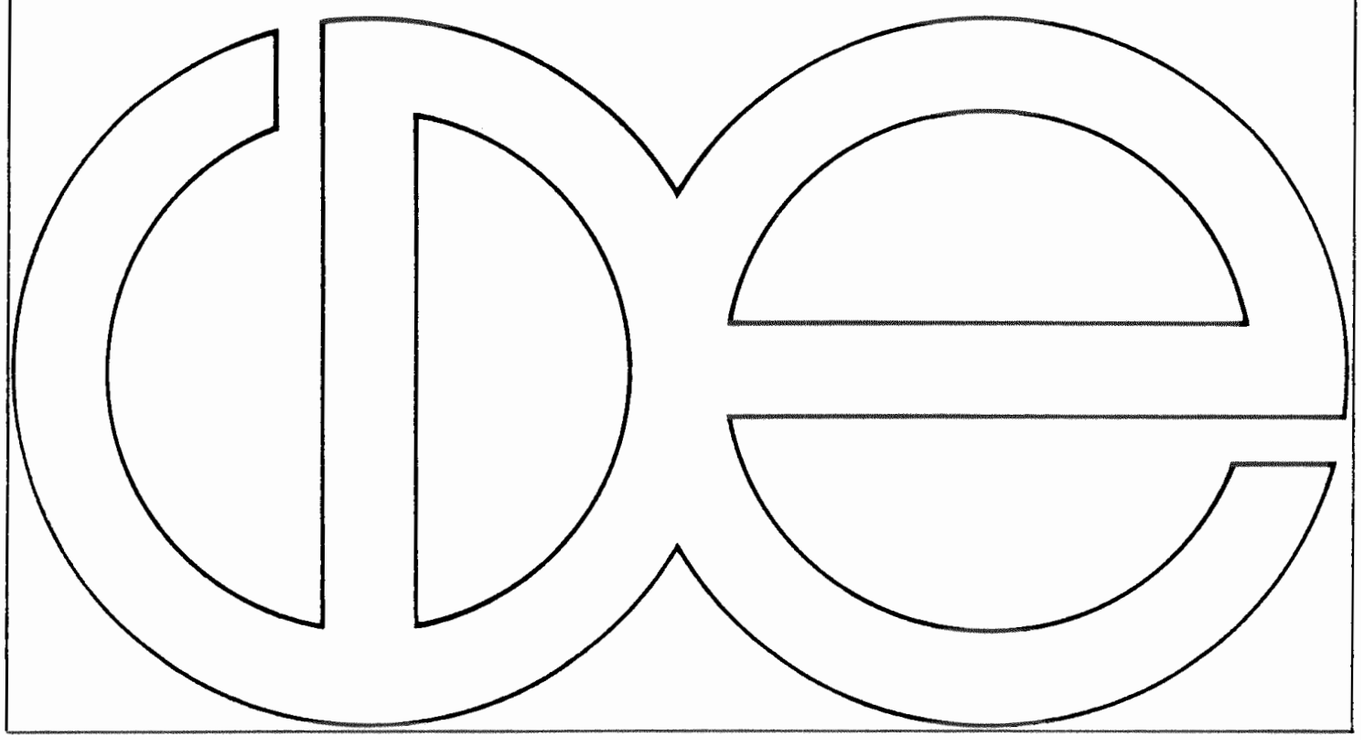
Center For Demography And Ecology
University of Wisconsin-Madison

METHODS FOR THE ANALYSIS OF EVENT HISTORY DATA

Alberto Palloni
Aage Sorensen

CDE Working Paper 86-36



INTRODUCTION

Methods for event history analysis are methods for the analysis of changes over time in discrete or categorical variables. These methods are increasingly being used by sociologists. Not only are appropiate data more available but there is also more awareness of the advantages of these methods for the analysis of change in discrete variables. It is the purpose of this chapter to review the now quite large and diverse literature on the topic.

The data suitable for event history analysis are observations on the types and timing of qualitative changes that a sample of units (individuals or more aggregated units of observation) undergo during a certain interval of time. In addition, information on fixed or variable characteristics of these units believed to be relevant for the types and timing of qualitative changes is also included and constitute an integral part of the analysis. The observable part of an event history is thus : (a) a sequence of states (outcomes) occupied by each unit over a full or partially measurable interval of time before transition to another state occurs; and (b) the sequence of values of accompanying characteristics -- some of which may undergo changes while others remain fixed for the duration of the interval. For example, a survey may ask respondents for the date of entry into first employment and record information on variables relevant for this event. The qualitative changes correspond to the transitions from the state of no-employment to the state of employment; the information on timing corresponds to the dates of entrance into a labor force and the date of first employment; the information on associated characteristics may be summarized by measures of levels of education at the time of entrance into the labor market (fixed over the period between entry into the labor force and first employment) and the type of work-related training that an individual experiences before entering the employed state (a characteristics that varies over time).

Event history analysis is the richest of the strategies available for the analysis of processes of change in qualitative variables. However, such processes are still often analyzed by sociologists using other methods. One reason is the lack of suitable data. Change in variables produced by causal processes, of course, have been inferred from cross-sectional data -- the main source of information for much sociological analysis. These inferences require that the causal processes that generate the observed outcomes can be assumed to be in equilibrium -- an untestable and often untenable assumption. A favorite design for the collection of longitudinal data in sociology remains the panel. Here, outcomes of change processes are observed at certain time intervals, but information on the timing of the changes is not available (unless the panel is combined with retrospective questioning). Lack of information on the timing of events at best

limits the identifiability of structural parameters in realistic continuous time models estimated from panel data. It often also leads to the application of regression or contingency table approaches developed for cross-sectional data. In these approaches time is used as just another causal variable without recognition of its special nature as the domain in which change takes place.

The use of conventional regression and contingency table approaches remains a methodological preference of sociologists even when event history information is available. They are a familiar and well-developed set of tools for statistical analysis. However, these approaches can severely misrepresent the structure of the processes since they provide descriptions of the associations over time among the variables but offer no possibility of identifying the parameters that govern these processes.

Increased awareness about these problems, the rapid development of methods for the study of event histories, and an orientation towards the collection of longitudinal (or retrospective) histories of events, has made possible more frequent and in-depth applications of event history analysis.

Event history analysis is used by sociologists to denote methods that, under different guiseses and names, have been in use in disciplines such as demography, biostatistics and engineering. As we shall show below life table analysis in demography has close similarity to event history analysis. In biostatistics and engineering, hazard rate analysis, survival analysis, or failure time analysis are different names for these methods. In econometrics, the techniques are often applied under the name of duration analysis. Different disciplines have emphasized somewhat different aspects of the techniques. Yet, the basic ideas are the same.

The full development of methods for the analysis of event history data and their adaptation to social science problems is of only recent origins. These methods serve to guide the formulation and allow the estimation of models representing the process that generates an event history. Methods for the analysis of event history address and resolve four fundamental problems: a) the possibly incomplete observation of the entire event history generated by a single underlying process, b) the requisite translation of an underlying continuous time process into one where changes are observed to occur at discrete times, c) the incorporation of possibly variable characteristics that affect the path of events followed by individuals, and d) the correct identification and separation of three components defining the process namely, the duration structure (the nature of the time dependence), the structure of effects of determinants or covariates, and the residual component of unmeasured heterogeneity.

The first problem is commonly referred to as censoring and occurs either because recording is initiated after the process has been set in motion (left censoring) or because observation is discontinued before the process has run its full course (right censoring). The second problem occurs when the plan of observation permits only an approximate registration of the timing of events. Also, there are situations when estimation procedures are easier to formulate and their implementation becomes more efficient with the aggregation of time even though the timing of events may be registered at finer levels. The third problem arises whenever important determinants of the occurrence of events are subject to change (and are thus themselves a process) even though these changes may be rightfully considered to be exogenous to the main process. The fourth problem is a fundamental one. It is only under very restricted conditions that one will be able to simultaneously identify the nature of time dependence embedded in the process and the set of effects of structural parameters without either of them being contaminated by the impact of unknown (or known but unmeasured) determinants.

In these chapter we attempt to clarify these four problems. The nature of this review and the limitations of space allow neither a very rigorous nor a very detailed treatment. The chapter aims solely to be a useful guide for those interested in the application of this class of methods. Its organization is simple. We first set up the fundamental language and objectives of event history analysis. We show its connection with the demographic life table approach for a simple death process, introduce the rudiments of estimation techniques, and present generalizations of the simple model. We then review estimation problems and techniques. It is here where the four fundamental problems identified before are studied. The paper closes with a summary of available software to implement estimation techniques.

The subjects that we review here have been studied in a wealth of detail by other researchers. A fundamental reference for social scientists is the insightful treatment (with extensions to the field of quantitative outcomes and combinations of qualitative and quantitative outcomes) by Tuma and Hannan (1984). The recent work on longitudinal analysis by Coleman (1981) is also an important reference. Well organized applied introductions to event history analysis with numerous applications can be found in Allison (1984), Carroll (1982), Teachman (1983), and Blossfeld et al. (1986). Most of the methodological advances that serve as foundations in the field are rigorously studied in the books by Kalbfleisch and Prentice (1980) and the more recent book by Cox and Oakes (1984). A useful although somewhat formal summary of recent advances is contained in the book by Miller (1981). Two other work of synthesis are highly recommended because of their thoroughness and accessibility (Elandt-Johnson & Johnson, 1980; Gross &

Clark, 1975). In addition to these, other references are cited throughout the paper to facilitate the search for additional methodological and applied contributions. For the sake of economy in the organization of references we cite, whenever it is possible, work that already contains important citations of the same author(s). Thus, the book by Cox and Oakes contains many references to the seminal work of Cox to which we seldom refer here. Similarly, much of the pioneering work done by Tuma and Hannan is cited in their book.

