CHAPTER 1
Introduction

WHAT IS SURVIVAL ANALYSIS?

Survival analysis is a class of statistical methods for studying the occurrence and timing of events. These methods are most often applied to the study of deaths. In fact, they were originally designed for that purpose, which explains the name survival analysis. That name is somewhat unfortunate, however, because it encourages a highly restricted view of the potential applications of these methods. Survival analysis is extremely useful for studying many different kinds of events in both the social and natural sciences, including the onset of disease, equipment failures, earthquakes, automobile accidents, stock market crashes, revolutions, job terminations, births, marriages, divorces, promotions, retirements, and arrests. Because these methods have been adapted—and sometimes independently discovered—by researchers in several different fields, they also go by several different names: event history analysis (sociology), reliability analysis (engineering), failure time analysis (engineering), duration analysis (economics), and transition analysis (economics). These different names don't imply any real difference in techniques, although different disciplines may emphasize slightly different approaches. Since survival analysis is the name that is most widely used and recognized, it is the name I use here.

This book is about doing survival analysis with the SAS System. I have also written an introduction to survival analysis that is not oriented toward a specific statistical package (Allison 1984), but I prefer the approach taken here. To learn any kind of statistical analysis, you need to see how it's actually performed in some detail. And to do that, you must use a particular computer program. But which one? Although I have performed survival analysis with many different statistical packages, SAS is the one I currently use for both research and teaching. One reason is its wide availability and

portability—I can be reasonably confident that my students will have access to SAS wherever they go or whatever machine they use. More important, I am convinced that SAS currently has the most comprehensive set of full-featured procedures for doing survival analysis. When I compare SAS with any of its competitors in this area, I invariably find some crucial capability that SAS has but that the other package does not. When you factor in SAS's extremely powerful tools for data management and manipulation, the choice is clear. On the other hand, no statistical package can do everything, and some methods of survival analysis are not available in SAS. I occasionally mention such methods, but the predominant emphasis in this book is on those things that SAS can actually do.

I don't intend to explain every feature of the SAS procedures discussed in this book. Instead, I focus on those features that are most widely used, most *potentially* useful, or most likely to cause problems and confusion. You should always consult the official documentation in the *SAS/STAT User's Guide, Version 6, Fourth Edition, Volume 1* and *Volume 2* or in the appropriate technical report.

WHAT IS SURVIVAL DATA?

Survival analysis was designed for longitudinal data on the occurrence of events. But what is an event? Biostatisticians haven't written much about this question because they have been overwhelmingly concerned with deaths. When you consider other kinds of events, however, it's important to clarify what is an event and what is not. I define an *event* as a qualitative change that can be situated in time. By a *qualitative change*, I mean a transition from one discrete state to another. A marriage, for example, is a transition from the state of being unmarried to the state of being married. A promotion consists of the transition from a job at one level to a job at a higher level. An arrest can be thought of as a transition from, say, two previous arrests to three previous arrests.

To apply survival analysis, you need to know more than just who is married and who is not married. You need to know *when* the change occurred. That is, you should be able to situate the event in time. Ideally, the transitions occur virtually instantaneously, and you know the exact times at which they occur. Some transitions may take a little time, however, and the exact time of onset may be unknown or ambiguous. If the event of interest is a political revolution, for example, you may know only the year in which it began. That's all right so long as the interval in which the event occurs is short relative to the overall duration of the observation.

You can even treat changes in *quantitative* variables as events if the change is large and sudden compared to the usual variation over time. A fever, for example, is a sudden, sustained elevation in body temperature. A stock market crash could be defined as any single-day loss of more than 20 percent in the market index. Some researchers also define events as occurring when a quantitative variable crosses a threshold. For example, a person is said to have fallen into poverty when income goes below some designated level. This practice may not be unreasonable when the threshold is an intrinsic feature of the phenomenon itself or when the threshold is legally mandated. But I have reservations about the application of survival methods when the threshold is arbitrarily set by the researcher. Ideally, statistical models should reflect the process generating the observations. It's hard to see how such arbitrary thresholds can accurately represent the phenomenon under investigation.

