Survival Analysis of Backward Recurrence Times

Paul D. Allison

Journal of the American Statistical Association, Vol. 80, No. 390 (Jun., 1985), 315-322.

Stable URL:
http://links.jstor.org/sici?sici=0162-1459%28198506%2980%3A390%3C315%3ASAOBRT%3E2.0.CO%3B2-R

Journal of the American Statistical Association is currently published by American Statistical Association.

Survival Analysis of Backward Recurrence Times

PAUL D. ALLISON*

Many surveys include questions that attempt to measure the time of the most recent occurrence of some event, for example, last visit to a physician. Although it is tempting to apply survival (failure-time) methods to such data, the conditions under which such applications are appropriate have not been apparent. In this article it is shown that standard methods may be applied when the data arise from certain well-known stochastic processes. Special procedures may be necessary if the models include duration dependence, however. The methods are illustrated by the estimation of regression models for data on residential mobility.

KEY WORDS: Forward recurrence times; Failure time analysis; Generalized gamma distribution; Open intervals; Regression analysis; Residential mobility.

1. INTRODUCTION

Surveys often include questions of the following sort: When did you last see a doctor? How long have you been working at your present job? When was the last time you and your spouse quarreled? How long have you been living in this house? All of these questions attempt to measure the length of time between the survey and the last occurrence of some event. Such questions are typically asked with no attempt to get information about earlier events, either because of limitations on the number or complexity of questions or because it is believed that respondents can accurately recall the timing and circumstances of only the most recent event.

Borrowing terminology from renewal theory (Cox 1962), I shall refer to such data as backward recurrence times. Though there are many possible uses for backward recurrence times, it is tempting to use them to measure the rate or hazard of event occurrence. If Jones published a paper last month, for example, whereas Smith's most recent paper appeared 10 years ago, it is natural to guess that Jones has a higher rate of publication. Or consider a more complicated example, which I shall pursue in detail in later sections: In 1968 a national sample of approximately 2,300 male heads of households was asked, "When did you move into this house (apartment)?" The aim is to estimate models in which the hazard of residential mobility depends on several covariates.

A natural approach to the analysis of these data is the body of methods known as survival analysis or failure time analysis (Kalbfleisch and Prentice 1980; Elandt-Johnson and Johnson 1980; Lawless 1982). A possible difficulty with that approach is that from one point of view, all observations are censored. That is, there are no completed intervals between events, only intervals that have been interrupted by the survey itself. When all observations are censored, it is easily shown that the likelihood function always reaches its maximum at a boundary of the parameter space, and hence maximum likelihood estimation is problematic. For example, if the data consist only of censored observations from an exponential distribution with parameter λ, the likelihood is $\exp(-\lambda T)$, where T is the sum of the censored times. Regardless of T, this reaches a maximum of 1 when $\lambda = 0$. As we shall see, however, this result is misleading because it ignores the fact that the interrupted intervals all began with the occurrence of an event.

Sørensen (1977) used backward recurrence times to estimate the rate parameter in a Poisson process, but his estimators are unnecessarily complicated by the fact that he did not consider likelihood-based estimation. And although Ginsberg (1979) used backward recurrence times to estimate the distribution function of a renewal process, his method required additional data beyond the recurrence times. There has also been considerable interest on the part of demographers in "open birth intervals" —the length of time between the last birth and the time of a survey. Srinivasan (1966, 1968, 1970) argued that such intervals are especially sensitive indicators of changes in natality patterns, and his work has inspired a number of empirical studies (e.g., Hastings and Robinson 1975). Interest appears to have fallen off in recent years, however, perhaps as a result of the criticisms of the method by Leridon (1969), Sheps et al. (1970), and Venkatacharya (1972). The most serious difficulty, noted by Sheps et al., is the severe truncation of the distribution of the open interval when the sample is stratified by number of previous births, a problem that I shall also discuss in Section 5.