FUNDAMENTALS OF EVENT HISTORY ANALYSIS.

A Simple Death Process

We start with the simplest of situations, one in which there are only two possible states that individuals can occupy, say 0 and 1. Further, flows can only occur in one direction from state 0 to state 1, e.g. state 1 is absorbing. This representation is appropriate when analyzing a dichotomous variable where change can only go in one direction, e.g. from life to death. A particular case of this representation is one where the state 0 indicates that no event has taken place and state 1 indicates that an event has occurred, and the events refer to changes in selected variables. Each individual is characterized by an underlying random variable T measuring the time it takes to transit from one state to the other. This random variable is completely characterized by the probability that the event or failure occurs during the small interval $(t, t + \Delta t)$, e.g. by the density function of T, $f(t)$:

$$f(t) = \lim_{\Delta t \to 0} (\text{Prob}(t \leq T \leq t + \Delta t) / \Delta t) \qquad (1)$$

A survival function characterizes the process and is given by S(t), the probability that the event has not yet occurred by time t:

$$S(t) = \int_t^{\infty} f(v)\, dv \qquad (2)$$

In this particular case, S(t) is the complement of the distribution function of the waiting time in state 0.

Finally, the risk or hazard that the event will occur during the small interval of time $(t, t+\Delta t)$, $\beta(t)$, is defined as:

$$\gamma(t) = \lim_{\Delta t \to 0} (\text{Prob}(t \leq T \leq t + \Delta t \mid T \geq t) / \Delta t) \qquad (3)$$

$\gamma(t)\,\Delta t$ can be interpreted as the conditional probability that the event will take place in the small interval $(t, t + \Delta t)$ given that no event has occurred before t. However, $\gamma(t)$ is not a probability, its range being $(0, \infty)$. Note also that the relations between (1), (2) and (3) imply that:

$$S(t) = S(0) \exp\left(-\int_0^t \gamma(v)\, dv\right) \qquad (4)$$

and that

$$f(t) = \gamma(t)\, S(t)$$

In most applications $S(0)$ is set equal to unity, e.g. individuals start the process in state 0 and the density $f(t)$ is proper, e.g. all individuals eventually fail or move into state 1.

Formula (4) shows that identification of any one of the three functions permits identification of the other two and, hence, a complete characterization of the process. This is a feature that will be preserved in the more complicated versions of this model that we introduce later.

Event history analysis focusses on models for and estimation of the hazard function. There are several reasons for this interest. First, $\gamma(t)$ corresponds to an intuitive notion of risk as the underlying generator of the process. It therefore permits more appealing and less ambiguous translations of theoretical propositions. Second, hazard functions are more sensitive to time dependent changes and hence can reveal finer features of the process much more clearly than the other two functions. Third, knowledge of the hazard facilitates the formulation of implications for other quantities of interest (waiting times in one state and, when applicable, the number of moves out of one state). Fourth, it is generally simpler to manipulate models when the risk is the quantity that is being parameterized.

The main alternative to directly modeling the hazard is to focus on the waiting time, T, using the so-called accelerated failure time models and log-linear regression models. Simple re-parameterizations of the hazard models provides this interpretation for most of the well known hazard models (see Kalbfleisch and Prentice, 1980, for examples). We shall discuss both representations in a later section.

Describing Event History Processes

If the observation plan permits recording of all the failure times for the individuals in the sample, it is exceedingly simple to calculate estimates of S(t), $\hat{f}$ (t) and $\hat{\rho}$(t) with no assumptions about the parametric form of any of these functions. If events are recorded at only certain time intervals, the estimators will be discrete functions of T. An estimate of S(t) can then be obtained dividing the number of events or failures that occurred after t by the number of individuals who had not experienced a failure by that time. From this a discrete analogue estimate for $\hat{f}$ (t) and $\hat{\rho}$(t) can be obtained.

Rarely, however, can one observe all the realizations of T in the sample. It is more common that a subset of them is right censored; that is, the corresponding value is longer than the length of observation. For example, in data collected as retrospective histories, someobservations will be right censored when the evnt of interest has notbeen experienced by the time of interview. With censored samples it is still possible to obtain estimates of the hazard without making any assumptions about its underlying form. Two important approaches to do this are presented below.

THE LIFE TABLE APPROACH The simple model contained in equation (4) underlies the single decrement life table which is routinely used in demography to study phenomena as diverse as mortality, fertility, nuptiality and migration.

In fact, the life table is a technique to provide a non-parametric estimate of the relevant functions. The classic life table describes the process of mortality. Its construction proceeds by first aggregating time into intervals (t_i, t_i+n). Frequently, an open ended interval is used to acommodate large values of T and some finer intervals are used to discriminate among smaller values. The second step is to count the number of events that took place in each interval and the number of cases censored in the interval. The third step is to construct a conditional probablity of failing in each of the intervals, (t_i):

$$Q(t_i) = \frac{D(t_i)}{N(t_i) - .5\, C(t_i)} \qquad (5)$$

where the numerator represents the number of events in the interval, $N(t_i)$ is the number of individuals who were at risk of failing, e.g those who had not experienced an event and had not been censored by time t_i, and $C(t_i)$ is the number of cases that were censored in the interval, e.g. individuals who were observed for a period longer than

t_i but not longer than t_i+n and had not failed. The adjustment in the denominator is correct if censoring occurs uniformly through the interval. The fourth step is to construct an estimator of S(t) by chaining together the values of $Q(t_i)$:

$$\hat{S}(t) = \prod_{j=0}^{i-1} [1 - Q(t_j)] \qquad (6)$$

The most common (discrete) estimator of $f(t)$ is obtained by assuming linearity of S(t) within each interval. Alternatively, one can assume that $f(t)$ is a step function taking on a constant value in each interval. The corresponding estimators are, respectively:

$$\hat{f}(t_i) = \frac{Q(t_i)}{n\,(1 - .5\,Q(t_i))}$$

and (7)

$$\hat{f}(t_i) = -n^{-1} \ln (1 - Q(t_i))$$

Despite its easy of computation and general applicability,life table estimates have been shown to be inconsistent when the size of the interval, n, is large by comparison to the rate at which events are occurring. Further, significant losses in precision may occur since the intervals are arbitrarely defined (Gehan, 1969; Crowley & Breslow, 1974). These disadvantages are resolved by an alternative procedure described below.

KAPLAN-MEIER ESTIMATOR OF S(t) Suppose that the times at which events are observed in a sample are organized in increasing order: $t_1 \leq t_2 \leq t_3 \ldots \leq t_d$ where d is the total number of observed events. Assume further that of $N(t_j)$ individuals who are observed to be at risk at t_j, $d(t_j)$ fail at t_j and $c(t_j)$ are censored in the interval $[t_j, t_{j+1})$. The Kaplan-Meier estimator of S(t) is defined as:

$$\hat{S}(t) = \prod_{(j:t_j \leq t)} \left[1 - \frac{d(t_j)}{N(t_j)}\right] \qquad (8)$$

with an associated estimator of the hazard rate given by the ratio of $d(t_j)$ to $N(t_j)$. These estimators have been shown to have good large sample properties (Kaplan & Meier, 1958; Efron, 1967; Breslow & Crowley, 1974; Peto et al, 1977). The life table estimates tend to Kaplan-Meier estimates whenever time intervals are small and/or the number of events is large. For small samples the Kaplan-Meier estimates of S(t) and the hazard are affected by biases but corrections do exist (Nelson, 1972; Aalen, 1978). The Kaplan-Meier estimate of S(t) is not defined for values of t which are longer than the longest censored time.

Multiple Types of Events

When the discrete variable of interest takes on more than two values, we need models with more than two states. It is useful to distinguish here between multiple origin states from which the individual moves, and multiple destination states. For example, in a study of differential mortality the multiple origin states could be different occupational categories and the various causes of deaths considered would be the multiple destination states.

Defining J origin states does not imply any new developments if flows among them are not allowed. In fact, one simply defines a set of hazards indexed by the origin state to which they correspond, $\mu_{j\cdot}(t)$, $j=1,\ldots J$. The dot serves to label for the only destination state in the model. Identification of the model would then be equivalent to identification of as many simple death processes as origin states are recognized.

The study of a model with one origin and several destination states may require new developments depending on the nature of the relation between the new types of events (Allison, 1984). The first possibility is that the occurrence of one type of event is wholly unrelated to the occurrence of the others so that the occurence of one type of event does not affect an individual's risk for another type of event. If this were the case, the model could again be treated as one containing several simple death processes and each of them identified one at a time indenpendently of the others. The second possibility is that the occurrence of one type of event affects the risks of occurrence of the others without, however, removing the individuals from being at risk of experiencing other events. This can be dealt with using information on the occurrence/non occurrence of one (or several) types of events as a characteristic or covariate affecting the hazard of interest. Since this characteristic will vary with time, we require estimation procedures to handle time varying covariates (discussed below), but no new model specification is necessary.

