficients in a logistic regression model.

whether a subject "died" during this period of time. In this hypothetical setting one might be tempted to analyze the end-of-study binary variable, death (yes = 1), using a logistic regression model. One should note that this binary variable is what we have defined as the censoring variable for the observation of time to death.[2] Suppose the value of the odds ratio for X, denoting gender (1 = male), is 2.0. This is interpreted to mean, under conditions where the odds-ratio approximates the relative risk, that the probability of death by the end of the study is 2 times higher for a male than for a female. A hazard ratio of 2 obtained from (4.4) means that, at any time during the study, the per-unit time rate of death among males is twice that of females. Thus, the hazard ratio is a comparative measure of survival experience over the entire time period, whereas the odds-ratio is a comparative measure of event occurrence only at the study endpoint. They are two different measures, and the fact that they may be of similar magnitude in an applied setting is, in a sense, irrelevant. Note that if one is able to observe the survival time for all subjects, logistic regression cannot be used at all.

In order to illustrate further the interpretation of the hazard ratio for a dichotomous covariate, survival times were created for a hypothetical cohort of 10,000 subjects, with 5,000 in each of two groups and a theoretical hazard ratio of 2.0. Subjects whose survival time exceeded 60 months were considered censored at 60 months. Each month the number at risk, the number of deaths, the estimated hazard rates, $h_k(t) = d_k(t)/n_k(t), k = 0,1$ and ratio, $\mathrm{HR}(t) = h_1(t)/h_0(t)$ were computed. These quantities are listed in Table 4.1 for the first 12 months and the last 13 months of this study. The hazard ratio is graphed for the entire study period in Figure 4.1. The average estimated hazard ratio, $\overline{\mathrm{HR}} = (1/58) \times \sum \mathrm{HR}(t)$, has been added to Figure 4.1. The hazard rates and their ratios indicate that, during each month of the 60 months of follow-up, the death rate for group 1 is approximately twice that seen in group 0. The scatter about 2.0 is due to the randomness in the number of deaths observed at each time.

---

[2] Even though the two regression models can, under certain conditions, yield similar coefficients, see Hosmer and Lemeshow (1989, Chapter 8), we are not suggesting that a logistic regression of the censoring variable be used in place of a proportional hazards regression of survival time. We assume the reader has a clear understanding of the interpretation of coefficients from a logistic regression model and use it only to explain the difference between the interpretation of a coefficient under the two regression models.

The increase in the scatter over time in Figure 4.1 is due to the fact that the number in the risk sets decreases over time. There were no deaths at 58 months in group 1, so the hazard rate for group 1 and the rate ratio cannot be estimated, at least using the same estimator used for the other months. By design of the example, all values of time greater than or equal to 60 months are censored, so the point estimate of each hazard rate is zero and the estimate of the hazard ratio is undefined.

Table 4.1 Partial Listing of the Number of Deaths, the Number at Risk and the Estimated Hazard Rate in Two Hypothetical Groups and the Estimated Hazard Ratio at Time t

t	$d_0(t)$	$n_0(t)$	$h_0(t)$	$d_1(t)$	$n_1(t)$	$h_1(t)$	HR(t)
1	109	5000	0.022	207	5000	0.041	1.9
2	216	4891	0.044	378	4793	0.079	1.79
3	190	4675	0.041	370	4415	0.084	2.06
4	162	4485	0.036	367	4045	0.091	2.51
5	165	4323	0.038	262	3678	0.071	1.87
6	178	4158	0.043	250	3416	0.073	1.71
7	153	3980	0.038	245	3166	0.077	2.01
8	160	3827	0.042	227	2921	0.078	1.86
9	153	3667	0.042	226	2694	0.084	2.01
10	142	3514	0.04	189	2468	0.077	1.9
11	120	3372	0.036	185	2279	0.081	2.28
12	149	3252	0.046	199	2094	0.095	2.07
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
48	28	708	0.04	10	92	0.109	2.75
49	30	680	0.044	5	82	0.061	1.38
50	27	650	0.042	7	77	0.091	2.19
51	26	623	0.042	8	70	0.114	2.74
52	21	597	0.035	3	62	0.048	1.38
53	23	576	0.04	5	59	0.085	2.12
54	25	553	0.045	5	54	0.093	2.05
55	22	528	0.042	2	49	0.041	0.98
56	23	506	0.045	5	47	0.106	2.34
57	22	483	0.046	3	42	0.071	1.57
58	25	461	0.054	0	39	0	*
59	20	436	0.046	2	39	0.051	1.12
60	0	416	0	0	37	0	*

* Estimator undefined.

In most applied settings, there will be too much variability in the pointwise estimators of the hazard rates, $d(t)/n(t)$, for a figure like Figure 4.1 to be particularly informative about the value of the hazard rate or to determine whether it is constant over time. More sophisticated methods are considered in Chapter 6.

Table 4.2 presents the results of fitting the proportional hazards model containing the dichotomous variable for IV drug use in the HMO-HIV+ study. The point estimate of the coefficient is $\hat{\beta} = 0.779$. Since IV drug use was coded as 1 = yes and 0 = no, we know from (4.3) and (4.4) that we can obtain the point estimator of the hazard ratio by exponentiating the estimator of the coefficient. In this example the estimate is

$$\hat{\text{HR}} = e^{0.779} = 2.18.$$

In the case of a dichotomous covariate coded zero and one, the hazard ratio depends only on the coefficient. Like the odds-ratio estimator

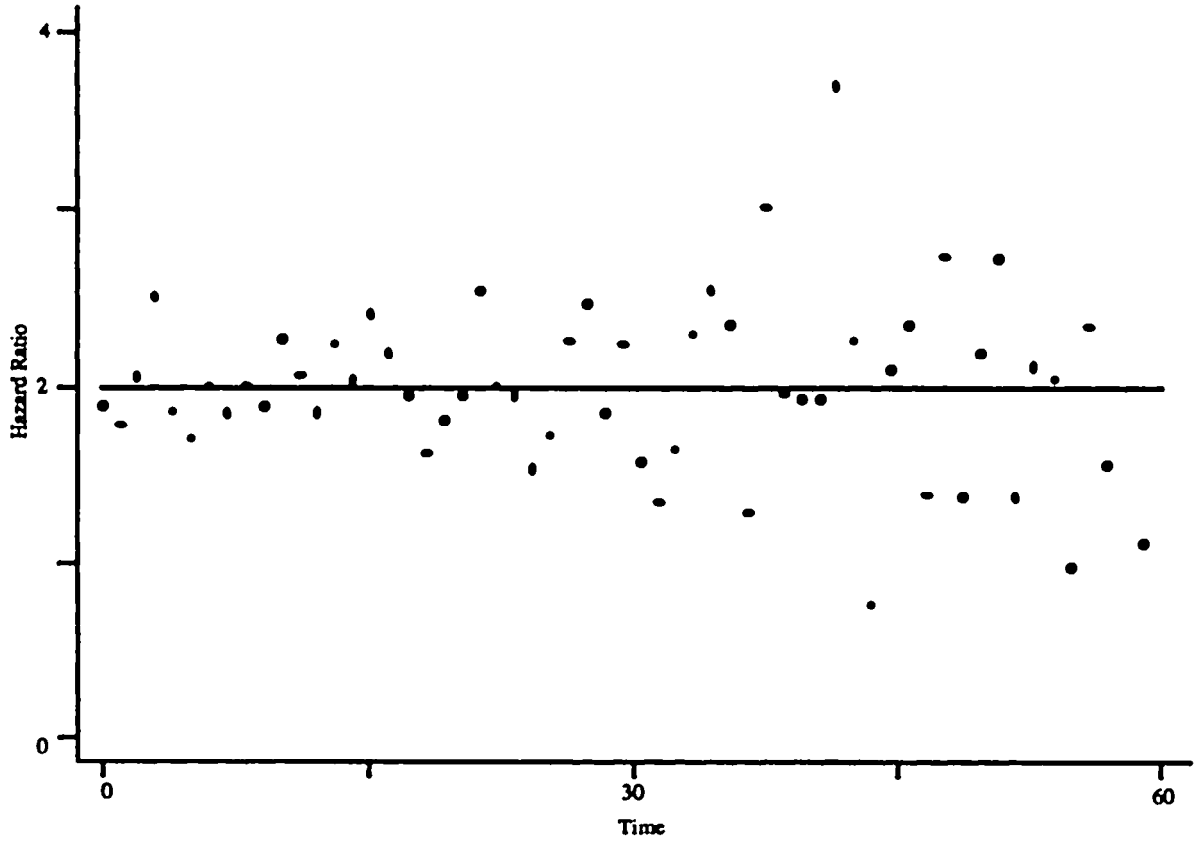


Figure 4.1 Graph of the estimated hazard ratios and the mean hazard ratio ($\overline{\text{HR}} = 2.0$) from Table 4.1.

Table 4.2 Estimated Coefficient, Standard Error, z-Score, Two-Tailed p-Value and 95% Confidence Interval for IV Drug Use from the HMO-HIV+ Study

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% Conf. Int.
DRUG	0.779	0.2422	3.22	0.001	0.304, 1.254

in logistic regression, the sampling distribution of the estimator of the hazard ratio is skewed to the right, so confidence interval estimators based on the Wald statistic and its assumption of normality may not have good coverage properties unless the sample size is quite large. Comparatively speaking, the sampling distribution of the estimator of the coefficient is better approximated by the normal distribution than the sampling distribution of the estimated hazard ratio. As a result, its Wald statistic-based confidence interval will have better coverage properties. In this case, we obtain the endpoints of a 95 percent confidence interval for the hazard ratio by exponentiating the endpoints of the confidence interval for the coefficient. In the current example these are

$$\exp\left[\hat{\beta} \pm 1.96\hat{\mathrm{SE}}\left(\hat{\beta}\right)\right] = \exp[0.779 \pm 1.96 \times 0.2422] = 1.355, 3.504.$$

Alternative confidence interval estimators have been studied, one of which is based on the partial likelihood. To date, this method has not been implemented in most software packages.

The interpretation of the estimated hazard rate of 2.18 is that subjects with a history of IV drug use die at about twice the rate of those without a history of IV drug use, throughout the study period. The confidence interval suggests that ratios as low as 1.4 or as high as 3.5 are consistent with the observed data, at the $\alpha = 0.05$ level.

As discussed in Chapter 3, the partial likelihood ratio test, the Wald test and the score test can be used to assess the significance of a coefficient. In the current example, the value of the partial likelihood ratio test is $G = 10.20$, with a p-value equal to 0.001. The Wald test statistic is $z = 3.22$, with a p-value also equal to 0.001. Both of these tests indicate that the coefficient for IV drug use is significant. One should note that the confidence interval for the hazard ratio does not include 1.0, another indicator of its significance. Most software packages provide output for the coefficients, but some, such as STATA, provide hazard ratios and/or coefficients.

Note that Table 4.2 contains no intercept term. This is the price one pays for choosing the semiparametric proportional hazards model. The intercept, were one present, would correspond to the log baseline hazard function, in this case drug group 0. The implication of this in practice is that we cannot, from the regression output of a proportional hazards model, reconstruct group-specific hazard rates. Only ratios can be estimated. If it is critical to have individual estimates of group-specific hazard rates, then one should use one of the fully parametric models discussed in Chapter 8.

Occasionally, the coded values for a dichotomous variable differ by more than 1 (e.g., +1, −1). In this case, it is not possible to obtain the estimator of the hazard ratio by simply exponentiating the estimator of the coefficient. One can always obtain the correct estimator by explicitly evaluating (4.2) and (4.3). If, as shown in (4.2) and (4.3), the two values are denoted as a and b, then the estimator of the hazard ratio is

$$\hat{\mathrm{HR}}(t,a,b,\beta) = e^{(a-b)\hat{\beta}}. \tag{4.5}$$

The endpoints of a $100(1-\alpha)$ percent confidence interval estimator for the hazard ratio can be obtained by exponentiating the endpoints of the confidence interval estimator for $(a-b)\beta$,

$$\exp\left[(a-b)\hat{\beta} \pm |a-b| z_{1-\alpha/2} \hat{\mathrm{SE}}\left(\hat{\beta}\right)\right], \tag{4.6}$$

where $|a-b|$ denotes the absolute value of $(a-b)$.

If a nominal scale covariate has more than two levels, denoted in general by K, we must model the variable using a collection of $K-1$ "design" (also known as "dummy" or "indicator") variables. The most frequent method of coding these design variables is to use *reference cell coding*. With this method, we choose one level of the variable to be the reference level, against which all other levels are compared. The resulting hazard ratios compare the hazard rate of each group to that of the referent group.

In Chapter 2, we considered an example in which age of subjects in the HMO-HIV+ study was categorized into four groups [20−29], [30−34], [35−39] and [40−54]. Our goal was to describe, qualitatively, how survival experience in the cohort changes with age, through plots of estimated survivorship functions and a log-rank test. We can continue along these same lines by fitting a proportional hazards model

Table 4.3 Coding of the Three Design Variables for the Age Groups in the HMO-HIV+ Study

Age Group	AGE_2	AGE_3	AGE_4
1:[20–29]	0	0	0
2:[30–34]	1	0	0
3:[35–39]	0	1	0
4:[40–54]	0	0	1

to these data. The estimated hazard ratios provide a convenient and easily interpreted summary measure of the comparative survival experience of the four groups.

The methods discussed in this example may be applied to any covariate with multiple groups. The coding for the three design variables based on the four age groups, using the youngest age group as the referent group, are presented in Table 4.3. The results of fitting a proportional hazards model using these three design variables are presented in Table 4.4.

The value of the partial likelihood ratio test for the overall significance of the coefficients is $G = 19.56$ and the p-value, computed using a chi-square distribution with three degrees-of-freedom, is less than 0.001. This suggests that at least one of the three older age groups has a hazard rate that is significantly different from the youngest age group. The p-values of the individual Wald statistics indicate that the hazard rate in each of the three older groups is significantly different from that in the youngest (or reference) age group.

Before we can use (4.2) and (4.3) to obtain estimators of the hazard ratios, we need the equation for the log-hazard function. The log-hazard function, ignoring the log baseline hazard function, for the model fit in Table 4.4 is

$$g(t, \text{AGE\_GRP}, \boldsymbol{\beta}) = \beta_1 \text{AGE\_2} + \beta_2 \text{AGE\_3} + \beta_3 \text{AGE\_4}.$$

The estimator of the hazard ratio comparing age group 2 to age group 1 is obtained by first calculating the difference in the estimators of the log-hazard functions, (4.2),

$$\left[g\left(t, \text{AGE\_GRP} = 2, \hat{\boldsymbol{\beta}}\right) - g\left(t, \text{AGE\_GRP} = 1, \hat{\boldsymbol{\beta}}\right)\right]$$
$$= \left(\hat{\beta}_1 1 + \hat{\beta}_2 0 + \hat{\beta}_3 0\right) - \left(\hat{\beta}_1 0 + \hat{\beta}_2 0 + \hat{\beta}_3 0\right) = \hat{\beta}_1 .$$

Exponentiating the result, we obtain

$$\widehat{\mathrm{HR}}(2,1) = e^{\hat{\beta}_1} .$$

We obtain the estimators of the other two hazard ratios by proceeding in a similar manner, and these are

$$\widehat{\mathrm{HR}}(3,1) = e^{\hat{\beta}_2}$$

and

$$\widehat{\mathrm{HR}}(4,1) = e^{\hat{\beta}_3} .$$

We calculate the value of the estimates in the example, shown in the second column of Table 4.5, by exponentiating the values of the coefficients, from Table 4.4.

When reference cell coding is used to create the design variables, the estimators of the hazard ratio comparing each group to the referent group are obtained by exponentiating the respective estimators of the coefficients.

We construct confidence interval estimators of the hazard ratios by exponentiating the endpoints of the confidence intervals for the individual coefficients. For example, the endpoints of the 95 percent confidence interval estimate for HR(2,1) shown in Table 4.5 are

$$\exp\left[\hat{\beta}_1 \pm 1.96\widehat{\mathrm{SE}}\left(\hat{\beta}_1\right)\right] = \exp[1.197 \pm 1.96 \times 0.4520] = 1.37, 8.01.$$

Similar calculations yield the endpoints for the other two confidence interval estimates.

The hazard ratios in Table 4.5 suggest: (1) subjects in their early thirties are dying at a rate which is about 3.3 times greater than subjects in their twenties, (2) subjects in their late thirties are dying at a rate

Table 4.4 Estimated Coefficients using Referent Cell Coding, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for Age Categorized into Four Groups from the HMO-HIV+ Study

Variable	Coeff.	Std. Err.	*z*	*P*>|z|	95% Conf. Int.
AGE_2	1.197	0.451	2.65	0.008	0.313, 2.081
AGE_3	1.313	0.459	2.86	0.004	0.414, 2.213
AGE_4	1.860	0.469	3.96	<0.001	0.941, 2.780

which is about 3.7 times greater than subjects in their twenties and (3) subjects 40 or older have a mortality rate that is approximately 6 times greater than subjects in their twenties.

Given the similarity of the hazard ratios comparing each of the two age groups [30–34] and [35–39] to the referent group, it would make sense to test whether the survival experience in these two groups differs. We can estimate their hazard ratio and determine whether it is different from 1.0. We do this by using the general approach in (4.2) and (4.3). The specific difference in the hazard functions for the two groups is

$$\begin{aligned}\left[g\left(t,\text{AGE\_GRP}=3,\hat{\boldsymbol{\beta}}\right)-g\left(t,\text{AGE\_GRP}=2,\hat{\boldsymbol{\beta}}\right)\right]&\\ =\left(\hat{\beta}_1 0+\hat{\beta}_2 1+\hat{\beta}_3 0\right)-\left(\hat{\beta}_1 1+\hat{\beta}_2 0+\hat{\beta}_3 0\right)&\\ =\hat{\beta}_2-\hat{\beta}_1\ .&\end{aligned}$$

The estimator of the hazard ratio is

$$\widehat{\text{HR}}(3,2)=e^{\left(\hat{\beta}_2-\hat{\beta}_1\right)},$$

and its estimate is $\exp(1.313-1.197)=1.123$. In order to obtain a confidence interval, we need an estimator for the variance of the difference between the two coefficients. The variance of the difference between two variables is

$$\hat{\text{Var}}(\hat{\beta}_2-\hat{\beta}_1)=\hat{\text{Var}}(\hat{\beta}_2)+\hat{\text{Var}}\left(\hat{\beta}_1\right)-2\hat{\text{Cov}}\left(\hat{\beta}_1,\hat{\beta}_2\right),$$

where $\hat{\text{Var}}$ denotes the estimator of the variance of the estimator in the parentheses and $\hat{\text{Cov}}$ denotes the estimator of the covariance of the two

Table 4.5 Estimated Hazard Ratios (HR) and 95% Confidence Intervals for Age Categorized into Four Groups from the HMO-HIV+ Study

Age Group	HR	95% Conf. Int.
1:[20–29]	1.00	
2:[30–34]	3.31	1.37, 8.01
3:[35–39]	3.72	1.51, 9.14
4:[40–54]	6.43	2.56, 16.12

estimators in the parentheses. These estimates may be obtained from most software packages by requesting the estimated covariance matrix of the estimated coefficients. Table 4.6 presents the covariance matrix for the estimated coefficients for the three age groups.

The estimated variances and covariance needed are $\hat{\text{Var}}(\hat{\beta}_1) = 0.2034$, $\hat{\text{Var}}(\hat{\beta}_2) = 0.2106$ and $\hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) = 0.1637$. The estimate of the variance of the difference in the two coefficients is

$$\hat{\text{Var}}(\hat{\beta}_2 - \hat{\beta}_1) = 0.2034 + 0.2106 - 2 \times 0.1637 = 0.0867,$$

and the estimated standard error is

$$\hat{\text{SE}}(\hat{\beta}_2 - \hat{\beta}_1) = 0.2945.$$

The endpoints of the 95 percent confidence interval estimate are

$$\begin{aligned} &\exp\left[(\hat{\beta}_2 - \hat{\beta}_1) \pm 1.96 \hat{\text{SE}}(\hat{\beta}_2 - \hat{\beta}_1)\right] \\ &\quad = \exp[(1.313 - 1.197) \pm 1.96 \times 0.2945] \\ &\quad = 0.63,\ 2.00. \end{aligned}$$

The confidence interval includes 1.0, indicating that the hazard rates for the two age groups may in fact be the same.

Instead of using the confidence interval, we could test the hypothesis of the equality of two coefficients via a Wald test. Many software packages allow the user to test whether specified contrasts of model coefficients are equal to zero. This is a convenient feature, especially when contrasts of interest are more complicated than simple differences. The Wald test for the contrast $\hat{\beta}_2 - \hat{\beta}_1$ is

$$z = \frac{\hat{\beta}_2 - \hat{\beta}_1}{\hat{\text{SE}}(\hat{\beta}_2 - \hat{\beta}_1)} = \frac{1.313 - 1.197}{0.2945} = 0.395,$$

and the two-tailed p-value computed from the standard normal distribution is 0.69. Since the p-value is large, greater than 0.05, we fail to reject the hypothesis that the two coefficients are equal and conclude that the death rates in the two age groups may not be different.

Table 4.6 Estimated Variances and Covariances for the Three Estimated Coefficients in Table 4.4

Variable (Coeff.)	AGE_2	AGE_3	AGE_4
AGE_2	0.2034	0.1637	0.1705
AGE_3	0.1637	0.2106	0.1666
AGE_4	0.1705	0.1666	0.2203

The test for a general contrast among the $K-1$ coefficients for a nominal scaled covariate with K levels is described as follows. Let the vector of estimators of the coefficients be denoted

$$\hat{\boldsymbol{\beta}}' = \left(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_{K-1}\right)$$

and the estimator of the covariance matrix be denoted $\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}\right)$. Let the vector of constants specifying the contrast be denoted

$$\mathbf{c}' = \left(c_1, c_2, \ldots, c_{K-1}\right),$$

where the sum of the constants is zero. The single degree-of-freedom Wald test for the contrast is

$$Q = \frac{\mathbf{c}'\hat{\boldsymbol{\beta}}}{\sqrt{\mathbf{c}'\hat{\mathbf{V}}\left(\hat{\boldsymbol{\beta}}\right)\mathbf{c}}}, \tag{4.7}$$

and the two-tailed p-value is obtained using the standard normal distribution. Most software packages will report the square of the Wald test and use the chi-square distribution to calculate the p-value. The equivalence of these two approaches follows from the fact that the distribution of the square of a $N(0,1)$ random variable follows a $\chi^2(1)$ distribution.

In the HMO-HIV+ study, it may be of interest to determine whether the average hazard ratio of the middle two age groups is equal to the hazard ratio for the oldest age group. The vector of constants for this contrast is $\mathbf{c}' = (0.5, 0.5, -1)$, the vector of estimated coefficients is given in Table 4.4, and the covariance matrix is shown in Table 4.6. We used STATA to perform the calculations, but other software packages, for example, SAS, could have been used. The value of the test statistic is $Q = 2.31$ with a p-value equal to 0.021. We conclude that the oldest age

group has a hazard rate that is significantly greater than the average rate of the middle two age groups.

The method of using a contrast to compare coefficients can be especially useful when trying to pool categories of a nominal scale covariate recorded with more levels than can be practically used. Practical considerations are of primary importance in deciding which categories to combine, but contrasts may be used to judge whether the hazard rates of clinically similar groups are statistically similar.

Referent cell coding is the most frequently used scheme for coding design variables; however, it is just one of many possible methods. An alternative is deviation from means coding. This type of design variable coding may be used when one simply needs an overall assessment of differences in hazard rates. To illustrate the method, we apply it to the four age groups in the HMO-HIV+ study. This coding is obtained by replacing the first row of zeros in Table 4.3 with a row in which each value is equal to −1. The resulting estimated coefficient for an age group estimates the difference between the log hazard of the group and the arithmetic mean of the log hazards. The exponentiated estimated coefficient provides the ratio of the hazard rate of the particular group to the geometric mean of the hazard rates of all K groups.

The results of fitting a proportional hazards model using the deviation from means coding are shown in Table 4.7. The value of the partial likelihood ratio test for the overall significance of the coefficients is identical to that obtained using reference cell coding and is $G = 19.56$ with a p-value, computed using a chi-square distribution with three degrees-of-freedom, less than 0.001. The value of 0.104 for the estimated coefficient of design variable AGE_2 is equal to the estimate of the difference between the log-hazard rate for age group 2, [30, 34], and the estimate of the mean log-hazard rate. The Wald statistic has a p-value of 0.589, indicating that the log-hazard rate for this age group may not differ significantly from the average log-hazard rate. The coefficient for group 4 is 0.768 and its Wald statistic has a p-value less than 0.001. Thus, the log-hazard rate for this age group is significantly larger than the average log-hazard rate.

