



Discrete-Time Methods for the Analysis of Event Histories

Author(s): Paul D. Allison

Source: *Sociological Methodology*, Vol. 13, (1982), pp. 61-98

Published by: American Sociological Association

Stable URL: <http://www.jstor.org/stable/270718>

Accessed: 15/08/2008 10:13

---

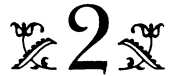
Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at <http://www.jstor.org/page/info/about/policies/terms.jsp>. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at <http://www.jstor.org/action/showPublisher?publisherCode=asa>.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

---

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact [support@jstor.org](mailto:support@jstor.org).



# DISCRETE-TIME METHODS FOR THE ANALYSIS OF EVENT HISTORIES

*Paul D. Allison*

UNIVERSITY OF PENNSYLVANIA

The history of an individual or group can always be characterized as a sequence of events. People finish school, enter the labor force, marry, give birth, get promoted, change employers, retire, and ultimately die. Formal organizations merge, adopt innovations, and go bankrupt. Nations experience wars, revolutions, and peaceful changes of government. It is surely the business of sociology to explain and predict the occurrence of such events. Why is it, for example, that some individuals try marijuana while others do not? Why do some people marry early while others marry late? Do educational

For helpful suggestions, I am indebted to Charles Brown, Rachel Rosenfeld, Thomas Santner, Nancy Tuma, and several anonymous referees.

enrichment programs reduce the likelihood of dropping out of school? What distinguishes firms that have adopted computerized accounting systems from those that have not? What are the causes of revolutions?

Perhaps the best form of data for answering questions like these is an event history. Quite simply, an event history is a record of when events occurred to a sample of individuals (Tuma and Hannan, 1978). If the sample consists of women of childbearing age, for example, each woman's event history might consist of the birthdates of her children, if any. If one is interested in the causes of events, the event history should also include data on relevant explanatory variables. Some of these, like race, may be constant over time while others, like income, may vary.

Although event histories are almost ideal for studying the causes of events, they also typically possess two features—censoring and time-varying explanatory variables—that create major difficulties for standard statistical procedures. In fact, the attempt to apply standard methods to such data can lead to serious bias or loss of information. These difficulties are discussed in some detail in the following pages. In the last decade, however, several innovative methods for the analysis of event histories have been proposed. Sociologists will be most familiar with the maximum-likelihood methods of Tuma and her colleagues (Tuma, 1976; Tuma and Hannan, 1978; Tuma, Hannan, and Groeneveld, 1979). Similar procedures have been developed by biostatisticians interested in the analysis of survival data (Gross and Clark, 1975; Elandt-Johnson and Johnson, 1980; Kalbfleisch and Prentice, 1980). A related approach, known as partial likelihood, offers important advantages over maximum-likelihood methods and is now in widespread use in the biomedical sciences (Cox, 1972; Kalbfleisch and Prentice, 1980; Tuma, present volume, Chapter 1).

Most methods for analyzing event histories assume that time is measured as a continuous variable—that is, it can take on any nonnegative value. Under some circumstances discrete-time models and methods may be more appropriate or, if less appropriate, highly useful.

First, in some situations events can only occur at regular, discrete points in time. For example, in the United States a change in party controlling the presidency only occurs quadrennially in the month of January. In such cases a discrete-time model is clearly more appropriate than a continuous-time model.

Second, in other situations events can occur at any point in time, but available data record only the particular interval of time in which each event occurs. For example, most surveys ask only for the year of a person's marriage rather than the exact date. It would clearly be inappropriate to treat such data as though they were continuous. Two alternative approaches are available, however. One is to assume that there is an underlying continuous-time model and then estimate the model's parameters by methods that take into account the discrete character of the data. The other approach is simply to assume that events can occur only at the discrete time points measured in the data and then apply discrete-time models and methods. In practice, these two approaches lead to very similar estimation procedures and, hence, both may be described as discrete-time methods.

Discrete-time methods have several desirable features. It is easy, for example, to incorporate time-varying explanatory variables into a discrete-time analysis. Moreover, when the explanatory variables are categorical (or can be treated as such), discrete-time models can be estimated by using log-linear methods for analyzing contingency tables. With this approach one can analyze large samples at very low cost. When explanatory variables are not categorical, the estimation procedures can often be well approximated by using ordinary least-squares regression. Finally, discrete-time methods are more readily understood by the methodologically unsophisticated.

For all these reasons, discrete-time methods for the analysis of event histories are often well suited to the sorts of data, computational resources, and quantitative skills possessed by social scientists. The aim of this chapter is to examine the discrete-time approach closely and compare it with continuous-time methods. Before undertaking this task, I shall

first discuss the problems that arise in the analysis of event histories and then summarize the continuous-time approach.

### *PROBLEMS IN ANALYZING EVENT HISTORIES*

Whether time is measured on a continuous or discrete scale, standard analytic techniques are not well suited to the analysis of event-history data. As an example of these difficulties, consider the study of criminal recidivism reported by Rossi, Berk, and Lenihan (1980). Approximately 430 inmates released from Maryland state prisons were followed up for one year after their release. The events of interest were arrests; the aim was to determine how the likelihood of an arrest depended on various explanatory variables.

Although the date of each arrest was known, Rossi and colleagues simply created a dummy variable indicating whether or not a person was arrested during the 12-month follow-up period. They then regressed this dummy variable on possible explanatory variables including age at release, race, education, and prior work experience. While this is not an unreasonable exploratory method, it is far from ideal. Aside from the well-known limitations of using a dummy dependent variable in a multiple regression (Goldberger, 1964), the dichotomization of the dependent variable is both arbitrary and wasteful of information. It is arbitrary because there was nothing special about the 12-month interval except that the study ended at that point. Using the same data, one might just as well compare those arrested before and after a 6-month dividing line. It is wasteful of information because it ignores the variation on either side of the cutoff point. One might suspect, for example, that a person arrested immediately after release had a higher propensity toward recidivism than one 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 multiple regression. But this strategy poses two new problems. First, the value of the dependent variable is unknown or "censored" for persons who experienced no ar-

rests during the one-year period. An ad hoc solution to this dilemma might be to exclude all censored observations and just look at those cases for whom an arrest is observed. But the number of censored cases may be large (47 percent were censored in this sample), and it has been shown that their exclusion can lead to large biases (Sørensen, 1977; Tuma and Hannan, 1978). An alternative ad hoc approach is to assign the maximum length of time observed, in this case one year, as the value of the dependent variable for the censored cases. Obviously this strategy underestimates the true value, and again substantial biases may result (Tuma and Hannan, 1978).

Even if none of the observations were censored, one would face another problem: how to incorporate explanatory variables that change in value over the observation period in the linear regression. In this study, for example, the individuals were interviewed at one-month intervals to obtain information on changes in income, marital status, employment status, and the like. It might seem reasonable to include 12 different income measures in the multiple regression, one for each month of follow-up. While this method might make sense for the person who is not arrested until the eleventh month, it is surely inappropriate for the person arrested during the first month after release; his or her later income 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 recidivism. It is also dangerous to use income during the month of arrest as a single independent variable. If income has an exogenous tendency to increase with time, this procedure could erroneously make it appear that higher incomes are associated with longer times to arrest. As Flinn and Heckman (1980) have shown, ad hoc efforts to introduce time-varying exogenous variables into regressions predicting duration usually have the unintended consequence of making those same variables endogenous.

These two problems—censoring and time-varying explanatory variables—typically occur in analyzing event his-

tories. Both problems have been solved by the continuous-time methods of maximum likelihood and partial likelihood reviewed in the next section.

### *CONTINUOUS-TIME METHODS*

In many event histories, each individual may experience multiple events of several different kinds. While both continuous-time and discrete-time methods can handle such data, the discussion is greatly simplified if we begin with a much more restricted kind of data. Specifically, let us assume that each individual experiences no more than one event and that all events are identical or at least can be treated as identical for purposes of analysis.

The setup for continuous-time models is as follows. We have a sample of  $n$  independent individuals ( $i = 1, \dots, n$ ), and we begin observing each individual at some natural starting point  $t = 0$ . In many cases, the appropriate starting point will be apparent. If the event of interest is divorce, the obvious starting point is the date of the marriage. In the recidivism example the natural starting point is the date of release from incarceration. Sometimes the choice of starting points is not so obvious, however, a problem that will be discussed further when we consider repeated events.