For many sociological applications, the third possibility is of greater

general interest. It corresponds with a situation where the occurrence of one type of event removes the individual from the set exposed to the risk of other types of events. This is the case of competing risks with several absorbing states. The clearest example occurs in the study of human mortality when individuals are simultaneously exposed to several causes of deaths. In other analyses, the researcher may be interested in discriminating between several types of failures or events rather than lumping them together in one single, homogeneous set. In the study of geographical mobility it would be natural to distinguish flows according to destination. In the study of marital stability it is of interest to consider separately marriages that are terminated by death of one of the spouses from those terminated by divorce. Research on job changes has made a distinction useful between shifts producing a gain in status and other job shifts (Sorensen & Tuma, 1981).

To deal with competing risks, a new random variable, K, is introduced defining the type of event experienced by the individual. The model will now depend on several hazards, one for each of the u competing risks considered. We will denote these hazards by $\gamma_{.k}(t)$ (k=1,...u), where the dot serves to label the (only) origin state. These hazards are defined by:

$$\gamma_{.k}(t) = \lim_{\Delta t \to 0} 1/\Delta t \ \text{Prob}\ (t \leq T \leq t+\Delta t, K=k \mid T \geq t) \qquad (9)$$

By the law of total probability it follows that

$$\gamma(t) = \sum_{k=1}^{u} \gamma_{.k}(t) \qquad (10)$$

Also, one can define

$$S^{o}_{.k}(t) = \exp\left[-\int_{o}^{t} \gamma_{.k}(v)dv\right] \qquad (11)$$

and

$$f^{o}_{.k}(t) = \gamma_{.k}(t)\, S(t) \qquad (12)$$

where

$$S(t) = \exp\left[-\int_{o}^{t} \sum_{k=1}^{u} \gamma_{.k}(v)dv\right]$$

Under conventional plans of observation, e.g. those recording type and timing of events for each non censored observation, one can retrieve estimates of $p_{\cdot k}(t)$ without further assumptions. Extensions of life table procedures (multiple decrement life tables) lead to estimation of conditional probabilities of experiencing the event k (under the assumption that events of other types and censored cases are uniformly distributed within an interval). One can then obtain discrete estimates of the event specific hazard in much the same way as indicated for the single event case (Chiang,1968). Similarly, Kaplan-Meier estimates of $S_{\cdot k}(t)$ and of $p_{\cdot k}(t)$ are obtained by focussing on the ordered sequence of failure times of type k, $\{t_{ki}\}$ and treating failures of types other than k as if they were censored observations. It should be noted, however, that the interpretation of $S_{\cdot k}(t)$ is not that given to a survival curve in the case of a single risk, e.g. as the complement of the distribution function of a waiting time (Tsiatis, 1975; Birnbaum, 1979).

Bidirectional Flows

A further generalization obtains by allowing bidirectional flows to occur among origin and destination states. This implies that we consider some (or all) destination states as non-absorbing. The resulting models will be representations of continous time Markovian or semi-Markovian processes which have an important history in mathematical sociology (Coleman, 1964). Recent treatments emphasizing the use of event history analysis for the estimation of the parameters of these processes are found in Coleman (1981) and Tuma & Hannan (1984). In these applications, event history analysis is used to estimate the transition rate $r_{jk}(t)$ from state j to state k at time t:

$$r_{jk}(t) = \lim_{\Delta t \to 0} 1/\Delta t \; p_{jk}(t,t+\Delta t) \qquad (13)$$

where $p_{jk}(t,t+\Delta t)$ is the probability of a move from j to k within the interval $(t,t+\Delta t)$. Note that $r_{jk}(t)$ is an origin-destination-specific hazard rate. From it one can retrieve the simpler hazards rates $h_{\cdot k}(t)$, the hazard of experiencing event k regardless of origin, and $h_{j\cdot}(t)$, the hazard of leaving state j regardless of destination.

Exploratory, non-parametric analysis of these models is again possible with life tables type of techniques or with direct application of Kaplan-Meier estimates. The trick is simply to deal separately with origin-destination-specific 'survival curves' and risks.

An important similarity exists between these models and the so-called increment-decrement techniques that have been recently developed in demographic analysis (Schoen and Land,1979; Land and Rogers,1982; Hannan,1982; El-Sayed Nour and Suchindran, 1982) to deal with aggregated types of data on migration flows, changes in marital status and fertility. The model underlying incremet-decrement tables does not differ from the ones presented here. The only difference are that, increment decrement analysis has adopted a different language (Hoem, 1983) and has not yet been extended to systematically incorporate the effects of covariates.

Repeated Events

Introducing multiple states, not all of which are absorbing, and bidirectional flows provides a natural framework for the analysis of repeated events. When the event is of a single type one has a counting process with transition rates $r_{jk}(t)$ where k=j+1. If the occurrence of several competing events does not remove individuals from the risk of experiencing them again, one has a counting process with competing events and rates $r^{l}_{jk}(t)$ for the lth occurrence of event of type k. A convenient but not always statistically efficient, representation of a process with repeated events is one which assumes a different cause-structure for each implied transition rate. Identification of the risks can then proceed by treating each type of event separately from the others, as in the case of multiple events and bidirectional flows. The only qualitatively different aspect of the model lies in the definition of t: it may refer to the duration since the prior occurrence rather than to chronological time elapsed since the origin of the process. This, however, does not preclude the simultaneous use of alternative representations of time in causal models such as age, cohort or others.

As in the case of multiple states and bidirectional flows, life tables techniques and Kaplan-Meier estimates can be used to provide nonparametric estimates of each of the risks and survival functions characterizing the process. All one needs to do is to deal separately with order-type-specific events.

ESTIMATION OF MODELS FOR EVENT HISTORY ANALYSIS.

In this section we review the model formulation and statistical inference strategies that are appropriate in event history analysis. We start with a brief introduction to maximum likelihood estimation which is followed by a discussion of censoring, and then proceed to examine strategies to impose structure on the models.

Maximum Likelihood Estimation

Most of the models for event history analysis can only be estimated by using maximum likelihood (ML) procedures. The resulting estimates have optimal large sample properties and permit tests of hypotheses regarding one or several variables. In particular, maximum likelihood is best suited to deal with the problem of censored data. Other methods of estimation (such as least squares) can be used in the absence of censoring, but no technique based on them has been developed to deal with censored data. Under certain conditions results obtained with ML procedures are exact. Under most conditions imposed by the nature of event history data and models, ML results are only approximate, e.g. they hold asymptotically, for very large samples, and when the functions involved satisfied certain regularity conditions. We shall not discuss these conditions further (except when reviewing the issue of censoring). However, it should be noted that results based on small sample may not be all that reliable, and that there may be cases in which the regularity conditions are not satisfied even when the sample is sufficiently large.

If the sample observations are independent, its likelihood function is just the product of the individual probabilities of the observations being what they are.[1] These probabilities are defined according to the postulated model and depend on the observed information. To obtain parameter estimates, one maximizes the logarithm of the likelihood function (setting its first derivatives equal to zero) and solves for the parameters. In only a few cases do the solutions have a closed form. For most cases one has to resort to numerical optimization methods (Goldfeld & Quandt, 1972). The estimates of the standard errors are obtained from a function of the matrix of second derivatives

---

1. Although individuals (or other aggregated entities) are the units of analysis, the likelihood function is usually calculated over types of events experienced by the units. Independence of these events, a subset of which may correspond to the same unit, is only claimed conditional on a series of suitably defined explanatory variables.

of the logarithm of the ML function (or suitable approximations to it). Tests of hypotheses can be conducted applying a variety of strategies. The most common are: a) standard normal approximations and the ratio of the estimated parameters to their estimated standard errors, and b) the likelihood ratio test. The results obtained using these alternative methods are consistent only if the ML function and its derivatives satisfy some regularity conditions (Cox & Hinkley, 1974; Mood et al 1974; LeCam, 1970; Moran, 1971; Goldfeld & Quandt,1974).

The Problem of Censoring

RIGHT CENSORING Right censoring occurs when the the waiting time for the occurrence of an event is longer than the period of observation. Results that are obtained by either discarding the censored cases or by assuming that in these cases the event occurred at the end of the period of observation, will be biased. The higher the proportion of cases that are censored, the more badly biased are the results (Tuma & Hannan, 1984; Sorensen, 1977; Kalbfleish & Prentice, 1980; Cox & Oakes, 1984; Elandt-Johnson & Johnson, 1980; Gross & Clark, 1975). The biases are explained by a very simple fact: even if censoring where random, it would be more likely for an observation to be censored if the corresponding hazard is lower and the waiting time longer. This implies that excluding censored cases leads to over estimates of the underlying risks as, in fact, is the case. Clearly, estimates of the effects of covariates will also be biased if only non censored cases are included in the analysis.