The coefficients in Table 4.7 are all positive, indicating that the average log-hazard rate falls between the log-hazard for age groups one and two. The estimated difference between the log hazard for the first age group and the average log-hazard rate is the negative of the sum of the coefficients in Table 4.7 and is −1.093. The easiest way to obtain an estimate of its standard error, Wald statistic, etc., is to make a small

Table 4.7 Estimated Coefficients Using the Deviation from Means Coding, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for Age Categorized into Four Groups from the HMO-HIV+ Study

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% Conf. Int.
AGE_2	0.104	0.192	0.54	0.589	–0.272, 0.481
AGE_3	0.221	0.206	1.07	0.285	–0.183, 0.624
AGE_4	0.768	0.209	3.67	<0.001	0.357, 1.178

change in the coding of the design variables and refit the model. We merely switch the row coded –1 with any other row. We do not recommend that hazard ratios be reported when using deviation from means coding, because the ratio cannot be interpreted in the same manner as the ratio from referent cell coding. The comparison is not a comparison of two distinct groups, but rather of one group to the geometric mean hazard rate of all groups combined.

Many other methods for coding design variables are possible. For example, coding that compares each group to the next largest group or each group to the average of the higher groups. These methods tend to be appropriate in special circumstances and will not be discussed further in this text. In general, the method of referent cell coding, perhaps followed by contrasts, should provide a useful and informative analysis in most circumstances.

4.3 CONTINUOUS SCALE COVARIATE

The interpretation of the coefficient for a continuous covariate is easier than that of a nominal scale variable in one sense, since indicator variables need not be introduced, but more difficult in another sense. Before we can use (4.2) and (4.3) to obtain an estimator of a hazard ratio we must do two things. First and foremost, we must verify that we have included the variable in its correct scale in the model. In this section we will assume that the log hazard is linear in the covariate of interest. Methods to assess the scale are discussed in Chapter 5. Second, we must decide what a clinically meaningful unit of change in the covariate is. Once these two steps are accomplished we may apply (4.2) and (4.3).

We illustrate the method using the HMO-HIV+ study and age as the covariate. The results of fitting a proportional hazards model containing age are shown in Table 4.8. The estimated coefficient in Table 4.8

gives the change in the log hazard for a 1-year change in age. Often a 1-year change in age is not of clinical interest. The HMO physicians conducting the study may be more interested in a 5-year change in age.

We obtain the correct change in the log-hazard function for a change of c units in a continuous covariate by using (4.2) and (4.3) with $a = x + c$ and $b = x$. This yields the following change in the log hazard:

$$\begin{aligned}\left[g(t, x+c, \beta) - g(t, x, \beta)\right] &= \left\{\ln\left[h_0(t)\right] + (x+c)\beta\right\} - \left\{\ln\left[h_0(t)\right] + x\beta\right\} \\ &= (x+c)\beta - x\beta \\ &= c\beta \end{aligned} \tag{4.8}$$

The change is simply equal to the value of the change of interest times the coefficient for a one-unit change. The estimator of the hazard ratio is

$$\hat{\mathrm{HR}}(c) = e^{c\hat{\beta}} \tag{4.9}$$

and the endpoints of a $100(1-\alpha)$ percent confidence interval estimator of the hazard ratio are

$$\exp\left[c\beta \pm z_{1-\alpha/2}|c|\hat{\mathrm{SE}}\left(\hat{\beta}\right)\right]. \tag{4.10}$$

Applying (4.9) and (4.10) for a 5-year change in age in the HMO-HIV+ study, we obtain an estimated hazard ratio of

$$\hat{\mathrm{HR}}(5) = e^{5\times 0.081} = 1.50$$

and the endpoints of a 95 percent confidence interval are

$$\exp[5\times 0.081 \pm 1.96\times 5\times 0.0174] = 1.264,\ 1.778.$$

Alternatively, we could have calculated the endpoints of the 95 percent confidence interval by multiplying the endpoints in Table 4.8 by 5 and then exponentiating. We suggest, for continuous covariates, that the hazard ratio for the clinically interesting unit of change, along with its confidence interval, be reported in any table of results. The unit of change should be indicated in the table heading or in a footnote.

Table 4.8 Estimated Coefficient, Standard Error, z-Score, Two-Tailed p-Value and 95% Confidence Interval for Age in the HMO-HIV+ Study

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% Conf. Int.
AGE	0.081	0.0174	4.67	<0.001	0.047, 0.116

The interpretation of an estimated hazard ratio of 1.5 is that the hazard rate increases by 50 percent for every 5-year increase in age and is independent of the age at which the increase is calculated. The independence of the increase in age is due to the fact that the log hazard was assumed to be linear in age and subtracts itself out of the calculation in (4.8). The confidence interval estimate suggests that an increase in the hazard rate of between 30 and 80 percent is consistent with the data.

In summary, we wish to emphasize that the interpretation of the estimated hazard ratio for a continuous covariate depends not only on the assumption of linearity in the log hazard but also on the basic premise of a proportional hazards model. Methods for checking these assumptions are considered in detail in Chapters 5 and 6, respectively.

4.4 MULTIPLE-COVARIATE MODELS

The primary asset of any regression model is its ability to include multiple covariates and thereby statistically adjust for possible imbalances in the observed data before making statistical inferences. This process of adjustment has been given different names in various fields of study. In traditional statistical applications it is called *analysis of covariance*, while in clinical and epidemiological investigations it is often called *control of confounding*. A statistically related issue is the inclusion of higher order terms in a model representing interactions between covariates. These are also called *effect modifiers*. The strengths and limitations of statistical adjustment and inclusion of interactions in generalized linear models apply when using the proportional hazards regression model to analyze survival time. In this section we discuss these issues and establish a set of basic guidelines that we employ when discussing model development in the next chapter.

The UMARU IMPACT Study (UIS) is introduced in Section 1.3 (Table 1.3) and provides some excellent examples for demonstrating the statistical issues involved in adjustment and interaction. The analyses presented in this section are in no way definitive. They are used

simply to demonstrate the interpretation of fitted proportional hazards models. Two variables collected on subjects in the UIS were age and IV drug use history. For demonstration purposes, we have recoded IV drug use history into a dichotomous variable, d, coded 0 = never and 1 = ever. We assume the log-hazard function is linear in the covariates and that our primary analysis goal is to estimate the hazard ratio associated with IV drug use, d.

Suppose we generate a proportional hazards model that contains only IV drug use. The log-hazard function of the model is

$$g(t,d,\theta_1) = \ln[h_0(t)] + d\theta_1 .$$

The difference in the log-hazard functions is

$$\begin{aligned}[g(t,d=1,\theta_1) - g(t,d=0,\theta_1)] &= \{\ln[h_0(t)] + 1\theta_1\} - \{\ln[h_0(t)] + 0\theta_1\} \\ &= \theta_1 . \end{aligned} \tag{4.11}$$

Suppose we generate a second model that contains both age and IV drug use. The log-hazard function for the larger model is

$$g(t,d,a,\boldsymbol{\beta}) = \ln[h_0(t)] + d\beta_1 + a\beta_2 , \tag{4.12}$$

where a denotes age. The adjusted log-hazard ratio is obtained from (4.2) and (4.12), comparing a subject of age a who has a history of IV drug use to one of the same age a who does not have a history of IV drug use, and is

$$\begin{aligned}&[g(t,d=1,a,\boldsymbol{\beta}) - g(t,d=0,a,\boldsymbol{\beta})] \\ &\quad = \{\ln[h_0(t)] + 1\beta_1 + a\beta_2\} - \{\ln[h_0(t)] + 0\beta_1 + a\beta_2\} \\ &\quad = \beta_1 + (a-a)\beta_2 \\ &\quad = \beta_1 . \end{aligned} \tag{4.13}$$

The results shown in (4.11) and (4.13) indicate that we have two estimators of the desired log-hazard ratio: (1) The so-called crude or unadjusted estimator $\hat{\theta}_1$ from (4.11), obtained from fitting the model that does not include age, and (2) the adjusted estimator $\hat{\beta}_1$ from (4.13), the coefficient of d obtained from fitting a model containing d and age. If the two estimators are similar, then adjustment for age was unneces-

sary, in a statistical sense. If the estimators are different, then adjustment was needed and the variable age is a confounder of the hazard ratio for d. The extent of adjustment, or difference, between $\hat{\theta}_1$ and $\hat{\beta}_1$ is a function of the difference in the distribution of age within the two IV drug use groups and the strength, $\hat{\beta}_2$, of the association between age and survival time.

Suppose that the model containing age, (4.12), is the correct model, and denote the average age of subjects with and without a history of IV drug use as $\bar{a}_1$ and $\bar{a}_0$, respectively. An approximation of the average log-hazard functions [see Fleming and Harrington (1991) page 134 for an exact expression] for the two drug use groups is

$$g(t,d=0,\beta)=\ln[h_0(t)]+\bar{a}_0\beta_2$$

and

$$g(t,d=1,\beta)=\ln[h_0(t)]+\beta_1+\bar{a}_1\beta_2.$$

Taking the difference between these two expressions, the crude or unadjusted log-hazard ratio is approximately

$$\hat{\theta}_1 \approx \hat{\beta}_1+(\bar{a}_1-\bar{a}_0)\hat{\beta}_2. \tag{4.14}$$

Thus, the crude estimator will be approximately equal to the adjusted estimator if the difference in the mean age of the two drug use groups is zero or if the coefficient for age is zero. The two estimators will differ if at least one of the two is large or both are moderate in size. We recommend that the percent change in the adjusted estimate be computed as a measure of the amount of adjustment. The percent change estimator, in general, is defined as

$$\Delta\hat{\beta}\% = 100\frac{\hat{\theta}-\hat{\beta}}{\hat{\beta}}, \tag{4.15}$$

where $\hat{\theta}$ denotes the crude estimator from the model that does not contain the potential confounder and $\hat{\beta}$ denotes the adjusted estimator from the model that does include the potential confounder. We discuss the use of (4.15) in model building in Chapter 5.

If we assume (4.14) is true, then the approximate percent change is

$$\Delta\hat{\beta}_1\% = 100\frac{\hat{\theta}_1 - \hat{\beta}_1}{\hat{\beta}_1} \simeq \frac{100(\bar{a}_1 - \bar{a}_0)\hat{\beta}_2}{\hat{\beta}_1}, \qquad (4.16)$$

an expression that isolates the two contributors to adjustment. In practice, one would evaluate only (4.15). The expressions in (4.14) and (4.16) are provided as a tool to assist in explaining why the crude and adjusted estimators could be different.

The results of fitting the two models to the UIS data are shown in Table 4.9. We note that AGE is missing for 5 subjects and DRUG is missing on an additional 18 subjects, so analyses have been restricted to the 605 subjects with complete data. The adjusted estimate for DRUG is 0.44 and the crude estimate is 0.32, a change of

$$\Delta\hat{\beta}_1\% = 100 \times \frac{0.32 - 0.44}{0.44} = -27\%.$$

The reasons the estimate changed are: (1) age is strongly associated with survival time, $p = 0.001$ in Table 4.9, and (2) the mean ages in the two drug use groups are different, 26.64 (never) and 31.05 (ever). Thus, both contributors to confounding on the right-hand side of (4.14) are large. In this example the right-hand side of (4.14) is equal to

$$0.32 = 0.44 - 0.026(31.05 - 26.64),$$

which is nearly identical to the crude estimate in Table 4.9.

A practical question is how large must the percent change in the coefficient be to indicate that we need to include the potential confounder in the model. There are no rules, only suggestions. In practice, we have found that a change greater than 15–20 percent indicates that adjustment is needed.

The ability of the proportional hazards regression model to provide correct adjusted estimates of log-hazard ratios depends on having fit the correct model. In practice, this means that the proportional hazards model is correct and that we have fit a model containing the correct covariates, all of which are scaled correctly. These issues are discussed in detail in Chapters 5 and 6.

The derivation of the adjusted estimator in (4.13) implicitly assumes that the log-hazard ratio is constant for all ages. If this is not the case, then the two variables are said to interact; in other words, age modifies the effect of IV drug use. We address this question by determining

Table 4.9 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for Two Models Fit to the UIS Data

Model	Variable	Coeff.	Std. Err.	z	P>|z|	95% Conf. Int.
Crude	DRUG	0.321	0.0948	3.39	0.001	0.135, 0.507
Adjusted	DRUG	0.439	0.1007	4.36	<0.001	0.242, 0.637
	AGE	-0.026	0.0078	-3.37	0.001	–0.042, –0.011

whether an interaction between the two variables (their product) contributes significantly to the model. The log-hazard function for the interactions model is obtained by adding the product of IV drug use and age to the model in (4.12) and is

$$g(t,d,a,\boldsymbol{\beta}) = \ln[h_0(t)] + d\beta_1 + a\beta_2 + (d \times a)\beta_3 \ . \qquad (4.17)$$

If we assume for the moment that (4.17) is the correct model, then the only way we can obtain the correct expression for the log-hazard ratio for IV drug use is to apply (4.2). The log-hazard ratio is

$$\begin{aligned}
&[g(t,d=1,a,\boldsymbol{\beta}) - g(t,d=0,a,\boldsymbol{\beta})] \\
&= \{\ln[h_0(t)] + 1\beta_1 + a\beta_2 + (1 \times a)\beta_3\} - \{\ln[h_0(t)] + 0\beta_1 + a\beta_2 + (0 \times a)\beta_3\} \\
&= \beta_1 + a\beta_3 \ . \qquad (4.18)
\end{aligned}$$

The implication of the result in (4.18) is that the log-hazard ratio for IV drug use depends on the age of the subject. Conversely, when there is no interaction (i.e., $\beta_3 = 0$), the log-hazard ratios in (4.13) and (4.18) are the same. In general, the reason we include interactions in a model is to better estimate the effects of the covariates since point and interval estimators obtained from (4.13) and (4.18) are different. This will happen only when the interaction term in (4.18) is statistically significant, as assessed via the partial likelihood ratio or an equivalent test. We recommend inclusion of interaction terms in a model only when they are statistically significant. We address this point in greater detail in the next chapter.

The primary goal of the UIS was to compare the effectiveness of two treatment interventions, "TREAT" in Table 1.3. Table 4.10 presents the results of fitting a proportional hazards model containing treatment, a second model containing treatment and age and a third

model containing these variables along with their interaction. For illustrative purposes, we use a larger level of statistical significance, $p \leq 0.15$, than we might choose to use in an actual model building application. The coefficient for treatment in the crude model is significant. When we add age to the model, the value of the partial likelihood ratio test comparing the new model to the model that contains treatment only is $G = 3.42$ with a p-value equal to 0.064, and the percent change in the coefficient for treatment is –1.5 percent. Thus, we conclude that age is associated with survival time but is not a confounder of the treatment effect. The partial likelihood ratio test comparing the age adjusted model to the interactions model is $G = 2.57$ with a p-value equal to 0.109. Thus, from a statistical significance point of view, the best model is the interactions model. However, it appears from that model that the significant treatment and age effects seen in the adjusted model have disappeared. The estimator of the log-hazard function for the interactions model, ignoring the log baseline hazard function, is

$$g\left(t, \text{TREAT}, \text{AGE}, \hat{\boldsymbol{\beta}}\right) = \hat{\beta}_1 \text{TREAT} + \hat{\beta}_2 \text{AGE} + \hat{\beta}_3 \text{TREAT} \times \text{AGE}.$$

The estimators of the log-hazard functions for the two treatment groups are

$$g\left(t, \text{TREAT} = 0, \text{AGE}, \hat{\boldsymbol{\beta}}\right) = \hat{\beta}_2 \text{AGE}$$

and

$$g\left(t, \text{TREAT} = 1, \text{AGE}, \hat{\boldsymbol{\beta}}\right) = \hat{\beta}_1 + \left(\hat{\beta}_2 + \hat{\beta}_3\right)\text{AGE}.$$

The coefficient for age, $\hat{\beta}_2 = -0.002$ in Table 4.10, is the slope in age of the log-hazard function for treatment group 0. The fact that it is not significant implies that age is unrelated to survival time in treatment group 0. The slope in age for treatment group 1 is the sum of the age and interaction coefficients, $\hat{\beta}_2 + \hat{\beta}_3 = -0.002 + (-0.023) = -0.025$ in Table 4.10, and its significance can only be tested using the method of contrasts discussed in Section 4.2. This results in a Wald statistic $z = 2.41$ with a p-value equal to 0.016, which is significant. We conclude, therefore, that there is evidence of a significant association between age and survival time in treatment group 1.

To obtain an estimator of treatment effect we can apply (4.2). The estimator of the difference in log-hazard functions is

$$\left[g\left(t,\text{TREAT}=1,\text{AGE},\hat{\boldsymbol{\beta}}\right)-g\left(t,\text{TREAT}=0,\text{AGE},\hat{\boldsymbol{\beta}}\right)\right]$$
$$=\left\{\ln[h_0(t)]+\hat{\beta}_1+\hat{\beta}_2\text{AGE}+\hat{\beta}_3\text{AGE}\right\}-\left\{\ln[h_0(t)]+\hat{\beta}_2\text{AGE}\right\}$$
$$=\hat{\beta}_1+\hat{\beta}_3\text{AGE}. \qquad (4.19)$$

The magnitude of this estimator depends on the age of the subject. The estimator of the coefficient for treatment, $\hat{\beta}_1$, would be the estimator of treatment effect for a subject with age zero years. If we had centered the age data by subtracting the mean, then the coefficient, $\hat{\beta}_1$, is the estimator of treatment effect for a subject of age equal to the mean. To display the results of fitting such a model, we recommend that a table be presented containing point and interval estimates of treatment effect for a few key values of age. The point estimator is

$$\widehat{\text{HR}}(\text{TREAT},\text{AGE})=\exp\left(\hat{\beta}_1+\hat{\beta}_3\text{AGE}\right),$$

and the confidence interval estimator is

$$\exp\left[\left(\hat{\beta}_1+\hat{\beta}_3\text{AGE}\right)\pm z_{1-\alpha/2}\widehat{\text{SE}}\left(\hat{\beta}_1+\hat{\beta}_3\text{AGE}\right)\right], \qquad (4.20)$$

where

$$\widehat{\text{SE}}\left(\hat{\beta}_1+\hat{\beta}_3\text{AGE}\right)=\left\{\widehat{\text{Var}}\left(\hat{\beta}_1\right)+\text{AGE}^2\widehat{\text{Var}}\left(\hat{\beta}_3\right)+2\text{AGE}\widehat{\text{Cov}}\left(\hat{\beta}_1,\hat{\beta}_3\right)\right\}^{0.5},$$

and the required estimated variances and covariance are obtained from the covariance matrix included in computer output. Table 4.11 contains values of the hazard ratio and associated 95 percent confidence

Table 4.10 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for Three Models Fit to the UIS Data

Model	Variable	Coeff.	Std. Err.	*z*	*P*>|z|	95% Conf. Int.
Crude	TREAT	–0.220	0.089	–2.46	0.014	–0.395, –0.045
Adjusted	TREAT	–0.223	0.089	–2.50	0.013	–0.398, –0.048
	AGE	–0.013	0.007	–1.84	0.066	–0.027, 0.001
Interaction	TREAT	0.523	0.474	1.10	0.271	–0.407, 1.453
	AGE	–0.002	0.010	–0.18	0.861	–0.022, 0.018
	TREAT×AGE	–0.023	0.015	–1.60	0.110	–0.052, 0.005

Table 4.11 Estimated Hazard Ratios (HR) and 95% Confidence Intervals for Treatment Effect in the UIS at Ages 25, 30, 35 and 40

Age	HR	95% Conf. Int.
25	0.944	0.723, 1.234
30	0.841	0.700, 1.011
35	0.749	0.617, 0.909
40	0.667	0.502, 0.886

interval for subjects of age 25, 30, 35 and 40 years.

The estimated hazard ratios in Table 4.11 are all less than one and decrease with age, indicating that the longer treatment period, TREAT = 1, is beneficial or protective for return to drug use and becomes increasingly beneficial the older the subject. The confidence intervals support a significant treatment effect for subjects 35 years and older.

Another form of an interactions model is one that contains continuous covariates that have been transformed, and one would like to estimate hazard ratios in the original measurement scale.

For example, suppose a log-hazard function contains both age and the square of age and we would like an estimate of the hazard ratio for a c year change in age. To obtain the correct expression for the difference in log-hazard functions for a c year change in age we must use (4.2) which yields

$$\begin{aligned}\left[g(t,\mathrm{AGE}+c,\boldsymbol{\beta})-g(t,\mathrm{AGE},\boldsymbol{\beta})\right] &= \left\{\ln\left[h_0(t)\right]+\beta_1(\mathrm{AGE}+c)+\beta_2(\mathrm{AGE}+c)^2\right\} \\ &\quad -\left\{\ln\left[h_0(t)\right]+\beta_1\mathrm{AGE}+\beta_2\mathrm{AGE}^2\right\} \\ &= \beta_1 c+\beta_2\left[2\mathrm{AGE}\times c+c^2\right],\end{aligned}$$

an expression which depends on both the change and the age at which the change is calculated. We obtain point and interval estimators of the hazard ratio by extending the result shown in (4.20). This yields the point estimator

$$\widehat{\mathrm{HR}}(\mathrm{AGE}+c,\mathrm{AGE}) = \exp\left[\hat{\beta}_1 c+\hat{\beta}_2\left(2\mathrm{AGE}\times c+c^2\right)\right], \qquad (4.21)$$

and the endpoints of the $100(1-\alpha)$ percent confidence interval are

$$\exp\left\{\hat{\beta}_1 c + \hat{\beta}_2(2\text{AGE} \times c + c^2) \pm z_{1-\alpha/2}\hat{\text{SE}}\left[\hat{\beta}_1 c + \hat{\beta}_2(2\text{AGE} \times c + c^2)\right]\right\}, \tag{4.22}$$

where

$$\hat{\text{SE}}\left[\hat{\beta}_1 c + \hat{\beta}_2(2\text{AGE} \times c + c^2)\right] = \left[\begin{array}{l} c^2\hat{\text{Var}}(\hat{\beta}_1) + (2\text{AGE} \times c + c^2)^2\hat{\text{Var}}(\hat{\beta}_2) \\ \qquad + 2c \times (2\text{AGE} \times c + c^2)\hat{\text{Cov}}(\hat{\beta}_1, \hat{\beta}_2) \end{array}\right]^{0.5}.$$

Expressions similar to (4.21) and (4.22) result when (4.2) is applied to other nonlinear transformations of a covariate.

Multiple variable proportional hazards regression models can be effective tools for sharpening estimates of hazard ratios for covariates. In the absence of interactions, one must be aware at each step in the model evaluation process of the amount of adjustment or confounding that is being controlled. If interactions are present in a model, then confounding is no longer an issue, as the estimate of effect depends on the value of other covariates and thus cannot be removed. In all cases, the interpretation depends on the assumption that the proposed model fits the data, the subject of Chapter 6.

4.5 INTERPRETATION AND USE OF THE COVARIATE-ADJUSTED SURVIVORSHIP FUNCTION

Methods for estimating the survivorship function following the fitting of a proportional hazards model were presented in Section 3.5. The key step presented in that section was the estimation of the baseline survivorship function, $\hat{S}_0(t)$, shown in (3.39). This estimator may be combined with the estimators of the coefficients in the model using (3.36) to obtain the estimator of the survivorship function, adjusting for the covariates, as follows:

$$\hat{S}(t, \mathbf{x}, \hat{\boldsymbol{\beta}}) = [\hat{S}_0(t)]^{\exp(\mathbf{x}'\hat{\boldsymbol{\beta}})}. \tag{4.23}$$

All software packages allow the user to request calculation of the estimator of the baseline survivorship function. The estimator may be used to derive other functions of survival time, for example, the estimator in (4.23), which is essential for graphical description of the results of the analysis and for other analyses, such as model assessment. We discuss graphical methods and estimation of quantiles and their interpretation in this section. Model assessment is discussed in Chapter 6.

We begin with the model containing IV drug use in the HMO-HIV+ study discussed in Section 4.2. In this section we use a dichotomous grouping variable, but the methods may be used with any nominal scale covariate. Table 4.2 presents the results of fitting the model. The estimator of the baseline survivorship function for this model is an estimator of the survivorship function for DRUG = 0. If we request that the baseline survivorship function be computed as part of the analysis, then the software evaluates (3.39), denoted

$$\hat{S}_0(t_i), \quad i = 1,2,\ldots,n, \tag{4.24}$$

for each subject in the study, regardless of their survival status or value of IV drug use. It follows from (3.43) that the estimator $\hat{S}_0(t)$ is constant between observed survival times. Thus, the estimated value for subjects who were censored is equal to the value at the largest observed survival time for which they were still at risk.