Assuming there is an observed starting point for each individual, the observation continues until time  $t_i$ , at which point either an event occurs or the observation is censored. Censoring means that the individual is not observed beyond  $t_i$ , either because the study ends at that point or because the individual is lost to follow-up for some reason.<sup>1</sup> For a discussion of alternative censoring mechanisms and their implications see Kalbfleisch and Prentice (1980) or Lagakos (1979). Virtually all models assume that censoring is independent of the occurrence

<sup>1</sup> Only right-censoring is considered in this chapter. Data are right-censored if the time of event occurrence is known only to be greater than some value. Much less common (but much more troublesome) are left-censored data in which the time of event occurrence is known only to be *less* than some value. For discussions of left-censoring, see Turnbull (1974, 1976).

of events—that is, individuals are *not* selectively withdrawn from the sample because they are more or less likely to experience an event. While this assumption is not always realistic, there does not yet appear to be a satisfactory alternative.

We define a dummy variable  $\delta_i$ , which equals 1 if the observation is uncensored and zero if censored. We also observe  $\mathbf{x}_i$ , a  $K \times 1$  vector of explanatory variables that are constant over time. The generalization to time-varying explanatory variables is considered later. The problem is to specify how the occurrence of an event depends on  $\mathbf{x}$ , the vector of explanatory variables. By far the most common approach is to define an unobservable or latent variable that controls the occurrence or nonoccurrence of the event, as well as the length of time until an event occurs (Tuma, Hannan, and Groeneveld, 1979; Kalbfleisch and Prentice, 1980). This variable, commonly called a *hazard rate*, is defined as follows. First let  $T$  be a random variable denoting the uncensored time of event occurrence (which may not be observed), and denote the hazard rate by  $\lambda(t)$ . Then

$$\lambda(t) = \lim_{\Delta \rightarrow 0} \Pr(t \leq T < t + \Delta \mid T \geq t) / \Delta \quad (1)$$

Here  $\lambda(t)$  may be thought of as the instantaneous probability that an event occurs at time  $t$ , given that it has not already occurred. It is not really a probability, however, since it may be greater than 1. Alternatively, it may be thought of as the expected number of events in a time interval that is 1 unit long.

We let  $\lambda(t)$  be a function of time to indicate that the probability of an event may vary with time. Sometimes, however, it is assumed that  $\lambda(t) = \lambda$ , a constant over time. This implies that  $T$ , the length of time until the occurrence of an event, has an exponential distribution. More generally, the function  $\lambda(t)$  completely determines the probability distribution of  $T$ . It can also be shown that

$$\lambda(t) = f(t) / [1 - F(t)] \quad (2)$$

where  $f(t)$  is the probability density for  $T$  and  $F(t)$  is the cumulative distribution function for  $T$ .



The next step is to express the hazard rate as a function of both time and the explanatory variables. The most widely used functional form is the so-called proportional hazards model,

$$\log \lambda(t, \mathbf{x}) = \alpha(t) + \boldsymbol{\beta}' \mathbf{x} \quad (3)$$

where  $\alpha(t)$  is an unspecified function of time and  $\boldsymbol{\beta}$  is a  $K \times 1$  vector of constants.<sup>2</sup> It is called the proportional hazards model because the ratio of the hazard rates for any two individuals at any point in time is a constant over time. The vector  $\boldsymbol{\beta}$  represents the effects of the explanatory variables on the instantaneous probability of an event. Thus if  $x_1$  has a positive coefficient  $\beta_1$ , an increase in  $x_1$  produces an increase in the likelihood that an event will occur. By assumption, these effects are constant over time in the proportional hazards model.

Special cases of this model are obtained by further specifying the function  $\alpha(t)$ . The simplest assumption is that  $\alpha(t) = \alpha$ , which again implies that  $T$  has an exponential distribution. If, however, it is assumed that

$$\alpha(t) = \alpha_0 + \alpha_1 \log t \quad (4)$$

one gets a Weibull distribution for  $T$  (Kalbfleisch and Prentice, 1980). Alternatively, the specification

$$\alpha(t) = \alpha_0 + \alpha_1 t \quad (5)$$

gives a Gompertz distribution for  $T$  (Tuma and Crockford, 1976). In general, specifying the function  $\alpha(t)$  is equivalent to specifying the probability distribution for  $T$ .

The problem, of course, is that we do not observe  $T$  for the censored cases. Nevertheless, the method of maximum likelihood (ML) allows one to make full use of the information available for these cases. The general likelihood equation for censored data is<sup>3</sup>

$$L = \prod_{i=1}^n [f(t_i)]^{\delta_i} [1 - F(t_i)]^{1-\delta_i} \quad (6)$$

In words: Each sample member contributes a factor that is the

<sup>2</sup> Natural logarithms are used throughout.

<sup>3</sup> This likelihood is only valid for right-censored data. See footnote 1.

density function for  $T$  if  $T$  is observed and 1 minus the cumulative distribution function if the individual is censored at  $t_i$ . The latter is the probability that an event occurs at some time beyond  $t_i$ . Substituting from (2) and also letting the functions depend on  $\mathbf{x}$  yields

$$L = \prod_{i=1}^n [\lambda(t_i, \mathbf{x}_i)]^{\delta_i} [1 - F(t_i, \mathbf{x}_i)] \quad (7)$$

Equation (2) implies that

$$F(t_i, \mathbf{x}_i) = 1 - \exp\left[-\int_0^{t_i} \lambda(u, \mathbf{x}_i) du\right] \quad (8)$$

which allows the likelihood function to be expressed entirely in terms of the hazard rate:

$$L = \prod_{i=1}^n [\lambda(t_i, \mathbf{x}_i)]^{\delta_i} \exp\left[-\int_0^{t_i} \lambda(u, \mathbf{x}_i) du\right] \quad (9)$$

Maximum-likelihood estimates are then obtained by substituting into (9) the appropriate expression for  $\lambda(t, \mathbf{x})$  and then choosing parameter estimates to maximize  $L$ . Usually the solution requires an iterative algorithm; for an example of one, see Kalbfleisch and Prentice (1980, pp. 55–56).

While ML estimation represents a great advance over ad hoc regression methods, it has the disadvantage of requiring that one specify the form of  $\alpha(t)$  in order to use the method. Typically there is insufficient theoretical or empirical basis for choosing among alternative specifications, yet results may vary considerably depending on the specification chosen. In an extremely influential paper, Cox (1972) proposed a method whereby  $\beta$  could be estimated without imposing restrictions on  $\alpha(t)$ . Later called partial likelihood (PL), this method has since been shown to be highly efficient (Efron, 1977) and has been widely used in the analysis of medical experiments. For further details, see Kalbfleisch and Prentice (1980).

Both ML and PL methods can be extended to handle multiple kinds of events and repeated events (Tuma, Hannan, and Groeneveld, 1979; Kalbfleisch and Prentice, 1980; Tuma, present volume, Chapter 1). The proportional hazards model can also be generalized to allow the explanatory variables to

change over time, simply by substituting  $\mathbf{x}(t)$  for  $\mathbf{x}$  in (3). Such a model can readily be estimated by the PL method (see, for example, Crowley and Hu, 1977). Time-dependent explanatory variables can also be incorporated into ML estimation, but this strategy often leads to rather cumbersome computational procedures. Tuma, Hannan, and Groeneveld (1979), for example, allow for such variables by dividing time into several disjoint intervals and assuming that the explanatory variables have constant values within each interval.

### *DISCRETE-TIME METHODS*

Although continuous-time models are usually plausible representations of the processes generating events, in practice time is always observed in discrete units, however small. When these discrete units are very small, relative to the rate of event occurrence, it is usually acceptable to ignore the discreteness and treat time as if it were measured continuously. When the time units are very large—months, years, or decades—this treatment becomes problematic.

In the case of continuous-time ML methods, the discrete character of the data can be all too easily ignored. The estimation procedure requires only that there be a numerical value for the time of event occurrence or censoring for each individual, and available programs have no way of distinguishing discrete data from (almost) continuous data. For Cox's PL method, on the other hand, the use of large intervals of time can lead to extremely difficult computational problems. These problems arise when the time intervals are large enough that more than one individual experiences an event in the same time interval. While such tied data can be handled in theory, the computational requirements can easily become so large as to outstrip currently available hardware. Approximate formulas that reduce the computational burden are in widespread use, but their adequacy has recently been called into question (Farewell and Prentice, 1980).

Among biostatisticians, these difficulties have been a major impetus to the development of discrete-time models and estimation methods that possess the virtues of the PL method.

As noted earlier, there are two general approaches to this problem. The simplest is to treat time as though it were truly discrete, an approach taken by Myers, Hankey, and Mantel (1973), Byar and Mantel (1975), Brown (1975), and Mantel and Hankey (1978). The alternative is to start with a continuous-time model, usually the proportional hazards model of (3), and then derive estimators of that model which are appropriate for data grouped into intervals. This approach has been used by Holford (1976, 1980), Thompson (1977), and Prentice and Gloeckler (1978). Whichever direction one takes, the results are remarkably similar. Accordingly, I shall make an integrated presentation of all the methods, noting any variations where appropriate.