For survival analysis, the best observation plan is prospective. You begin observing a set of individuals at some well-defined point in time, and you follow them for some substantial period of time, recording the times at which the events of interest occur. It's not necessary that every individual experience the event. For some applications, you may also want to distinguish different kinds of events. If the events are deaths, for example, you might record the cause of death. Unlike deaths, events like arrests, accidents, or promotions are repeatable; that is, they may occur two or more times to the same individual. While it is definitely desirable to observe and record multiple occurrences of the same event, you need specialized methods of survival analysis to handle these data appropriately.

You can perform survival analysis when the data consist *only* of the times of events, but a common aim of survival analysis is to estimate causal or predictive models in which the risk of an event depends on covariates. If this is the goal, the data set must obviously contain measurements of the covariates. Some of these covariates, like race and sex, may be constant over time. Others, like income, marital status, or blood pressure, may vary with time. For time-varying covariates, the data set should include as much detail as possible on their temporal variation.

Survival analysis is frequently used with *retrospective* data in which people are asked to recall the dates of events like marriages, child births, promotions, etc. There is nothing intrinsically wrong with this as long as you recognize the potential limitations. For one thing, people may make substantial errors in recalling the times of events, and they may forget some events entirely. They may also have difficulty providing accurate information on time-dependent covariates. A more subtle problem is that the sample of people who are actually interviewed may be a biased subsample of those who may have been at risk of the event. For example, people who have died or

moved away will not be included. Nevertheless, although prospective data are certainly preferable, much can be learned from retrospective data.

WHY USE SURVIVAL ANALYSIS?

Survival data have two common features that are difficult to handle with conventional statistical methods: *censoring* and *time-dependent covariates* (sometimes called time-varying explanatory variables). Consider the following example, which illustrates both these problems. A sample of 432 inmates released from Maryland state prisons was followed for one year after release (Rossi et al. 1980). The event of interest was the first arrest. The aim was to determine how the occurrence and timing of arrests depended on several covariates (predictor variables). Some of these covariates (like race, age at release, and number of previous convictions) remained constant over the one-year interval. Others (like marital status and employment status) could change at any time during the follow-up period.

How do you analyze such data using conventional methods? One possibility is to perform a logit (logistic regression) analysis with a dichotomous dependent variable: arrested or not arrested. But this analysis ignores information on the timing of arrests. It's natural to suppose that people who are arrested one week after release have, on average, a higher propensity to be arrested than those who are not arrested until the 52nd week. At the least, ignoring that information should reduce the precision of the estimates.

One solution to this problem is to make the dependent variable the length of time between release and first arrest and then estimate a conventional linear regression model. But what do you do with the persons who were not arrested during the one-year follow-up? Such cases are referred to as *censored*. A couple of obvious ad-hoc methods exist for dealing with censored cases, but neither method works well. One method is to discard the censored cases. That method might work well if the proportion of censored cases is small. In our recidivism example, however, fully 75 percent of the cases were not arrested during the first year after release. That's a lot of data to discard, and it has been shown that large biases may result. Alternatively, you could set the time of arrest at one year for all those who were not arrested. That's clearly an underestimate, however, and some of those ex-convicts may *never* be arrested. Again, large biases may occur.

Whichever method you use, it's not at all clear how a time-dependent variable like employment status can be appropriately incorporated into either the logit model for the occurrence of arrests or the linear model for the timing of arrests. The data set contains information on whether each person was working full time during each of the 52 weeks of follow-up. You

could, I suppose, estimate a model with 52 indicator (dummy) variables for employment status. Aside from the computational awkwardness and statistical inefficiency of such a procedure, there is a more fundamental problem that all the employment indicators for weeks *after* an arrest might be *consequences* of the arrest rather than causes. In particular, someone who is jailed after an arrest is not likely to be working full time in subsequent weeks. In short, conventional methods don't offer much hope for dealing with either censoring or time-dependent covariates.

By contrast, all methods of survival analysis allow for censoring, and many also allow for time-dependent covariates. In the case of censoring, the trick is to devise a procedure that combines the information in the censored and uncensored cases in a way that produces consistent estimates of the parameters of interest. You can easily accomplish this by the method of maximum likelihood or its close cousin, partial likelihood. Time-dependent covariates can also be incorporated with these likelihood-based methods. Later chapters explain how you can usefully apply these methods to the recidivism data.