Censored event histories are sometimes used to estimate means and other moments of the waiting time distribution. For example, Current Population Surveys registers duration of unemployement until the survey week for unemployed respondents and one might want to use this information to estimate the mean duration of unemployment. But the distribution of waiting times for censored events will, as argued above, differ from the distribution for all events. In fact, in the simplest case of exponentially distributed waiting times (corresponding to a constant hazard), the mean of the censored events will be exactly twice the true mean (Sorensen, 1977). In the general case, the observed mean of the censored events will depend on both the mean and the variance of the underlying waiting time distribution. If neglected, very misleading inferences from censored duration data can result. For a general discussion of this issue as it applies to curation of unemployment see the work of Salant (Salant, 1977).

One can assume that for each individual i there is a waiting time Ti for the occurrence of the event of interest and an underlying censoring time Ci. Right censoring for individual i will occur if Ti $\gg$ Ci. The mechanism producing censoring is crucial for the use of ML procedures. In general, ML are viable only when the censoring mechanism produces

independent censoring times, e.g. when the random variables Ci are independent between themselves and independent of the waiting times Ti. More generally, what an independent censoring mechanism requires is that the Ci's should not be related to the risk of experiencing the event or, equivalently, that individuals censored at time C be representative of all other individuals (with the same values in the covariates) who survived up to C. Dependence could be introduced, for example, if individuals who are at higher risks of experiencing the event stand a higher chance of being censored.

When censoring times are fixed in advance we speak of censoring of type I. When censoring depends on a prefixed number of failures having occurred in the sample, we speak of censoring of type II. By and large, the plans of observation available in social sciences do not correspond to either type. In most cases we have a fixed period of observation for all individuals. Censoring times can then be treated as if generated by an independent mechanism. This assumption will be violated when individuals withdraw from observation before the study ends (the censoring time is no longer dependent on the prefixed length of the study). If the analyst suspects that withdrawing may be related to the risks of experiencing the event, a different strategy has to be followed. The latter requires to consider withdrawal as an event on its own right, a special case of a competing event. Only the remaining censored cases may then be treated as if their censoring times had been generated by an independent mechanism.

To make these ideas more concrete and to introduce the general formulation of the likelihhood used in event history analysis, we now study a quite general, multistate, situation. Suppose we are interested in the first occurrence of an event that implies a change from state j to state k. Let $r_{jk}(t,\theta,Z)$ be the corresponding risk and $S_{j.}(t,\theta,Z)$ the associated survivor function for state j. We are assuming that the process depends on parameters contained in the vector θ and covariates contained in the vector Z. Our interest is in drawing inferences about the parameters in θ. Assume further that the censoring time for individual i is a random variable C_i with survival function $G_i(c)$ and density function $g_i(c)$. Under a regime of independent censoring, the probability that an individual experiences the event in the time interval $(t_i, t_i+\Delta t)$ and is not censored is given by:

$$A_i = r_{jk}(t_i, \theta, Z_i) * S_{j.}(t_i, \theta, Z_i) * G_i(t) \qquad (14)$$

If the observation for individual i had been censored in the interval $(t_i, t_i+\Delta t)$, the probability is given by:

$$B_i = g_i(t_i) * S_{j.}(t_i, \theta, Z_i) \qquad (15)$$

Thus, the probability of having a particular occurrence (event or censoring) for individual i can be written as:

$$L_i = (A_i)^{\delta_i} * (B_i)^{(1-\delta_i)} \qquad (16)$$

where δ_i equals 1 if the event is observed and 0 if the observation is censored. L_i can be written in an alternative form:

$$L_i = \Gamma_i * (r_{jk}(t_i, \theta, Z_i))^{\delta_i} * (S_{j.}(t_i, \theta, Z_i)) \qquad (17)$$

where Γ_i collects all the censoring information conveyed by $g_i(c)$ and $G_i(c)$. If Γ_i does not depend on θ, we speak of non informative censoring. Otherwise we speak of informative censoring. Under non informative censoring, Γ_i can be treated as a constant and one can simply work with the other part of the likelihood. Most types of analysis in event history deal with this 'partial' likelihood rather than with the 'full' likelihood. If censoring is informative, one can still work with the partial likelihood but then the estimates will have higher standard errors.

LEFT CENSORING Left censoring occurs when the observation begins after the process has been initiated. For example, in the study of job changes one may observe an individual in a first job knowing neither the timing of the arrival at this job nor the conditions under which it took place, e.g. the values that important characteristics had at the time the shift towards first job took place. If the risk for the event of interest were constant, nothing would be lost by assuming that the process started at the time the observation begun. This will rarely be the case, however.

In some instances the observation may begin before the process starts but the actual time when individuals become at risk for the event is unknown. The hazard will then be zero for a period until individuals become at risk. For example, in the analysis of nuptiality the waiting times until entry into first marriage are measured from birth rather than from the time the individuals enter into a marriage market. It may be possible to find a reasonable approximation for the true starting time -- in the case of marriage, say age 18. Such approximations should be used whenever possible. Alternatively, one could introduce threshold parameters in the models and estimate the starting time from

the data. However, such threshold parameters violate regularity conditions of the likelihood and are therefore not included as a free parameters to estimate (Kalbfleisch & Prentice, 1980).

Left censoring can create one or two problems (Tuma & Hannan, 1984). The first is a selection problem that results because the distribution of individuals by states or the distribution of important covariates at the time the observation starts is likely to be different than what they were at the beginning of the process. This is the same problem that one faces when drawing causal inferences with selected samples (Maddala, 1983; Amemiya, 1982; Heckman, 1979). The second problem is lack of knowledge about the values of the characteristics at the beginning of the process for those individuals (or events) that are observed. This latter problem is especially severe if the hazard depends on duration since the prior event, since this duration is unknown for left censored events.

At the present time, general solutions for left censoring are either very complicated to implement or depend on unrealistic assumptions (Flinn & Heckman, 1982a). The first problem may be resolved by calculating individual likelihoods conditional on the event(s) of interest not having occurred up to the time the observation starts. The second difficulty can be resolved by assuming an initial distribution of relevant characteristics and then calculating their distribution at the time the observation starts assuming that the process is well represented by the model tested. In most situations, however, we lack the knowledge to select between alternative distributions of the initial conditions.

In what follows we will assume that there is no left censoring or that left censored observations have been deleted from the sample. This is not a satisfactory strategy but simplifies the presentation of the discussion.

Formulating event history models.

So far we have not dicussed strategies that make explicit the nature of time dependence or the form in which covariates may be entered in the model. We first discuss the problem of time dependence and then present alternatives ways of formulating models incorporating exogenous covariates.

THE TIME STRUCTURE OF THE MODEL The life table and Kaplan-Meier estimators permit a representation that does not depend on assumptions about the functional form of the hazard. This may be advantageous when there is no theoretical guidance as to the nature of the hazard. In

other situations there could be strong reasons to suspect that its time profile should be constant, monotonically increasing or decreasing, or non-monotonic. In these cases it is more efficient and desirable for testing purposes to attempt to retrieve the time structure of the hazard. Two issues are of importance here. The first refers to the nature of time dependence and the second to the distinction between continuous and discrete time formulations.

The nature of the time structure of the hazard As we have used it so far, the random variable T represents calendar time measured since a suitable defined origin of the process. This time dimension serves to account for unknown or unmeasured processes. Apart from cases where the time structure is induced by unmeasured heterogeneity (discussed below), there may be theoretical reasons to explicitly define it. Thus there are instances where the effects of T may be captured by age, others where they are measured by using duration elapsed since the occurrence of a significant event, and others still where they can be captured by chronological time. For example, the application of the life table to the study of mortality is based on an explicit dependence of the risk on the age of the individual. The latter is only a proxy for more difficult to measure processes that affect mortality such as the capability to adapt to a new environment or biological deterioration. A similar role appears to play the 'age' of an organization in the study of 'mortality' of formal organizations (Freeman et al., 1983) or the 'age' of a government coalition in the study of cabinet durability (Cioffi-Revilla, 1984; Palloni & Franzosi, 1986, unpublished).

Duration dependence is frequently introduced in the study of repeatable events as it is thought that the time elapsed since the last occurrence (or other prior occurrences) is an accurate proxy for transformations that affect the risk of a new event. For example, some economic theories suggest that the longer an individual stays unemployed, the higher the risk of accepting an employment offer due to decreased reservation wages (Flinn & Heckman, 1982 a,b).

In many situations the nature of time dependence may be explored using nonparametric approaches. In fact, simple algebraic transformations of $S(t)$ and $p(t)$ (see Section 2) can be used to gauge the plausibility of some parametric representations. Thus, a plot of $\ln(-\ln S(t))$ against $\ln t$ that is roughly linear with a unit slope provides indications of a death process with constant rate. One which is roughly linear on $\ln t$ but with a slope different than one supports the idea of a death process with time varying rate following a Weibull distribution. It is a good practice to explore the nature of the data with nonparametric approaches prior to the use of fully parametric forms (Elandt-Johnson & Johnson,1980;Gross & Clark, 1985; Miller, 1981).