The notation for discrete-time models is similar to that for continuous time. It is assumed that time can take on only positive integer values ( $t = 1, 2, 3, \dots$ ) and that we observe a total of  $n$  independent individuals ( $i = 1, \dots, n$ ) beginning at some natural starting point  $t = 1$ . The observation continues until time  $t_i$ , at which point either an event occurs or the observation is censored. Censoring means that the individual is observed at  $t_i$  but not at  $t_i + 1$ .<sup>4</sup> As usual, it is assumed that the time of censoring is independent of the hazard rate for the occurrence of events. The variable  $\delta_i$  is set equal to 1 if  $i$  is uncensored; otherwise it is zero. Also observed is a  $K \times 1$  vector of explanatory variables  $\mathbf{x}_{it}$ , which may take on different values at different discrete times.<sup>5</sup>

<sup>4</sup> If there is an underlying continuous-time process generating the data, this definition of censoring implicitly assumes that data are censored at the endpoint of the interval corresponding to  $t_i$ . This assumption preserves the consistency and asymptotic normality of ML estimates even if censoring actually occurred at some point *within* the interval corresponding to  $t_i + 1$  (Prentice and Gloeckler, 1978). For an alternative treatment of censoring for grouped data, see Thompson (1977).

<sup>5</sup> It is assumed that at each point in discrete time there is one and only one value for each of the explanatory variables. If the data do not conform to this assumption, some ad hoc approximation or adjustment is necessary. It may happen, for example, that an explanatory variable is measured at every fifth discrete-time unit. One might then assume that the value in that time unit also holds for the four surrounding time units. Alternatively, one could use linear interpolation to generate values for the explanatory variable in the unmeasured time units. If, on the other hand, an explanatory variable is measured more than once in each time unit, a simple average should suffice.

We begin by defining a discrete-time hazard rate

$$P_{it} = \Pr[T_i = t \mid T_i \geq t, \mathbf{x}_{it}] \quad (10)$$

where  $T$  is the discrete random variable giving the uncensored time of event occurrence. This is just the discrete-time analog of the hazard rate defined in (1). It is also the conditional probability that an event occurs at time  $t$ , given that it has not already occurred.

The next step is to specify how this hazard rate depends on time and the explanatory variables. The most popular choice (Cox, 1972; Myers, Hankey, and Mantel, 1973; Byar and Mantel, 1975; Brown, 1975; Thompson, 1977; Mantel and Hankey, 1978) is the logistic regression function

$$P_{it} = 1/[1 + \exp(-\alpha_t - \boldsymbol{\beta}'\mathbf{x}_{it})] \quad (11)$$

which can also be written in logit form:

$$\log[P_{it}/(1 - P_{it})] = \alpha_t + \boldsymbol{\beta}'\mathbf{x}_{it} \quad (12)$$

Note that  $\alpha_t$  ( $t = 1, 2, \dots$ ) is just a set of constants. Leaving them unspecified is analogous to leaving the function  $\alpha(t)$  unspecified in the proportional hazards model (3).

Although the logistic regression model is a somewhat arbitrary choice, it does have several virtues: it constrains  $P_{it}$  to lie in the unit interval for any values of  $\boldsymbol{\beta}$  and  $\mathbf{x}$ ; it is computationally convenient; and it implies that there are sufficient statistics. On the other hand, if one assumes that the data are really generated by the continuous-time proportional hazards model (3), it has been shown (Holford, 1976; Prentice and Gloeckler, 1978) that the corresponding discrete-time hazard function is given by

$$P_{it} = 1 - \exp[-\exp(\alpha_t + \boldsymbol{\beta}'\mathbf{x}_{it})] \quad (13)$$

where the coefficient vector  $\boldsymbol{\beta}$  is identically equal to  $\boldsymbol{\beta}$  in the proportional hazards model (3). Equation (13) may be solved to yield the so-called complementary log-log function:

$$\log[-\log(1 - P_{it})] = \alpha_t + \boldsymbol{\beta}'\mathbf{x}_{it} \quad (14)$$

The fact that  $\beta$  is the same in (3) and (13) implies that discrete-time estimates based on (13) are also estimates of the underlying continuous-time model. It also follows that the coefficient vector of the complementary log-log model is invariant to the length of the time intervals.<sup>6</sup> This property is not shared by the logistic model.<sup>7</sup> In practice, however, the difference between the two models is likely to be trivial. The smaller the time interval the smaller that difference will be because as the interval width becomes smaller, the logistic model converges to the proportional hazards model (Thompson, 1977).

Some special cases of these models are obtained by imposing restrictions on the set of constants  $\alpha_t$ . For example:

$$\alpha_t = \alpha \quad (15)$$

$$\alpha_t = \alpha_0 + \alpha_1 t \quad (16)$$

or

$$\alpha_t = \alpha_0 + \alpha_1 \log t \quad (17)$$

Mantel and Hankey (1978) propose that  $\alpha_t$  be expressed as a polynomial in  $t$ . It is also possible to generalize the models so that the effects of the explanatory variables can themselves vary with time, simply by substituting  $\beta_t$  for  $\beta$  in (11) or (13).

How can these models be estimated? For the logit model (11), Cox (1972) proposed a PL estimator analogous to that for

<sup>6</sup> Singer and Spilerman (1976) and Flinn and Heckman (1980) have warned that discrete-time analyses can lead to inferences that are sensitive to the arbitrary choice of interval length. Nevertheless, they also observe that discrete-time models derived from continuous-time models do not suffer from this defect. If the aim of an analysis is to estimate fundamental structural parameters that can be compared with estimates for other periods, populations, and data collection procedures, then invariance to interval length should be a major consideration. But if the goal is merely to identify the explanatory variables that have significant effects or to gauge the relative importance of different variables for a single population, the advantages of interval invariance are slight—analyses based on the logit model or the complementary log-log model will nearly always yield the same qualitative conclusions.

<sup>7</sup> Myers, Hankey, and Mantel (1973) proposed a version of the logistic model that is invariant to the length of the interval, but their model has the disadvantage of constraining  $\alpha_t$  to be the same for all  $t$ .

continuous-time data. As previously noted, this method becomes extremely demanding computationally if events occur to many individuals during the same unit of time. Fortunately, conventional ML estimation of models (11) and (13) is possible in discrete time without any restrictions on  $\alpha_t$ . The construction of the likelihood function merits examination in some detail since it has important conceptual and computational implications.

For either (11) or (13), the likelihood of the data may be written as

$$L = \prod_{i=1}^n [\Pr(T_i = t_i)]^{\delta_i} [\Pr(T_i > t_i)]^{1-\delta_i} \quad (18)$$

which is analogous to the continuous-time likelihood in (6). Each of the probabilities in (18) can be expressed as a function of the hazard rate. Using elementary properties of conditional probabilities, it can be shown that

$$\Pr(T_i = t) = P_{it} \prod_{j=1}^{t-1} (1 - P_{ij}) \quad (19)$$

$$\Pr(T_i > t) = \prod_{j=1}^t (1 - P_{ij}) \quad (20)$$

Substituting (19) and (20) into (18) and taking the logarithm yields the log-likelihood function

$$\begin{aligned} \log L = & \sum_{i=1}^n \delta_i \log \{P_{it_i}/(1 - P_{it_i})\} \\ & + \sum_{i=1}^n \sum_{j=1}^{t_i} \log (1 - P_{ij}) \end{aligned} \quad (21)$$

At this point one can substitute the appropriate regression model for  $P_{it}$  (either Equation 11 or 13) and then proceed to maximize  $\log L$  with respect to  $\alpha_t$  ( $t = 1, 2, \dots$ ) and  $\boldsymbol{\beta}$ . Most investigators stop at this point. A little further manipulation, however, leads to something more familiar. If we define a dummy variable  $y_{it}$  equal to 1 if person  $i$  experiences an event at

time  $t$ , otherwise zero, then (21) can be rewritten as

$$\begin{aligned} \log L = & \sum_{i=1}^n \sum_{j=1}^{t_i} y_{it} \log \{P_{ij}/(1 - P_{ij})\} \\ & + \sum_{i=1}^n \sum_{j=1}^{t_i} \log (1 - P_{ij}) \end{aligned} \quad (22)$$

But this is just the log likelihood for the regression analysis of dichotomous dependent variables (Cox, 1970; Nerlove and Press, 1973; Hanushek and Jackson, 1977). This identity implies that discrete-time hazard rate models can be estimated by using programs for the analysis of dichotomous data—an approach to estimation first noted by Brown (1975) but generally ignored by others.