We can compute an estimate of the survivorship function for DRUG = 1 by using the previously calculated value of the baseline survivorship function and evaluating

$$\hat{S}\left(t_i, \text{DRUG} = 1, \hat{\beta}_1 = 0.779\right) = \left[\hat{S}_0(t_i)\right]^{\exp(0.779)}, \quad i = 1,2,\ldots,100, \tag{4.25}$$

where the value of the coefficient for DRUG is obtained from Table 4.2. The graphs of the two estimated survivorship functions, (4.24) and (4.25), are shown in Figure 4.2. The plot has been drawn with steps connecting the points rather than straight lines to emphasize the fact that the estimator is constant between observed survival times. It follows from (4.24) and (4.25) that each function has been plotted at exactly the same $n = 100$ values of time. The shape of the two curves is a consequence of the proportional hazards assumption. The ratio of the hazards at each point in Figure 4.2 is forced to be equal to 2.18 = exp(0.779).

We presented a plot in Chapter 2, Figure 2.7, that is similar in appearance to Figure 4.2. There is an important distinction. The two curves in Figure 2.7 are based on separate, nonparametric Kaplan–Meier estimators. The Kaplan–Meier estimator uses only the data in each group and does not assume the hazards are proportional. The distinction between the curves in Figure 2.7 and Figure 4.2 is analogous to the distinction between a plot of the observed cumulative percent distribution as compared to a plot of the cumulative distribution function based on an assumption of normality (i.e., using the observed sample mean and variance). If the data are nearly normally distributed, the two curves will look alike, but the latter curve will be "smoother" (due to the normality assumption) than the former curve.

The difference between the curves in Figures 2.7 and 4.2 can be most clearly seen between 15 and 56 months. The lower curve in Figure 2.7 is constant between 15 and 56 months. No survival times in the IV drug use present group were observed in this interval and the largest observed time was a censored observation at 56 months. The lower curve in Figure 4.2, however, has jumps at each observed survival time,

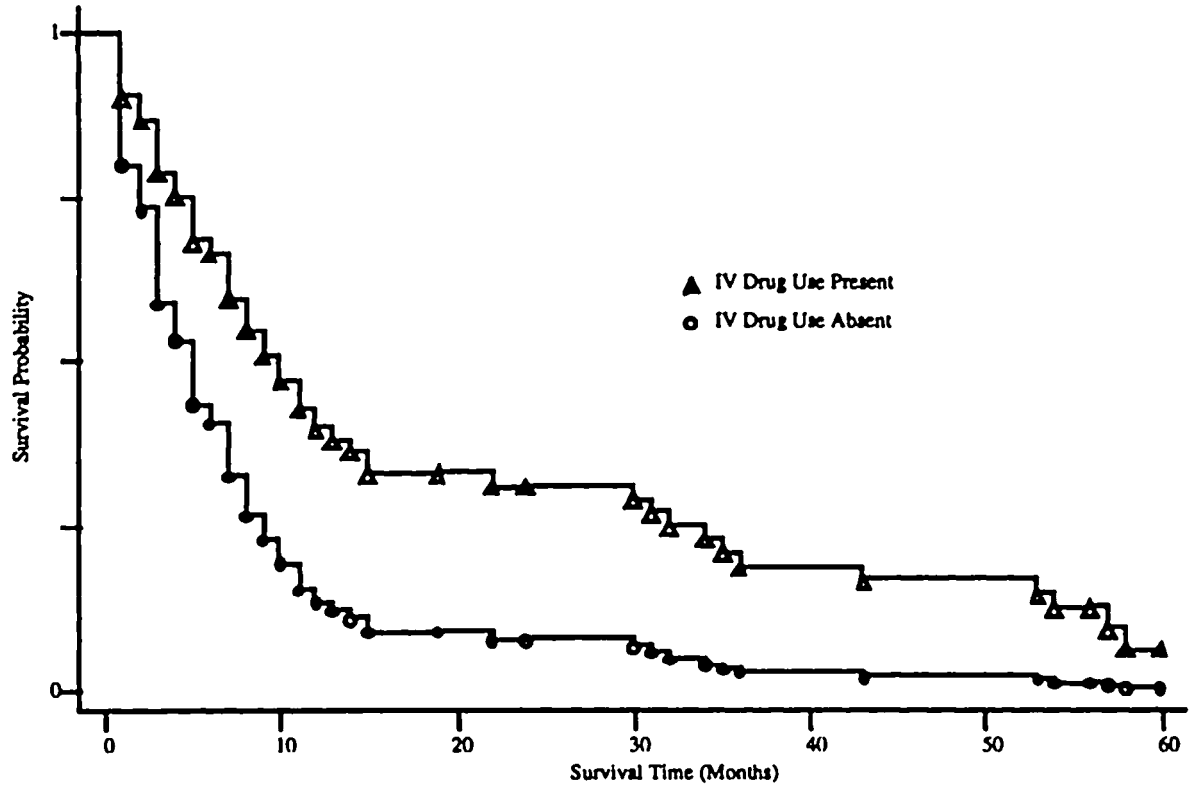


Figure 4.2 Graph of the estimated proportional hazards model survivorship function for IV drug use present (Δ) and absent (o) for the HMO-HIV+ study. Points are plotted at each observed time for both curves.

in the DRUG = 0 group between 15 and 56 months. Furthermore, the lower curve in Figure 4.2 does not end at 56 months with the value at 15 months, as was the case in Figure 2.7. The curve takes a downward jump as the hazard ratio is 2.18 at each point.

Another way one can think of the curves plotted in Figure 4.2 is to consider them as being like "fitted" or "predicted" regression lines. Here the "prediction" is on the survivorship probability scale. In this example, there is an implicit model-based extrapolation present in Figure 4.2. The lower curve, in the interval from 15 to 56 months, predicts or estimates the survivorship experience if: (1) the estimate of the baseline survivorship function correctly describes the survivorship experience in the IV drug use absent group, and (2) the proportional hazards model is correct.

The situation in Figure 4.2 is analogous to using linear regression to model weight as a function of height in males and females. It is likely that the shortest subjects are female and the tallest subjects are male. Once a model has been fit, the software may be used to graph the fitted model over the entire observed range of heights. A point on the line for females in the range of heights only observed for males is a prediction that depends on the unverifiable assumption that the fitted model is correct for females as tall as the tallest males. We have extrapolated the model beyond the observed range of data. The same type of extrapolation can occur in plots of survivorship functions.

As noted, the extrapolation in Figure 4.2 is in an interval between observed values. A more serious extrapolation problem would have occurred if the largest observed time in the IV drug use present group had been 12 months. These extrapolation issues suggest that one must give careful consideration to what points are used when plotting a covariate-adjusted survivorship function. A more conservative plot than Figure 4.2 is shown in Figure 4.3 where points are plotted only for observed values of time. The plot in Figure 4.2 has 200 plotted values while there are 100 plotted values in Figure 4.3.

The plot in Figure 4.3 is constant for the IV drug use present group between 15 and 56 months and thus better reflects the observed data. Figure 4.3, in conjunction with the analysis in Table 4.2, illustrates the significantly poorer survival experience of the IV drug use present group.

If the observed range of survival times is comparable for each group, we recommend using a plot like Figure 4.2 as it uses all the data and reflects best the fitted model and its assumptions. However, if there are clinically important differences in the observed range of survival

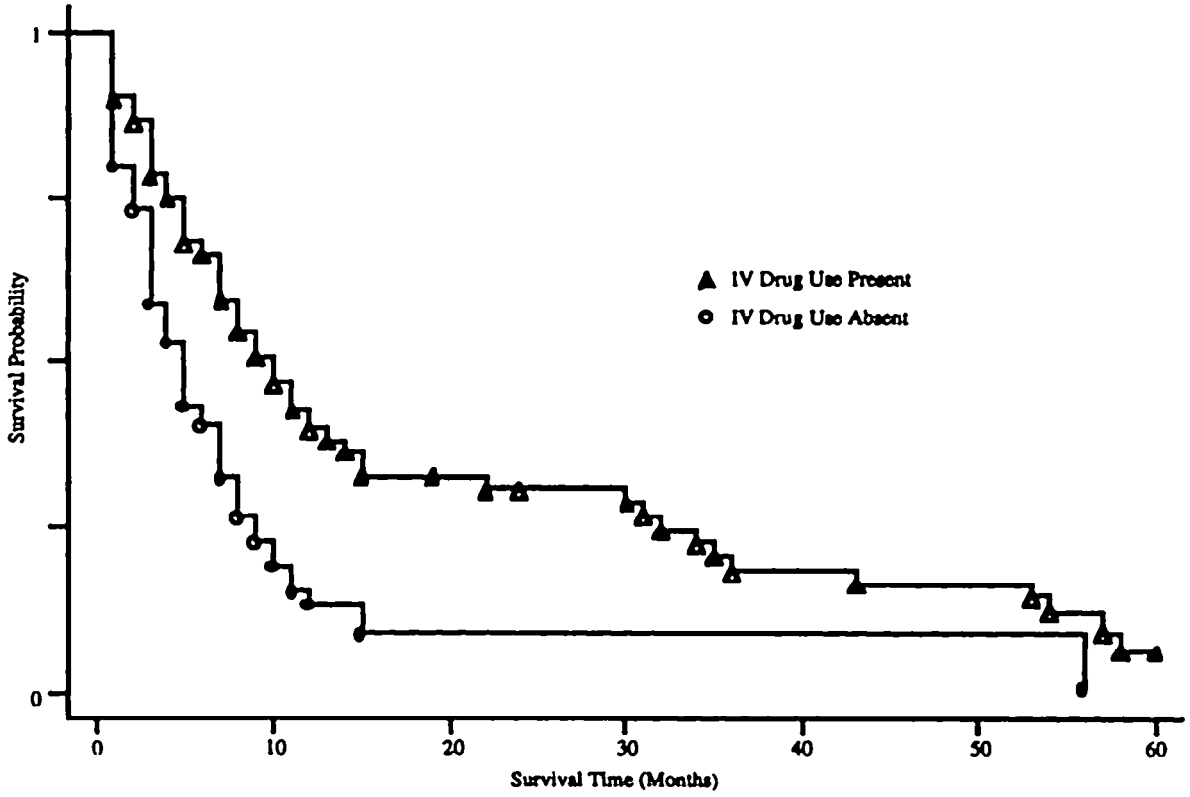


Figure 4.3 Graph of the estimated proportional hazards survivorship function for IV drug use present (Δ) and absent (o) for the HMO-HIV+ study. Points are plotted at observed times for each group.

times, then we recommend using a plot like Figure 4.3. One must use caution when reading the literature, as it may be difficult to determine whether plotted data involve inappropriate extrapolation of the fitted model. The best approach in practice is to provide results from both a thorough univariate analysis of survival experience in subgroups of special interest, as well as results from a regression analysis.

As noted in the previous section, the strength of regression modeling is the ability to adjust statistically for possible imbalances in the observed data. As an example of a more complicated model, suppose we fit a proportional hazards model containing age and IV drug use. The goal is to present survivorship functions for the two drug use groups, controlling for age. We must give some thought to what we mean by *controlling for age*.

We would like the estimated survivorship function to use the covariates in the same way that covariates are controlled for in a linear regression. In linear regression, a point on the regression line (or plane for a multiple variable model) is the model-based estimate of the conditional mean of the dependent variable among subjects with values of the

covariates defined by the point. The analogy to linear regression can help our thinking, but the situation is a bit more complicated in a proportional hazards regression analysis. Since the model does not contain an intercept, we do not have a fully parametric hazard function and thus the model cannot predict an individual point estimate of the conditional "mean" survival time. The estimated survivorship function in (4.23) is the proportional hazards model-based estimator of the conditional statistical distribution of survival time. The word "conditional" here means restricting observation to a cohort with covariate values equal to values specified. Pursuing this notion further, suppose we were able to follow an extremely large cohort of subjects for 5 years. Suppose also that the cohort is large enough that we can perform a fully stratified analysis and compute the Kaplan–Meier estimator of the survivorship function for each possible set of values of the covariates, such as, IV drug users who are 40 years old. If the proportional hazards model is correct, then the estimator in (4.23) and the Kaplan–Meier estimator should be similar, within statistical variation. One may use this estimator, (4.23), to describe survival time graphically and to compute estimates of quantiles, such as the median, in the same way the Kaplan–Meier estimator was used in Chapter 2.

The most frequent use of estimated survivorship functions in applied settings is to provide curves, like those in Figure 4.2 or 4.3, which may be used to compare groups visually and control for other model covariates. If the model does not contain grouping variable by covariate interactions, then the resulting survivorship functions are in a sense "parallel" in a way similar to lines with the same slope in a linear regression. In practice, one would choose one set of "typical" values of the other covariates. For a continuous covariate like age, we usually choose the common value to be the mean, median or another central value. In the HMO-HIV+ study, the mean age is 36.02 and the median age is 35. Thus, 35 seems like a good value to use for age in this example. If we merely fit the proportional hazards model containing age and IV drug use and request computation of the baseline survivorship function, then the program would estimate survivorship experience for IV drug use absent and, although biologically impossible, age equal to zero years. To obtain the estimate of the two age-adjusted survivorship functions, we would have to evaluate the expression in (4.23) using the coefficients from the fitted model with $(\text{DRUG} = 0, \text{AGE} = 35)$ and $(\text{DRUG} = 1, \text{AGE} = 35)$. This approach, while algebraically correct, can cause unwanted round-off and computational error in some situations. We would like to avoid computations that involve exponentiating large

Table 4.12 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Model Containing IV Drug Use and Age Centered at 35 Years in the HMO-HIV+ Study

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% Conf. Int.
DRUG	0.941	0.2555	3.68	<0.001	0.441, 1.442
AGE_C	0.092	0.0185	4.95	<0.001	0.055, 0.128

positive or negative numbers. One way to do this is to center continuous covariates. In the current example, we fit the model using $\text{AGE\_C} = \text{AGE} - 35$, and the results are shown in Table 4.12. These results are identical to ones which would have been obtained if we had used age uncentered, with the only difference being in the baseline survivorship function. When we center age, not only are our results computationally more accurate, but the estimate of the baseline survivorship function corresponds to IV drug use absent and age equal to 35 years, the zero value of the two covariates in the model. To obtain the second estimated survivorship function we compute

$$\hat{S}\left(t_j, \text{DRUG} = 1, \text{AGE\_C} = 0, \hat{\boldsymbol{\beta}}\right) = \left[\hat{S}_0\left(t_j\right)\right]^{\exp(1\times 0.941)}, \quad j = 1,2,\ldots,100 .$$

Graphs of the two estimated survivorship functions, plotted at observed values of time in each group, are shown in Figure 4.4. In this example, we chose these points to plot because of the absence of data in the interval from 15 to 56 months in the IV drug present group. The curves in this graph provide proportional hazards estimates of the survivorship experience of two cohorts of 35-year-old subjects differing in their IV drug use. Each point on the two curves depends on the actual observed survival times and the proportional hazards assumption.

The shapes of the curves in Figure 4.5 are determined by the choice of age equal to 35 as the center or "zero" value and the proportional hazards assumption. In order to illustrate the parallelism in the plots for any value of age and the effect of the choice of the center, four sets of curves have been plotted in Figures 4.5a–4.5d. The plots use age equal to 30, 35, 40 and 45 years, respectively. Since these plots have been prepared to demonstrate the effect of centering and the proportional hazards assumption, the two curves in each plot use all observed survival times. The basic parallelism is present, as the hazard ratio at each point in each of the four plots is 2.18. The progressive steepness in the plots

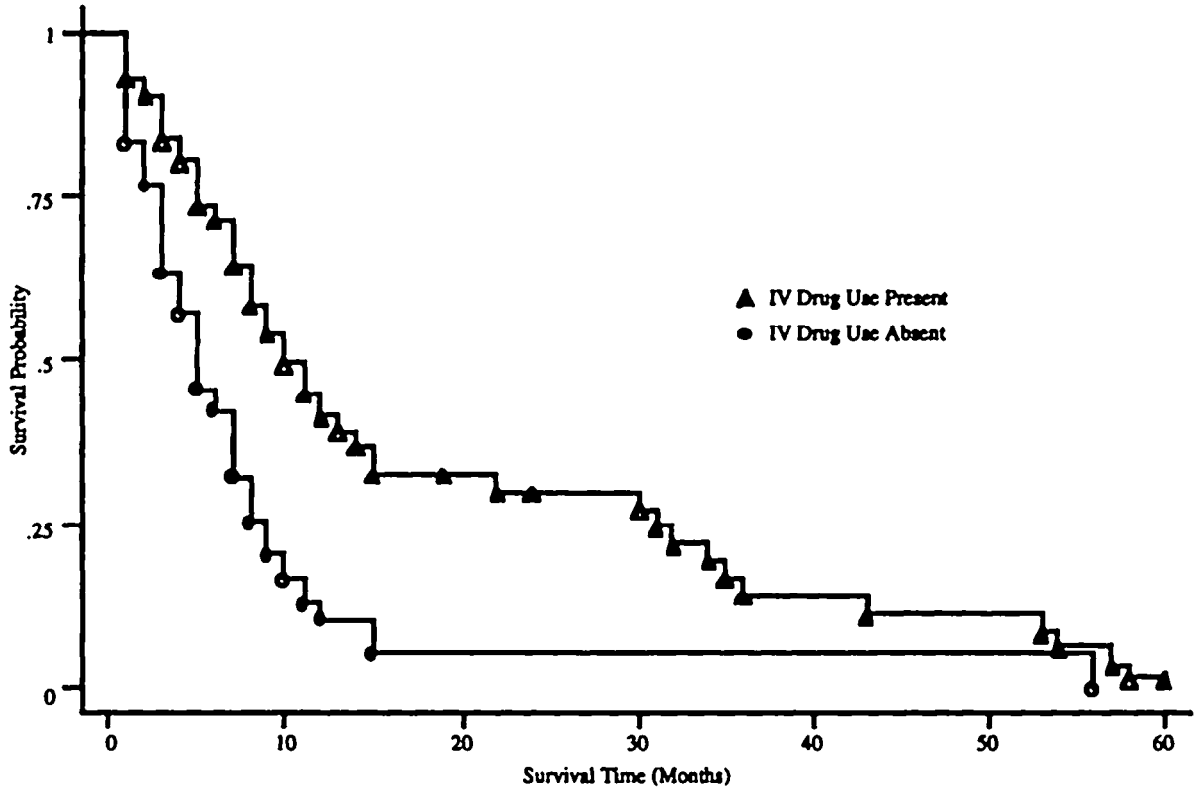


Figure 4.4 Graph of the age-adjusted estimated proportional hazards survivorship function for IV drug use present (Δ) and absent (o) for the HMO-HIV+ study. Points are plotted at observed survival time for each group.

is due to the fact that age is inversely related to survival, has a positive coefficient in the log-hazard function and is increasing from Figure 4.5a to 4.5d. For simply demonstrating the effect of IV drug use on survival controlling for age, any one of the four plots could have been used. Since the median age was 35, the plot in Figure 4.4 corresponds to Figure 4.5b, but with the noted difference in plotted values.

When the fitted model is even moderately complex, it may be difficult to decide what combination of covariate values best represents the middle of the data. We recommend that continuous covariates be centered to avoid potential numerical problems. In such complex situations, plots based on values of a quantity called the *risk score* are frequently used in practice. The risk score is the value of the linear portion of the proportional hazards model. In a model containing p covariates, the estimator is

$$\hat{r}\left(\mathbf{x},\hat{\boldsymbol{\beta}}\right)=\sum_{k=1}^{p}\hat{\beta}_k x_k .$$

For ease of notation, we denote the value of the estimator of the risk score for the jth subject as

$$\hat{r}_j = \hat{r}\left(\mathbf{x}_j, \hat{\boldsymbol{\beta}}\right) = \sum_{k=1}^{p} \hat{\beta}_k x_{jk} \ , \tag{4.26}$$

and, to be in agreement with notation used in Chapter 3, its exponentiated value as

$$\hat{\theta}_j = e^{\hat{r}_j} \ . \tag{4.27}$$

Most software packages will provide calculated values of either or both (4.26) and (4.27). The baseline survivorship function corresponds to a risk score of zero which may or may not be of clinical interest. Typically, one can obtain the values of the quartiles of the risk score from a descriptive statistics routine and obtain the estimated survivorship function from (4.23) by evaluating

$$\hat{S}\left(t_j, \hat{r}_q, \hat{\boldsymbol{\beta}}\right) = \left[\hat{S}_0\left(t_j\right)\right]^{\exp\left(\hat{r}_q\right)}, \quad j = 1, 2, \ldots, n \ , \tag{4.28}$$

where the $\hat{r}_q$, $q = 25, 50, 75$, correspond to the empirical quartiles of the risk score.

This procedure may be modified when we wish to graph the estimated survivorship functions for a grouping variable, controlling for a risk score based on the remaining covariates. In this setting, we subtract out the contribution of the grouping variable to the risk score, calculate the median value of what remains, and then add back in the contribution of the grouping variable when calculating the estimator of the survivorship function. Suppose the grouping variable is dichotomous and is the first of the p covariates in the model. The modified risk score, obtained by removing the effect of the grouping variable, is

$$\hat{rm}_j = \hat{r}_j - \hat{\beta}_1 x_{j1}, \quad j = 1, 2, \ldots, n \, .$$

If we denote the median of the modified risk scores as $\hat{rm}_{50}$, then the estimates of the survivorship functions for the two groups at this median are

$$\hat{S}\left(t_j, x_1 = 0, \hat{rm}_{50}, \hat{\boldsymbol{\beta}}\right) = \left[\hat{S}_0\left(t_j\right)\right]^{\exp\left(\hat{rm}_{50} + \hat{\beta}_1(0)\right)} \tag{4.29}$$

and

$$\hat{S}\left(t_j, x_1 = 1, \hat{rm}_{50}, \hat{\boldsymbol{\beta}}\right) = \left[\hat{S}_0\left(t_j\right)\right]^{\exp\left(\hat{rm}_{50} + \hat{\beta}_1(1)\right)} \tag{4.30}$$

for $j = 1, 2, \ldots, n$ subjects. Before plotting the graphs of (4.29) and (4.30), one should examine the range of observed survival times in the two groups for biologically important gaps.

The data from the UIS may be used to provide examples of plotting survivorship functions from a more complex model. This model has been chosen for demonstration purposes only. We have not attended to a number of important modeling details and, as a result, this model should not be construed as being the final model for assessing treatment effect. Table 4.13 presents the results of fitting a model containing treatment, age, history of IV drug use (0 = never, 1 = previous or recent) and the number of previous drug treatments. We centered age at 30 years and the number of previous drug treatments at 3. The results in Table 4.13 support a significant treatment effect after adjusting for the other variables in the model. The estimated hazard ratio for treat-

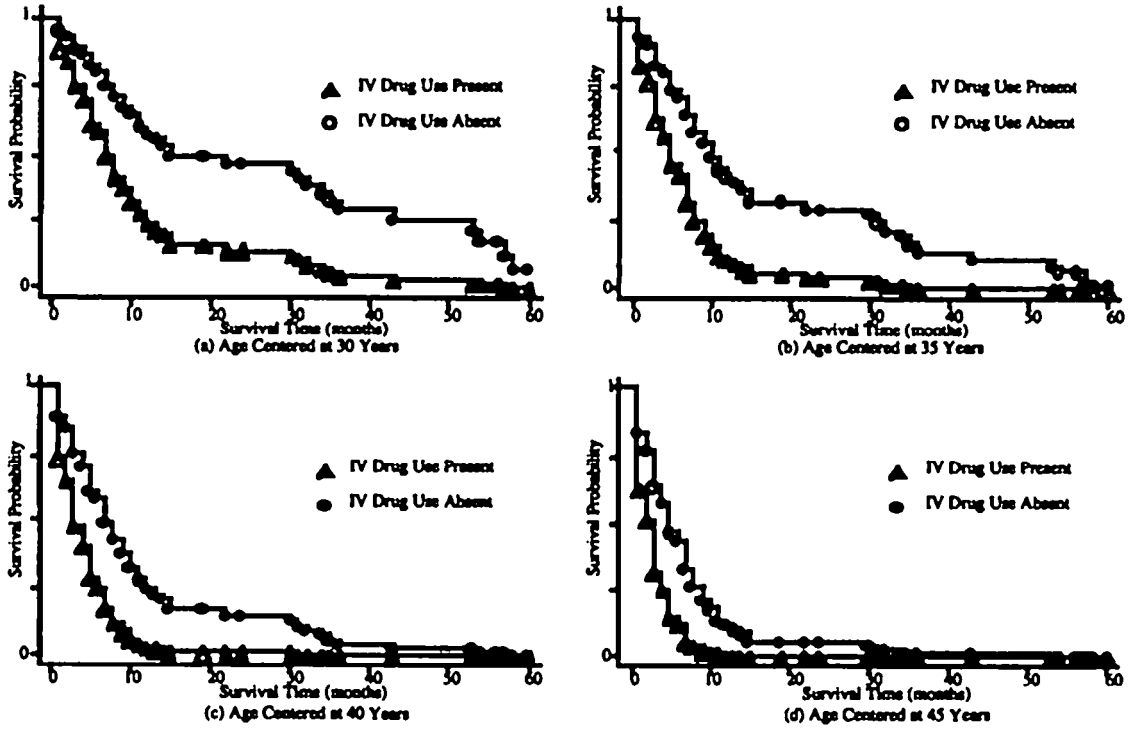


Figure 4.5 Graphs of the age-adjusted estimated proportional hazards survivorship function for IV drug use present (Δ) and absent (o) at four different ages in the HMO-HIV+ study.

ment and the 95 percent confidence interval are $\hat{HR} = \exp(-.227) = 0.80$ and (0.666, 0.954), respectively. The estimate supports a protective effect for the longer treatment, with about a 20 percent reduction in the hazard rate for returning to drug use.