APPROACHES TO SURVIVAL ANALYSIS

One of the confusing things about survival analysis is that there are so many different methods: life tables, Kaplan-Meier estimators, exponential regression, log-normal regression, proportional hazards regression, competing risks models, and discrete-time methods, to name only a few. Sometimes these methods are complementary. Life tables have a very different purpose than regression models, for example, and discrete-time methods are designed for a different kind of data than continuous-time methods. On the other hand, it frequently happens that two or more methods may seem attractive for a given application, and the researcher may be hard pressed to find a good reason for choosing one over another. How do you choose between a log-normal regression model (estimated with the LIFEREG procedure) and a proportional hazards model (estimated with the PHREG procedure)? Even in the case of discrete-time versus continuous-time methods, there is often considerable uncertainty as to whether time is best treated as continuous or discrete. One of the aims of this book is to help you make intelligent decisions about which

method is most suitable for your particular application. SAS/STAT software contains six procedures that can be used for survival analysis. Here's an overview of what they do:

LIFETEST	is primarily designed for univariate analysis of the timing of events. It produces life tables and graphs of survival curves (also called survivor functions). Using several methods, this procedure tests whether survival curves are the same in two or more groups. PROC LIFETEST also tests for associations between event times and time-constant covariates, but it does not produce estimates of parameters.
LIFEREG	estimates regression models with censored, continuous-time data under several alternative distributional assumptions. PROC LIFEREG allows for several varieties of censoring, but it does not allow for time-dependent covariates.
PHREG	uses Cox's partial likelihood method to estimate regression models with censored data. The model is somewhat less restrictive than the models in PROC LIFEREG, and the estimation method allows for time-dependent covariates. PROC PHREG handles both continuous-time and discrete-time data.
LOGISTIC AND PROBIT	are designed for general problems in categorical data analysis, but they are effective and flexible in estimating survival models for discrete-time data with time-dependent covariates.
GENMOD	estimates the same discrete-time survival models as LOGISTIC and PROBIT. The procedure is also good for estimating the piecewise exponential model (described in Chapter 4).

Although not discussed in official documentation, all of these procedures can be used to estimate *competing risks* models that allow for multiple kinds of events, as described in Chapter 6, "Competing Risks." Only PROC PHREG has special capabilities for handling *repeated* events like arrests or hospitalizations. See Chapter 8 for more details.

WHAT YOU NEED TO KNOW

I have written this book for the person who wants to analyze survival data using SAS, but who knows little or nothing about survival analysis. The book should also be useful if you are already knowledgeable about survival analysis and simply want to know how to do it with SAS.

I assume that you have a good deal of practical experience with ordinary least-squares regression analysis and that you are reasonably familiar with the assumptions of the linear model. There is little point in trying to estimate and interpret regression models for survival data if you don't understand ordinary linear regression analysis.

You do not need to know matrix algebra, although I sometimes use the vector notation $\boldsymbol{\beta}\mathbf{x} = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$ to simplify the presentation of regression models. A basic knowledge of limits, derivatives, and definite integrals is helpful in following the discussion of hazard functions, but you can get by without those tools. Familiarity with standard properties of logarithms and exponentials is essential, however. Chapters 4 and 5 each contain a more technical section that you can skip without loss of continuity. (I note this at the beginning of the section). Naturally, the more experience you have with SAS/STAT and the SAS DATA step, the easier it will be to follow the discussion of SAS statements. On the other hand, the syntax for most of the models considered here is rather simple and intuitive, so don't be intimidated if you are a SAS neophyte.

COMPUTING NOTES

Most of the examples in this book were executed on a FastData 486 machine running at 33 MHz with 8 MB of memory. I initially used Release 6.04 SAS software running under the DOS operating system, but I later switched to Release 6.08 under the Windows operating system. The examples requiring Release 6.10 were run on a Power Macintosh 7100/80 with 16 MB of memory using a preproduction release of SAS. Occasionally, I report comparisons of computing times for different procedures, options, sample sizes, and program code. These comparisons should only be taken as rough illustrations, however, since computing time is heavily dependent on the software and hardware configuration.

8