In practice, the procedure amounts to this: Each discrete time unit for each individual is treated as a separate observation or unit of analysis. For each of these observations, the dependent variable is coded 1 if an event occurred to that individual in that time unit; otherwise it is coded zero. Thus if an individual experienced an event at time 5, five different observations would be created. For the fifth observation, the dependent variable would be coded 1. For the other four observations, the dependent variable would be coded zero. The explanatory variables for each of these new observations would be assigned whatever values they had at that particular unit of time. Lagged values could also be included. To estimate the constants  $\alpha_t$  ( $t = 1, 2, \dots$ ), a dummy independent variable would be created for each of the possible time units less 1. To estimate models like (16) or (17), which impose restrictions on  $\alpha_t$ , time itself (or some function of time) may be included as an explanatory variable.

The final step is to pool these observations and compute ML estimates of the appropriate regression model, either (11) or (13), for a dichotomous dependent variable. To my knowledge, the only publicly available program for estimating the complementary log-log model is GLIM (Baker and Nelder,



1978).<sup>8</sup> The logistic regression model can be estimated by using any ML logit program. Moreover, if the explanatory variables are all categorical (or can be treated as such) the logit hazard model can be estimated from a contingency table by using log-linear programs like ECTA. Finally, a linear model of the form  $P_{it} = \alpha_t + \beta' \mathbf{x}_{it}$  is often a good approximation to the models in (11) or (13), which suggests that ordinary least squares with a dummy dependent variable be applied to the pooled time units. While this approach has well-known limitations (Goldberger, 1964; Nerlove and Press, 1973), it can be quite useful as an exploratory method.

#### AN EXAMPLE

To illustrate the discrete-time methods, I have analyzed a set of event histories in which the event of interest is a change of employers. The sample consists of 200 male biochemists who received Ph.D.'s in the late 1950s or early 1960s, and who at some point in their careers held positions as assistant professors in graduate university departments. For a detailed description of the sample see Long, Allison, and McGinnis (1979). They were observed for a maximum of 5 years beginning in the first year of their first positions as assistant professors. Table 1 shows the number of biochemists who changed employers in each of the 5 years of observation. Of the 200 cases, 129 did not change employers during the observation period and are considered censored.

We begin with a simple model in which the hazard rate varies in each of the 5 years but does not depend on explanatory variables. It can be shown that, in this case, the ML estimate of the hazard rate is obtained by taking the ratio of the number changing employers in year  $t$  to the number at risk in

<sup>8</sup> GLIM is a general-purpose program for fitting linear, log-linear, logistic, probit, and other models. It was developed by the Working Party on Statistical Computing of the Royal Statistical Society. Further details are available from The GLIM Coordinator, NAG(USA) Inc., 1250 Grace Court, Downers Grove, IL 60515, or directly from The GLIM Coordinator, NAG Central Office, 7 Banbury Road, Oxford OX2 6NN, U.K.

TABLE 1  
Distribution of Year of Employer Change for 200 Biochemists

Year	Number Changing Employers	Number at Risk	Estimated Hazard Rate
1	11	200	0.055
2	25	189	0.132
3	10	164	0.061
4	13	154	0.084
5	12	141	0.085
>5	<u>129</u>		
Total	200	<u>848</u>	

year  $t$ . The third column of Table 1 gives the number at risk of changing employers in each of the 5 years. In year 1, all 200 cases are at risk. In year 2, the number at risk is diminished by 11, the number who changed employers in year 1. (These 11 persons could obviously change employers again, but this analysis is restricted to the first such change.) The estimated hazard rate for each year is shown in the last column of Table 1. Since the number at risk steadily diminishes, it is possible for the hazard rate to increase even when the number who change employers declines. The hazard rate in year 3, for example, is greater than the hazard rate in year 1 even though more persons changed employers in year 1.

These calculations assumed that, within each year, everyone's hazard rate is the same. We now turn to models in which the unobserved hazard rate depends on explanatory variables. Five independent variables were considered. Two of them described the employing institution and are assumed to be constant over time: a measure of the prestige of the employing department (Roose and Andersen, 1970) and a measure of federal funds allocated to the institution for biomedical research. Three variables describing individual biochemists were measured annually: cumulative number of published articles; number of citations made by other scientists to each individual's life work; and academic rank coded as 1 for associate professor and zero for assistant professor. (Although all the biochemists began the observation period as assistant professors, some were later promoted.) The aim is to find out how the

probability of changing employers depends on these five variables.

To implement the method described above, the first step was to create a separate observation for each year that each person was observed, up to the year in which an employer change occurred. Thus persons who changed employers in year 1 contributed 1 person-year each; those who changed jobs in year 3 contributed 3 person-years. Censored individuals—those who were still with the same employer in the fifth year—contributed the maximum of 5 person-years. For the 200 biochemists, there were a total of 848 person-years. From Table 1, it can be seen that this total may be obtained by summing the number at risk in each of the 5 years.

The dependent variable for each person-year was coded 1 if the person changed employers in that year; otherwise it was coded zero. The independent variables were simply assigned the values they took on in the given person-year. Taking the 848 person-years as a single sample, several logit models were fit by the ML method using the program GLIM (Baker and Nelder, 1978). Results for two of the models are shown in Table 2.

Model 1 imposes the constraint that  $\alpha_t = \alpha$ . This is equivalent to assuming that each person's hazard rate does not change autonomously over time—any changes must occur in response to changes in the explanatory variables. Table 2 gives coefficient estimates and large-sample  $t$  statistics for the null hypothesis that each coefficient is equal to zero. The coefficient estimates are like unstandardized regression coefficients in that they depend on the metric of each independent variable. For our purposes, it is more instructive to focus on the  $t$  statistics that are metric-free and give some indication of relative importance.

The results for model 1 indicate that three of the variables have a significant impact on the hazard rate for changing employers. Specifically, persons with many citations are more likely to change employers. Persons employed by institutions receiving high levels of funding are less likely to change employers. And, finally, associate professors are less likely to

TABLE 2  
Estimates for Logit Models Predicting the Probability of an Employer Change  
in 848 Person-Years

Independent Variables	Model 1			Model 2		
	$\hat{\beta}$	$t$	OLS $t$	$\hat{\beta}$	$t$	OLS $t$
Prestige of department	0.045	0.21	0.22	0.056	0.26	0.25
Funding	-0.077	-2.45	-2.34	-0.078	-2.47	-2.36
Publications	-0.021	-0.75	-0.86	-0.023	-0.79	-0.91
Citations	0.0072	2.44	2.36	0.0069	2.33	2.23
Rank	-1.4	-2.86	-2.98	-1.6	-3.12	-3.26
Year 1				-0.96	-2.11	-2.07
Year 2				-0.025	-0.06	0.18
Year 3				-0.74	-1.60	-1.54
Year 4				-0.18	-0.42	-0.38
Constant	4.95			2.35		
Chi square		461.9			452.5	
DF		842			838	

change employers than assistant professors. Prestige of department and number of publications seem to make little difference.

Model 2 relaxes the constraint imposed in model 1 by allowing the hazard rate to be different in each of the 5 years even when other variables are held constant. This was accomplished by creating a set of four dummy variables, one for each of the first 4 years of observation. Coefficient estimates and test statistics are shown in Table 2. The coefficient for each dummy variable gives the difference in the logarithm of the odds of changing employers in that particular year and the log odds of changing employers in year 5, net of other variables. No clear pattern emerges from these coefficients, although there is some tendency for the hazard rate to increase with time. Note that the introduction of the four dummy variables makes little difference in the estimated effects of the other independent variables. This will not always be the case, however.

By comparing the fit of models 1 and 2, one can test the null hypothesis that the hazard rate for changing employers does not vary with time, net of other variables. For each logit

model fitted, one gets a likelihood-ratio chi-square statistic with an associated number of degrees of freedom. The values for models 1 and 2 are shown at the bottom of Table 2. If one model is a special case of another model, their fit may be compared by using the difference in their likelihood-ratio statistics. Under the null hypothesis that there is no difference, the difference between the likelihood-ratio statistics will have a chi-square distribution. The associated degrees of freedom will be the difference in degrees of freedom for the two models. In this case, the difference is 9.4 with 4 degrees of freedom, which is just below the critical value for the 0.05 level of significance. Thus the evidence is marginal that the hazard rate varies with time. Note that while the *difference* in likelihood-ratio statistics has an approximate chi-square distribution, each statistic by itself does not have an approximate chi-square distribution (Haberma, 1978, p. 341). Hence the fact that the likelihood-ratio statistics for models 1 and 2 are extremely low relative to their degrees of freedom should not be taken as evidence either for or against the models.