In this example, the equation for the estimated risk score for the jth subject is

$$\hat{r}_j = -0.227 \times \text{TREAT}_j - 0.031 \times (\text{AGE\_C}_j) + 0.343 \times \text{DRUG}_j + 0.031 \times (\text{NDRGTX\_C}_j)$$

and the modified estimated risk score is

$$\hat{rm}_j = \hat{r}_j - (-0.227 \times \text{TREAT}_j) .$$

The median value of the modified risk score is 0.1588 and the equations for the estimators of the modified risk score-adjusted survivorship functions obtained from (4.29) and (4.30) are

$$\hat{S}(t_j, \text{TREAT} = 0, \hat{rm}_{50}, \hat{\boldsymbol{\beta}}) = [\hat{S}_0(t_j)]^{\exp(0.1588)} \tag{4.31}$$

and

$$\hat{S}(t_j, \text{TREAT} = 1, \hat{rm}_{50}, \hat{\boldsymbol{\beta}}) = [\hat{S}_0(t_j)]^{\exp(0.1588-0.227)} . \tag{4.32}$$

Since the observed range of survival times in the two treatment groups is comparable, we chose to use all observed survival times to plot (4.31) and (4.32), which are shown in Figure 4.6. Since the analysis is based on a large number of subjects (593 of the 628 subjects had complete data on the four covariates), the use of separate plotting symbols has been suppressed to avoid an unnecessarily cluttered graph.

The two curves in Figure 4.6 reflect both the use of the median modified risk score and the assumption of proportional hazards. Use of other quantiles of the modified risk score would shift the curves to the left or right in a manner similar to Figure 4.5. The longer times until return to drug use for the longer duration treatment group are illustrated in the graph.

We can use adjusted estimated survivorship functions, such as those shown in Figures 4.5 and 4.6, to estimate the adjusted median survival

Table 4.13 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Model Containing Treatment (TREAT), Age–30 (AGE_C), IV Drug Use (DRUG) and Number of Prior Drug Treatments–3 (NDRGTX_C) from the UIS, *n* = 593

Variable	Coeff.	Std. Err.	*z*	*P*>|*z*|	95% Conf. Int.
TREAT	–0.227	0.0916	–2.48	0.013	–0.407, –0.048
AGE_C	–0.031	0.0079	–3.87	<0.001	–0.046, –0.015
DRUG	0.343	0.1043	3.29	0.001	0.138, 0.547
NDRGTX_C	0.031	0.0080	3.87	<0.001	0.015, 0.047

times in the same manner as described in Section 2.3. If the graph is not too complicated, we can use the graphical approach illustrated in Figure 2.6. However, since this is not likely to be accurate enough in most applied settings, we determine the estimator as

$$\hat{t}_{50} = \min\left\{t : \hat{S}\left(t, \mathbf{x}, \hat{\boldsymbol{\beta}}\right) \leq 0.50\right\}, \tag{4.33}$$

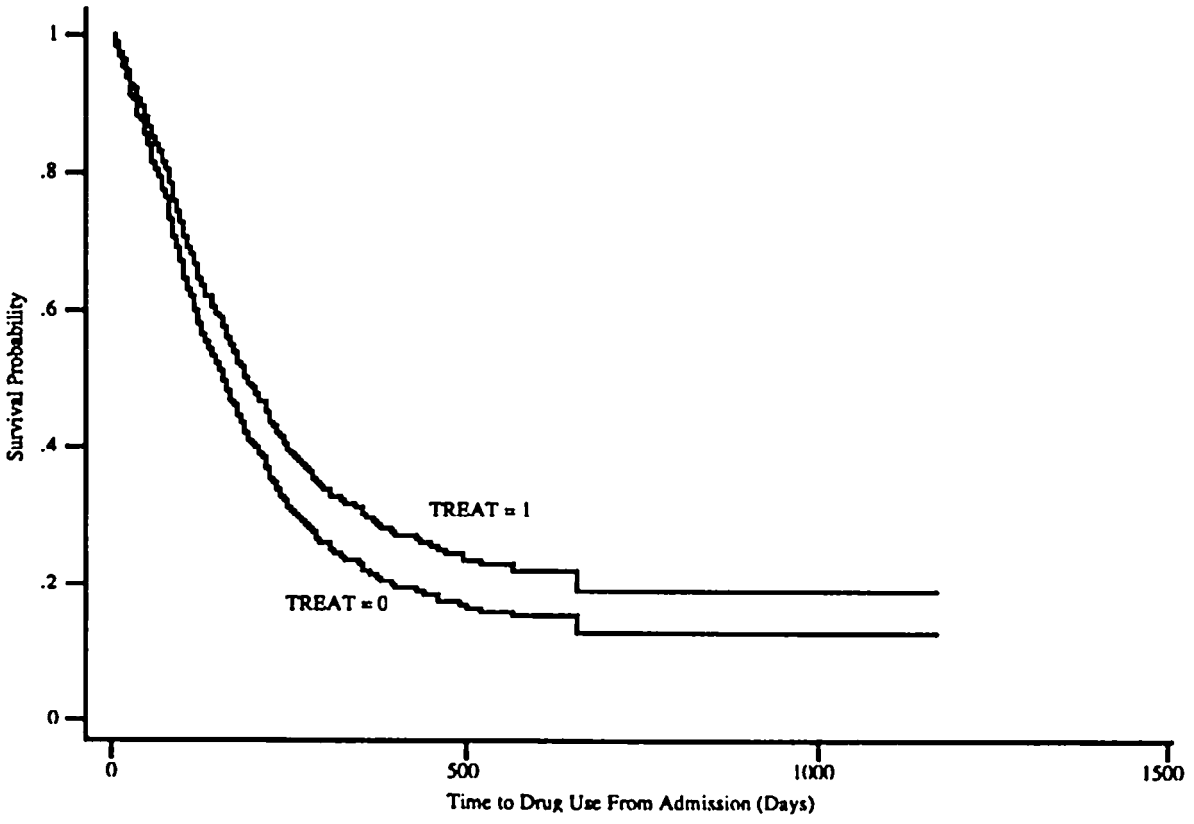


Figure 4.6 Graphs of the modified risk score-adjusted estimated survivorship function for treatment in the UIS.

where the estimator of the survivorship function in (4.33) is any one of the adjusted estimators. Application of (4.33) to the estimates graphed in Figure 4.6 yields adjusted estimated median time to return to drug use of 159 days and 190 days for the short and long treatments, respectively.

Confidence interval estimators for the covariate-adjusted estimator of the median survival time are discussed in the next section. In general, the methods are a bit more complex than those discussed to this point in the text. The methods have not been implemented in standard statistical software packages, but one could write a routine to calculate the confidence interval estimator.

The covariate-adjusted survivorship function in (4.23) is the one most frequently used in applied settings. An alternative estimator, first proposed by Makuch (1982) and further studied by Thomsen, Keiding and Altman (1991), is discussed and recommended by Marubini and Valsecchi (1995, Chapter 6). The estimator, called the "direct adjusted survival curve" by Makuch, is obtained by averaging the individual estimated survivorship functions over all subjects at each observed time. The average may be computed overall or within subgroups of subjects. As shown in Marubini and Valsecchi (1995), the direct adjusted estimator is easy to compute for a simple model containing one or two dichotomous covariates, but a more complicated model requires special programming. Evaluation of the direct estimator requires calculating an average over all subjects in specified groups at each observed value of time. In order to illustrate the method, we computed the directly adjusted estimator for the model shown in Table 4.12. It is presented in Figure 4.7, along with the estimator in (4.23) for the IV drug use groups present and absent shown in Figure 4.4. The two sets of curves in Figure 4.7 certainly convey the same message regarding the effect of treatment and, in this case, would yield similar estimates of the median time to drug relapse in each drug group. The difference between the two estimators is most easily explained if we focus on a single value of time, say 24 months. The value of the upper covariate-adjusted estimated survivorship function, (4.23), is just $\hat{S}_0(24)$, that is, the survivorship function for age equal to 35 years and IV drug use absent. The direct adjusted estimate for this group at 24 months is

$$\tilde{S}(24, \text{DRUG} = 0) = \frac{1}{51}\sum_{j=1}^{51}\hat{S}\left(24, \text{DRUG}_j = 0, \text{AGE}_j - 35\right), \quad (4.34)$$

where

$$\hat{S}\left(24, \text{DRUG}_j = 0, \text{AGE}_j - 35\right) = \left[\hat{S}_0(24)\right]^{\exp\left(0 \times \hat{\beta}_1 + (\text{AGE}_j - 35) \times \hat{\beta}_2\right)}.$$

The mean on the right-hand side of (4.34) is the average restricted to the 51 subjects in the HMO-HIV+ study who had no history of IV drug use. The point on the lower covariate-adjusted curve is

$$\hat{S}(24, \text{DRUG} = 1, \text{AGE\_C} = 0) = \left[\hat{S}_0(24)\right]^{\exp\left(1 \times \hat{\beta}_1\right)},$$

and the corresponding point on the directly-adjusted curve is

$$\tilde{S}(24, \text{DRUG} = 1) = \frac{1}{49} \sum_{j=1}^{49} \hat{S}\left(24, \text{DRUG}_j = 1, \text{AGE}_j - 35\right), \quad (4.35)$$

where

$$\hat{S}\left(24, \text{DRUG}_j = 1, \text{AGE}_j - 35\right) = \left[\hat{S}_0(24)\right]^{\exp\left(1 \times \hat{\beta}_1 + (\text{AGE}_j - 35) \times \hat{\beta}_2\right)}.$$

The mean on the right-hand side of (4.35) is the average restricted to the 49 subjects in the HMO-HIV+ study who had a history of IV drug use. In order to obtain the graph of the direct-adjusted survivorship functions in Figure 4.7, the expressions in (4.34) and (4.35) must be computed for each observed value of time.

Marubini and Valsecchi (1995) state that the directly adjusted estimator better takes into account the variability in the observed values of the covariates. The difference between the two adjusted curves is due to the fact that the survivorship function is a non-linear function of the covariate (i.e., age) in the example. Which of the two adjusted curves, covariate or direct, one should use in practice depends on the goal of the analysis. If the purpose is to provide a figure to be used to compare survivorship under different levels of a nominal scale covariate, the more easily calculated covariate-adjusted curve is likely to be adequate in most applied settings.

Model-based estimation of the survival probability at a fixed time point is another application of the estimator in (4.23). For example, we may use the results in Table 4.12 to obtain an estimator of the probability of survival to 18 months for a subject with specified age and history of IV drug use. This is analogous to using a fitted linear regression

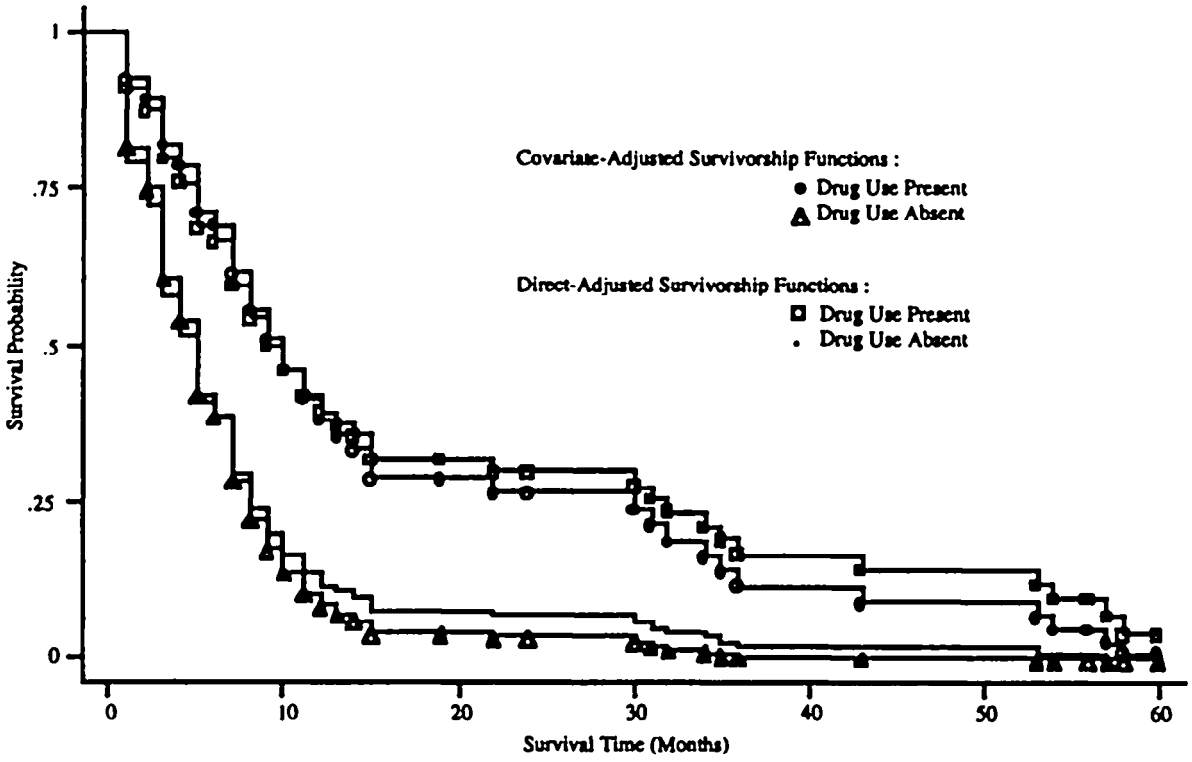


Figure 4.7 Graphs of the covariate age-adjusted and direct age-adjusted estimated survivorship function for IV drug use present and absent for the HMO-HIV+ study.

model to predict the outcome in a new subject. Suppose our new subject is 42 years old and has a previous history of IV drug use. The predicted 18-month survival probability for this subject is

$$\hat{S}(18, \text{DRUG} = 1, \text{AGE\_C} = 42 - 35) = \left[\hat{S}_0(18)\right]^{\exp(1\times 0.941 + 7\times 0.092)}$$
$$= [0.3239]^{\exp(1\times 0.941 + 7\times 0.092)} = [0.3239]^{4.879} = 0.0041,$$

where the value of $\hat{S}_0(18)$ is obtained from a tabulation of the estimated baseline survivorship function. Thus the fitted model[3] predicts less than a 1 percent chance of survival to 18 months. We discuss the confidence interval estimator for the predicted survival probability in the next section. The entire survivorship function may be obtained using the same

[3] We wish to remind the reader that the HMO-HIV+ study data are hypothetical and, as such, fitted models should not be used to draw substantive, real-world conclusions.

method discussed above for covariate-adjusted survival curves (i.e., considering (4.23) with the stated covariate values as a function of time).

The estimated survivorship function, (4.23), can be an effective tool to describe the results of a regression analysis of survival time. We wish to reemphasize the importance of giving careful thought to the plotted range of the curve and estimates of survival probabilities. It is all too easy, with current statistical software, to present graphs and predictions that may inappropriately extrapolate the fitted model.

4.6 CONFIDENCE INTERVAL ESTIMATION OF THE COVARIATE-ADJUSTED SURVIVORSHIP FUNCTION

In this section we present a confidence interval estimator for the covariate-adjusted survivorship function. The method and resulting formula are not difficult to understand as they follow from the methods discussed in Chapter 2 for the Kaplan–Meier estimator. However, application requires that one use matrices and matrix calculations, and these require greater computing expertise than is required in other sections.

Andersen, Borgan, Gill and Keiding (1993) present an estimator [equation (7.2.33), page 506] of the variance of the log of the covariate-adjusted survivorship function. Their estimator is identical to one presented by Marubini and Valsecchi (1995, in the Appendix to Chapter 6). As noted in Chapter 2, better coverage properties are obtained if a confidence interval for the survivorship function is based on the log-log transformation of the function. An expression for a variance estimator for this further transformation is given by Andersen, Borgan, Gill and Keiding (1993) following (7.2.33). The equation for the variance estimator for a fixed set of the p covariates, denoted $\mathbf{x}_0$, is

$$\hat{\mathrm{V}}\mathrm{ar}\left\{\ln\left[-\ln\left(\hat{S}\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right)\right)\right]\right\} = \hat{A}\left(t,\hat{\boldsymbol{\beta}}\right) + \hat{B}\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right), \tag{4.36}$$

where

$$\hat{A}\left(t,\hat{\boldsymbol{\beta}}\right) = \frac{1}{\left\{\ln\left[\hat{S}_0(t)\right]\right\}^2} \sum_{t_{(i)} \le t} \frac{d_i}{\left[\hat{C}\left(t_{(i)},\hat{\boldsymbol{\beta}}\right)\right]^2},$$

$$\hat{B}\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{x}_0 - \bar{\mathbf{x}}(t)\right)' \hat{\mathrm{V}}\mathrm{ar}\left(\hat{\boldsymbol{\beta}}\right)\left(\mathbf{x}_0 - \bar{\mathbf{x}}(t)\right),$$

$$\hat{C}\left(t_{(i)},\hat{\boldsymbol{\beta}}\right)=\sum_{j\in R(t_{(i)})}e^{\mathbf{x}_j'\boldsymbol{\beta}},$$

and

$$\bar{\mathbf{x}}(t)=\frac{1}{\ln\left[\hat{S}_0(t)\right]}\sum_{t_{(i)}\le t}d_i\left\{\frac{\sum_{j\in R(t_{(i)})}\mathbf{x}_j e^{\mathbf{x}_j'\boldsymbol{\beta}}}{\left[\hat{C}\left(t_{(i)},\hat{\boldsymbol{\beta}}\right)\right]^2}\right\}.$$

We note that d_i represents the number of subjects with survival time equal to $t_{(i)}$ and $\hat{\mathrm{Var}}\left(\hat{\boldsymbol{\beta}}\right)$ denotes the estimator of the covariance matrix of the estimated coefficients. The endpoints of a $100(1-\alpha)$ percent confidence interval for the log-log function are

$$\ln\left[-\ln\left(\hat{S}(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}})\right)\right]\pm z_{1-\alpha/2}\hat{\mathrm{SE}}\left\{\ln\left[-\ln\left(\hat{S}(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}})\right)\right]\right\}, \qquad (4.37)$$

where $\hat{\mathrm{SE}}\{\ }$ in (4.37) denotes the estimator of the standard error and, in this case, is the positive square root of the estimator in (4.36). If we denote the lower and upper endpoints obtained from (4.37) as $l\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right)$ and $u\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right)$, then the lower and upper endpoints of the confidence interval estimator of the survivorship function are obtained in a manner similar to (2.8) and are

$$\exp\left\{-\exp\left[u\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right)\right]\right\} \text{ and } \exp\left\{-\exp\left[l\left(t,\mathbf{x}_0,\hat{\boldsymbol{\beta}}\right)\right]\right\}. \qquad (4.38)$$

The expressions in (4.37) and (4.38) may be used to obtain a confidence interval for an individual predicted survival probability or to provide a pointwise confidence band for a covariate-adjusted survivorship function. In the previous section, we used the fitted model shown in Table 4.12 for the HMO-HIV+ study to predict the 18-month survival time for a 42-year-old subject with a history of prior drug use. The model-based prediction was 0.004. We use (4.37) and (4.38) to obtain the endpoints of a 95 percent confidence interval, giving us the interval (0.00005, 0.04894). The interpretation of this interval is that a predicted probability in this interval would be consistent with the ob-

served data. Stated another way, the subject has at best a 4.9 percent chance of surviving 18 months.[4]

We obtain pointwise confidence bands for the survivorship function by evaluating (4.38) at each observed survival time. These can be plotted, along with the estimated survivorship function, similar to the plot in Figure 2.5. The Hall and Wellner joint confidence bands discussed in Chapter 2 have not been extended to the covariate-adjusted survivorship function.

We plotted, in Figure 4.4, age- (equal to 35) adjusted survivorship functions for both levels of the IV drug use variable. Confidence bands for each function are obtained through definition of the fixed covariate $\mathbf{x}_0$. These have been calculated and are shown separately in Figures 4.8a and 4.8b. The confidence bands for the IV drug use groups absent and present in (4.37) and (4.38) are

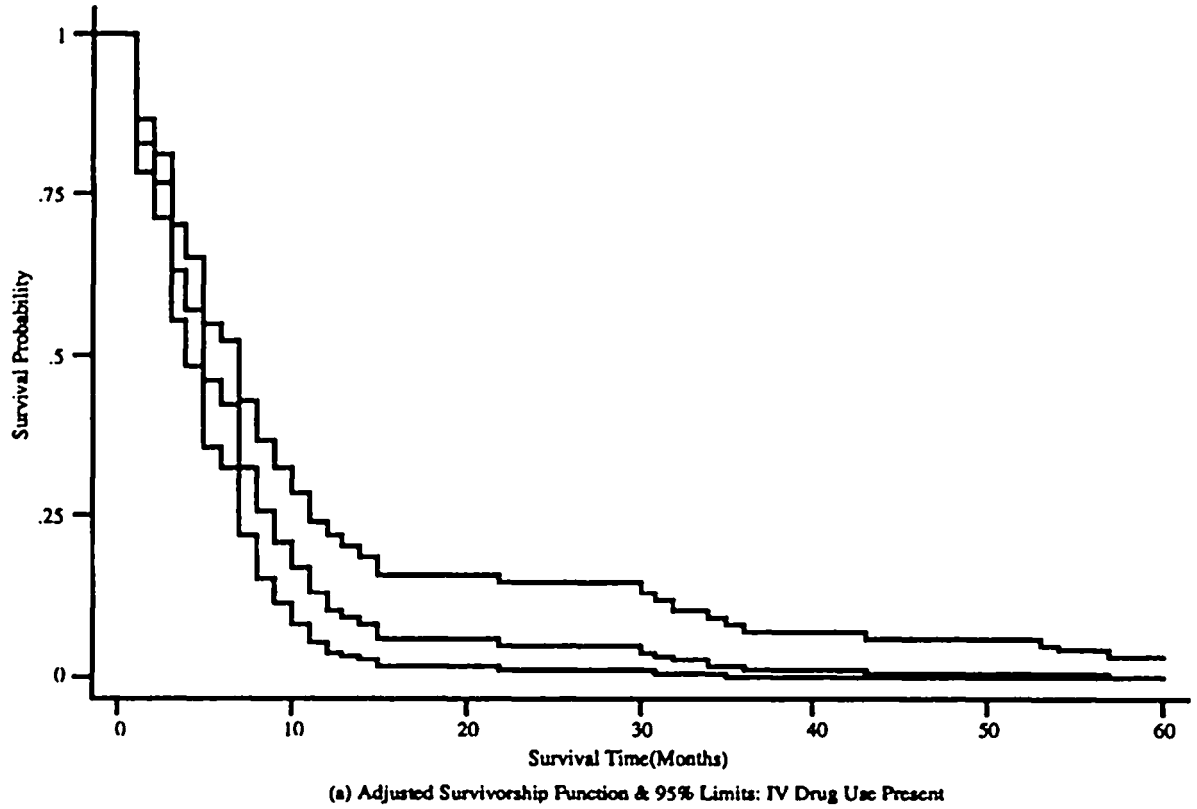
$$\mathbf{x}_0' = (\text{DRUG} = 0, \text{AGE\_C} = 0) \text{ and } \mathbf{x}_0' = (\text{DRUG} = 1, \text{AGE\_C} = 0),$$

respectively. The graphs have been drawn in a manner similar to Figure 4.2, using each observed survival time. The general shape of the confidence bands is similar to those presented in Chapter 2, although they are wider and more skewed for longer survival times.