Continuous and discrete time models Regardless of the source of time dependence, one can distinguish models for events that can only occur at discrete intervals from models for events that can occur continuously through time. For the most part social scientists deal with the latter type of events. However, plans of observation for event histories are never precise enough to record the exact timing of the events. As a consequence, although the models are more adequately formulated in continuous time, the data can be organized only in discrete time units. This can lead to misinterpretations and inconsistent estimates due to the lack of correspondence between what the estimates reflect and what they represent in the models (Kalbfleish & Prentice,1980; Petersen,1983). As a consequence it is a healthy practice to emphasize formulations in continuous time and,if needed, translate them into discrete versions to ensure correspondence between what is estimated and what is represented in the model.

THE NATURE OF THE COVARIATES Simple representations of a process may assume that the occurrence of events does not depend on past history, except for dependence on the state being occupied at the initiation of an episode or spell and on the duration elapsed since the previous event (Markov and Semi-Markov processes). Such assumptions tend to be rather simplistic and one may wish to account in some way for past history in order for the representations to hold. The process is then modeled as a simple one conditioned on the values of relevant covariates and on the history of the process up to the time of initiation of the episode of interest. This can be done by resorting to parts of the process as explanatory variables for the paths of events that occur subsequently. For example, in the analysis of repeated events, the number of occurrences prior to a event of higher order, or the timing and type of states occupied prior to its occurrence may be used as explanatory variables.

In addition to covariates that reflect part of the past history of the process, others may be relevant. They may be continuous or discrete and may be either fixed or variable over the duration of the process. Apart from the the complications that time varying covariates create (see below), the only real problem in selecting suitable covariates emerges when drawing the line between exogenous from endogenous ones. In some cases this is an obvious decision. In some others it may depend on theoretical considerations. This is especially the case for the initial conditions of the process, e.g. for the time at which the process begins for an individual and for the particular state occupied at the origin (Tuma & Hannan 1984; Flinn & Heckman, 1982a;Cox & Oakes,1984).

For simplicity we assume below that the vector Z of covariates contains only exogenous variables (of a continuous or discrete nature and of the

fixed or time-varying type).

Alternative Models for Event Histories

In what follows we will assume that the object of analysis is a simple event or failure which is dependent on time and certain exogenous covariates. At the cost of complicating the notation without introducing new elements, the procedures can be extended to the consideration of processes with competing risks and, more generally, incorporating multiple states and bidirectional flows.

FULLY PARAMETRIC MODELS There are two general families of models to represent the effects of time and of covariates. The first is the so called proportional hazard models in which the effect of the covariates is to act multiplicatively on the risk:

$$\lambda(t,Z,\beta) = \lambda_0(t) * \Omega(\beta*Z)$$
$$S(t,Z,\beta) = [S_0(t)]^{\Omega(\beta*Z)} \qquad (19)$$

where $\lambda_0(t)$ is a baseline hazard that is increased (or decreased) by the effects of the covariates, and $S_0(t)$ is the corresponding baseline survival function. In most applications $\Omega(y) = \exp(y)$, a function that guarantees non negative values for the hazards. If the dependence of $\lambda_0(t)$ on time is specified, we obtain a fully parametric model. The most common specifications are the exponential model, where the hazard is assumed constant in time, the Weibull model where the hazard is specified as $\lambda_0(t) = \gamma\rho(\rho t)^{\rho-1}$, and the Gompertz model where the hazard is $\gamma \exp(\gamma t)$. However, the dependence of the hazard on time can also be left unspecified resulting in a partially parametric model. This important class of models are discused below.

The second type of models are the so called accelerated failure time models. They postulate multiplicative effects both on the hazard and on the waiting times:

$$\lambda(t,Z,\beta) = \lambda_0(t*\Omega(\beta*Z)) * \Omega(\beta*Z) \qquad (20)$$

Although applications of accelerated failure times exist in social sciences (Vanderhoeft, 1984; Coale & McNeil, 1972), it is the proportional hazards family that has received most attention.
Proportional hazards and accelerated failure time representations are identical when the waiting times are distributed as an exponential or a Weibull (Kalbfleisch & Prentice, 1980; Cox & Oakes, 1984).

These two types of models explicitly define the hazard. Other models

are built defining instead the waiting time, T, for the event to occur. Since this is a positive valued random variable, a sensible practice is to model the natural logarithm of T. Its distribution is sufficiently skewed to require non normal distributional assumptions for the errors. A very general representation, conditional on a set of explanatory variables, is the following log linear model:

$$\ln T = \alpha + \beta * Z + \sigma * W \qquad (18)$$

where Z is a (column) vector of covariates, W is an error term and σ is a scale parameter; α and β, a row vector, represent the constant and covariate effects respectively. Two of the most common assumptions for the distribution of W are: a) W is an extreme value distributed variable and $\sigma=1$. This occurs if T has an exponential distribution with parameter $\mu=\exp(-(\beta * Z+\alpha))$; b) W is an extreme value distributed variable but $\sigma=1$. This occurs if T has a Weibull distribution with parameters $\rho=1/\sigma$ and $\mu=\exp(-(\alpha+\beta/\sigma))$. Other commonly used distributions for W are standard normal (when T is log normal), logistic (when T is log logistic).

It should be noted that while the representation (18) is instructive for the understanding of the structure of the models, estimation in the presence of censoring must take place through ML estimation of the hazard version of these models.

PARTIALLY PARAMETRIC MODELS If the analyst is more interested in recovering the effects of the covariates than in understanding the duration structure of the process or if there is no guidance about the nature of the latter, the best choice is a partially parameterized proportional hazard model. It has been shown by Cox in a seminal paper (Cox, 1972) and more formally demonstrated by others (Efron, 1977; Breslow, 1975; Oakes, 1977) that if the ratio of the hazards corresponding to individuals with different values in the covariates is independent of time, it is possible to retrieve quite efficient estimates of β without in anyway making explicit the nature of the nuisance function, $\mu_o(t)$.

In fact, assume that Z_i were dichotomous assuming values 0 and 1. Two individuals having the same value in the remaining covariates can be shown to have relative risks given by $\exp(\beta_1)$, a factor not involving $\mu_o(t)$. Suppose that the ordered failure times in the sample of interest are $t_1 < \ldots < t_d$, where the subscripts are indeces indentifying individuals with vectors of covariates being respectively $Z_1,\ldots Z_d$, and d is the total number of failures observed (events of the type being analysed). One can first argue that intervals of time within which no events have occurred can in no way add information on β (Cox, 1972; Kalbfleish & Prentice, 1980; Coleman, 1981). The argument then

proceeds by calculating for each failure time t_j the probability that the failure should have occurred to the individual with covariates Z_j. Notice that in addition to the individual failing at t_j, all the individuals whose failure (or censoring) times exceed t_j are members of a set still exposed to the risk of failing at t_j. Given such risk set and given that a failure occurred at t_j, the probability that it was experienced by the individual with covariates Z_j is given by:

$$P_j = \frac{\lambda_0(t_j) \exp(\beta * Z_j)}{\sum_{k \in R(t_j)} (\lambda_0(t_k) \exp(\beta * Z_k))} \quad (21)$$

where $R(t_j)$ is a set of labels containing all those who were at risk at t_j. The qdantity $\lambda_0(t)$ cancels out from numerator and denominator. What remains provides sufficient information on the parameters in β. The product of P_j over all j (failure times) gives the so called partial likelihood.

Several points should be emphasized. First, the resulting partial likelihood not only leaves out accounting of the occurrence of no events in the inter-failure intervals. It also leaves out information about the total number and sequence of failures that have occurred before each failure time j. Second, as most likelihood functions in event history analysis, it also leaves out information on the censoring mechanisms. Although Cox's partial likelihood has neither a marginal nor a conditional probabilistic interpretation, its maximization produces consistent and efficient estimates of β (Efron, 1977; Kalbfleish & Prentice,1980; Oakes,1977, Tsiatis,1981).[2] In particular, it provides estimates of the effects of covariates that, when contrasted with those estimated using (the correct) representation of the baseline risk, compare rather well (Cox & Oakes, 1984; Tuma & Hannan, 1984). The efficiency of small sample estimates, however, is notoriously reduced.

In the argument used above, it was assumed that no ties existed in the failure times. If they do, corrections to the likelihood can be applied (Peto, 1972; Breslow, 1974; Efron, 1977; Oakes, 1981). Further, if there are ties between censored and failure times, little is lost by assuming that the former occur inmediately after the latter. If ties

---

2. It has been shown, however, that the likelihood proposed by Cox corresponds to the marginal likelihood of ranks (Kalbfleish & Prentice, 1980).

are very frequent, more efficient estimates are obtained by using discrete models.

There are two additional issues of importance in the estimation of a proportional hazard model. The first regards the possible retrieval of the baseline hazard. Even though the investigator may not be interested in drawing inferences for it, predictions cannot be made witout some knowledge about it. Breslow (Breslow, 1974) has proposed a method to derive the baseline hazard. This method is based on an aggregation of time into intervals defined by the observed inter-failure times. Other procedures, however, are available (Cox, 1972; Kalbfleisch & Prentice, 1980; Oakes, 1972).