The next step was to test whether the effects of the independent variables changed over time. Instead of doing this for all the variables simultaneously, I estimated five different models, each of which allowed the coefficients for one of the explanatory variables to vary with year. In each case, this was accomplished by adding four new variables to model 2. These variables were constructed by multiplying an explanatory variable by each of the four dummy variables representing the four different years. The logic behind this approach is essentially the same as that for analysis of covariance with dummy variables (Blalock, 1979, pp. 534–538). One can test the null hypothesis that the coefficients do not vary over time by taking the difference between the likelihood-ratio chi square for each of these five models and the likelihood-ratio chi square for model 2. Since none of these differences approached statistical significance, the detailed results are not reported here. Thus these data offer little evidence that the effects of the explanatory variables change over time.

All these models were reestimated under the comple-

mentary log-log specification of (14) by using the program GLIM. Results were virtually identical. Earlier I also suggested that a simple linear function might serve as an approximation to the logit or complementary log-log functions and that such a model could be estimated by ordinary least squares (OLS) with a dummy dependent variable. Models 1 and 2 were again reestimated with that procedure. That is, OLS regressions were run on the 848 person-years with the dependent variable coded 1 if a move occurred in that person-year and otherwise coded zero. The  $t$  statistics for these regressions are given in Table 2, next to the  $t$  statistics for the logit model. The results are remarkably similar.<sup>9</sup> Of course, the fact that least-squares regression worked well for these data is no guarantee that it will do so for other data sets. For guidance as to when least squares will give a good approximation, see Goodman (1975). Perhaps the best approach is to use least squares as an exploratory method. Then, when one has identified a small number of models that are serious contenders, reestimate the models by using ML with either the logit or complementary log-log specification.

### *PROBLEMS WITH THE DISCRETE-TIME APPROACH*

We have seen that discrete-time models for event histories may be estimated by treating each individual's history as a set of independent observations, one for each observed time unit. While this strategy has conceptual and computational advantages, it also raises two serious questions about the legitimacy and practicality of the procedure. Both issues are discussed in this section.

<sup>9</sup> The least-squares coefficients are not reported because they are not readily comparable to the logit coefficients. An approximate comparison is possible by using the adjustment formula  $b/[\bar{p}(1 - \bar{p})]$ , where  $b$  is the least-squares coefficient and  $\bar{p}$  is the marginal proportion of sample units coded 1 on the dependent variable. This adjusted value can be interpreted as an estimate of the corresponding coefficient in the logit model. To make least-squares coefficients comparable to the coefficients in the complementary log-log model (13), the adjustment formula is  $-b/[(1 - \bar{p})\log(1 - \bar{p})]$ . Both formulas are obtained from the first term in a Taylor series expansion of (11) or (13).

Is it legitimate to treat the multiple time units for each individual as though they were independent? In the example just considered, we started with just 200 observations but ended up analyzing 848 observations. To some, this may appear to be an artificial inflation of the sample size, leading to test statistics that are misleadingly high. To answer this objection, let us suppose that either Model (11) or (13) for the discrete-time hazard rate is a true description of how the data were generated. If that is the case, then the derivations in (18) through (22) show that the estimation procedure proposed here is indeed the ML estimator for the corresponding model. Thus, under weak regularity conditions, the estimates possess the well-known properties of being consistent, asymptotically efficient, and asymptotically normally distributed. Moreover, the estimated standard errors will be consistent estimators of the true standard errors. The fact that the ML estimator can be obtained by treating all the time units for all individuals as though they were independent is merely an incidental convenience.

All this depends of course on the truth of the original model. Unfortunately, these models tend to be unrealistic in one key respect. Consider, for example, the complementary log-log model of (14), which was

$$\log[-\log(1 - P_{it})] = \alpha_t + \boldsymbol{\beta}' \mathbf{x}_{it} \quad (14)$$

This model implicitly asserts that the variables in the vector  $\mathbf{x}$  exhaust all the sources of individual variation in the hazard rate. Obviously this will rarely be the case. By analogy with linear models, a more plausible specification would include a random disturbance term  $\epsilon_{it}$  such that

$$\log[-\log(1 - P_{it})] = \alpha_t + \boldsymbol{\beta}' \mathbf{x}_{it} + \epsilon_{it} \quad (23)$$

where  $\mathbf{x}_{it}$  and  $\epsilon_{it}$  are assumed to be independent for all  $i$  and  $t$ . This complication would be relatively innocuous were it not for the fact that, for a given individual,  $\epsilon_t$  will almost surely be correlated with  $\epsilon_{t+1}$ ,  $\epsilon_{t+2}$ , and so on. In other words, the unobserved sources of variation in the hazard rate will have some stability over time. Although the exact consequences are difficult to derive, this fact will almost certainly create difficulties for the discrete-time estimator. In particular, it will no longer

be correct to treat the multiple time units for each individual as though they were independent. By analogy with ordinary least-squares regression, one would expect this dependence among the observations to lead to inefficient coefficient estimates and estimated standard errors that are biased downward.

While this problem must be taken seriously, it should not lead one to abandon the discrete-time approach in favor of continuous-time methods. In fact, the continuous-time models suffer from precisely the same difficulty, although it is somewhat less conspicuous. For example, the continuous-time proportional hazards model (3) also assumes that the hazard rate is completely determined by the measured explanatory variables. Again it would be more appropriate to modify the model by introducing a disturbance term, so that

$$\log \lambda(t) = \alpha(t) + \boldsymbol{\beta}' \mathbf{x}(t) + \epsilon(t) \quad (24)$$

where  $\epsilon(t)$  may be regarded as a continuous-time stochastic process. While this complication does not induce dependence among the original observations, it does imply that the distribution of the time until event occurrence is different from what it would be without the random disturbance. If this disturbance is ignored in the estimation procedure, it can again lead to bias, inefficiency, and inconsistent standard errors (Fennell, Tuma, and Hannan, 1977; Chamberlain, 1979). Moreover, for analogous models the consequences should be virtually identical for discrete-time and continuous-time estimators. This should not be surprising since the mathematical construction and justification for the continuous-time and discrete-time ML estimators are essentially the same. Again the fact that the discrete-time estimator can be obtained by treating time units as separate, independent observations is incidental.

In some cases, it is possible to modify the estimation procedure to allow explicitly for a random disturbance. Tuma (forthcoming), for example, proposed the continuous-time model

$$\log \lambda(t, \mathbf{x}) = \alpha + \boldsymbol{\beta}' \mathbf{x} + \epsilon$$

where  $\exp(\epsilon)$  has a gamma distribution with a mean of 1 and a constant variance  $\sigma^2$ . She then obtained ML estimators for  $\alpha$ ,  $\boldsymbol{\beta}$ ,



and  $\sigma^2$ . Flinn and Heckman (1980) proposed a general model for the introduction of a random disturbance. Similar modifications can be made for discrete-time models and estimation methods. For example, a disturbance term may be added to a constant-rate version of (11) to yield<sup>10</sup>

$$P_{it} = [1 + \exp(-\alpha - \boldsymbol{\beta}' \mathbf{x}_i)]^{-1} + \epsilon_i \quad (25)$$

Weighted least-squares and ML estimators for this and similar models are discussed in Allison (1980). The methods of Chamberlain (1980) could also be adapted to provide discrete-time estimators of such models.

It appears, then, that the possibility of dependence among the time units confers no special disadvantage on discrete-time estimators. The discrete-time approach may become impractical, however, when the time units are small relative to the total period of observation. Consider, for example, the biochemists' employment histories. If we had used person-days instead of person-years, the effective sample size would have been more than 300,000. And this was from an original sample of only 200. The problem is even worse when one is estimating models that allow the hazard rate to be an arbitrary function of time. For such models, a dummy variable must be included for each point in discrete time (less 1). With many time points, the number of dummy variables becomes impossibly large.<sup>11</sup>

<sup>10</sup> While it may seem unusual to add the disturbance term *after* the logistic function has been applied to the linear combination of the explanatory variables, this construction is far more tractable than the alternative of adding the disturbance directly to the linear combination. Moreover, the model in (25) is merely a generalization of the well-known beta binomial model used, for example, by Heckman and Willis (1977).

<sup>11</sup> For large numbers of dummy variables, Prentice and Gloeckler (1978) suggest a version of the Newton–Raphson algorithm that reduces computation. Note also that estimation of models in which  $\alpha_t$  is an arbitrary function requires the exclusion of data pertaining to time units in which no events occurred to any individuals. If no biochemists had changed employers in year 4, for example, no person-years for year 4 could be included in the pooled sample. If those person-years were mistakenly included, the coefficient for the dummy variable for year 4 would tend toward minus infinity and the iterative algorithm would not converge. A consequence of this fact is that, no matter how small the time intervals, the maximum number of dummy variables is the number of individuals in the sample.