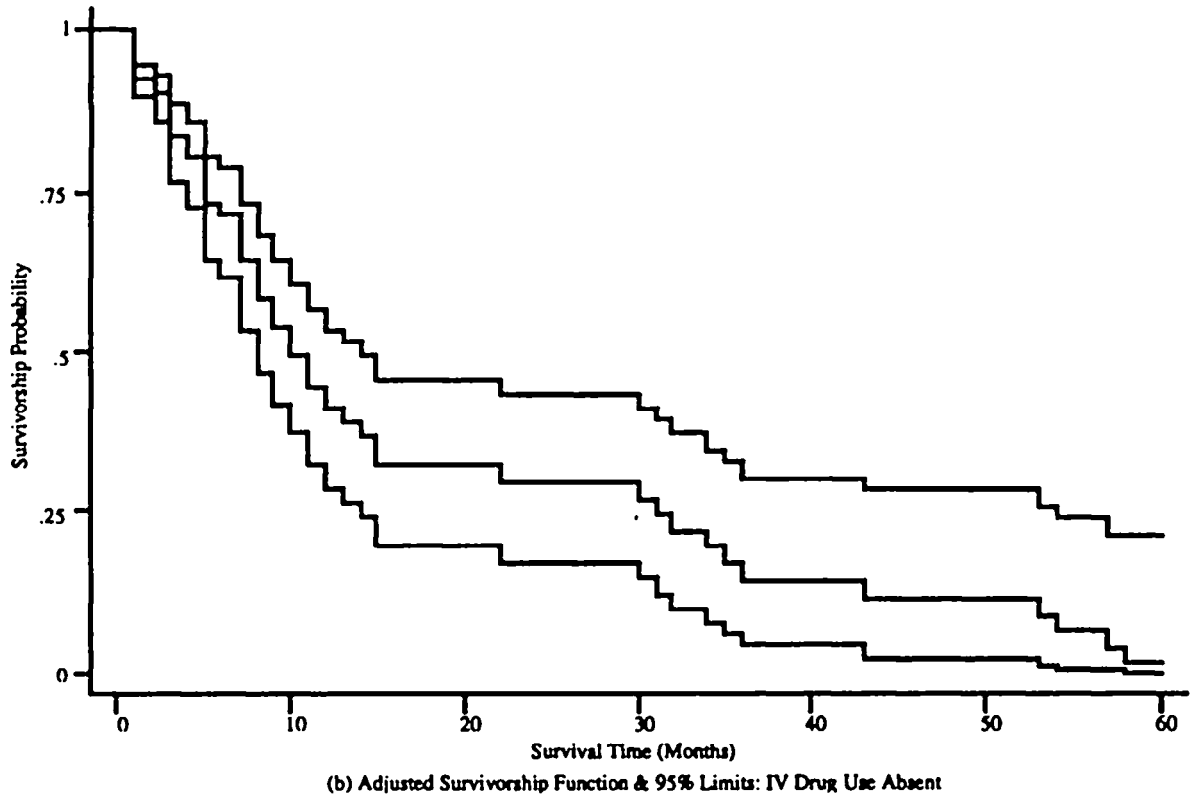
We can also use the graphs to obtain a confidence interval estimate of the median survival time (or other quantile) by following the same method described in Chapter 2. A more accurate determination may be obtained by applying (4.33) to the lower and upper confidence bands, respectively. Applying (4.33) to the actual values generating the curves in Figures 4.8a and 4.8b yields age-adjusted median survival times and 95 percent confidence limits for IV drug use present of 5 (4, 7) and for IV drug use absent of 10 (8, 14). These confidence intervals are essentially an extension of the Brookmeyer-Crowley method discussed in Chapter 2. We noted in Chapter 2 that the Brookmeyer–Crowley method assumes that there are no tied survival times. However, there are a number of ties in the HMO-HIV+ study and, as a result, the presented confidence interval estimate should be interpreted cautiously, as the ef-

---

[4] One should not infer from this statement or our presentation of the methods in this text that we advocate the use of fitted survival time models and possible subsequent predictions as tools for individual patient/subject decision-making. This is a difficult and sensitive subject, and a statistical model should be one small part of a much larger discussion.



(a) Adjusted Survivorship Function & 95% Limits: IV Drug Use Present



(b) Adjusted Survivorship Function & 95% Limits: IV Drug Use Absent

Figure 4.8 Graphs of age-adjusted estimated survivorship function and 95% pointwise confidence limits for IV drug use present and absent for the HMO-HIV+ study.

fect of ties on the coverage properties of the interval has not been studied.

In applied settings, we recommend that adjusted point and interval estimates of median survival time be used for descriptive purposes only. One should avoid the temptation to use the confidence intervals to draw inferential conclusions about the equality of median survival times. This hypothesis should be tested via the partial likelihood ratio test for the significance of the coefficient for the grouping variable in the fitted proportional hazards model.

In Section 4.5 we presented a covariate-adjusted survivorship function that used the median risk score for a single curve and a modified risk score when adjusted survivorship functions for two groups were compared. Unfortunately, we cannot directly employ the variance estimator and confidence interval in (4.37) and (4.38), since the median value of the risk score may not correspond to a single fixed set of covariates. One solution is to rerun the analysis adjusting for the set of covariate values yielding a risk score that came closest to the median. However, it is possible, with a complex model, that several sets of covariate values will have risk scores equally close to the median value. In this case, one could choose the set that seems clinically closest to a middle set of values. The adjusted survival curves obtained from using the actual median risk score and the ones obtained from the set of covariates with risk score nearest the median should be quite similar and adequate for descriptive analyses. Again, this is an issue only if one wants to add confidence bands to the plot of a risk score-adjusted survivorship function.

EXERCISES

1. Using all the data from the WHAS (i.e., ignore cohort), with length of follow-up as the survival time variable and status at last follow-up as the censoring variable, do the following:

(a) Fit the proportional hazards model containing sex and estimate the hazard ratio, pointwise and with a 90 percent confidence interval. Interpret the point and interval estimates.

(b) Add age to the model fit in 1(a). Is age a confounder of the effect of sex? Explain the reasons for your answer.

(c) Is there a significant interaction between age and sex? (Use $\alpha = 0.10$ for this problem).

(d) Using the model fit in 1(c) estimate the hazard ratio, pointwise and 90 percent confidence interval, for gender at age 50, 60, 65, 70 and 80.

(e) Using the model fit in 1(c), estimate (pointwise and with a 90 percent confidence interval) the hazard ratio for a 10-year increase in age for each gender.

(f) Using the model fit in 1(c), compute, and then graph, the estimated survivorship functions for 65-year-old males and females. Interpret the survivorship experience presented in this graph.

(g) Using the graph in 1(f), estimate the median survival time for 65-year-old males and females.

2. Repeat problem 1, parts (b)–(d), with age broken into four groups at its quartiles. In part (d) estimate hazard ratios for each age group.

3. Using the data from the WHAS (i.e., ignore cohort), with length of follow-up as the survival time variable and status at last follow-up as the censoring variable, do the following:

(a) Fit the proportional hazards model containing age centered at 65 years, sex, peak cardiac enzymes centered at 650, cardiogenic shock complications, left heart failure complications and MI order and obtain the estimated baseline survivorship function. (Note: In this problem, ignore the possible sex $\times$ age interaction investigated in problems 1 and 2.) Estimate hazard ratios (via point estimates and 95 percent confidence intervals) for each variable in the model.

(b) Using the methods for the modified risk score, compute and graph estimated survivorship functions for subjects with and without cardiogenic shock complications. Use the estimated survivorship functions to estimate the median survival time.

CHAPTER 5

Model Development

5.1 INTRODUCTION

In any applied setting, performing a proportional hazards regression analysis of survival data requires a number of critical decisions. It is likely that we will have data on more covariates than we can reasonably expect to include in the model, so we must decide on a method to select a subset of the total number of covariates. When selecting a subset of covariates, we must consider such issues as clinical importance and adjustment for confounding, as well as statistical significance. Once we have selected the subset, we must determine whether the model is "linear" in the continuous covariates and, if not, what transformations are suggested by the data and clinical considerations. Which interactions, if any, should be included in the model is another important decision. In this chapter we discuss these and other practical model development issues.

The end use of the estimated regression model will most often be a summary presentation and interpretation of the factors that have influenced survival. This summary may take the form of a table of estimated hazard ratios and confidence intervals and/or estimated covariate-adjusted survivorship functions. Before this step can be taken, we must critically examine the estimated model for adherence to key assumptions (e.g., proportional hazards) and determine whether any subjects have an undue influence on the fitted model. In addition, we may calculate summary measures of goodness-of-fit to support our efforts at model assessment. Methods for model assessment are discussed and illustrated in Chapter 6.

The methods available to select a subset of covariates to include in a proportional hazards regression model are essentially the same as those

used in any other regression model. In this chapter we present three methods for selecting a subset of covariates. Purposeful selection is a method that is completely controlled by the data analyst, while stepwise and best subsets selection of covariates are statistical methods. These approaches to covariate selection have been chosen since use of one or more of them will yield, in the vast majority of model building applications, a subset of statistically and clinically significant covariates.

A word of caution: statistical software for fitting regression models to survival data is, for the most part, easy to use and provides a vast array of sophisticated statistical tools and techniques. One must be careful, therefore, not to lose sight of the problem and end up with the software prescribing the model to the analyst rather than the other way around.

Regardless of which method is used for covariate selection, any survival analysis should begin with a thorough bivariate analysis of the association between survival time and all important covariates. These methods are discussed in detail in Chapter 2. For categorical covariates, this should include Kaplan–Meier estimates of the group-specific survivorship functions, point and interval estimates of the median and/or other quantiles, survival time and use of one or more of the significance tests to compare survivorship experience across the groups defined by the variable. For descriptive purposes, continuous covariates could be broken into quartiles, or other clinically meaningful groups, and the methods for categorical covariates could then be applied. Alternatively, point and interval estimates of the hazard ratio for a clinically relevant change in the covariate could be used in conjunction with the significance level of the partial likelihood ratio test. These results should be displayed using the tabular conventions of the scientific field.

5.2 PURPOSEFUL SELECTION OF COVARIATES

Purposeful selection of covariates begins with a multivariable model that contains all variables significant in the bivariate analysis at the 20–25 percent level, as well as any other variables not selected with this criterion, but which are judged to be of clinical importance. If there are adequate data to fit a model containing all study covariates, this full model could be the beginning multivariable model. The rationale for choosing a relatively modest level of significance is based on recommendations for linear regression by Bendel and Afifi (1977), for discriminant analysis by Costanza and Afifi (1979), and for change in coefficient modeling in epidemiology by Mickey and Greenland (1989).

Use of this level of significance should lead to the inclusion, in the preliminary multivariable model, of any variable that has the potential to be either an important confounder, or is statistically significant. Following the fit of the initial multivariable model, we use the p-values from the Wald tests of the individual coefficients to identify covariates that might be deleted from the model. Some caution should be taken at this point not to reduce the size of the model by deleting too many seemingly nonsignificant variables at one time. The p-value of the partial likelihood ratio test should confirm that the deleted covariate is not significant. This is especially important when a nominal scale covariate with more than one design variable has been selected for deletion, since we typically make a rough guess about overall significance based on the significance levels of the individual coefficients of the design variables. Following the fitting of the reduced model, we assess whether or not removal of the covariate has produced an "important" change in the coefficients of the variables remaining in the model. In general, we use a value of about 20 percent as an indicator of an important change in a coefficient. If the variable excluded is an important confounder, it should be added back into the model. This process continues until no covariates can be deleted from the model.

At this point, we recommend that any variable excluded from the initial multivariable model be added back into the model to confirm that it is neither statistically significant nor an important confounder. We have encountered situations in practice where a variable had a bivariate test p-value that exceeded 0.8 but it became highly significant when added to a multivariable model. At the conclusion of this step we have the "preliminary main effects model."

The next step is to examine the scale of the continuous covariates in the preliminary main effects model. A number of techniques are available, all of which are designed to determine whether the data support the hypothesis that the effect of the covariate is linear in the log hazard and, if not, which transformation of the covariate is linear in the log hazard. The simplest method is to replace the covariate with design variables formed from the quartiles or other cutpoints that may have been used in the bivariate descriptive analysis. The estimated coefficients for the design variables are plotted versus the midpoints of the groups and, at the midpoint of the first group, a point is plotted at zero. If the correct scale is linear in the log hazard, then the polygon connecting the points should be nearly a straight line. If the polygon departs substantially from a linear trend, its form may be used to suggest a transformation of the covariate. The advantage of the quartile method is that it does not

require any special software. The disadvantage is that it is not powerful enough to detect subtle, but often important, deviations from a linear trend.

Another approach is to use the method of fractional polynomials, developed by Royston and Altman (1994), to suggest transformations. We wish to determine what value of x^p yields the best model for the covariate. In theory, we could incorporate the power, p, as an additional parameter in the estimation procedure. However, this would greatly increase the complexity of the estimation problem. Royston and Altman propose replacing full maximum likelihood estimation of the power by a search through a small but reasonable set of possible values. We will provide a brief description of the method and later demonstrate its use, along with the other methods, in an example.

The method of fractional polynomials may be used with a multivariable proportional hazards regression model, but, for sake of simplicity, we describe the procedure using a model with a single continuous covariate. The hazard function for the proportional hazards regression model shown in (3.7) is

$$h(t,x,\beta) = h_0(t)e^{x\beta},$$

and the log-hazard function, which is linear in the covariate, is

$$\ln\left[h(t,x,\beta)\right] = \ln\left[h_0(t)\right] + x\beta.$$

One way to generalize this log-hazard function is to specify it as

$$\ln\left[h(t,x,\beta)\right] = \ln\left[h_0(t)\right] + \sum_{j=1}^{J} F_j(x)\beta_j \ .$$

The functions $F_j(x)$ are a particular type of power function. The value of the first function is $F_1(x) = x^{p_1}$. In theory, the power, p_1, could be any number, but in most applied settings we would try to use something simple. Royston and Altman (1994) propose restricting the power to be among those in the set $\wp = \{-2,-1,-0.5,0,0.5,1,2,3\}$, where $p_1 = 0$ denotes the log of the variable. The remaining functions are defined as

$$F_j(x)=\begin{cases} x^{p_j}, p_j \neq p_{j-1} \\ F_{j-1}(x)\ln(x), p_j = p_{j-1} \end{cases}$$

for $j=2,\ldots,J$ and restricting powers to those in $\wp$. For example, if we chose $J=2$ with $p_1=0$ and $p_2=-0.5$, then the log-hazard function is

$$\ln\left[h(t,x,\boldsymbol{\beta})\right]=\ln\left[h_0(t)\right]+\ln(x)\beta_1+\frac{1}{\sqrt{x}}\beta_2.$$

As another example, if we chose $J=2$ with $p_1=2$ and $p_2=2$, then the log-hazard function is

$$\ln\left[h(t,x,\boldsymbol{\beta})\right]=\ln\left[h_0(t)\right]+x^2\beta_1+x^2\ln(x)\beta_2.$$

The model is quadratic in x if $p_1=1$ and $p_2=2$. Again, we could allow the covariate to enter the model with any number of functions, J; but in most applied settings an adequate transformation may be found if we use $J=1$ or 2. Implementation requires, for $J=1$, fitting 8 models, that is, $p_1\in\wp$. The best model is the one with the largest log partial likelihood. The process is repeated with $J=2$ by fitting the 64 models obtained from all possible pairs of powers, that is, $(p_1,p_2)\in\wp\times\wp$, and the best model is again the one with the largest log partial likelihood. The relevant question is whether either of the two best models is significantly better than the linear model. Let $L(1)$ denote the log partial likelihood for the linear model, that is, $J=1$ and $p_1=1$, and $L(p_1)$ denote the log partial likelihood for the best $J=1$ model and $L(p_1,p_2)$ denote the log partial likelihood for the best $J=2$ model. Royston and Altman (1994) suggest, and verify with simulations, that each term in the fractional polynomial model contributes approximately 2 degrees-of-freedom to the model, effectively one for the power and one for the coefficient. Thus, the partial likelihood ratio test comparing the linear model to the best $J=1$ model,

$$G(1,p_1)=-2\left\{L(1)-L(p_1)\right\},$$

is approximately distributed as chi-square with 1 degree-of-freedom under the null hypothesis of linearity. The partial likelihood ratio test comparing the best $J=1$ model to the best $J=2$ model,

$$G[p_1,(p_1,p_2)]=-2\{L(p_1)-L(p_1,p_2)\},$$

is approximately distributed as chi-square with 2 degrees-of-freedom under the null hypothesis that the second function is equal to zero. Similarly, the partial likelihood ratio test comparing the linear model to the best $J=2$ model is distributed approximately as chi-square with 3 degrees-of-freedom. Note that to keep the notation simple, we have used p_1 to denote the best power both when $J=1$ and as the first of the two powers for $J=2$. These are not likely to be the same numeric value in practice.

In an applied setting, the partial likelihood ratio tests are used to choose which of the three forms of the covariate is best. In general, we recommend that, if a more complicated model is selected for use, it should provide a statistically significant improvement over a simpler model, and the transformations should make clinical sense.

The only software package that has fully implemented the method of fractional polynomials is STATA. In addition to the method as described above, STATA's fractional polynomial routine offers the user considerable flexibility in expanding the set of powers searched; however, in most settings the default set of values should be adequate.

Graphical methods to check the scale of covariates may be performed in most software packages. The most easily used of these are similar to residual methods from linear regression; see Ryan (1997). A complete discussion of residuals is provided in Chapter 6. The reader wishing to know the details of residual construction is welcome to read Section 6.2 before proceeding, but it is not necessary for the purpose of using them in plots to assess the scale of a covariate. The components of the residual for the ith subject are the value of the censoring variable, c_i, and the estimated cumulative hazard $\hat{H}_i=\hat{H}\left(t_i,\mathbf{x}_i,\hat{\boldsymbol{\beta}}\right)$; see (3.41), and these are used to calculate the martingale residuals, defined as

$$\hat{M}_i=c_i-\hat{H}_i.$$

Therneau, Grambsch and Fleming (1990) suggest fitting a model that excludes the covariate of interest. The results are used to calculate $\hat{M}_i$ and to generate smoothed values (e.g., the lowess smooth). These are then plotted versus the values of the excluded covariate, and the shape of the plot, and especially the smooth, provides an estimate of the functional form of the covariate in the model. Grambsch, Therneau and Fleming (1995) expand on their earlier work and suggest that one begin with a fit of the model containing all covariates. They demonstrate in

simulations and examples that a plot of the log of the ratio of a smoothed c to a smoothed $\hat{H}$ versus the covariate has greater diagnostic power than their earlier proposed method. Both of these plots are illustrated in the example in this chapter. Descriptions of applications of these and other related methods may be found in Therneau (1995).

Gray (1992) suggests that spline functions may be used as a way of modeling a continuous covariate without meeting stringent assumptions of a linear scale. Ryan (1997) discusses the construction and use of spline functions in linear regression. Harrell et al. (1996) demonstrate the use of spline functions in a variety of modeling settings, including the proportional hazards model. Since spline functions are not readily available in most software packages, they will not be discussed further or used in the example.

The final step in the variable selection process is to determine whether interactions are needed in the model. In this setting, an interaction term is a new variable that is the product of two covariates in the model. There may be special considerations that dictate that a particular interaction term or terms be included in the model, regardless of the statistical significance of the coefficient(s). If this is the case, these interaction terms and their component terms should be added to the main effects model and the larger model fit before proceeding with a statistical evaluation of other possible interactions. However, in most settings, there will be insufficient clinical theory to justify automatic inclusion of interactions.

The selection process begins by forming a set of biologically plausible interaction terms from the main effects in the model. The significance of each separate interaction is assessed by adding it to the main effects model and using the partial likelihood ratio test. All interactions significant at the 5 percent level are then added jointly to the main effects model. Wald statistic p-values are used as a guide to selecting interactions that may be eliminated from the model, but significance should be checked by the partial likelihood ratio test.

Several important points should be kept in mind when selecting interaction terms. Since the reason for including interactions is to improve inferences and obtain a more realistic model, we feel that all interaction terms should be statistically significant at usual levels of significance, such as 5 or 10 percent, and perhaps as low as 1 percent in some settings. Inclusion of nonsignificant interactions in a model will needlessly increase standard error estimates, thus unnecessarily widening confidence interval estimates of hazard ratios.

When an interaction term is added to a model, large changes in the coefficients of the corresponding main effects are likely to occur. However, changes in the main effect coefficients induced by interaction terms are not relevant and, as a result, do not indicate confounding. When interaction terms are present, the corresponding main effect terms do not, in most cases, estimate hazard ratios of interest. In addition, when there is statistically significant interaction, we include the corresponding main effect terms in the model regardless of their statistical significance. We are interested in examining how the main effect and interaction terms combine to estimate hazard ratios.

At this point we have a preliminary model. Our next step would be to assess its fit and adherence to key assumptions. These methods are discussed in the next chapter.

We illustrate the method of purposeful selection of covariates using the data from the UIS. These data were introduced in Section 1.3 and the variables are defined in Table 1.3. Recall that the goal of the study was to compare the effectiveness of two treatments (of different lengths) for the prevention of return to drug use at two different sites. At this point we will not consider the covariate, length of stay, for inclusion in the model since it is related to the outcome variable, time to drug use as measured from admission date. We will use it when we consider extending the proportional hazards model to include time-dependent or varying covariates in Chapter 7.

A modification that is sometimes used in a clinical trial setting where there is a clear "treatment" variable is to exclude the treatment variable from the variable selection process. The treatment variable is then added to the preliminary main effects model containing all of the variables associated with outcome, irrespective of treatment. The rationale for this approach is that one obtains an estimate of the additional effect of treatment, adjusting for other covariates. This approach is in contrast to modeling in epidemiological studies where "treatment" would be the risk factor of interest. In these settings, selection of variables may be based on the change in the coefficient (estimate of effect) of the risk factor variable. Thus, rather than being the last variable to enter, the risk factor enters the model first. What this points out is that one must have clear goals for the analysis and proceed thoughtfully using a variety of statistical tools and methods. The variable selection methods discussed may be an integral part of this analysis. In the example, we include the treatment variable among those in the first multivariable model.

The results of the bivariate analysis of each covariate in relation to time to drug relapse are presented in Table 5.1 for discrete covariates

Table 5.1 Estimated Median Time to Drug Use with 95% Brookmeyer–Crowley Confidence Intervals, Log-Rank Test and Partial Likelihood Ratio Test *p*-Values for Categorical Covariates in the UIS (n = 628)

Variable	Category	Median Time to Drug Use (95% CIE)	Log-Rank Test *p*-Value	Partial Likelihood Ratio Test *p*-Value
HERCOC	Both	150 (106,196)	0.047	0.051
	Heroin	142 (110, 184)		
	Cocaine	183 (148, 226)		
	Neither	181 (154, 220)		
IVHX	Never	194 (171, 228)	<0.001	<0.001
	Previous	170 (130, 226)		
	Recent	147 (115, 168)		
RACE	White	152 (124, 174)	0.007	0.006
	Other	193 (164, 232)		
TREAT	Short	130 (113, 154)	0.009	0.010
	Long	190 (175, 226)		
SITE	A	156 (131, 174)	0.124	0.121
	B	198 (159, 231)		
AGE	20 - 27	154 (121, 198)	0.282	0.282
	28 - 32	148 (123, 180)		
	33 - 37	162 (121, 207)		
	38 - 56	189 (162, 242)		
BECKTOTA	0 - <10	211 (166, 245)	0.229	0.229
	10 - <15	169 (124, 208)		
	15 - <25	168 (136, 192)		
	25 - <55	147 (106, 187)		
NDRUGTX	0 - 1	170 (142, 227)	0.002	0.002
	2 - 3	177 (162, 207)		
	4 - 6	127 (106, 183)		
	7 - 40	123 (106, 184)		

and in Table 5.2 for continuous covariates. All variables, except age categorized in four groups, are significant at the 20 percent level and therefore are candidates for inclusion in the multivariable model. The discrete forms, in Table 5.1, of the continuous covariates in Table 5.2 are presented primarily for descriptive purposes. If a variable is significant with either coding scheme, the variable should be added to the list

Table 5.2 Estimated Hazard Ratio for Time to Drug Relapse with 95% Confidence Intervals, Wald Test and Partial Likelihood Ratio Test *p*-Values for Continuous Covariates in the UIS (n = 628)

Variable	Change	Hazard Ratio for Change (95% CIE)	Wald Test *p*-Value	Partial Likelihood Ratio Test *p*-Value
AGE	5 years	0.94 (0.87, 1.01)	0.074	0.072
BECKTOTA	10 points	1.12 (1.02, 1.22)	0.020	0.021
NDRUGTX	5 treatments	1.16 (1.08, 1.25)	<0.001	<0.001

for inclusion in the multivariable model and used in its continuous form. We assess the correct scale of the variable following the fitting of the preliminary main effects model.

Before we fit the multivariable model, we note the close agreement in Table 5.1 between the significance levels of the partial likelihood ratio test and the log-rank test. This is as expected since, for a discrete covariate, the score test is algebraically related to the log-rank test and the performance of the score test is quite similar to the partial likelihood ratio test. The implication is that the log-rank test is an acceptable choice for purposes of covariate selection for the initial multivariable model.

Table 5.3 presents the results of fitting the multivariable proportional hazards model containing all variables significant at the $p < 0.20$ level in the bivariate analysis. This analysis includes 575 subjects for whom complete information is available on all covariates. Examining the *p*-values for the Wald statistics with the goal of trying to simplify the model, we note that none of the design variables for previous heroin or cocaine use is significant. The coefficient for intervention site is also not significant but, due to its importance in the study design, we keep it in the model. We next fit a model excluding previous heroin or cocaine use. Table 5.4 presents the results of fitting this reduced model. The partial likelihood ratio test comparing the models in Tables 5.3 and 5.4 is $G = 1.39$ which, with 3 degrees-of-freedom, has a *p*-value of 0.71, supporting our decision to remove the variable. The maximum change in the coefficient for any variable remaining in the model is 18.5 percent for the design variable for recent IV drug use, IVHX_3. This is not judged to be an important enough change to warrant inclusion of the heroin and cocaine use design variables in the model, so we proceed with the simpler model.