The second issue regards the legitimacy of the assumption about proportionality of hazards. This is not an assumption that can be trivially justified and its appropiatness is as relevant as the correct specification and measurement of covariates. There are several strategies that can be followed to check whether or not proportionality holds. First, general test for the analysis of residuals may be useful (Cox & Snell, 1968; Crowley and Hu, 1977; Kay, 1977). These tests, however, have been found to be somewhat unsatisfactory (Lagakos, 1981). In some cases it is possible to empirically examine nonparametric estimates of the survival functions for subgroups characterized by different values of the covariates (the latter have to be discrete variables or categorized continuous variables). Plots of log(-log(S(t)) that are parallel to each other reinforces the idea of proportional hazards. These tests may suggest stratification of the sample into subgroups within which the proportionality assumption is closer to reality.

Non proportionality generally emerges when: a) there is an interaction between some of the covariates in the model and time, b) there are important covariates that are time varying rather than fixed, and c) the model has been misspecified. To test the existence of interaction between time and covariates it is necessary to have a method that permits the inclusion of time varying covariates. The same type of methods allow estimation of the models whith time varying covariates. Finally, if model misspecification is due to left out covariates, one needs to draw from the theory on inference with underlying heterogeneity. Both issues are dealt with after the next section.

Discrete Models for the Analysis of Event Histories

Although rare, there are instances in which the events of interest occur at discrete times. More common are situations in which either the times of failure are not exactly recorded or only the number of events and censored cases within discrete intervals is known. Finally, there are other examples characterized by continuously occurring events

with a fine (but not exact) recording of failure times and exhibiting a large number of ties. The latter may be induced in part by a large number of observations and in part by high rates of occurrence. In these three cases it is convenient to translate the proportional hazard model into a discrete version. We will briefly review a few alternatives strategies that permit estimation of these models.

A DISCRETE HAZARD MODEL Suppose that events are recorded as occurring in a number of exogenously defined disjoint intervals. If the underlying risk follows a proportional hazard model, then the hazard contribution of a failure that occurs in the ith interval and that is characterized by covariates Z_i is given by:

$$1-(1-\gamma_i)^{\exp(\beta * Z_i)}$$

where γ_i is the cumulated baseline risk within the interval. If one assumes that censoring occurs only prior to the end of the intervals, a simple likelihood function obtains. The estimates of the discrete versions of the risk, α_i, are $\log(-\log(1- d_i/N_i))$ where d_i and N_i are respectively the number of failures in the i^{th} interval and the number of individuals who were exposed to the risk at the beginning of the interval. Similarly, estimates of the effects of covariates can be obtained. This procedure is recommended when the intervals are neither too numerous nor too fine so that the number of events contained in them is sufficiently large. If there are too many intervals, there will be too many nuisance parameters to estimate. If the intervals are too fine, the numerical optimization methods can be quite unstable and the variance of the estimates large.

A LOG LINEAR VERSION OF A HAZARD MODEL Assume as before that the failure times are grouped in a number of exogenously defined discrete intervals. Further assume that: a)the hazard is a step function that remains constant within each of the intervals, and b)failures and censoring are assumed to occur in the middle of each interval. If all the covariates are discrete, it can be shown that the likelihood function reduces to the likelihood function of a log linear model. One can then retrieve estimates of the covariates and of the hazards within each interval using a program for log linear analysis suitably modified to take into account the exposure contributed by all those who failed or were censored within the intervals (Laird & Olivier, 1981; Trussell & Hammerslough, 1983).

If some covariates are continuous, the approach presented above cannot be used but a discrete formulation can still be applied. However, estimates have to be retrieved through especially tailored optimization procedures rather than by ready made computer packages (Menken et al,

1981; Palloni & Millman, 1986; Petersen, 1986)

In both cases care has to be taken to define the intervals. If the assumption made about the exact timing of failures and censoring is not correct, small inconsistencies will result (Petersen, 1983). The smaller the interval, the lesser the inconsistencies will be. However, it should be recalled that smaller intervals may lead to low frequencies of events and hence to potential instability of the numerical procedures and/ or to inefficiency of the estimates so obtained.

A LOGISTIC VERSION OF A HAZARD MODEL To facilitate estimation through the use of ready made estimation programs, one could formulate a discrete model where instead of directly defining the risks one defines the odds of failure as if they followed a logistic pattern:

$$\frac{q_{ji}}{1-q_{ji}} = \exp\,(-\beta_j * Z_i)$$

where q_{ji} is the probability of failing in interval j for individual i. Including a unit variable in the vector Z, leads to a model in which the odds are proportional to each other. The estimate of the constant is an estimate of a baseline for the odds. Thus, the parameters that one retrieves do not correspond to estimates of effects on the hazards. The discrepancies, however, are small when the intervals are small or when the underlying risks are low. The advantage of this model is that its parameters can be estimated using any of the computer packages that have been designed to do analysis of discrete data (see Section 4). The estimation can be done separately for each interval or jointly, constraining β_j to be the same (Allison, 1982; 1984; Guilkey & Rindfuss, 1983, unpublished).[3] This latter feature allows testing of models that are more general than the ones included in the proportional hazards model. In fact, the hypothesis can be tested that the causal process may be different across time intervals. The latter may be so because: a) different covariates are included depending on the time interval being examined, and b) because the parameters β_j are different across

---

3. It is important to notice, however, that using a package designed to estimate a logistic model, will not provide correct estimates of the baseline risks. In fact, the estimates are obtained as if the individuals censored within the interval had been censored right before the end of it. If this is unrealistic, simple corrections will produce more accurate estimates.

time intervals,e.g. the assumption of proportionality is violated.).

Other procedures to estimate discrete versions of a proportional hazards model have been formulated by several authors (Cox, 1972;Breslow, 1974;Kalbfleish & Prentice, 1980). In these formulations the likelihood function does not reduce to one that can be easily maximized with standard software.

Time Varying Covariates

So far we have assumed that the covariates included in the model are fixed. This may be a wrong representation in many cases. Provided that the relevant time varying covariates are exogenous to the process (that is, their changes are not induced by the process itself), one can quite easily formulate models incorporating them. The likelihood function for the observed failures is very similar to the one presented above except for the fact that the vector Z may contain a number of covariates indexed by time. It is particulary simple to incorporate categorical or discrete time varying covariates. The strategy then is to censor spells each time that a covariate changes. Care must be taken to keep track of the correct total spell length in models incorporating duration dependence (Petersen, 1986).

Time varying covariates may present difficulties. One complication is that for the likelihood to be maximized one requires that the covariates be defined at the time a failure occurs. This requirement is very difficult to be satisfied with plans of observations that do not record the values of covariates continuously. To resolve this problem one must resort to interpolation of values or to actually modeling the process that generates them (Tuma & Hannan, 1984). It has been demostrated, however, that the estimates of the covariates are extremely sensitive to the procedures one uses to assign values to the covariates at points when the actual information is missing (Flinn & Heckman, 1982a).

Another complication is of a computational nature: even when there are very few observations and very few time varying covariates, the numerical algorithms that can be implemented to estimate the parameters are costly and time consuming (Allison, 1984; Cox & Oakes, 1984; Kalbfleish & Prentice, 1980). The only alternative that appears to be feasible is to sacrifice some efficiency by aggregating time and to proceed with models for discrete data described previously. Resorting to this type of solution, however, does not permit to get around the first complication mentioned above.

Estimation in the Presence of Unobserved Heterogeneity

The models that we presented above do not explicitly incorporate the presence of disturbances of any type. When no covariates are relevant, the models imply the existence of a single risk for all individuals as if all of them were homogeneous. When covariates are introduced, one assumes that the risks, conditioned on the values of the covariates, are the same. Such assumption is likely to be unrealistic in most applications in social sciences. It would be desirable to be able to incorporate the presence of disturbances. However, the theory of inference for proportional hazards models in the presence of heterogeneity faces considerable problems and is as yet little developed. We first discuss the nature of the problem and then review some of the solutions that have been proposed.

THE NATURE OF THE PROBLEM We start with a simple example in which no covariates are present. Suppose that the individual risks are time invariant. Suppose further that the population is initially divided into two subgroups one of which exhibits higher risks than the other. As times passes the composition of the population in terms of these subgroups will change and the proportion of individuals with lower risks will increase relative to the one with higher risks. As a consequence, the aggregated risks will appear to decrease rather than remain constant. In this case, the omission of the characteristic defining the subgroups leads to erroneously identify a negative duration dependence where there is none. This mechanism was long ago argued to acount for rates of job changes that appeared to decline over time (Blumen et al, 1955). The same idea is at the base of the explanations given to account for the 'cross- over' of the mortality curves of the US Black and White populations (Vaupel et al, 1979) and it is a good candidate to explain declining risks of re-employment in unemployment spells (Heckman & Willis, 1977).

Other examples illustrate somewhat different types of biases. Thus, if the individual risks exhibited positive duration dependence but the levels of the risks were different across individuals, the aggregated risk would exhibit a weaker positive duration dependence, e.g. would rise less steeply with time. In fact, the individual risks could be so distributed that the aggregated rate would appear to be constant (Heckman & Singer, 1982b, 1984a).

In all these examples the structure of time dependence of the risk is misidentified due to the omission of relevant covariates. A more general situation occurs when the investigator postulates models that include both a time structure and a set of covariates that is incomplete to characterize the phenomenon. Since the distribution of individual risks, will change over time due to the effects of the left out characteristics, one will attribute to the time dependence effects

that pertain to the covariates.