The demographic approach has been almost exclusively focused on the mean of the open interval. This article, however, is principally concerned with hazard functions, specifically the conditions under which it is possible to estimate hazard functions from data on backward recurrence times. The major finding is that in some situations it is possible to use standard methods with little or no modification, treating time as though it ran backwards from the time of the survey until the time of the event. Inference is also possible in other situations, but special procedures may be necessary.

Note that my objective is not to argue unequivocally for the use of backward recurrence times, but rather to clarify the conditions under which they can and cannot be legitimately used to estimate hazard functions. There are important limitations to the use of such data, and there are many areas of application in which such usage would be inappropriate.

A crucial requirement for estimating hazard functions from backward recurrence times is that the event be repeatable. If the event is not repeatable, the hazard is necessarily zero in the

* Paul D. Allison is Associate of the Population Studies Center and Associate Professor, Department of Sociology, University of Pennsylvania, Philadelphia, PA 19104. This work was supported in part by National Institute of Justice Grant 82-IJ-CX-0018. Samuel H. Preston originally suggested the use of backward recurrence times to estimate hazard models and convinced me that it was possible despite my initial skepticism. His subsequent advice and encouragement were invaluable. For helpful suggestions, I am also indebted to Thomas Santner, Bruce Turnbull, and Ronald Schoenberg. Michael Strong and Stephen Taber provided programming assistance. Ross Prentice and Vern Farewell generously provided me with a copy of their computer program.

Journal of the American Statistical Association
June 1985, Vol. 80, No. 390, Applications

interval between the event and the survey; hence the length of that interval gives no information about the hazard prior to the event. Repeatable events are typically modeled as a stochastic process (Gail et al. 1980; Prentice et al. 1981; Lawless 1982), and I shall rely heavily on results from the stochastic process literature. Such models are necessarily more complicated than those for single events and pose a formidable array of specification decisions. The strategy here is to start with very simple models and then introduce more realistic complications one by one.

In Section 2, I consider the Poisson process, which is characterized by a constant hazard. Later sections then generalize this model in several different directions. Section 3 allows the hazard to vary with time from the origin of the process, and Section 4 allows variation with duration since the previous event. Sections 2–4 also apply the theoretical results to the analysis of data on residential mobility. In Section 5, I consider the problem of estimating models in which the hazard for event occurrence depends on the number of previous events. Section 6 extends some of the results of earlier sections to the case in which individuals alternate between two states. In Section 7, I discuss *forward* recurrence times, indicating that although most of the results of earlier sections apply to prospective data, some do not.

2. A CONSTANT-HAZARD MODEL

For events such as residence changes, which are repeatable and undifferentiated, it is commonplace to model the occurrence of events as a point process. Of course the simplest and best-known point process is the Poisson process, which I shall consider in this section. I begin with some notation that closely follows that of Prentice et al. (1981) and will be used in subsequent sections.

Let $N(t) = \{n(u) \colon 0 \leqslant u \leqslant t\}$ be a point process such that $n(u)$ counts the number of events in $[0, u)$. Note that $N(t)$ is equivalent to the set of random event times $T_1, T_2, \ldots, T_{n(t)}$ in $[0, t)$. Let z denote a $p \times 1$ vector of covariate values that, for now, are assumed to be constant over time. I define the hazard or intensity function to be

$$\lambda\{t \mid N(t), z\} = \lim_{h \to 0} \Pr\{n(t + h) - n(t) = 1 \mid N(t), z\}/h. \quad (2.1)$$

Equivalently,

$$\lambda\{t \mid N(t), z\} = \lim_{h \to 0} \Pr\{T_{n(t)+1} < t + h \mid N(t), z\}/h. \quad (2.2)$$

The models examined in Sections 2–5 are obtained by imposing various restrictions on this hazard function.

Now suppose that we observe n individuals ($i = 1, \ldots, n$) whose event histories are independent realizations of the point process just defined. For all individuals, the process is interrupted at some time τ, which is independent of $N(t)$. Let t_i be the earliest known time such that the half-open interval (t_i, τ) contains no events; and let $\delta_i = 1$ if an event is known to occur at t_i, 0 otherwise. For each individual, then, the data consist of (u_i, δ_i, z_i), where $u_i = \tau - t_i$ is the backward recurrence time and z_i is the covariate vector. We assume that nothing is known about the individual's event history (sample path) before t_i or after τ.