Table 5.3 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Proportional Hazards Model Containing Variables Significant at the 20% Level in the Bivariate Analysis for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
AGE	–0.029	0.008	–3.53	<0.001	–0.045, –0.013
BECKTOTA	0.008	0.005	1.68	0.094	–0.001, 0.018
NDRUGTX	0.028	0.008	3.42	0.001	0.012, 0.045
HERCO_2	0.065	0.150	0.44	0.663	–0.228, 0.359
HERCO_3	–0.094	0.166	–0.57	0.572	–0.418, 0.231
HERCO_4	0.028	0.160	0.18	0.861	–0.286, 0.342
IVHX_2	0.174	0.139	1.26	0.208	–0.097, 0.446
IVHX_3	0.281	0.147	1.91	0.056	–0.007, 0.569
RACE	–0.203	0.117	–1.74	0.082	–0.432, 0.026
TREAT	–0.240	0.094	–2.54	0.011	–0.425, –0.055
SITE	–0.102	0.109	–0.94	0.348	–0.317, 0.112

Log-likelihood = –2640.0305.

Examining the p-values for the Wald statistics in Table 5.4, we find that other than SITE (which, for practical reasons, will stay in the model) and BECKTOTA (which is marginally significant), the only non-significant variable is one of the pair of design variables IVHX_2 and IVHX_3, which together describe previous IV drug use. Two possible modeling strategies are: (1) Keep the design variables intact using all three codes, or (2) collapse the categories for "never" and "previous" to create a binary variable coded as "not recent" versus "recent." The

Table 5.4 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Reduced Proportional Hazards Model for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
AGE	–0.028	0.008	–3.45	0.001	–0.044, –0.012
BECKTOTA	0.008	0.005	1.60	0.110	–0.002, 0.077
NDRUGTX	0.028	0.008	3.35	0.001	0.012, 0.044
IVHX_2	0.196	0.137	1.43	0.153	–0.073, 0.465
IVHX_3	0.333	0.120	2.78	0.006	0.098, 0.568
RACE	–0.209	0.116	–1.81	0.071	–0.436, 0.018
TREAT	–0.232	0.094	–2.47	0.013	–0.415, –0.048
SITE	–0.099	0.109	–0.92	0.359	–0.312, 0.113

Log-likelihood = –2640.7278.

Table 5.5 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Reduced Proportional Hazards Model for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	P>|z|	95% CIE
AGE	–0.026	0.008	–3.25	0.001	–0.042, –0.010
BECKTOTA	0.008	0.005	1.70	0.090	–0.001, 0.018
NDRUGTX	0.029	0.008	3.54	<0.001	0.013, 0.045
IVHX_3	0.256	0.106	2.41	0.016	0.047, 0.464
RACE	–0.224	0.115	–1.95	0.051	–0.450, 0.001
TREAT	–0.232	0.093	–2.48	0.013	–0.416, –0.049
SITE	–0.087	0.108	–0.80	0.422	–0.298, 0.124

Log-likelihood = –2641.7294.

second choice yields a simpler model and may be preferred if the non-significant design variable does not confound the associations of the remaining variables in the model. Table 5.5 presents the results of fitting the model with IV drug use recoded as "not recent" versus "recent." The partial likelihood ratio test comparing the models in Tables 5.4 and 5.5 is $G = 2.00$ which, with one degree-of-freedom, yields a p-value of 0.157. The maximum percent change in a coefficient is –23.1 percent for the new binary variable, IVHX_3, but this change is uninterpretable since the reference group is different in the two models. Arguments could be given for the use of either the three-code version or the collapsed two-code version of the IV drug use variable. We will use the binary variable as it yields a simpler model and going from three to two codes has not changed the coefficients for any of the other variables, most notably treatment.

The next step in the modeling process is to examine the scale of the three continuous variables in the model: AGE, BECKTOTA and NDRUGTX. The first method we illustrate is the use of design variables. This approach to scale selection involves, for each of the three continuous variables, replacing the variable in the model with three design variables formed using the cutpoints shown in Table 5.1. Table 5.6 presents a summary of the resulting coefficients and group midpoints. The second step is to graph the coefficients against the group midpoints. These are shown in Figure 5.1.

The plots of the coefficients for age and especially Beck score support an assumption of linearity in the log hazard. The shape of the plot for number of previous drug treatments is more complicated. Analysis

Table 5.6 Estimated Coefficients for the Three Design Variables Formed from the Cutpoints Shown in Table 5.1 for the Variables AGE, BECKTOTA and NDRUGTX, in the UIS (n = 575)

AGE		BECKTOTA		NDRUGTX	
Midpoint	Coeff.	Midpoint	Coeff.	Midpoint	Coeff.
24.0	0.000	5.0	0.000	0.5	0.000
30.5	0.036	12.5	0.047	2.5	–0.070
35.5	–0.209	20.0	0.098	5.0	0.259
47.5	–0.391	40.0	0.216	23.5	0.399

of the Wald statistics shows that the coefficient for the second group is not significant, and the Wald test of the equality of the coefficients for the third and fourth groups is $z = 1.02$ with a p-value of 0.312, also not significant. This suggests that an alternative coding possibility is to form a binary covariate using the median (three previous treatments) as the cutpoint. The results of fitting this model are encouraging, in that the coefficient for the new binary variable was highly significant and the fitted model had a log partial likelihood that was only slightly smaller than that from the model with the three design variables. We defer discussing this model until we explore the use of the method of fractional polynomials for examining the scale of the three continuous covariates.

The results of the fractional polynomial analysis for age and Beck score confirm what was observed in the plot of the design variables in Figures 5.1a and 5.1b. An assumption of linearity in the log hazard seems quite reasonable for these two variables, so the computer output will not be presented.

The analysis of number of previous drug treatments (NDRUGTX) suggested that the log hazard is not linear. Table 5.7 presents the fractional polynomial results. The table contains four rows, and each row corresponds to a particular parametrization of the number of previous drug treatments. The first row represents a model containing all covariates in Table 5.5 except NDRUGTX, that is, the coefficient is set equal to zero. The model represented in the second row is the one shown in Table 5.5, as noted by the power of 1 in the last column. The significance level reported in the third column of the second row is for the partial likelihood ratio test of NDRUGTX entering the model as a linear term, that is,

$$G = [5294.497 - 5283.459] = 11.038$$

and p = 0.00099. The best power when NDRUGTX enters the model with a single, $J=1$, term is $p_1=0.5$ (i.e., the square root of NDRUGTX). The approximate partial likelihood ratio test comparing the use of $p_1=1$ to $p_1=0.5$ is

$$G=[5283.459-5283.088]=0.371,$$

and the reported p-value is $\Pr[\chi^2(1)\geq 0.371]=0.543$. From this we conclude that a model using the square root of NDRUGTX is no better than a model using NDRUGTX as a linear term. The best powers when NDRUGTX enters the model with two terms, $J=2$, is described by $(p_1=-1,\ p_2=-1)$. The interpretation is that the two terms are x^{-1} and $(x^{-1})\ln(x)$. Since NDRUGTX is equal to zero for some subjects, the software fits the model using $x=(\text{NDRUGTX}+1)/10$. The partial likelihood ratio test of this model versus the linear model is

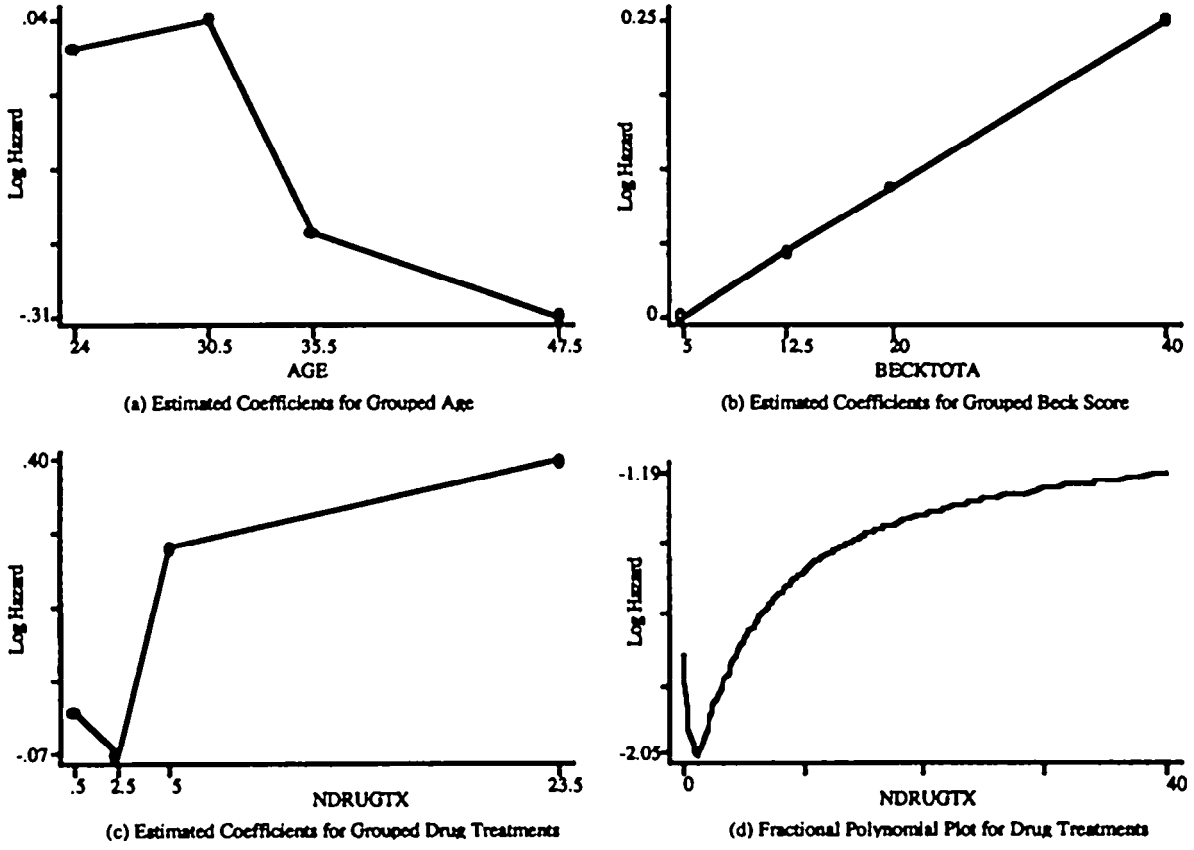


Figure 5.1 Graphs of estimated coefficients versus group midpoints for (a) AGE, (b) BECKTOTA, (c) NDRUGTX, and (d) the best two-term fractional polynomial model (–1,–1) for NDRUGTX.

Table 5.7 Summary of the Use of the Method of Fractional Polynomials for Number of Previous Drug Treatments for the UIS (n = 575)

	−2×Log-like.	G for Model vs Linear	Approx. p-Value	Powers
Not in model	5294.497			
Linear	5283.459	0.000	0.001*	1
J = 1 (2 df)	5283.088	0.371	0.543+	0.5
J = 2 (4 df)	5276.543	6.916	0.038#	−1, −1

* Compares linear model to model without NDRUGTX.
+ Compares the best J = 1 model to one with NDRUGTX linear.
# Compares the best J = 2 model to the best J = 1 model.

$$G = [5283.459 - 5276.543] = 6.916,$$

and its significance is $p = \Pr[\chi^2(3) \geq 6.916] = 0.075$. The partial likelihood ratio test of the best $J = 1$ model to the best $J = 2$ model is

$$G = [5283.088 - 5276.543] = 6.545$$

with $p = \Pr[\chi^2(2) \geq 6.545] = 0.038$. This test has 2 degrees-of-freedom since, when J is increased from 1 to 2, two additional terms (power and coefficient) are added to the model. To aid in the interpretation of the best two-term model, its graph is presented in Figure 5.1d. Even though the vertical scales are different in Figures 5.1c and 5.1d, there is a striking similarity in their shape, suggesting that the drop in the log-hazard function for a few previous drug treatments may be an important finding. This point is discussed in more detail in Chapter 6.

The two residual-based plots discussed earlier may be used as an alternative or adjunct to the method of fractional polynomials. The plots are shown in Figure 5.2 for age, in Figure 5.3 for the Beck score, and in Figure 5.4 for number of previous drug treatments. Each figure contains two plots. The top plot (a) is of the residuals and their smooth from a model that excludes the covariate of interest. The bottom plot (b) is of the log of the ratio of smoothed censor to smoothed cumulative hazard, called the expected in the figure headings. Since the scales of the two plots are different, plot (b) tends to overemphasize the shape, but the shapes in the two plots are consistent with each other.

For the sake of clarity we describe in some detail the steps used to produce the two plots in Figure 5.2. For Figure 5.2a, we fit a model containing the covariates in the model shown in Table 5.5, excluding

AGE. We requested that the values of $\hat{M}_i$, the martingale residuals, be calculated and saved. Figure 5.2a is a scatterplot of the $\hat{M}_i$ and their lowess smooth versus age.

To construct Figure 5.2b, we began by fitting the model in Table 5.5, including AGE, and requested that the martingale residuals be calculated and saved, also denoted as $\hat{M}_i$ for ease of notation. These residuals were used to calculate $\hat{H}_i = c_i - \hat{M}_i$, where c is the censoring variable. The values of c_i were plotted versus age and a lowess smooth was calculated and saved, denoted c_{ism}. The values of $\hat{H}_i$ were plotted versus age and a lowess smooth was calculated and saved, denoted $\hat{H}_{ism}$. The smoothed values were used to calculate

$$f_i = \ln\left(\frac{c_{ism}}{\hat{H}_{ism}}\right) + \hat{\beta}_{\text{AGE}} \times \text{AGE}_i,$$

where $\hat{\beta}_{\text{AGE}} = -0.026$ from Table 5.5. Figure 5.2b is a plot of the pairs (f_i, AGE_i) connected by straight lines. The plots in Figure 5.3 and Figure 5.4 were obtained in an identical manner. The size of the plotting symbol for $\hat{M}_i$ in Figure 5.2a, 5.3a and 5.4a has been reduced to emphasize the smoothed values.

The smoothed values in Figures 5.2a and 5.2b are nearly straight lines, supporting our treatment of age as linear in the model. The plots in Figures 5.3a and 5.3b demonstrate the instability of smoothed values in areas where there are not many values. One subject had a Beck score of 54 and was censored at 621 days, and the next smallest Beck score was 43. Thus, one subject is causing the downturn seen in both parts of Figure 5.3. After eliminating the effect in the plot of this one subject, the smoothed values are nearly a straight line and support treating the Beck score as linear in the model.

The plots in Figure 5.4 show the same decline and then rise in the log hazard for number of previous drug treatments that was observed in Figure 5.1. The graphs clearly illustrate the nonlinear behavior of number of previous drug treatments in the model. However, even the most experienced analyst would be hard-pressed to come up with a parametric function describing this shape. An advantage of the method of fractional polynomials is that it suggests the functional form for nonlinearly scaled continuous covariates.

Table 5.8 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Proportional Hazards Model Using the Best Two-Term Fractional Polynomial Model for Number of Previous Drug Treatments for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
AGE	–0.028	0.008	–3.46	0.001	–0.044, –0.012
BECKTOTA	0.009	0.005	1.84	0.066	–0.001, 0.019
NDRUGFP1	–0.523	0.124	–4.20	<0.001	–0.767, –0.279
NDRUGFP2	–0.195	0.048	–4.04	<0.001	–0.289, –0.100
IVHX_3	0.259	0.108	2.39	0.017	0.047, 0.470
RACE	–0.242	0.116	–2.10	0.036	–0.468, –0.016
TREAT	–0.211	0.094	–2.25	0.024	–0.395, –0.027
SITE	–0.105	0.109	–0.97	0.335	–0.319, 0.109

Log-likelihood = –2638.272.

In summary, thoughtful model development should include the use of both the graphical methods described and the method of fractional polynomials to assess the scale of continuous covariates.

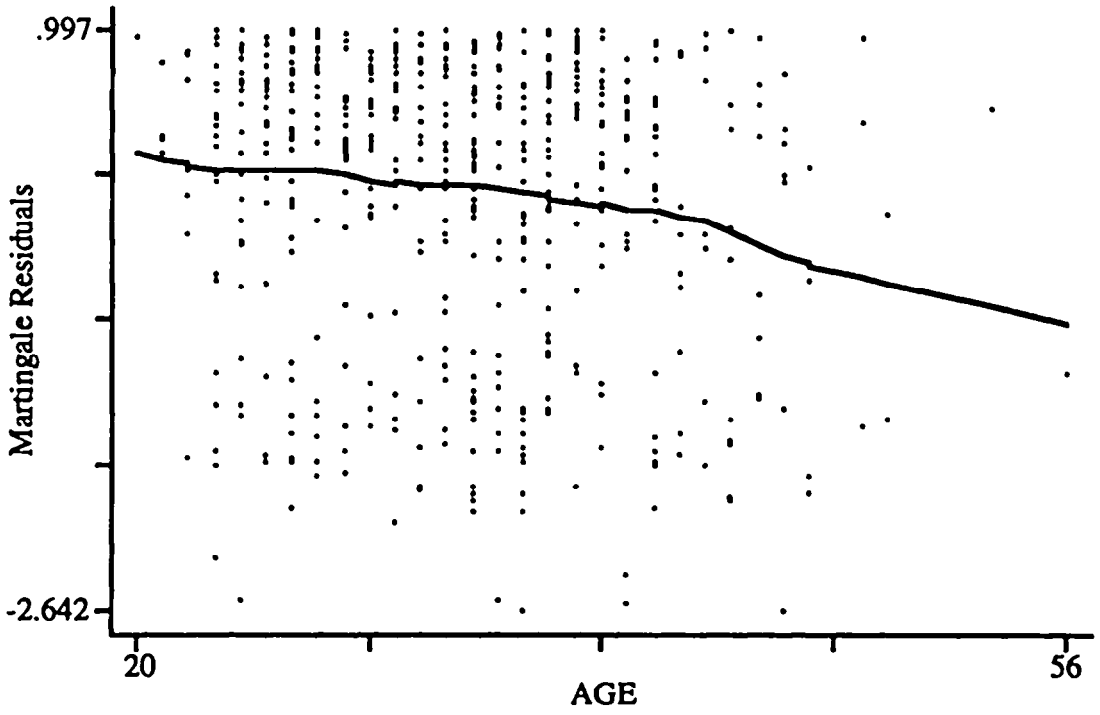
The final decision as to what scale to use comes down to a choice between a model with a single linear term, a binary variable using the median as the cutpoint, and the best two-term fractional polynomial model. The model with a single square root term is no better than the model with a single linear term. As noted in the discussion of the use of design variables, the model using a binary covariate with the median as the cutpoint was slightly better than the linear model, (log-likelihood = –2644.61 using a linear term versus a log-likelihood = –2643.58 using the binary coding). Given the simplicity and ease of interpretation of the binary coding, this model is the better of these two. The best fractional polynomial model is considerably more complicated than the binary model. However, consultation with the study team confirmed that the drop and rise in the log-hazard function seen in Figures 5.1c and 5.1d is not only plausible but of considerable interest. Thus, we will proceed using the scaling of NDRUGTX as selected by the method of fractional poly-nomials. Table 5.8 presents the results of fitting this preliminary main effects model. In this model

$$\text{NDRUGFP1} = [(\text{NDRUGTX}+1)/10]^{-1}$$

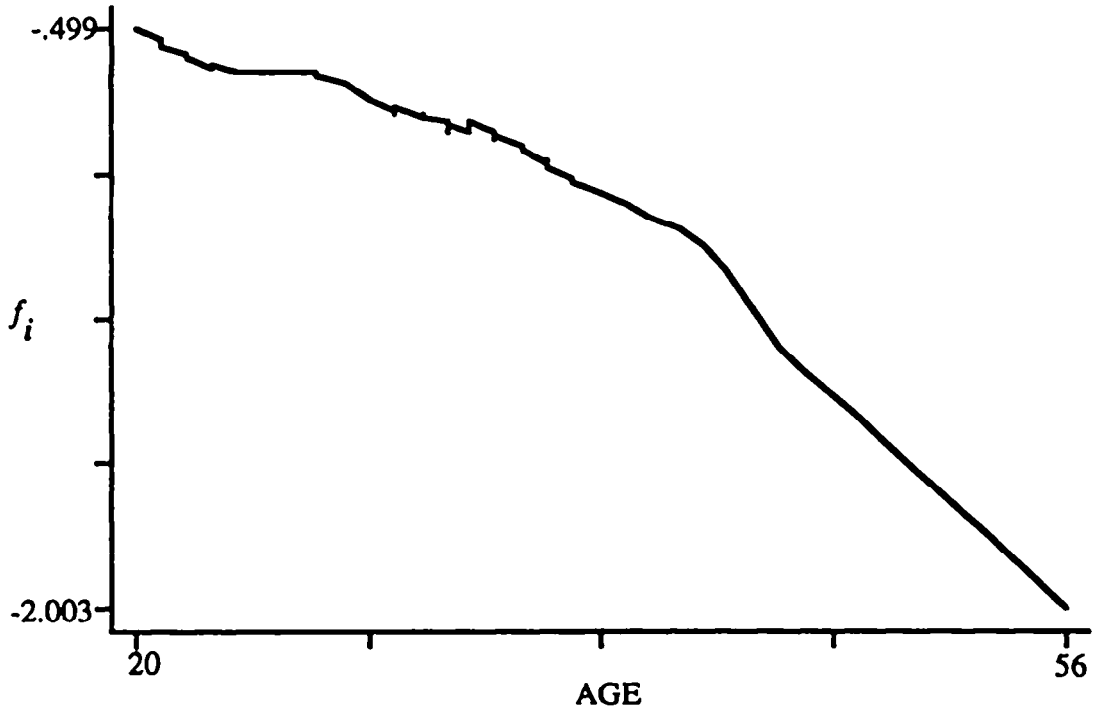
and

$$\text{NDRUGFP2} = [(\text{NDRUGTX}+1)/10]^{-1} \times \ln[(\text{NDRUGTX}+1)/10].$$

Figure 5.2a contains residuals from the model excluding AGE and Figure 5.2b the log of the ratio of smoothed values.



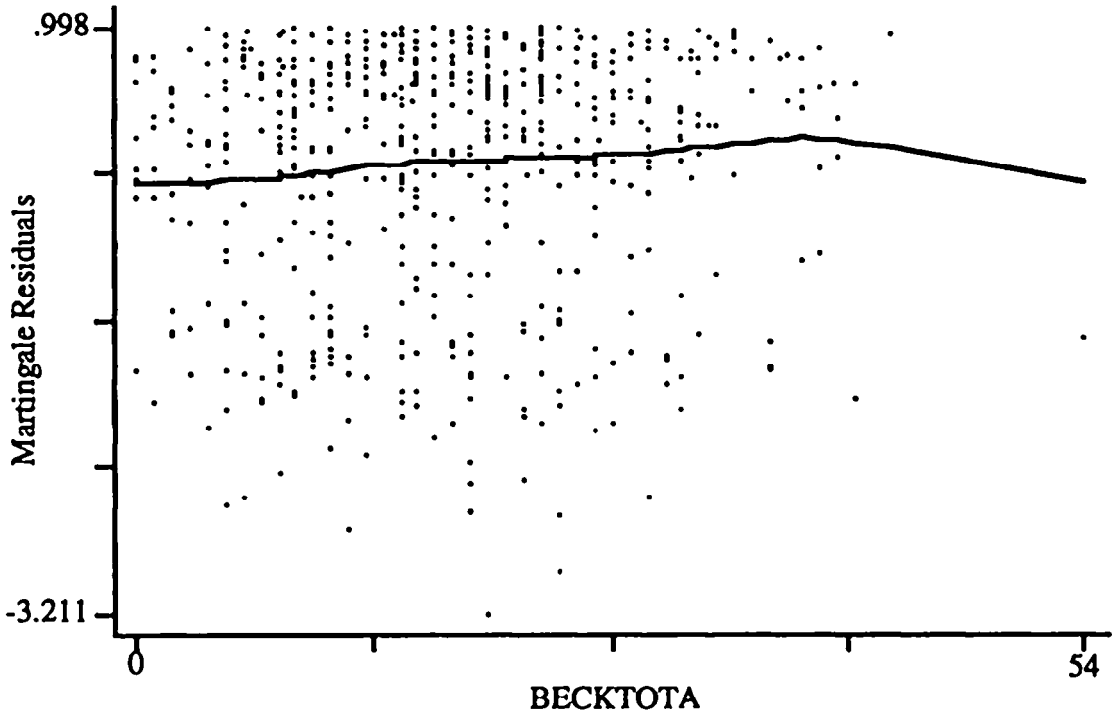
(a) Martingale Residuals and Lowess Smoothed Residuals.