One can always avoid these problems by aggregating the data into larger intervals of time, but that tactic necessarily discards some information. While it might seem clearly preferable to use continuous-time ML or PL methods, the situation is not quite that simple. There are several cases in which, even with very small time units, the discrete-time ML estimator compares favorably in cost with the continuous-time ML and PL estimators.

One such case occurs when the explanatory variables are constant over time and the baseline hazard function  $\alpha(t)$  is also constant over time. Under these assumptions, the multiple time units for each individual can be regarded as a set of independent trials, all with the same probability of a "success" (success being the occurrence of an event). The probability of a success depends in turn on explanatory variables according to Function (11) or (13). The upshot is that one can use computational methods for *grouped* binomial data rather than computing over each dichotomous observation separately. If a grouped estimation procedure is used, the computation time is invariant to the length of the time intervals, so that no information need be lost. In fact, if the intervals are made small enough, the discrete-time ML estimator of  $\beta$  will coincide almost exactly with the continuous-time ML estimator. Most maximum-likelihood logit programs will handle grouped binomial data (sometimes called replicated data). The program GLIM (Baker and Nelder, 1978) will do so for both the logit model (11) and the complementary log-log model (13). In practice, each individual is treated as a single observation with a dependent variable composed of two parts: the number of trials and the number of successes. If, for example, an individual was observed for a total of eight time units, the number of trials would be eight. If the individual experienced an event in the eighth time unit, the number of successes would be one. If the individual were censored at that point, the number of successes would be zero. Thus all individuals will either have 1 or 0 for the number of successes.

In a time-independent model, therefore, continuous-time and discrete-time estimators can be equally inexpensive to compute, no matter how small the discrete time unit. On the other hand, when explanatory variables change in value over

time—and do so at frequent intervals—continuous-time and discrete-time estimators can *all* become very expensive. There is no getting around the fact that such data sets contain a great amount of information, and the complete use of that information will be costly. Thus if one makes monthly observations on 1,000 persons over a 5-year period, the effective number of observations for an event-history analysis is much closer to 60,000 than it is to 1,000. It is not yet clear whether any of the three estimators discussed here (PL, discrete ML, or continuous ML) is more computationally efficient than the others in this situation. The PL estimator may be preferable, however, since it ignores all data pertaining to time units in which no events occurred to any individual.

Small time intervals also pose little difficulty for the discrete-time estimator when all the explanatory variables are categorical (or can be treated as such). Consider a case in which there are observations on 500,000 person-days with five dichotomous, time-varying explanatory variables. Suppose also that the baseline hazard function is a constant—that is,  $\alpha_t = \alpha$ . In effect, one has six dichotomous variables: the five explanatory variables and a sixth variable indicating the occurrence or non-occurrence of an event in each person-day. Instead of analyzing these data at the individual level, it is much more efficient to form the 64-cell contingency table and analyze *it* since the cost of the analysis then depends on the number of cells in the table and not on the number of person-days. It is well known that logit models like (11) are special cases of log-linear models for contingency tables. Hence standard log-linear techniques (Goodman, 1972; Fienberg, 1977) can be readily applied.<sup>12</sup> A time-dependent hazard rate can also be introduced by including time itself as another categorical variable. The number of categories for time should not be too large, however, or the number of cells in the table will become impractically large. Still, for the 64-cell table just considered, the addition of a

<sup>12</sup> The program GLIM will also estimate complementary log-log models for contingency tables. Note also that the contingency table approach is closely related to the analysis of continuation odds proposed by Fienberg and Mason (1978).

10-category time variable would produce a table of only 640 cells, a quite manageable number.

### *MULTIPLE KINDS OF EVENTS*

To this point we have assumed that events are only of one kind. It is often useful to distinguish among different kinds of events, however, especially when there is reason to believe that effects of the explanatory variables differ among different kinds of events. For example, the biochemists' employer changes studied earlier could be divided into voluntary and involuntary changes. It would be reasonable to expect that the prestige of the department would have a negative effect on the hazard rate for voluntary moves but a positive effect on the hazard rate for involuntary moves.

The continuous-time hazard rate models can easily be generalized to allow for multiple kinds of events (Tuma, Hannan, and Groeneveld, 1979; Kalbfleisch and Prentice, 1980). This generalization is accomplished by defining a separate hazard rate for each different kind of event, an approach sometimes referred to as modeling "competing risks." Suppose, for example, that there are  $m$  different kinds of events ( $j = 1, \dots, m$ ), and let  $J$  be a random variable indicating which of the  $m$  events occurs. The hazard rate<sup>13</sup> for event  $j$  is then defined as

$$\lambda_j(t) = \lim_{\Delta \rightarrow 0} \Pr(t \leq T < t + \Delta, J = j \mid T \geq t) / \Delta \quad (26)$$

It follows that the overall hazard rate  $\lambda(t)$ , defined in (1), is just

$$\lambda(t) = \sum_j \lambda_j(t) \quad (27)$$

Next we express the dependency of each of the  $m$  hazard rates on the explanatory variables. The most common specification

<sup>13</sup> These event-specific hazard rates are formally equivalent to the transition rates or transition intensities defined for continuous-time semi-Markov processes (Coleman, 1964; Tuma, Hannan, and Groeneveld, 1979). In this context, however, there is only a single origin state that is not explicitly defined.

is, again, the proportional hazards model

$$\log \lambda_j(t) = \alpha_j(t) + \boldsymbol{\beta}_j' \mathbf{x}(t) \quad (28)$$

Note that both  $\alpha_j$  and  $\boldsymbol{\beta}_j$  are allowed to differ across different kinds of events.

Maximum-likelihood estimation of the  $\boldsymbol{\beta}_j$ 's involves nothing new, owing to the fact that the likelihood function factors into a separate component for each different kind of event. This implies that each  $\boldsymbol{\beta}_j$  can be estimated separately. Moreover, the estimation procedure for each  $\boldsymbol{\beta}_j$  is the same as if there were only a single kind of event; events other than  $j$  are treated as though the individual were censored at the time the event occurred. Thus the ML estimation procedures already discussed are appropriate. The same conclusions apply to PL estimation. For further details, see Kalbfleisch and Prentice (1980).

While it is also straightforward to develop discrete-time models for multiple kinds of events, the results are somewhat different. As with the continuous-time model, the first step is to define a discrete-time hazard rate for each different kind of event:

$$P_{tj} = \Pr(T = t, J = j \mid T \geq t) \quad (29)$$

It follows that  $P_t = \sum_j P_{tj}$  is the overall hazard rate defined in (10). Before specifying the dependence of  $P_{tj}$  on the explanatory variables, it is instructive to examine the likelihood function based on (29). Suppose that individual  $i$  experiences event  $j_i$  at time  $t_i$  or else is censored at  $t_i$ . As usual, we set  $\delta_i = 1$  if an event occurred at time  $t_i$  and otherwise set it at zero. The likelihood function can then be shown to be

$$L = \prod_{i=1}^n [P_{t_i, j_i} / (1 - P_{t_i})]^{\delta_i} \prod_{k=1}^{t_i} (1 - P_k) \quad (30)$$

The important thing about (30) is that, unlike the likelihood function for the continuous-time model, the discrete-time likelihood cannot be factored into separate components for each of the  $m$  kinds of events. Hence ML estimation must be done simultaneously for all kinds of events.

While this conclusion does not depend on the function relating  $P_{tj}$  to  $\mathbf{x}_t$ , some specifications are much more tractable than others.<sup>14</sup> Here I shall consider a generalization of the logistic model (11). Specifically, let

$$P_{tj} = \exp[\alpha_{jt} + \boldsymbol{\beta}'_j \mathbf{x}_t] / (1 + \sum_l \exp[\alpha_{lt} + \boldsymbol{\beta}'_l \mathbf{x}_t]) \quad (31)$$

$$j = 1, \dots, m$$

which reduces to (11) when  $m = 1$ . Substituting (31) into (30) and taking the logarithm yields the log likelihood for a *multinomial* logit problem in which all observed time units for all individuals are treated as separate, independent observations (see, for example, Hanushek and Jackson, 1977). Thus, just as the single-event model can be estimated with a binomial logit program, the multiple-event model can be estimated by using a multinomial logit program.

When all the independent variables are categorical, estimation of the multinomial logit model is most easily accomplished by log-linear methods (Fienberg, 1977; Goodman, 1970). As before, the procedure is first to break up each individual's event history into a set of discrete time units. The dependent variable is a polytomy having  $m + 1$  categories. (The extra category is for the nonoccurrence of any of the  $m$  events.) Pooling all the time units, the next step is to form a contingency table for all the categorical variables. Finally, one fits log-linear models corresponding to logit models for the  $m + 1$  category dependent variable.<sup>15</sup>

The necessity of estimating discrete-time models simultaneously for all kinds of events may seem to be a serious defect. Suppose, for example, that the event of interest is a death resulting from cancer. For the continuous-time model, one can

<sup>14</sup> It is possible to construct a generalization of the complementary log-log specification in (13) for multiple kinds of events, but such a model cannot be estimated by existing computer programs.