STRATEGIES TO DEAL WITH HETEROGENEITY A general proportional hazard model for a single hazard in the presence of unmeasured heterogeneity can be written as follows:

$$\gamma(t,Z,\beta,\xi) = \gamma(t)\exp(\beta * Z)\,\xi \qquad (22)$$

where the first component captures the time structure of the model, the second reflects the effects of measured covariates, and the third is an error component. A crucial restriction in what follows is that neither Z nor ξ are allowed to vary with time (Heckman & Singer, 1984a, 1984b). The problem that the component ξ engenders is essentially one of identifyability: if it is not included, the observed data may not permit to correctly capture either the time structure or the set of covariate effects because they are hopelessly confounded. If the error component is included by making some assumptions about its distribution, the question becomes whether the identified time structure and set of effects would have been different had one changed such distribution.

There are several strategies to control for the effects of the heterogeneity component. We will classify them according to the type of data that they require and to their parametric nature.

With data on single events (the jth episode of a repeated event or the only episode of an unrepeatable event) there are two strategies. The first one is the so called random effects model. It revolves around the formulation of a distribution for ξ that dependes on a certain number of parameters. It can be shown that, once the distribution is especified, model (22) can be manipulated to yield a likelihood function that does not depend on ξ but depends on the parameters determining the time structure of the model, the set of covariates effects, β, and the parameters governing the distribution of ξ. The manipulation procedure is euphemistically referred to as 'integrating out' the heterogeneity component. The second strategy is nonparametric, in that one does not need to assume a well defined distribution for ξ but only the existence of a finite set of values for it ('support points '). The estimation procedures then yield as a result -- in addition to estimates of the time structure and β -- estimates of the number of support points and the sample distribution around them (Heckman & Singer, 1982a; Heckman & Singer, 1982b).

If what is available is a sample of individuals that have experienced an event repeatedly, it is possible to eliminate the effects of ξ. One strategy requires that ξ be the same across each episode for an individual but possibly different across individulas (fixed effect model) (Judge et al, 1984). The other strategy requires that the

structure of time duration, conditioned on ξ, be defined by a member of the exponential family of waiting times. If so, it is possible to use multiple spells to cancel out the effects of ξ (Chamberlain, 1982).

The parametric strategy with single episodes has been extensively reviewed (Vaupel & Yashin, 1982) and applied in a variety of areas: for the analysis of employment events (Flinn & Heckman, 1982b), of mortality (Manton et al, 1981), of social mobility (Spilerman, 1972), of job shifts (Tuma & Hannan, 1984), and of fertility (Newman, 1981, unpublished). However, its application is marred by several complications. The first is of a computational nature: the procedure requires numerical integration in most cases and can increase enormously the time and computational costs. The second problem is much more serious, since it refers to the sensitivity of the estimates of β to the distributional assumptions made about ξ. It has been shown with concrete examples (Heckman & Singer,1982a; Trussell & Richards, 1985) that the estimated values of β can vary wildly denpending on which distribution is used. Since in social sciences there hardly are any instances in which we could claim to have a justification to choose between alternative distributions, this is a major stumbling block: if we leave out the heterogeneity component we are probably biasing our results. If it is introduced with parametric strategies, the results may be biased owing to misidentification of the correct error distribution.

The nonparametric strategy, based on work done on nonparametric ML procedures (Laird, 1978), was introduced very recently (Heckman & Singer, 1982a) and more experimentation is required to pass judgment on its utility. However, recent studies have revealed that its apparent advantage is diminished by some major limitations. The first one is rooted in the estimation procedures: it is quite difficult to numerically maximize the likelihood function when a finite number of support points is introduced regardless of what type of numerical techniques one uses (Expectation Maximization or conventional gradient methods). The second one is that, although the set of covariates effects can be recovered quite well with a fixed duration structure, changes in the latter lead to very different estimates of both the set of covariate effects and the underlying mixing distribution (Trussell & Richards, 1985). Satisfactory identification of the latter occurs, however, only when the underlying distribution producing heterogeneity is discrete. Finally, no statistical theory has been formulated for the nonparametric estimators derived from the procedures suggested by Heckman and Singer.

POTENTIAL SOLUTIONS The foregoing considerations suggest that in order to avoid misleading conclusions, the selection of a duration structure and of a mixing distribution has to be done with great caution. On the one hand, with single spell data, complete identification of the

parameters in (22) is impossible unless some restrictions are imposed apriori. Recent work on the area (Elbers & Ridders, 1982; Heckman & Singer, 1984b) has explored the identifyiability of families of models that depend on flexible duration structures and on mixing distributions that have more structure than those of the non parametric approach.

On the other hand, multiple spell data convey more information than single spell data and may permit the removal of restrictive assumptions. The problem here is that information on multiple spells may be of lower quality than that that for single spells. Yet, exploitation of these type data deserves as much attention as the introduction of refinements for the study of single spell data with heterogeneity components.

AVAILABLE SOFTWARE FOR THE ESTIMATION OF MODELS.

This is a brief summary of computer programs that are available for estimation of the models proposed in the previoussection. The listing is limited to ready-made packages that can be directly used to estimate hazard models. We do not describe general optimization programs that can be modified to estimate hazard models. Most of the problems generated by dealing with heterogeneity require use of the latter.

Two special programs have been written for analysis of event history data. One of them is available from Kalbfleisch and Prentice and a listing is published in the Appendix to their book. It is a program to estimate the partially parametric (Cox's) proportional hazards models with time varying covariates. The other is the program called RATE (Tuma,1979). This program permits the introduction of time dependent covariates and the estimation of linear and log-linear models for the hazard with or without time dependecy (Gompertz models). It is also possible, but a bit cumbersome, to estimate the partially parametric proportional hazards model. A new version, soon to be released, should increase the variety of models that can be estimated and also allow for easy variable transformations.

GLIM (Baker & Nelder, 1978) is a general purpose program which fits general linear models. Logistic models can be very easily estimated. With some manipulations, it also possesses the capability for producing estimation of the exponetial, Weibull, and log-logistic hazard models.

BMDP: Recent releases directly incorporate a program to estimate the partially parametric proportional hazards model with time varying covariates and corrections for the occurrence of ties. A program to obtain both Life-Table and Kaplan-Maier estimates of the Survivor functions is also available. BMPD also contains general algorithms that can be utilized for maximization of likelihood functions with or without exact derivatives (P3R). These algorithms may be used to estimate all the main hazard models, including models with time varying covariates, using a special subroutine developed by Petersen (1986).

LOGLIN: This program was developed by Olivier and Neff (Olivier & Neff, 1976). It allows estimation of the hazard models with discrete covariates and permits exact accounting of exposure. It also permits to combine categories of variables. It is inflexible for creating new variables (data must be prepared before hand) and does not allow too much flexibility in specifying different models for different intervals. GLIM is more efficient in this regard.

SPSS: Recent releases permit estimation of the survivor function and tests for differences between survivor functions. Hazard models cannot be estimated.

Literature Cited

Aalen, O.O. 1978. Nonparametric inference for a family of counting processes. *Annals of Statistics*,6,pp.701-726.

Allison, P.D. 1982. Discrete time methods for the analysis of event histories. In *Sociological Methodology*.Ed.S.Leinhardt, pp.61-98.San Francisco:Jossey-Bass.

Allison, P.D. 1984.*Event history analysis*. Quantitative Applications in the Social Sciences. Sage Publications.

Amemiya,T. 1984. Tobit models:a survey. *Journal of Econometrics*, 24,pp.3-61.

Baker, R.J., Nelder,J.A. 1978.*The GLIM System Manual, Release 3*. London: Numerical Algorrithms Groups.

Birnbaum, Z.W. 1979. On the mathematics of competing risks. *Vital and Health Statistics* Series 2-Number 77. DHEW Publication Publication No(PHS)70-1351.

Blossfeld,H.P.,Hammerle,A.,Mayer,K.U. 1986.*Ereignisanalyse: Statistische Theorie und Anwendung in den Wirtschafts- und Sozialwissenschaften*. Frankfurt: Campus Verlag.

Blumen,I., Kogan,M., McCarthy,P.J. 1955. *The industrial mobility as a probability process*. Cornell Studies in Industrial and Labor Relations No 6. Ithaca: Cornell University Press.

Breslow,N.E., Crowley,J. 1974. A large sample study of the life table and product limit estimates under random censorship. *Ann.Stat.*,2,pp.437-453

Breslow,N.E. 1974.Covariance analysis of censored survival data. *Biometrics*,30,pp.89-100.

Breslow,N.E., 1975. Analysis of survival data under the proportional hazards model. *Int. Statist. Rev.*,pp.45-58.

Carroll,G.R. 1982.*Dynamic analysis of discrete dependent variables: a didactic essay: ZUMA-Bericht No 1982/08.*

Chamberlain,G. 1982. On the use of panel data. In *Longitudinal Studies of the Labor Market*. Eds. J.J. Heckman,B.Singer. New York: Academic Press.(forthcoming).

Chiang,C.L. 1968. *Introduction to Stochastic Processes in*

Biostatistics.New York: Wiley.

