CHAPTER 1. INTRODUCTION

In virtually every area of the social sciences, there is great interest in events and their causes. Criminologists study crimes, arrests, convictions, and incarcerations. Medical sociologists are concerned with hospitalizations, visits to a physician, and psychotic episodes. In the study of work and careers, much attention is given to job changes, promotions, layoffs, and retirements. Political scientists are interested in riots, revolutions, and peaceful changes of government. Demographers focus on births, deaths, marriages, divorces, and migration.

In each of these examples, an event consists of some qualitative change that occurs at a specific point in time. One would not ordinarily use the term "event" to describe a gradual change in some quantitative variable. The change must consist of a relatively sharp disjunction between what precedes and what follows.

Because events are defined in terms of change over time, it is increasingly recognized that the best way to study events and their causes is to collect *event history* data. In its simplest form, an event history is a longitudinal record of when events happened to a sample of individuals or collectivities. For example, a survey might ask respondents to give the dates of their marriages, if any. If the aim is to study the causes of events, the event history should also include data on possible explanatory variables. Some of these variables, such as race, may be constant over time, while others, such as income, may vary.

Although event histories are ideal for studying the causes of events, they typically possess two features—censoring and time-varying explanatory variables—that create major problems for standard statistical procedures such as linear regression. In fact, the attempt to apply standard methods can lead to severe bias or loss of information. Over the past 40 years, however, several innovative approaches have been developed to accommodate these two peculiarities of event history data. In fact, there is no single method of event history analysis but rather a collection of related methods that sometimes compete and sometimes complement one another.

This monograph will survey these methods with an eye to those approaches that are most useful for the kinds of data and questions that are typical in the social sciences. In particular, the focus will be on regression methods in which the occurrence of events is dependent on one or more explanatory variables. Although much attention will be given to the statistical models that form the basis of event history analysis, consideration will also be given to such practical concerns as data management, cost, and the

availability of computer software. Before turning to these methods, let us first examine the difficulties that arise when more conventional procedures are applied.

Problems in the Analysis of Event Histories

To appreciate the limitations of standard methods when applied to event history data, it is helpful to look at a concrete example. A study of recidivism reported by Rossi, Berk, and Lenihan (1980) followed 432 inmates for one year after they were released from Maryland state prisons. The events of interest were arrests; the aim was to determine how the likelihood of an arrest depended on several explanatory variables.

Although the date of each arrest was known, Rossi et al. simply created a dummy (1, 0) variable indicating whether or not an individual was arrested at any time during the 12-month follow-up period. This dummy variable was the dependent variable in a linear regression with several explanatory variables including age at release, race, education, and prior work experience. While this is a passable exploratory method, it is far from ideal. Aside from the well-known problems in the use of ordinary least squares with a dummy dependent variable (Long, 1997), dichotomizing the dependent variable is arbitrary and wastes information. It is arbitrary because there was nothing special about the 12-month dividing line except that the study ended at that point. Using the same data, one might just as well compare those arrested before or after the six-month mark. It wastes information because it ignores the variation on either side of the dividing line. One might suspect, for example, that someone arrested immediately after release had a higher propensity toward criminal activity than someone arrested 11 months later.

To avoid these difficulties, it is tempting to use the length of time from release to first arrest as the dependent variable in a linear regression. But this strategy poses new problems. First, the value of the dependent variable is unknown or "censored" for persons who were not arrested at all during the one-year period. If the number of censored cases were small, it might be acceptable simply to exclude them. But 74% of the cases were censored in this sample, and it has been shown that exclusion of censored cases can produce large biases (Sørensen, 1977; Tuma & Hannan, 1978). An alternative solution might be to assign the maximum length of time observed—in this case, one year—as the value of the dependent variable for the censored cases. But this obviously underestimates the true value and, again, substantial bias may result.

Even if none of the observations was censored, one would still face another problem: how to include explanatory variables that change in value

over the observation period. In this study, for example, individuals were interviewed monthly during the follow-up year to obtain information on changes in income, marital status, employment status, and the like. Although awkward, it might seem reasonable to include 12 different income measures in the regression model, one for each month of follow-up. This might make sense for the person who is not arrested until the 12th month, but it is surely inappropriate for the person arrested during the first month after release; his income after the first month should be irrelevant to the analysis. Indeed, the person may have been incarcerated during the remainder of the follow-up period so that income then becomes a consequence rather than a cause of arrests. In short, there is simply no satisfactory way of incorporating time-varying explanatory variables in a linear regression predicting time of an event.

These two problems—censoring and time-varying explanatory variables—are quite typical of event history data. Censoring is the more common difficulty because often the explanatory variables are measured only once. Nevertheless, it is increasingly common to find longitudinal data sets with measurements of many variables at regular intervals. For most kinds of events, such data are essential to get accurate estimates of the effects of variables that change over time.

An Overview of Event History Methods

Event history data are by no means unique to the social sciences, and many of the most sophisticated approaches have been developed in other disciplines. This is a source of great confusion for the novice because similar and sometimes identical ideas are often expressed in quite different ways, typically in substantive contexts that are unfamiliar to social scientists. It is helpful, then, to begin with a brief historical and comparative survey of this expansive body of methods.

From demography we get the earliest, best-known, and still widely used method for analyzing event history data—the life table. It is not a method I shall discuss in this monograph, however, because it is amply treated in standard demography texts (e.g., Preston, Heuveline, & Guillot, 2000) and because it does not involve regression models with explanatory variables. It should be noted, however, that one of the most influential regression methods—Cox's (1972) partial likelihood method—was inspired by the fundamental ideas behind the life table.