<sup>15</sup> Log-linear programs like ECTA that constrain the *sum* of a set of parameters to be zero will not directly yield estimates of the parameters defined in (31). In particular, ECTA will estimate an additional parameter pertaining to the *nonoccurrence* of any of the  $m$  events. To get estimates of the parameters defined in (31), it is only necessary to subtract this extra parameter from the remaining parameter estimates. For a related approach based on weighted least squares, see Johnson and Koch (1978).

treat a death due to any other cause as though the individual were censored at that point in time and then proceed to apply the usual methods for analyzing censored data for a single kind of death. It would appear that for the discrete-time model, on the other hand, one must carefully specify each different kind of death before estimating the appropriate multinomial logit model. This is indeed a cumbersome procedure, but there is a simpler alternative which is analogous to that for continuous-time methods. Suppose one were to treat all noncancer deaths as though the individual were censored at the beginning of the interval in which the death occurred. This tactic would amount to discarding all time units in which noncancer deaths occurred. In all the remaining time units, either a cancer death or no death will have occurred. Then one could apply the methods for single kinds of events described in earlier sections. It can be shown that this procedure yields a *conditional* ML estimator that possesses the usual ML properties of consistency and asymptotic normality (see Andersen, 1980, for a proof of a special case). Unlike the multinomial logit estimator, however, it is not fully efficient (Andersen, 1973). Nevertheless, the standard error estimates generated by this procedure are consistent estimates of the true standard errors.

### REPEATED EVENTS

Until now it has been assumed that each individual in the sample experiences no more than one event. While that assumption simplifies matters greatly, one often finds data on such repeatable events as births, marriages, changes of employer, or changes of residence, especially when observation continues over a long period of time. Continuous-time methods have been generalized to handle repeated events (Tuma, Hannan, and Groeneveld, 1979; Gail, Santner, and Brown, 1980; Kalbfleisch and Prentice, 1980), and similar generalizations can be made for discrete-time methods.

Before examining these generalizations, it is essential to clear up any confusion about the origin of the time scale. Although it has been assumed so far that each individual is observed at times  $t = 1, 2, 3, \dots, t_i$ , it is not always clear what

time is time 1. For many data sets, one may have to choose among several possible starting times. Moreover, observation may begin long after the natural starting time, and it may begin at different times for different individuals. Suppose, for example, that a panel of persons is interviewed annually for five consecutive years and that the event of interest is a change of residence. One way to fix the starting time is to let the first interview year be time 1. Although that strategy simplifies matters, it ignores the fact that the year of the first interview is likely to be an arbitrary point in the individual's life having little to do with processes that affect residence change. It is surely more plausible to assume that rates of residence change vary systematically with either the individual's age or the length of time since the last residence change. Actually, one could easily construct a model in which the hazard rate depends on both age and length of time since the last residence change simply by including both these time scales as independent variables (Tuma, 1976). The important things to remember are that the starting time should be carefully chosen and the choice should be based on expectations as to how rates of event occurrence vary with time.

The choice of starting time is *always* problematic when the data contain repeated events. In general, the method of analysis for repeated events is not much different than for single events: Simply break up each individual's event history into a set of discrete time units that are treated as independent observations, and construct a dichotomous dependent variable indicating whether or not an event occurred in each time unit.<sup>16</sup>

<sup>16</sup> It is assumed that the time interval is sufficiently short that no more than one event occurs in any discrete time unit. If this assumption is not satisfied, there are two alternatives. One is to treat intervals with more than one event as if only one event occurred. This approach is especially appropriate for the complementary log-log model of (13) because that model actually predicts the probability that *at least* one event occurs in an interval. When the number of time units with more than one event is large, however, such a procedure can discard considerable information. The alternative is to assume that the number of events in an interval has a Poisson distribution, conditional on the explanatory variables. The Poisson distribution is implied by several continuous-time models for recurrent events (Allison, 1977). Maximum-likelihood estimation of regression models for Poisson-distributed variables can be accomplished with the program GLIM (Baker and Nelder, 1978).



The difference is that now each individual may contribute more than one event to the sample. As a consequence, it becomes necessary to consider more complex models for the dependence of the hazard rate on time and on the individual's previous event history.

Some additional notation is needed to discuss such models. Let  $T_k$  ( $k = 1, 2, 3, \dots$ ) be a set of random variables denoting the time at which the  $k$ th event occurs to some individual, and let  $t_k$  be the realized value of  $T_k$ . The discrete-time hazard rate for the  $k$ th event is then defined as

$$P_k(t) = \Pr(T_k = t \mid T_k \geq t, T_1 = t_1, T_2 = t_2, \dots, T_{k-1} = t_{k-1}) \quad (32)$$

which is identically equal to zero if  $t \leq t_{k-1}$ . The next step is to let this hazard rate depend on explanatory variables. A very simple model is

$$P_k(t) = G[\boldsymbol{\beta}' \mathbf{x}(t)] \quad (33)$$

where  $G$  could be either the logistic function in (11) or the inverse of the complementary log-log function as in (13). This model says that the hazard rate changes with time only through changes in the explanatory variables.

If we now wish to introduce an explicit dependence of the hazard rate on time, we must choose between using the same starting time for all events or "resetting the clock" each time an event occurs. The former approach is symbolized by

$$P_k(t) = G[\alpha(t) + \boldsymbol{\beta}' \mathbf{x}(t)] \quad k = 1, 2, \dots \quad (34)$$

while the latter is represented by

$$P_k(t) = G[\alpha(t - t_{k-1}) + \boldsymbol{\beta}' \mathbf{x}(t)] \quad k = 1, 2, \dots \quad (35)$$

As usual,  $\alpha(t)$  is some function that may be further specified or left arbitrary. While it usually makes more sense to choose the specification in (35), one could nevertheless allow for dependence on both the initial starting time and the time since last event. In all cases, the models are easily estimated by including appropriate independent variables to represent the different time scales.

One limitation of Models (34) and (35) is the assumption that the processes affecting the occurrence of the first event are the same as those for the second, third, and later events. To allow for differences, one can simply rewrite (35) as

$$P_k(t) = G[\alpha_k(t - t_{k-1}) + \beta'_k \mathbf{x}(t)] \quad k = 1, 2, \dots \quad (36)$$

There are two different ways to estimate this model. One is to do a separate analysis for each event  $k$ , eliminating from the pooled sample all time units after  $t_k$  or before (and including)  $t_{k-1}$ . A second approach is to construct a set of dummy variables representing the different events in the sequence. Thus the dummy variable for the  $k$ th event is coded 1 for all time units between  $t_{k-1} + 1$  and  $t_k$ , inclusive, and otherwise coded zero. One then forms the products between these dummy variables and the explanatory variables in the model. Again the procedure is analogous to that for analysis of covariance using dummy variables (Blalock, 1979, pp. 534–538). The advantage of this method is that it allows one to test explicitly whether the  $\alpha$ 's and  $\beta$ 's differ for different events in the sequence. Moreover, one can allow the  $\beta$ 's to differ across different events while forcing the  $\alpha$ 's to be the same, or vice versa.

One further limitation of all these models is the implicit assumption that, conditional on the explanatory variables, the time at which an individual's  $k$ th event occurs is independent of the previous event history. Obviously this is an unrealistic assumption. The easiest way to relax the assumption is to introduce additional independent variables representing the dependency of the hazard rate on the individual's previous history (Tuma, 1976). For example, one could estimate a model of the form

$$P_k(t) = G[\alpha(t - t_{k-1}) + \beta'_k \mathbf{x}(t) + (k - 1)\gamma_1 + (t_{k-1} - t_{k-2})\gamma_2] \quad (37)$$

where  $\gamma_1$  and  $\gamma_2$  are fixed constants. This model says that the hazard rate for the  $k$ th event depends on the number of previous events ( $k - 1$ ) and the length of time between the last two events ( $t_{k-1} - t_{k-2}$ ). One could specify much more complex kinds of dependence on previous history but, in most cases, relatively simple models should suffice.