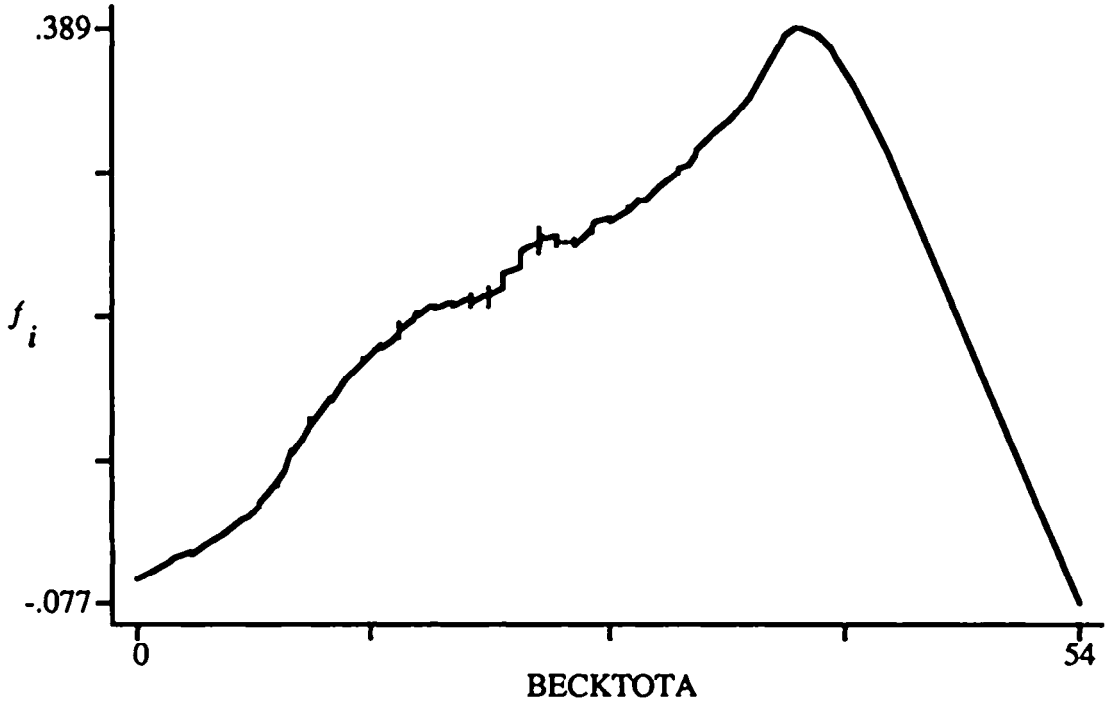
(b) Log(Smoothed Censor/Smoothed Expected) + β_{AGE}×AGE.

Figure 5.2 Plots of two residual-based methods for selecting the scale of AGE in the UIS ($n = 575$).

Figure 5.3a contains residuals from the model excluding BECKTOTA and Figure 5.3b the log of the ratio of smoothed values.



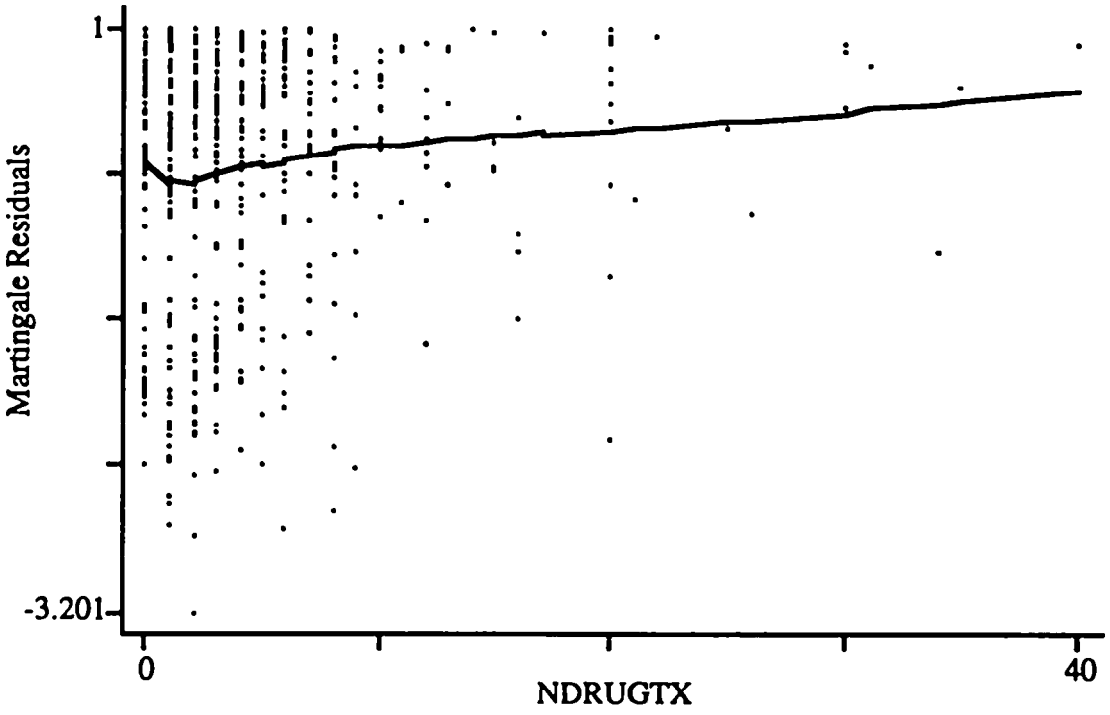
(a) Martingale Residuals and Lowess Smoothed Residuals.



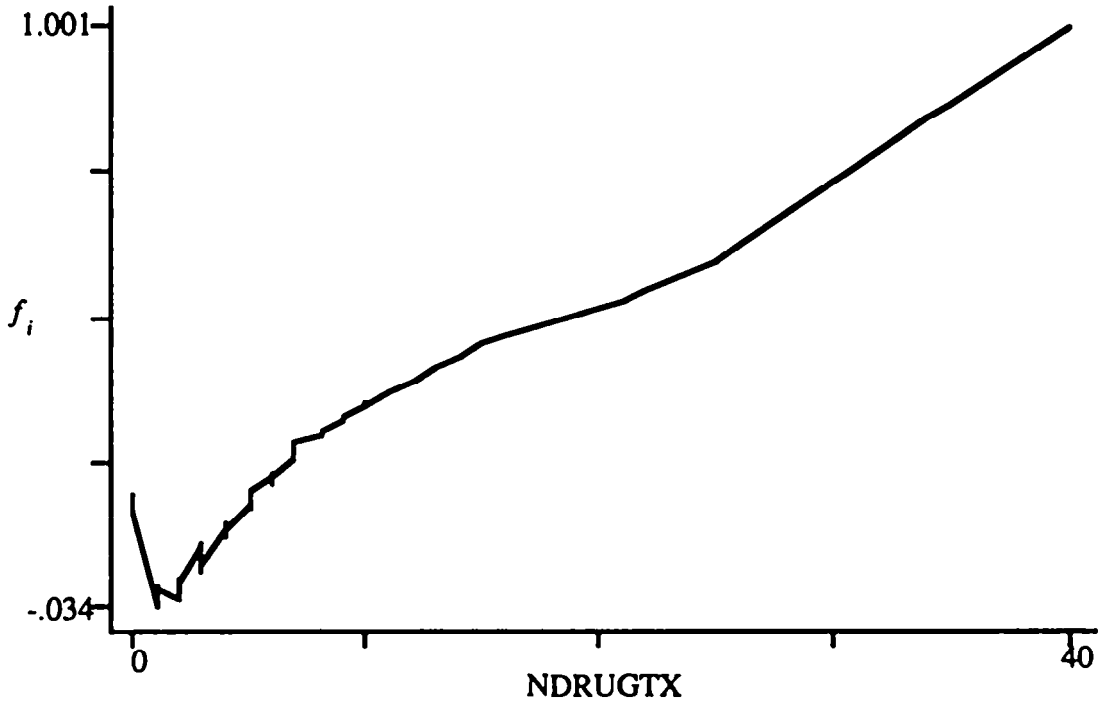
(b) Log(Smoothed Censor/Smoothed Expected) + $\beta_{BECKTOTA}$×BECKTOTA.

Figure 5.3 Plots of two residual-based methods for selecting the scale of BECKTOTA in the UIS ($n = 575$).

Figure 5.4a contains residuals from the model excluding NDRUGTX and Figure 5.4b the log of the ratio of smoothed values.



(a) Martingale Residuals and Lowess Smoothed Residuals.



(b) Log(Smoothed Censor/Smoothed Expected) + $\beta_{NDRUGTX}$×NDRUGTX.

Figure 5.4 Plots of two residual-based methods for selecting the scale of NDRUGTX in the UIS ($n = 575$).

Table 5.9 Interaction Variables, Degrees-of-Freedom (df) and *p*-Values for the Partial Likelihood Ratio Test for the Addition of the Interaction to the Model in Table 5.8

Interaction	Variables	df	*p*-Value
AGE	BECKTOTA	1	0.989
	NDRUGTX	2	0.028
	IVHX_3	1	0.460
	RACE	1	0.896
	TREAT	1	0.190
	SITE	1	0.028
BECKTOTA	NDRUGTX	2	0.316
	IVHX_3	1	0.241
	RACE	1	0.649
	TREAT	1	0.354
	SITE	1	0.912
NDRUGTX	IVHX_3	2	0.392
	RACE	2	0.746
	TREAT	2	0.214
	SITE	2	0.640
IVHX_3	RACE	1	0.568
	TREAT	1	0.534
	SITE	1	0.385
RACE	TREAT	1	0.310
	SITE	1	0.001
TREAT	SITE	1	0.247

The next step in the model building process is to add the design variables for heroin or cocaine use back into the model to be sure that they are neither significant in their own right nor confounders of the other main effects. The partial likelihood ratio test for the inclusion of heroin or cocaine use in the model is $G = 1.67$ which, with 3 degrees-of-freedom, yields a p-value of 0.644. The maximum percent change in a coefficient was less than 20 percent for all main effects in Table 5.8. We therefore conclude that heroin or cocaine use is not required in the model.

The final step in the model building process is the consideration of interaction terms. This step begins with the creation of a list of plausible interactions formed from the main effects in Table 5.8. Consultation with the study team determined that any pair of variables in the preliminary main effects model could generate a clinically plausible inter-

Table 5.10 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Preliminary Interactions Proportional Hazards Model for the UIS (n = 575)

Variable	Coeff.	Std. Err.	*z*	*P*>|z|	95% CIE
AGE	–0.054	0.028	–1.94	0.053	–0.109, 0.001
BECKTOTA	0.010	0.005	2.01	0.044	0.000, 0.020
NDRUGFP1	–0.674	0.644	–1.05	0.294	–1.938, 0.589
NDRUGFP2	–0.172	0.252	–0.68	0.496	–0.667, 0.322
IVHX_3	0.229	0.108	2.13	0.034	0.018, 0.441
RACE	–0.488	0.135	–3.62	<0.001	–0.752, –0.224
TREAT	–0.242	0.095	–2.56	0.010	–0.427, –0.057
SITE	–1.119	0.546	–2.05	0.040	–2.190, –0.049
AGEXSITE	0.026	0.017	1.60	0.111	–0.006, 0.059
RACEXSITE	0.863	0.248	3.48	0.001	0.376, 1.349
AGEXNDRUGFP1	0.002	0.019	0.08	0.933	–0.036, 0.040
AGEXNDRUGFP2	–0.002	0.008	–0.26	0.796	–0.017, 0.013

Log-likelihood = –2627.424

action. These are added, one at a time, to the preliminary main effects model. Table 5.9 presents the two variables forming the interaction, the degrees-of-freedom and the *p*-value for the partial likelihood ratio test comparing the models with and without the interaction. Thus Table 5.9

Table 5.11 Estimated Coefficients, Standard Errors, *z*-Scores, Two-Tailed *p*-Values and 95% Confidence Intervals for the Preliminary Final Proportional Hazards Model for the UIS (n = 575)

Variable	Coeff.	Std. Err.	*z*	*P*>|z|	95% CIE
AGE	–0.041	0.010	–4.18	<0.001	–0.061, –0.022
BECKTOTA	0.009	0.005	1.76	0.078	–0.001, 0.018
NDRUGFP1	–0.574	0.125	–4.59	<0.001	–0.820, –0.329
NDRUGFP2	–0.215	0.049	–4.42	<0.001	–0.310, –0.119
IVHX_3	0.228	0.109	2.10	0.036	0.015, 0.441
RACE	–0.467	0.135	–3.47	0.001	–0.731, –0.203
TREAT	–0.247	0.094	–2.62	0.009	–0.432, –0.062
SITE	–1.317	0.531	–2.48	0.013	–2.359, –0.275
AGEXSITE	0.032	0.016	2.02	0.044	0.001, 0.064
RACEXSITE	0.850	0.248	3.43	0.001	0.365, 1.336

Log-likelihood = –2630.418

contains all possible pairs of variables. The interaction terms are formed as the arithmetic product of the pair of variables. The interactions involving the number of previous drug treatments are formed using the two terms obtained from the method of fractional polynomials. Three interactions are identified as being significant, $p < 0.05$: age and number of previous drug treatments, age and site and race and site. These interactions were added to the preliminary main effects model in Table 5.8, and the resulting fitted model is shown in Table 5.10.

The p-values for the Wald tests suggest that the interaction between age and number of previous drug treatments may not be important in the larger interactions model. We note that the two coefficients for this interaction are of approximately the same magnitude, but with opposite signs. This suggests that these two variables may be highly colinear. To explore this, we fit the model containing only AGEXNDRUGFP1. The Wald statistic for its coefficient was significant ($p = 0.013$). However, the Wald statistic for AGEXSITE in this model was not significant ($p = 0.107$). To further explore the interactions with age, we fit a model containing only AGEXSITE. In this model, the Wald statistic for the coefficient was signifcant ($p = 0.044$). Thus it appears that we have a choice between two models, one containing only AGEXNDRUGFP1, the other containing only AGEXSITE. Since the latter model is simpler and easier to interpret, we define our preliminary final model as the one presented in Table 5.11. The model will not be identified as the final model until its fit and adherence to model assumptions nas been checked. Before this important topic is considered in detail in Chapter 6, we consider stepwise and best subsets selection of covariates, two statistical methods for selection of main effect variables.

5.3 STEPWISE SELECTION OF COVARIATES

Covariates may be selected for inclusion in a proportional hazards regression model using stepwise selection methods that operate in an identical manner to those used in other regression models, such as linear or logistic regression. The statistical test used as a criterion is most often the partial likelihood ratio test. However, the score test and Wald test are also often used by software packages. From the conceptual point of view, it does not matter which test is actually used. However, the partial likelihood ratio test is the best of the three tests and should be used when there is a choice.

We assume familiarity with stepwise methods from either linear or logistic regression, thus the presentation here will not be detailed. Detailed descriptions of stepwise selection of covariates may be found in Hosmer and Lemeshow (1989), Chapter 4, for logistic regression and in Ryan (1997), Chapter 7, for linear regression.

We begin by describing the full stepwise selection process, which consists of forward selection followed by backward elimination. The forward selection process adds to the model the covariate that is most statistically significant among those not in the model. The backward elimination process checks each covariate in the model for continued significance. Two variations of the full stepwise procedure available in most software packages are to use forward selection only or backward elimination only.

Most software packages that have implemented stepwise selection of covariates for the proportional hazards model treat all the covariates available for selection as if they were continuous. This implies that to consider nominal scale covariates with more than two levels correctly, one must create and include in the list of covariates the individual design variables. An additional problem is that the individual design variables are not considered as a unit, and the program may select a subset of them. When this occurs, there has been an implicit recoding of the covariate and the user must make sure that the recoding makes clinical sense or must add the unselected design variables into the model when it is examined in more detail. We will return to this point in the example. The stepwise procedure will be described, as it is currently implemented by default, using single degree-of-freedom tests for entry and removal of covariates.

Step 0: Assume that there are p possible covariates, denoted x_j, $j = 1,2,...,p$. This list is assumed to include continuous covariates as well as all design variables for nominal scaled covariates. Thus, for example, a particular x_j might stand for age or for the design variable for IV drug use at level 2. At step 0 the partial likelihood ratio test and its p-value for the significance of each covariate is computed by comparing the log partial likelihood of the model containing x_j to the log partial likelihood of model zero (i.e., the model containing no covariates). This test statistic is

$$G^{(0)}(j) = -2\left[L^{(0)}(j) - L(0)\right], \quad j = 1,2,\ldots,p, \tag{5.1}$$

where $L(0)$ is equal to the log partial likelihood of model zero, the no covariate model, and $L^{(0)}(j)$ is equal to the log partial likelihood of the model containing covariate x_j. The test's significance level is

$$p^{(0)}(j) = \Pr\left[\chi^2(1) \geq G^{(0)}(j)\right] . \tag{5.2}$$

Evaluation of (5.1) and (5.2) requires fitting p separate proportional hazards models. The parenthesized superscript in (5.1) and (5.2) denotes the step, and j indexes the particular covariate. The candidate for entry into the model at step 1 is the most significant covariate and is denoted by x_{e_1}, where

$$p^{(0)}(e_1) = \min_j \left[p^{(0)}(j)\right] . \tag{5.3}$$

For the variable x_{e_1} to be entered into the model, its p-value must be smaller than some pre-chosen criterion for significance, denoted p_E. If the variable selected for entry is significant (i.e., $p^{(0)}(e_1) < p_E$), then the program goes to step 1; otherwise it stops.

Step 1: This step begins with variable x_{e_1} in the model. Then $p-1$ new proportional hazards models (each including one remaining variable along with x_{e_1}) are fit, and the results are used to compute the partial likelihood ratio test of the fitted two-variable model to the one-variable model containing only x_{e_1},

$$G^{(1)}(j) = -2\left[L^{(1)}(j) - L\left(x_{e_1}\right)\right], \quad j = 1,2,\ldots,p \text{ and } j \neq e_1, \tag{5.4}$$

where the values of the deviance in (5.4) are –2 times the partial log likelihoods of the respective models. The p-value for the test of the significance of adding x_j to the model containing x_{e_1} is

$$p^{(1)}(j) = \Pr\left[\chi^2(1) \geq G^{(1)}(j)\right] . \tag{5.5}$$

The variable selected as the candidate for entry at step 2 is x_{e_2} where

$$p^{(1)}(e_2) = \min_{j \neq e_1} \left[p^{(1)}(j)\right] . \tag{5.6}$$

If the selected covariate x_{e_2} is significant, $p^{(1)}(e_2) < p_E$, then the program goes to step 2; otherwise it stops.

Step 2: This step begins with both x_{e_1} and x_{e_2} in the model. During this step, two different evaluations occur. The step begins with a backward elimination check for the continued contribution of x_{e_1}. That is, does x_{e_1} still contribute to the model after x_{e_2} has been added? This is essentially an evaluation of (5.4) and (5.5) with the roles of the two variables reversed. From an operational point of view, we choose a different significance criterion for this check, denoted p_R. We choose this value such that $p_R > p_E$ to eliminate the possibility of entering and removing the same variable in an endless number of successive steps. Assume the variable entered at step 1 is still significant.

The program fits $p-2$ proportional hazards models (each including one remaining variable along with x_{e_1} and x_{e_2}) and computes the partial likelihood ratio test and its p-value for the addition of the new covariate to the model, namely

$$G^{(2)}(j) = -2\left[L^{(2)}(j) - L\left(x_{e_1}, x_{e_2}\right)\right], \quad j = 1,2,\ldots,p \text{ and } j \neq e_1, e_2$$

and

$$p^{(2)}(j) = \Pr\left[\chi^2(1) \geq G^{(2)}(j)\right].$$

The covariate x_{e_3} selected for entry at step 3 is the one with the smallest p-value, that is,

$$p^{(2)}(e_3) = \min_{j \neq e_1, e_2} \left[p^{(2)}(j)\right].$$

The program proceeds to step 3 if $p^{(2)}(e_3) < p_E$; otherwise it stops.

Step 3: Step 3, if reached, is similar to step 2 in that the elimination process determines whether all variables entered into the model at earlier steps are still significant. The selection process then followed is identical to the selection part of earlier steps. This procedure is followed until the last step, step S.

Step S: At this step one of two things may happen: (1) all the covariates are in the model and none may be removed or (2) each covariate not in the model has $p^{(S)}(j) > p_E$. At this point, no covariates are selected for entry and none of the covariates in the model may be removed.

The number of variables selected in any application will depend on the strength of the associations between covariates and survival time and the choice of p_E and p_R. Due to the multiple testing that occurs, it is nearly impossible to calculate the actual statistical significance of the full stepwise process. Research in linear regression by Bendel and Afifi (1977) and in discriminant analysis by Costanza and Afifi (1979) indicates that use of $p_E = 0.05$ excludes too many important covariates and that one should choose a level of significance of 15 percent. In many applications it may make sense to use 25–50 percent to allow more variables to enter than will ultimately be used and then narrow the field of selected variables using $p < 0.15$ to obtain a multivariable model for further analysis. An unavoidable problem with any stepwise selection procedure is the potential for the inclusion of "noise" covariates and the exclusion of important covariates. One must always examine the variables selected and excluded for basic scientific plausibility.

The model at this point is likely to contain continuous covariates, and these should be examined carefully for linearity using the previously discussed methods. The next step is to see if there are any interactions which significantly improve the model. The procedure for stepwise selection is to use as candidate variables a list of plausible interactions among the main effects previously identified during the initial stepwise model building. One must begin with a model containing all the main effects, and the final model is selected using usual levels of statistical significance.

As an example of stepwise selection we consider covariates in the UIS. The list of candidate variables includes: age, Beck score, number of previous drug treatments, race, treatment, site, three design variables for previous heroin or cocaine use and two design variables for previous IV drug use, for a total of 11 covariates. The exact order of variable selection will depend on whether one uses the partial likelihood ratio test, the score test or the Wald test. The results presented in Table 5.12 were obtained using the partial likelihood ratio test. The variables selected are the same as those selected by the score test.

For illustrative purposes, the results presented in Table 5.12 were obtained using entry and removal p-values of $p_E = 0.5$ and $p_R = 0.8$. There were a total of 10 steps, counting step 0. At step 0, the variable with the smallest p-value was the number of previous drug treatments, NDRUGTX, with $p = 0.004$. Since this value is smaller than $p_E = 0.5$, the variable enters the model at step 1. At step 1 AGE has the smallest p-value with $p = 0.0064$ and it is smaller than $p_E = 0.5$, so AGE enters the model at step 2. At step 2, both AGE and NDRUGTX have p-values

to remove which are less than $p_R = 0.8$ and thus remain in the model. Among the variables not in the model, the design variable for IV drug use at level 3, IVHX_3 (a recent user), has the smallest p-value and it is less than the criteria for entry into the model. Then the program goes to step 3, where the three-variable model is fit. All the p-values to remove are less than 0.8 and no variables are taken from the model. The variable with the smallest p-value for entry is treatment, TREAT, with $p = 0.0107$, which is less then 0.5. The program then goes to step 4 and fits the four-variable model.

This process of fitting, checking for continued significance, and selection continues until step 9. At this step, each of the nine variables in the model has a p-value to remove which is less than 0.8, and the p-values to enter for the two variables not in the model exceed 0.5. Therefore, the program terminates the selection process at step 9.

We use the results in Table 5.12 with a significance level of 0.15 to identify the preliminary main effects model by proceeding sequentially to the next step, as long as the smallest p-value for entry is less than 0.15. The first time it exceeds 0.15 is at step 6. At this step, the potential variable for inclusion is the design variable for previous IV drug use at level 2, IVHX_2 (previous user). Using the 15 percent rule, we would take as our model the one fit at step 6 which contains NDRUGTX, AGE, IVHX_3, TREAT, RACE and BECKTOTA. Inclusion of only IVHX_3 implies a recoding of previous IV drug use to 1 = recent, 0 = not recent. If we were performing variable selection for the first time, we might

Table 5.12 Results of Stepwise Selection of Covariates, p-Value for Entry to the Right and p-Value to Remove to the Left of the Solid Line in Each Row for the UIS (n = 575). Columns are in Order of Entry.

Step	NDRUGTX	AGE	IVHX_3	TREAT	RACE	BECKTOT	IVHX_2	HERC_3	SITE	HERC_2	HERC_4
0	0.0004	0.0759	0.001	0.011	0.005	0.035	0.883	0.058	0.294	0.044	0.434
1	0.0004	0.0064	0.010	0.009	0.013	0.036	0.832	0.109	0.426	0.113	0.832
2	<.0001	0.0064	0.001	0.007	0.016	0.046	0.997	0.037	0.327	0.048	0.618
3	0.0008	0.0005	0.001	0.011	0.058	0.102	0.090	0.275	0.790	0.481	0.577
4	0.0005	0.0004	0.001	0.011	0.081	0.109	0.101	0.232	0.631	0.431	0.476
5	0.0010	0.0007	0.003	0.015	0.081	0.085	0.148	0.258	0.382	0.528	0.526
6	0.0009	0.0011	0.007	0.015	0.064	0.085	0.179	0.208	0.420	0.528	0.450
7	0.0016	0.0005	0.003	0.016	0.093	0.102	0.179	0.279	0.357	0.537	0.545
8	0.0013	0.0004	0.014	0.014	0.098	0.087	0.238	0.280	0.348	0.676	0.908
9	0.0014	0.0004	0.027	0.012	0.069	0.095	0.213	0.273	0.348	0.688	0.968

choose the model at step 7 to avoid having to recode this variable. If we do this, the next phase of the model building process would be the steps we went through during the purposeful selection of covariates. Recall that that process suggested the same set of covariates.