While the life table has been in use since the 18th century, it was not until the late 1950s and early 1960s that more modern methods for event history analysis were actively pursued. In the biomedical sciences, the substantive problem that called for such methods was the analysis of survival data and,

indeed, much of the literature on event history methods goes under the name of survival analysis. For example, an experiment may be performed in which laboratory animals are exposed to different doses of some substance thought to be toxic or palliative. The experimenter then observes how long the animals survive under each of the treatment regimens. Thus the event is the death of the animal. Censoring occurs because the experiment is usually terminated before all the animals die. Biostatisticians have produced a prodigious amount of literature on the most effective ways to analyze such data (for a bibliography, see Klein & Moeschberger, 2010). These methods have become standard practice in the analysis of data on the survival of cancer patients.

Meanwhile, engineers were facing similar problems in analyzing data on the breakdown of machines and electronic components. The methods they developed—which go by the name of "reliability" analysis or "failure time" analysis—are quite similar in spirit but slightly different in orientation from those of the biostatisticians (Nelson, 2004).

Social scientists were somewhat late to the game and for several years were unaware of the developments in biostatistics and engineering. Nevertheless, a vigorous tradition of applying the theory of Markov processes to social science data emerged in the late 1960s and early 1970s (see Singer & Spilerman, 1976). A turning point in this approach came with Tuma's (1976) introduction of explanatory variables into continuous-time Markov models, an innovation that effectively bridged the gap between the sociological approach and what had already been done in biostatistics and engineering. Economists have also made important contributions to this literature (e.g., Lancaster, 1992).

In the remainder of this chapter, I will delineate some of the major dimensions distinguishing different approaches to the analysis of event history data. In some cases, these dimensions effectively differentiate methods developed in biostatistics, engineering, and the social sciences. Other dimensions cut across the different disciplines. These dimensions serve as the organizing basis for the rest of this monograph.

Distributional versus regression methods. Much of the early work on event history analysis can be described as the study of the distribution of the time until an event or the time between events. This is the main task of life table analysis, for example. Similarly, in applications of Markov processes to social science phenomena, a principal focus has been on the distribution of individuals across different states. More recently, all the major disciplinary traditions have focused on regression models in which the occurrence of an event depends on a linear function of explanatory variables. As already noted, we will deal almost exclusively with regression models here.

Repeated versus nonrepeated events. Because the events of greatest interest to biologists are deaths, it is not surprising that biostatistical work has emphasized methods for single, nonrepeatable events. Social scientists, on the other hand, have emphasized the study of events like job changes and marriages that can occur many times over the lifetime of an individual. It might seem natural, then, for this monograph to focus on repeatable events. On the other hand, models for repeatable events tend to be more complicated and also raise a number of difficult statistical questions. Moreover, mastery of the methods for single events is essential for a full understanding of the more complex models. Accordingly, we shall spend a substantial amount of time on the simpler case of nonrepeated events.

Single versus multiple kinds of events. In many cases, it is expedient to treat all the events in an analysis exactly the same. Thus, a study of job terminations may not distinguish one termination from another. A life table may treat all deaths alike. In other cases, however, it may be desirable to distinguish different kinds of events. In the study of job terminations, it may be crucial to separate voluntary from involuntary terminations. In a study of the effectiveness of a cancer treatment, it is obviously important to distinguish deaths due to cancer and deaths from other causes. To accommodate different kinds of events, biostatisticians have developed methods for "competing risks" and demographers have developed multiple decrement life tables. The generalizations of Markov models developed by Tuma and Groeneveld (1979) also allow for multiple kinds of events. Again, however, the introduction of multiple kinds of events leads to complications that are best postponed until methods for single kinds of events are well understood.

Parametric versus nonparametric methods. Biostatisticians have tended to favor nonparametric methods that make few if any assumptions about the distribution of event times. Engineers and social scientists, on the other hand, have gravitated toward models that assume that the time until an event (or the times between events) comes from very specific distributional families, the most common being the exponential, Weibull, and Gompertz distributions. A major bridge between these two approaches is the proportional hazards model of Cox (1972) which can be described as semiparametric or partially parametric. It is parametric insofar as it specifies a regression model with a specific functional form; it is nonparametric insofar as it does not specify the exact form of the distribution of event times. In this sense, it is roughly analogous to linear models that do not specify any distributional form for the error term.

Discrete versus continuous time. Methods that assume that the time of event occurrence is measured exactly are known as "continuous-time" methods. In practice, time is always measured in discrete units, however small. When these discrete units are very small, it is usually acceptable to treat time as if it were measured on a continuous scale. On the other hand, when the time units are large—months, years, or decades—it is more appropriate to use discrete-time methods (also known as grouped-data methods). While continuous-time methods have predominated in the event history literature, there is also a sizable body of work devoted to discrete-time methods, especially in biostatistics (Brown, 1975; Prentice & Gloeckler, 1978; Mantel & Hankey, 1978; Holford, 1980; Laird & Olivier, 1981). Because discrete-time methods are particularly easy to understand and implement, they serve as a useful introduction to the basic principles of event history analysis.

Computing

Most major statistical packages have commands or procedures for doing some kind of survival analysis. For this book, I have used both SAS (release 9.3) and Stata (release 12) to do the analyses that are reported in the tables. These are the two packages that I use on a daily basis, and they both have excellent capabilities for doing survival analysis. Computer code for all the analyses can be found at http://www.statisticalhorizons.com/resources/books. The data sets used for these analyses can be found at www.statisticalhorizons.com/resources/data-sets.