### CONCLUSION

Problems of censoring and time-varying explanatory variables are major impediments to the application of standard analytic techniques to longitudinal data on the occurrence of events. While these problems have largely been solved by the development of maximum-likelihood and partial-likelihood methods, the emphasis has been on data in which the timing of events is precisely measured. In the social sciences, however, much of the available event-history data contains only the information that events fell within certain intervals of time. For such data, it is often desirable to use discrete-time methods. In practice, these methods have considerable intuitive appeal and are relatively easy to apply. The essence of the methods is to break up each individual's event history into a set of discrete time units in which an event either did or did not occur. Pooling these time units over all individuals, one then obtains maximum-likelihood estimators for binary regression models. While this strategy has the flavor of an ad hoc approach, the resulting estimators are true maximum-likelihood estimators of models that are exact analogs to those for continuous-time data. In fact, some of the discrete-time models can be derived by assuming an underlying continuous-time model. The methods can be readily extended to the analysis of repeated events and multiple kinds of events.

### REFERENCES

ALLISON, P. D.

1977 "The reliability of variables measured as the number of events in an interval of time." In K. F. Schuessler (ed.), *Sociological Methodology 1978*. San Francisco: Jossey-Bass.

1980 "Introducing a disturbance into logit and probit regression models." Unpublished manuscript, Cornell University.

ANDERSEN, E. B.

1973 *Conditional Inference and Models for Measuring*. Copenhagen: Mentalhygienjisk Forlag.

- 1980 *Discrete Statistical Models with Social Science Applications*. Amsterdam: North-Holland.
- BAKER, R. J., AND NELDER, J. A.  
1978 *The GLIM System*. Oxford: Numerical Algorithms Group.
- BLALOCK, H. M., JR.  
1979 *Social Statistics*. 2nd rev. ed. New York: McGraw-Hill.
- BROWN, C. C.  
1975 "On the use of indicator variables for studying the time dependence of parameters in a response-time model." *Biometrics* 31:863–872.
- BYAR, D. P., AND MANTEL, N.  
1975 "Some interrelationships among the regression coefficient estimates arising in a class of models appropriate to response-time data." *Biometrics* 31:943–947.
- CHAMBERLAIN, G. D.  
1979 "Heterogeneity, omitted variable bias, and duration dependence." Discussion Paper 691. Harvard Institute of Economic Research, Harvard University.  
1980 "Analysis of covariance with qualitative data." *Review of Economic Studies* 47:225–238.
- COLEMAN, J. S.  
1964 *Introduction to Mathematical Sociology*. New York: Free Press.
- COX, D. R.  
1970 *The Analysis of Binary Data*. London: Methuen.  
1972 "Regression models and life-tables." *Journal of the Royal Statistical Society, Series B*, 34:187–202.
- COX, D. R., AND HINKLEY, D. B.  
1974 *Theoretical Statistics*. London: Chapman & Hall.
- CROWLEY, J., AND HU, M.  
1977 "Covariance analysis of heart transplant survival data." *Journal of the American Statistical Association* 72:27–36.
- EFRON, B.  
1977 "The efficiency of Cox's likelihood function for censored data." *Journal of the American Statistical Association* 72:557–565.
- ELANDT-JOHNSON, R. C., AND JOHNSON, N. L.  
1980 *Survival Models and Data Analysis*. New York: Wiley.
- FAREWELL, V. T., AND PRENTICE, R. L.  
1980 "The approximation of partial likelihood with emphasis on case-control studies." *Biometrika* 67:273–278.

- FENNELL, M. L., TUMA, N. B., AND HANNAN, M. T.  
 1977 "Quality of maximum likelihood estimates of parameters in a log-linear rate model." Technical Report 59. Laboratory for Social Research, Stanford University.
- FIENBERG, S. E.  
 1977 *The Analysis of Cross-Classified Categorical Data*. Cambridge, Mass.: M.I.T. Press.
- FIENBERG, S. E., AND MASON, W. M.  
 1978 "Identification and estimation of age-period-cohort models in the analysis of discrete archival data." In Karl Schuessler (ed.), *Sociological Methodology 1979*. San Francisco: Jossey-Bass.
- FLINN, C. J., AND HECKMAN, J. J.  
 1980 "Models for the analysis of labor force dynamics." Discussion Paper 80-3. Chicago: Economics Research Center, NORC.
- GAIL, M. H., SANTNER, T. J., AND BROWN, C. C.  
 1980 "An analysis of comparative carcinogenesis experiments based on multiple times to tumor." *Biometrics* 36:255-266.
- GOLDBERGER, A. S.  
 1964 *Econometric Theory*. New York: Wiley.
- GOODMAN, L. A.  
 1970 "The multivariate analysis of qualitative data: Interactions among multiple classifications." *Journal of the American Statistical Association* 65:226-256.  
 1972 "A modified multiple regression approach to the analysis of dichotomous variables." *American Sociological Review* 37:28-46.  
 1975 "The relationship between modified and usual multiple-regression approaches to the analysis of dichotomous variables." In D. R. Heise (ed.), *Sociological Methodology 1976*. San Francisco: Jossey-Bass.
- GROSS, A. J., AND CLARK, V. A.  
 1975 *Survival Distributions: Reliability Applications in the Biomedical Sciences*. New York: Wiley.
- HABERMAN, S. J.  
 1978 *Analysis of Qualitative Data*. Vol. 1. New York: Academic Press.
- HANUSHEK, E. A., AND JACKSON, J. E.  
 1977 *Statistical Methods for Social Scientists*. New York: Academic Press.

- HECKMAN, J. J., AND WILLIS, R. J.  
1977 "A beta-logistic model for the analysis of sequential labor force participation by married women." *Journal of Political Economy* 85:27–58.
- HOLFORD, T. R.  
1976 "Life tables with concomitant information." *Biometrics* 32:587–597.  
1980 "The analysis of rates and of survivorship using log-linear models." *Biometrics* 36:299–305.
- JOHNSON, W. D., AND KOCH, G. G.  
1978 "Linear models analysis of competing risks for grouped survival times." *International Statistical Review* 46:21–51.
- KALBFLEISCH, J. D., AND PRENTICE, R. L.  
1980 *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- LAGAKOS, S. W.  
1979 "General right censoring and its impact on the analysis of survival data." *Biometrics* 35:139–156.
- LONG, J. S., ALLISON, P. D., AND MCGINNIS, R.  
1979 "Entrance into the academic career." *American Sociological Review* 44:816–830.
- MANTEL, N., AND HANKEY, B.  
1978 "A logistic regression analysis of response-time data where the hazard function is time dependent." *Communications in Statistics—Theory and Methods* A7:333–347.
- MYERS, M. H., HANKEY, B. F., AND MANTEL, N.  
1973 "A logistic-exponential model for use with response-time data involving regressor variables." *Biometrics* 29: 257–269.
- NERLOVE, M., AND PRESS, S. J.  
1973 *Univariate and Multivariate Log-Linear and Logistic Models*. Santa Monica: Rand.
- PRENTICE, R. L., AND GLOECKLER, L. A.  
1978 "Regression analysis of grouped survival data with application to breast cancer data." *Biometrics* 34:57–67.
- ROOSE, K. D., AND ANDERSEN, C. J.  
1970 *A Rating of Graduate Programs*. Washington, D.C.: American Council on Education.
- ROSSI, P. H., BERK, R. A., AND LENIHAN, K. J.  
1980 *Money, Work and Crime: Some Experimental Results*. New York: Academic Press.

- SINGER, B., AND SPILERMAN, S.  
 1976 "Some methodological issues in the analysis of longitudinal surveys." *Annals of Economic and Social Measurement* 5:447-474.
- SØRENSEN, A. B.  
 1977 "Estimating rates from retrospective questions." In D. R. Heise (ed.), *Sociological Methodology 1977*. San Francisco: Jossey-Bass.
- THOMPSON, W. A., JR.  
 1977 "On the treatment of grouped observations in life studies." *Biometrics* 33:463-470.
- TUMA, N. B.  
 1976 "Rewards, resources and the rate of mobility: A nonstationary multivariate stochastic model." *American Sociological Review* 41:338-360.  
 "Effects of labor market structure on job-shift patterns." Forthcoming In J. J. Heckman and B. Singer (eds.), *Longitudinal Studies of the Labor Market*.
- TUMA, N. B., AND CROCKFORD, D.  
 1976 "Invoking RATE." Unpublished program manual.
- TUMA, N. B., AND HANNAN, M. T.  
 1978 "Approaches to the censoring problem in analysis of event histories." In K. F. Schuessler (ed.), *Sociological Methodology 1979*. San Francisco: Jossey-Bass.
- TUMA, N. B., HANNAN, M. T., AND GROENEVELD, L. D.  
 1979 "Dynamic analysis of event histories." *American Journal of Sociology* 84:820-854.
- TURNBULL, B. W.  
 1974 "Nonparametric estimation of a survivorship function with doubly censored data." *Journal of the American Statistical Association* 69:74-80.  
 1976 "The empirical distribution function with arbitrarily grouped censored and truncated data." *Journal of the Royal Statistical Society, Series B*, 38:290-295.