Cioffi-Revilla,C. 1984.The political reliability of Italian governments. *American Political Science Review*,78,pp.318-337.

Coale,A.J., McNeil,D. 1972. The distribution by age of the frequency of first marriages in a female cohort.*Journal of the American Statistical Association*,67,pp.743-749.

Coleman,J.S. 1964. *Introduction to Mathematical Sociology*.New York: Free Press.

Coleman,J.S. 1981. *Longitudinal Data Analysis*.New York: Basic Books.

Cox,D.R.,Snell,E.J. 1968. A general definition of residuals (with discussion). *J.R.Stat.Soc.*,30,pp.248-275.

Cox,D.R. 1972. Regression models and life tables (with discussion). *J.R.Stat.Soc.*,B,34,pp.187-220.

Cox,D.R.,Hinkley,D.V. 1974. *Theoretical Statistics*.London: Chapman and Hall.

Cox,D.R.,Oakes,D. 1984. *Analysis of Survival Data*.London: Chapman and Hall.

Crowley,J.,Hu,M. 1977. Covariance analysis of heart transplant data. *Journal of American Statistical Association*,72,pp.27-36.

Efron,B. 1967. The two sample problem with censored data. *Proceedings of the Fifth Berkeley Symposium in Mathematical Statistics*,IV. New York: Prentice-Hall.

Efron,B. 1977. Efficiency of Cox's likelihood function for censored data. *Journal of the American Statistical Association*,72, pp.557-565.

Elandt-Johnson,R.C., Johnson,N.L. 1980. *Survival Models and Data Analysis*.New York:Wiley.

Elbers,C.,Ridder,G. 1982. True and spurious duration dependence: the identifiability of the proportional hazards model.*Review of Economic Studies*,49,pp.403-411.

Flinn,C.,Heckman,J.J. 1982a. New methods for analyzing individual event histories. In *Sociological Methodology*.

Ed. S.Leinhardt,pp.99-140. San Francisco: Jossey-Bass.

Flinn,C.,Heckman,J.J. 1982b. New methods for analyszing structural models of labor force dynamics. *Journal of Econometrics*, 18,pp.115-168.

Freeman,J.H.,Carroll,G.R.,Hannan,M.T. 1983. The liability of newness: age dependence in organizational death rates. *American Sociological Review*,48,pp.692-710.

Gehan,E. 1969. Estimating survivor functions from the life table. *Journal of Chronic Diseases*,21,629-644.

Goldfeld,S.,Quandt,R. 1972. *Nonlinear Methods in Econometrics*. Amsterdam:North-Holland.

Gross,A.,Clark,V. 1975.*Survival Distributions: Reliability Applications in the Biomedical Sciences*.New York: Wiley.

Heckman,J.J.,Willis,R. 1977. A beta logistic model for the analysis of sequential labor force participation of married women. *Journal of Political Economy*,85,pp.27-58.

Heckman,J.J. 1979. Sample selection bias as a specification error. *Econometrica*,47,pp.153-161.

Heckman,J.J.,Singer,B. 1982a. Population heterogeneity in demographic models. In *Multidimensional Mathematical Demography*.Eds. K.C. Land, A.Rogers,pp.567-601. New York: Academic Press.

Heckman,J.J.,Singer,B. 1982b.The identification problem in econometric models for duration data. In *Advances in Econometrics*.Ed. W.Hildebrand, pp.39-76. Cambridge University Press.

Heckman,J.J.,Singer,B. 1984a. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*,52,pp.271-320.

Heckman,J.J.,Singer,B. 1984b. The identifiability of the proportional hazard model. *Review of Economic Studies*,LI,pp.231-241.

Hoem,J.M. 1983. *Multistate Mathematical Demography Should Adopt the Notions of Event History Analysis*.Research Report in Demography,No 10.Department of Statistics,University of Stockholm

Judge,G.G.,Griffiths,W.E.,Carter Hill,R.,Lutkepohl,H.,Chao Lee,T. 1984 *The Theory and Practice of Econometrics*.New York: Wiley.

Kalbfleish,J.D.,Prentice,R.L. 1980. *The Statistical Analysis of Failure Time Data*.New York: Wiley.

Kaplan,E.L.,Meier,P. 1958. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*,53,pp.457-481.

Kay,R. 1977. Proportional hazard regression models and the analysis of censored survival data.*J.R.Stat.Soc.*,C,26,pp.227-237.

Lagakos,S. 1981. The graphical evaluation of explanatory variables in proportional hazard regression models.*Biometrika*,68, pp.93-98.

Laird,N. 1978.Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association*,73,pp.805-811.

Laird,N.,Olivier,D. 1981.Covariance analysis od censored survival data using log-linear analysis techniques.*Journal of the American Statistical Association*,76,pp.231-240.

Le Cam,L. 1970.On the assumptions used to prove asymptotic normality of maximum likelihood estimates.*Ann.Math.Stat*,41,802-828.

Maddala,G.S., 1983. *Limited Dependent and Qualitative Variables in Econometrics*.Cambridge University Press.

Menken,J.,Trussell,J.,Stempel,D.,Babakol,O. 1981.Proportional hazards life table models:an illustrative analysis of sociodemographic influences on marriage dissolution in the United States. *Demography*,18,pp.181-200.

Miller,R.G., 1981.*Survival Analysis*.New York: Wiley.

Moran,P.A.P. 1971.Maximum likelihood estimation in non-standard conditions. *Proc.Camb.Phil.Soc.*,70,pp.441-450.

Oakes,D. 1972.Contribution to discussion of paper by D.R.Cox *J.R.Statist.Soc.*,B,34,pp.208.

Oakes,D. 1977.The asymptotic information in censored survival data *Biometrika*,64,p.441-448.

Oakes,D. 1981.Survival times:aspects of partial likelihood(with discussion). *Int.Stat.Rev*,49,pp.199-233.

Olivier,D.,Neff,R. 1976.*LOGLIN 1.0: User's guide. Harvard School of Public Health.*

Palloni,A.,Millman,S. 1986.Effects of interbirth intervals and breastfeeding on infant and early childhood. *Population Studies* (forthcoming).

Petersen,T., 1983. *Time aggregation bias in continuous time hazard rate models for analysing duration data.*Center for Demography and Ecology, University of Wisconsin,Working Paper No 83-45.

Petersen,T., 1986. Estimating fully parametric failure time distributions with time-dependent covariates by the method of maximum likelihood. *Sociological Methods and Research* , 14,pp. 219-246.

Peto,R. 1972.Contribution to the discussion of paper by D.R.Cox. *J.R.Statis.Soc.*,B,34,pp.205-207.

Peto,R.,Pike,M.C.,Armitage,P.,Breslow,N.E.,Cox,D.R.,et al. 1977.Design and analysis of randomized clinical trials requiring prolonged observation of each patient.Part 2. *Br.J.Cancer*,35,pp.1-39.

Salant,S.W.1977. Search theory and duration data: a theory of sorts. *Quarterly Journal of Economics*.91,pp.39-57.

Schoen,R.,Land,K.C. 1979.A general algorithm for estimating a Markov-generated increment-decrement life table with applications to marital status patterns. *Journal of the American Statistical Association*,74,pp.761-776.

Sorensen,A.B. 1977. Estimating rates from retrospective questions. In *Sociological Methodology*,ed. D.Heise,pp.209-223. San Francisco: Jossey-Bass.

Sorensen,A.B.,Tuma, N.B.1981. Labor market structures and the rate of job shifts. In *Research in Social Stratification and Mobility*, ed. D.Treiman,1,pp.361-84.

Spilerman,S. 1972.Extensions of the mover-stayer model. *American Journal of Sociology*,78,pp.599-626.

Teachman,J.D. 1983.Analyzing social processes: life tables and proportional hazards models. *Social Science Research*,12,pp.263- 301.

Trussell,J.T.,Hammerslough,C. 1983.A hazards model analysis of the

covariates of infant and child mortality in Sri Lanka. *Demography*,20,pp.1-26.

Trussell,J.T.,Richards,T. 1985.Correcting for unobserved heterogeneity in hazards models using the Heckman-Singer procedure. In *Sociological Methodology*, ed. N.Tuma,pp.242-276. San Francisco:Jossey-Bass.

Tsiatis,A. 1975.A nonidentifiability aspect of the problem of competing risks.*Proceedings of the National Academy of Sciences*,72,pp.20-22.

Tsiatis,A. 1981.A large sample study of Cox's regression model.*Ann. Statist*,9,pp.93-108.

Tuma,N.B. 1979. *Invoking Rate*. Department of Sociology, Stanford University.

Tuma,N.B.,Hannan,M.T. 1984. *Social Dynamics: Models and Methods* New York: Academic Press.

Vanderhoeft,C. 1984.Accelerated failure time models:an application to current status breast-feeding data from Pakistan.*Genus*,????, pp.135-157.

Vaupel,J.,Manton,K.,Stallard,E. 1979.The impact of heterogeneity in individual frailty on the dynamics of mortality.*Demography*, 16,pp.439-454.

Vaupel,J.,Yashin,A.I., 1982.*The deviant dynamics of death in heterogeneous populations*.International Institute for Applied Systems Analysis, Working paper No 82-47.