At this point, we examine the scale of the continuous covariates in the model following the same procedure illustrated in the previous section. This analysis yields the same model as in Table 5.8 if we add SITE for the same reasons it was included when we discussed purposeful selection. Stepwise selection of interactions proceeds using as candidate variables the interactions listed in Table 5.9. At step 0, the model contains all the main effects, the model in Table 5.8. Since all the stepwise selection programs perform single degree-of-freedom selection tests, there are a total of 27 individual interaction terms to choose from at step 0, since the process of checking the scale of the continuous covariates has led to transforming NDRUGTX into two nonlinear terms.

For illustrative purposes we discuss the variables selected using 0.15 as the significance level for entry. We use a smaller level of significance for entry as we wish to include in the model only those interactions that are significant, since confounding is not an issue when selecting interactions. Three interactions were identified as being important: At step 1, the RACE by SITE interaction entered the model with $p = 0.001$; at step 2, the AGE by NDRUGFP2 interaction entered the model with $p = 0.006$; and at step 3, the AGE by SITE interaction entered the model

Table 5.13 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for the Preliminary Final Proportional Hazards Model for the UIS (n = 575)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
AGE	–0.054	0.012	–4.53	<0.001	–0.077, –0.031
BECKTOTA	0.010	0.005	2.07	0.038	0.001, 0.020
RACE	–0.483	0.135	–3.59	<0.001	–0.747, –0.219
TREAT	–0.222	0.094	–2.37	0.018	–0.406, –0.039
SITE	–0.278	0.122	–2.28	0.023	–0.517, –0.039
IVHX_3	0.234	0.108	2.17	0.030	0.023, 0.445
NDRUGFP1	–0.838	0.160	–5.25	<0.001	–1.151, –0.525
NDRUGFP2	–0.229	0.049	–4.70	<0.001	–0.325, –0.134
RACEXSITE	0.897	0.247	3.63	<0.001	0.412, 1.382
AGEXNDRUGFP1	0.007	0.003	2.77	0.006	0.002, 0.012

Log-likelihood = –2628.739.

with $p = 0.113$. Using the 5 percent level of significance, we would choose the model at step 2 containing all the main effects, the RACE by SITE interaction and, surprisingly, the AGE by NDRUGFP2 interaction.

At this point we have a somewhat complicated model to sort out. Subsequent analyses (details are not presented) reveal that: (1) there is a significant interaction of AGE with either one of the two fractional polynomial variables for number of previous drug treatments, (2) it does not seem to matter which of the two fractional polynomial variables AGE interacts with as both models have almost the same log partial likelihood and (3) the AGE by SITE interaction is significant only if the AGE by NDRUGFP1 or the AGE by NDRUGFP2 interaction is not included in the model (see Table 5.11).

Thus it appears that there are two possible models, each with the same eight main effect terms and two interaction terms: (1) the model in Table 5.11 containing the AGE × SITE and RACE × SITE interactions and (2) the model in Table 5.13 containing the RACE × SITE and AGE × NDRUGFP1 interactions. We defer deciding which of the two models to use for inferential purposes until after we have examined each for adherence to model assumptions, goodness-of-fit, and influential observations. From a practical point of view, we favor the model in Table 5.11 as it does not include any interaction terms involving fractional polynomials, making it easier to interpret. However, if nothing changes, the estimate of the effect of treatment is about the same in both models, so from that point of view we could use either model.

5.4 BEST SUBSETS SELECTION OF COVARIATES[1]

In the previous section we discussed stepwise selection of covariates. The advantage of stepwise selection is that most analysts are familiar with its use from other regression modeling settings and it is available in most major software packages. A disadvantage is that the procedure considers only a small number of the total possible models that can be formed from the covariates. The method of best subsets selection provides a computationally efficient way to screen all possible models.

The conceptual basis for best subsets selection of covariates in a proportional hazards regression is the same as in linear regression. The

---

[1] Implementation of the methods in this section requires matrix calculations not automatically performed by software packages. If one is familiar with simple matrix algebra and the software package has matrix capabilities, then they are relatively easy to perform.

procedure requires a criterion to judge a model. Given the criterion, the software screens all models containing q covariates and reports the covariates in the best, say 5, models for $q = 1,2,\ldots,p$, where p denotes the total number of covariates.

Software to implement best subsets normal errors linear regression is readily available and has been used to provide best subsets selection capabilities for non-normal errors linear regression models such as logistic regression, see Hosmer and Lemeshow (1989, Chapter 4). There are three requirements to use the method described by Hosmer and Lemeshow: (1) It must be possible to obtain estimates of the coefficients of the model containing all p covariates from a weighted linear regression where the dependent variable is of the form

$$\mathbf{x}'\hat{\boldsymbol{\beta}} + \hat{\text{weight}} \times (\hat{\text{residual}}),$$

(2) the weight must be an easily computed function of the variance of the residual and (3) both weight and residual must be easily computed functions of the estimated coefficients and covariates. Only requirement 1 is satisfied by the proportional hazards regression model when it is fit using the partial likelihood. The difficulty is that even though the partial likelihood, see (3.17), is a product of n terms, the terms are not independent of each other. Each "subject" may contribute information to more than one term in the product, that is, "subjects" appear in every risk set until they fail or are censored. Thus Hosmer and Lemeshow's method may not be used to perform best subsets proportional hazards regression. We do not want to dwell on this point, but feel that it is important to explain why this well-known and easily used approach is not appropriate in this setting.

Kuk (1984) described how best subsets selection in a proportional hazards regression model may be performed with a normal errors linear regression best subsets program if the program allows input of the data via a covariance matrix. Kuk's method is related to a general method described by Lawless and Singhal (1978), which requires special software. We will illustrate Kuk's method using BMDP9R, but any best subsets linear regression program which permits a covariance matrix as data input can be used.

The computational steps one must perform to use Kuk's method are as follows:

(1) Fit the proportional hazards model containing all p covariates. This model must contain all the design variables for nominal scale

covariates coded at more than two levels. As was the case in stepwise selection, these related design variables will be considered as distinct binary variables in the best subsets selection method.

(2) Obtain the estimated covariance matrix of the estimated coefficients, denoted as $\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})$, and obtain its inverse, denoted

$$\mathbf{I}(\hat{\boldsymbol{\beta}}) = \left[\hat{\mathrm{Var}}(\hat{\boldsymbol{\beta}})\right]^{-1}.$$

This matrix is the observed information matrix. If the program does not provide the observed information matrix, then one must compute its value.

(3) Compute the $p \times 1$ matrix $\mathbf{B} = \mathbf{I}(\hat{\boldsymbol{\beta}})'\hat{\boldsymbol{\beta}}$ and the 1×1 matrix $\mathbf{C} = \hat{\boldsymbol{\beta}}'\mathbf{I}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{\beta}}$.

(4) Use the matrices computed in steps 2 and 3 to form the $(p+1)\times(p+1)$ matrix

$$\mathbf{A} = \begin{bmatrix} \mathbf{I}(\hat{\boldsymbol{\beta}}) & \mathbf{B} \\ \mathbf{B}' & (n-p)+\mathbf{C} \end{bmatrix}.$$

[There is a small mistake in Kuk (1984) in that he adds $(n-p-1)$ to C.]

(5) Verify that the matrix A is correct. This may be done by performing linear regression with covariance matrix input, A, declaring the $(p+1)$st variable as the dependent variable and assigning the names used in fitting the proportional hazards regression model in step 1 to the first p variables. The estimated coefficients and estimated standard errors of the estimated coefficients obtained from the linear regression output should be equal to those computed in step 1 from the proportional hazards regression model. The mean residual sum-of-squares should be equal to $n-1$. [Another small mistake in Kuk (1984) is that he states that this mean square is equal to 1.0.]

(6) Use a best subsets linear regression program with data set up as in step 5. Select Mallow's C [Mallows (1973)] as the criterion for best subsets [see Ryan (1997, Chapter 7) for a discussion of the use of Mallow's C in normal errors linear regression modeling].

The result of using these computational tricks is that best subsets are chosen using the values of multivariable Wald tests obtained after fitting the full p variable proportional hazards model, that is, the model fit in step 1. We will apply best subsets selection to the 11 possible main ef-

Table 5.14 Five Best Models Identified Using Mallow's C. Model Covariates, Mallow's C, the Wald Test for the Excluded Covariates, Its Degrees-of-Freedom and p-Value for the UIS (n = 575)

Model	Model Covariates	C	W	df	p
1	AGE, BECK, NDRUGTX, IVHX_3, RACE, TREAT	5.06	4.06	5	0.541
2	AGE, BECK, NDRUGTX, IVHX_2, IVHX_3, RACE, TREAT	5.21	2.21	4	0.697
3	AGE, BECK, NDRUGTX, HER_3, IVHX_3, RACE, TREAT	5.48	2.48	4	0.648
4	AGE, BECK, NDRUGTX, IVHX_2, IVHX_3, TREAT	5.93	4.93	5	0.424
5	AGE, NDRUGTX, IVHX_2, IVHX_3, RACE, TREAT	5.94	4.94	5	0.423

fect variables used in the UIS. For sake of illustration, consider a possible model that excludes 4 variables: the three design variables for heroin or cocaine use and SITE. The significance of the excluded variables may be assessed by the partial likelihood ratio test comparing the full 11-variable model to the 7-variable model containing AGE, BECKTOTA, NDRUGTX, IVHX_2, IVHX_3, RACE and TREAT. An equivalent test is the multivariable Wald test for the coefficients of the 4 excluded variables obtained following the fit of the full 11-variable model. The value of this 4 degrees-of-freedom Wald test is 2.41 and the value of Mallow's C for this 7-variable model is $5.21 = 2.21 + (11 - 2 \times 4)$.

In order to establish the relationship between Mallow's C and the Wald statistic in general, denote the value of the multivariable Wald test for q variables excluded from the full p variable model by W_q. The multivariable Wald test was described in Chapter 3 and is distributed as chi-square with degrees-of-freedom equal to the number of coefficients hypothesized to be equal to zero. The value of Mallow's C reported by BMDP9R, in step 6 above, is

$$C = W_q + (p - 2q) \ . \tag{5.7}$$

As in linear regression, good models will be ones with small values of C. Under the hypothesis that the coefficients for the q variables excluded from the model are zero, the mean of the Wald test is approximately q. Thus, the mean of Mallow's C is approximately $p - q$, the number of variables in the model. This is the same reference standard for Mallow's C used in normal errors linear regression.

Table 5.14 reports a summary of the five best models obtained by performing the six-step procedure with the UIS data. The best model

contains the same six covariates identified using both purposeful and stepwise selection methods. The second best model is the same as one alternative model identified by the stepwise method. The remaining three models suggest other possible sets of covariates. The values of Mallow's C are relatively homogeneous across the five models. This is also the case for the Wald test p-values. The covariates selected for these models suggest that any good model is going to contain age, number of previous drug treatments, a binary variable for recent IV drug use, and treatment. There are four other covariates suggested.

Since no one model appears to be superior to the other four, one possible strategy is to fit a multivariable model containing all eight covariates used in the five models in Table 5.14. Following the fit of this model, we would proceed as in purposeful selection to try and reduce the size of the model. Based on models fit in the section on purposeful selection of covariates, this process would return us to model 1 in Table 5.14. The next steps in model development are the same as those described and illustrated in the section on purposeful selection of covariates: assessment of the scale of continuous covariates and identification of interactions.

An alternative method for best subset selection is to mimic the approach used in stepwise selection and choose as best models those in which the covariates in the model are significant. Selection of covariates thus proceeds by inclusion rather than exclusion. The best models containing $p-q$ covariates are those with the largest values of a test of the significance of the model. Theoretically, one could use any one of the three equivalent tests: partial likelihood ratio, Wald or score test. The SAS package, PROC PHREG, has implemented this selection method using the score test. Models identified are, for each fixed number of covariates, the ones with the largest value of the score test.

The problem with using the score test for model significance is that it is difficult to compare models of different sizes since the score test tends to increase with the number of covariates in the model. One possible solution is to use the values of the score test to approximate the value of Mallow's C in (5.7). Let the score test for the model containing all p covariates be denoted S_p and the score test for the model containing a particular set of $p-q$ covariates be denoted S_{p-q}. The value of the score test for the exclusion of the q covariates from the full p variable model is approximately $S_q = S_p - S_{p-q}$. Since the Wald and score tests are equivalent, this suggests that an approximation to Mallow's C for a fitted model containing $p-q$ covariates is

$$C = S_q + (p - 2q). \tag{5.8}$$

We note that if covariate selection had been based on the partial likelihood ratio test instead of the Wald and score test, the value of C in (5.7) would be equal to the value in (5.8).

As an example, consider model 1 in Table 5.14, with $p-q = 11-5=6$ covariates in the model. The value of the score test for the significance of the 11-covariate model is $S_{11} = 49.35$, and the value of the score test for the significance of the 6-covariate model is $S_6 = 45.52$. The approximation to the score test for the addition of the 5 covariates to the 6-covariate model is

$$S_5 \cong S_{11} - S_6 = 49.35 - 45.52 = 3.83.$$

The value of the approximation to Mallow's C is

$$C = 3.83 + (11 - 2 \times 5) = 4.83$$

and the correct value from Table 5.14 is 5.06. The approximation is close, but certainly not perfect. Of more practical interest is what models would be selected as best using (5.8) in conjunction with the values of the score tests provided by SAS in PROC PHREG. The best five models using this approach are summarized in Table 5.15.

The results in Table 5.15 are quite similar to those in Table 5.14. The three best models are the same and the fourth best model in Table 5.15 is the same as the fifth best model in Table 5.14. Thus it appears that the approximation in (5.8) provides a useful way to rank order models containing different numbers of covariates when models have

Table 5.15 Five Best Models Identified Using the Score Test Approximation to Mallow's C. Model Covariates, Approximate Mallow's C and the Approximate Score Test for the Excluded Covariates for the UIS (n = 575)

Model	Model Covariates	C	S_q
1	AGE, BECK, NDRUGTX, IVHX_3, RACE, TREAT	4.83	3.83
2	AGE, BECK, NDRUGTX, IVHX_2, IVHX_3, RACE, TREAT	5.00	2.00
3	AGE, BECK, NDRUGTX, HER_3, IVHX_3, RACE, TREAT	5.20	2.20
4	AGE, NDRUGTX, IVHX_2, IVHX_3, RACE, TREAT	5.43	4.43
5	AGE, NDRUGTX, IVHX_3, RACE, TREAT	5.94	6.52

been selected using the score test for model significance.

One point that must be kept firmly in mind when using procedures such as stepwise or best subsets selection to identify possible model covariates is that the results should be taken as suggestions for models to be examined in more detail. One cannot rule out the possibility that these methods may reveal new and interesting associations, but the collection of covariates must make clinical sense to the researchers. The statistical selection procedures suggest, but do not dictate, what the model might be.

5.5 NUMERICAL PROBLEMS

The software available in the major statistical packages for fitting the proportional hazards model is easy to use and, for the most part, contains checks and balances that warn the user of impending numerical disasters. However, there are certain configurations of data that cause numerical difficulties that may not produce a suitable warning to the user. The problem of *monotone likelihood* described by Bryson and Johnson (1981) is one such problem. This problem in a survival analysis is similar to the occurrence of a zero frequency cell in a two by two contingency table or when the distributions of a continuous covariate is completely separated by the binary outcome variable in logistic regression. The problem occurs in a proportional hazards regression when the rank ordering of the covariate and the survival times are the same. That is, at each observed survival time the subject who fails has the largest (smallest) value of one of the covariates among the subjects in the risk set.

To illustrate the problem, we created a hypothetical data set containing 100 observations of survival time in days, truncated at one year with approximately 30 percent of the observations censored. We created a dichotomous covariate whose value is equal to one if the observed

Table 5.16 Estimated Coefficient, Standard Error, z-Score, Two-Tailed p-Value and 95% Confidence Intervals for a Proportional Hazards Model Containing a Monotone Likelihood Covariate (n =100)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
x	37.08	9.7E6	0.00	1.00	−1.92E7, 1.92E7

Log-likelihood = −209.74.

Table 5.17 Estimated Coefficients, Standard Errors, z-Scores, Two-Tailed p-Values and 95% Confidence Intervals for a Proportional Hazards Model Containing Two Highly Correlated Continuous Covariates (n =100)

Variable	Coeff.	Std. Err.	z	$P>\|z\|$	95% CIE
x1	18.00	41.44	0.43	0.66	–63.2, 99.2
x2	–17.72	41.44	–0.43	0.66	–98.9, 63.5

Log-likelihood = –228.19.

survival time was less than the median and zero otherwise. The results of fitting the proportional hazards model are shown in Table 5.16, where the notation "9.7E6" means 9.7×10^6.

The estimated coefficient and its standard error are unreasonably large. The software also required 25 iterations to obtain this value. As in the case of logistic regression, any implausibly large coefficient and standard error is a clear indication of numerical difficulties. In this case, a graph of the covariate versus time would indicate the problem.

The example in Table 5.16 is a simple one since it involves a single covariate. In practice, the situation is likely to be more complex, with a combination of multiple covariates inducing the same effect. Bryson and Johnson (1981) show that certain types of linear combinations (e.g., a simple sum of the covariates) may yield monotone likelihood. In these situations the problem will manifest itself with unreasonably large coefficients and standard errors.

Extreme collinearity among the covariates is another possible problem. Most software packages contain diagnostic checks for highly correlated data, but clinically implausible results may be produced before the program's diagnostic switch is tripped. The results of fitting a proportional hazards model when the relationship between the two covariates is $x_2 = x_1 + u$, where u is the value of a uniformly distributed random variable on the interval (0, 0.01), are shown in Table 5.17. The correlation between the covariates is effectively 1.0, yet the program prints a result. Similar results were obtained until $u \sim U(0,0.0001)$, at which point one of the covariates was dropped from the model by the program.

The bottom line is that it is ultimately the user of the software who is responsible for the results of an analysis. Any analysis producing "large" effect(s) or standard error(s) should be treated as a "mistake" until the involved covariate(s) are examined critically.

EXERCISES

For all exercises in this section involving analyses from the WHAS, use survival time defined by LENFOL, censoring defined by FSTAT, and data from all cohorts (i.e., ignore YEAR).

1. An important step in any model building process is assessing the scale of continuous variables in the model. The two continuous variables, AGE and CPK, in the WHAS present a challenge. Use the methods discussed in this chapter to assess the scale of AGE when it is the only covariate in a proportional hazards model. Repeat this process for CPK. In this problem, pay particular attention to the effect that a few subjects with either small or large values of the covariate can have on the methods for assessing the scale of a covariate.

2. Using the methods for model building discussed in this chapter, find the best model for estimating the effect of the covariates on long-term survival following hospitalization for an acute myocardial infarction in the WHAS. This process should include the following steps: variable selection, assessment of the scale of continuous variables and selection of interactions.

3. Present the results of the model selected in problem 2 in a table or tables that are suitable for publication in an applied journal. This presentation should include estimates of hazard ratios, with confidence intervals.

Note: Save any work done for problems 2 and 3 as there is a problem in Chapter 6 dealing with the assessment of fit of this model.

CHAPTER 6

Assessment of Model Adequacy

6.1 INTRODUCTION

Model-based inferences depend completely on the fitted statistical model. For these inferences to be "valid" in any sense of the word, the fitted model must provide an adequate summary of the data upon which it is based. A complete and thorough examination of a model's fit and adherence to model assumptions is just as important as careful model development.

The goal of statistical model development is to obtain the model which best describes the "middle" of the data. The specific definition of "middle" depends on the particular type of statistical model, but the idea is basically the same for all statistical models. In the normal errors linear regression model setting, we can describe the relationship between the observed outcome variable and one of the covariates with a scatterplot. This plot of points for two or more covariates is often described as the "cloud" of data. In model development we find the regression line, plane or hyperplane that best fits/splits the cloud. The notion of "best" in this setting means that we have equal distances from observed points to fitted points above and below the surface. A "generic" main effects model with some nominal covariates, which treats continuous covariates as linear, may not have enough tilts, bends or turns to fit/split the cloud. Each step in the model development process is designed to tailor the regression surface to the observed cloud of data.

In most, if not all, applied settings the results of the fitted model will be summarized for publication using point and interval estimates of clinically interpretable measures. Examples of summary measures include the mean difference in linear regression, the odds ratio in logistic regression and the hazard ratio for the proportional hazards regression

model. Since any summary measure is only as good as the model it is based on, it is vital that one evaluate how well the fitted regression surface describes the data cloud. This process is generally referred to as *assessing the adequacy of the model*; like model development, it involves a number of steps. Performing these in a thorough and conscientious manner will assure that the inferential conclusions based on the fitted model are the best and most valid possible.

The methods for assessment of a fitted proportional hazards model are essentially the same as for other regression models, and we assume some experience with these, particularly with logistic regression [see Hosmer and Lemeshow (1989, Chapter 5)]. Requirements for model assessment are: (1) methods for testing the assumption of proportional hazards, (2) subject-specific diagnostic statistics that extend the notions of leverage and influence to the proportional hazards model and (3) overall summary measures of goodness-of-fit.

6.2 RESIDUALS

Central to the evaluation of model adequacy in any setting is an appropriate definition of a residual. As we discussed in Chapter 1, the fact that the outcome variable is time to some event and the observed values may be incomplete or censored is what sets a regression analysis of survival time apart from other regression models. In earlier chapters we suggested that the semiparametric proportional hazards model is a useful model for data of this type and we described why and how it may be fit using the partial likelihood. This combination of data, model and likelihood make definition of a residual much more difficult in modeling survival time than is the case with other statistical models.

Consider a logistic regression analysis of a binary outcome variable. In this setting, values of the outcome variable are "present" ($y=1$) or "absent" ($y=0$) for all subjects. The fitted model provides estimates of the probability that the outcome is present (i.e., the mean of Y). Thus, a natural definition of the residual is the difference between the observed value of the outcome variable and that predicted by the model. This form of the residual also follows as a natural consequence of characterizing the observed value of the outcome as the sum of a systematic component and an error component. The two key assumptions in this definition of a residual are: (1) the value of the outcome is known and (2) the fitted model provides an estimate of the "mean of the dependent variable" or systematic component of the model. Since assumption 2

and, more than likely, assumption 1 are not true when using the partial likelihood to fit the proportional hazards model to censored survival data, there is no obvious analog to the usual "observed minus predicted" residual used with other regression models.

The absence of an obvious residual has lead to the development of several different residuals, each of which plays an important role in examining some aspect of the fit of the proportional hazards model. Most software packages provide access to at least one of these residuals. Only two packages, SAS and S-PLUS, have full residual analysis capabilities at this time. This situation is likely to change as other packages update and modify their proportional hazards routines.

We assume, for the time being, that there are p covariates and that the n independent observations of time, covariates and censoring indicator are denoted by the triplet $(t_i, \mathbf{x}_i, c_i)$, $i = 1,2,\ldots,n$, where $c_i = 1$ for uncensored observations and is zero otherwise. Schoenfeld (1982) proposed the first set of residuals for use with a fitted proportional hazards model and packages providing them refer to them as the "Schoenfeld residuals." These are based on the individual contributions to the derivative of the log partial likelihood. This derivative for the kth covariate is shown in (3.21) and is repeated here as

$$\frac{\partial L_p(\beta)}{\partial \beta_k} = \sum_{i=1}^{n} c_i \left\{ x_{ik} - \frac{\sum_{j \in R(t_i)} x_{jk} e^{x_j'\beta}}{\sum_{j \in R(t_i)} e^{x_j'\beta}} \right\}$$

$$= \sum_{i=1}^{n} c_i \left\{ x_{ik} - \bar{x}_{w_i k} \right\}, \tag{6.1}$$

where

$$\bar{x}_{w_i k} = \frac{\sum_{j \in R(t_i)} x_{jk} e^{x_j'\beta}}{\sum_{j \in R(t_i)} e^{x_j'\beta}} . \tag{6.2}$$

The estimator of the Schoenfeld residual for the ith subject on the kth covariate is obtained from (6.1) by substituting the partial likelihood estimator of the coefficient, $\hat{\beta}$, and is

$$\hat{r}_{ik} = c_i \left(x_{ik} - \hat{\bar{x}}_{w_i k} \right), \tag{6.3}$$