The variable δ_i indicates whether or not t_i is left censored. Note, however, that if t_i is left censored, then u_i is right censored. Despite this ambiguity we shall continue to refer to such censoring as left censoring. Left censoring can occur for a variety of reasons. If the events are residence changes, it may happen that some individuals have lived in the same house since birth. For such individuals, u_i is the person's age at time τ and $\delta_i = 0$. Censoring also occurs when individuals cannot recall the exact time of the most recent event, but only that it happened more than, say, 10 years ago. Even when respondents do report the time of the most recent event, it is sometimes desirable to treat all event times earlier than a certain value as censored at that value. This may happen either because respondents' recall is not trusted beyond a certain point in the past or because the specified model is not trusted beyond a certain point in the past. As an example of the latter situation, a model that assumes a constant hazard may be plausible over a relatively short interval, but not over a longer interval. As usual, censoring times are assumed to be independent of event times.

I turn now to the Poisson process, specified as

$$\lambda\{t \mid N(t), z\} = \lambda(z), \quad (2.3)$$

which asserts that the hazard is constant over time for each individual but may vary across individuals as a function of the covariates. When there are no covariates, it is well known that U_i—the random variable denoting the backward recurrence time—has a possibly censored exponential distribution with parameter λ (Feller 1971; Sørensen 1977). Allowing for dependence on covariates is a straightforward extension. Thus the problem of estimating $\lambda(z)$ from (u_i, δ_i, z_i) reduces to the standard problem of regression analysis of a censored exponential variate (Zippin and Armitage 1966; Glasser 1967). Note that this approach is equivalent to treating τ as the origin and letting time run backwards.

I applied this method to data on residential mobility. As part of the Michigan Panel Study on Income Dynamics (Survey Research Center 1972), a national sample of 4,802 families in the U.S. was interviewed for the first time in 1968. In that year respondents were asked, "When did you move into this house (apartment)?" The analysis is restricted to 2,297 U.S.-born males who were "heads of households" and between the ages of 30 and 59 (inclusive) in 1968. The distribution of responses for this subsample is shown in Table 1.

In the data supplied by the Survey Research Center, moves prior to 1964 were grouped into multiyear intervals of varying length. This created problems that were most easily dealt with by treating all moves in 1963 or earlier as censored at 1964. As noted earlier, an additional advantage of such deliberate censoring is that it increases the plausibility of the assumption that the hazard is constant over the period of observation.

More than 400 variables were coded for the 1968 interview, many of which are plausible determinants of residential mobility. The vast majority of these variables did not realize their 1968 values until *after* the most recent move occurred, however. To facilitate a causal interpretation, covariates were limited to those determined prior to 1964, the earliest year at which an uncensored move could occur. This was a severe restriction, since the only variables that unambiguously satisfied that criterion were race, region of birth, father's education, parents'

Table 1. Year of Most Recent Move for Respondents Interviewed in 1968

Year	Number Who Moved	Number of Person-Years at Risk	Estimated Probability of Moving
1968	103	818[a]	.126
1967	229	2,194	.104
1966	216	1,965	.110
1965	200	1,749	.114
1964	178	1,549	.115
<1963	1,371		
Total	2,297		

[a] Although all respondents were at risk in 1968, they were not at risk for the entire year.

economic status when respondent was a child, and number of siblings. Also included were respondent's education and first occupation, even though there may have been a few cases in which education was completed or first occupation entered after 1964.

The model specified that $\lambda(z) = \exp(\beta z)$, where β is a $1 \times p$ vector of coefficients. This gives rise to the likelihood

$$\prod_{i=1}^{n} \exp[\delta_i \beta z_i - u_i \exp(\beta z_i)], \tag{2.4}$$

which is standard for exponential regression. Since the data are actually discrete, it would be desirable to estimate a model obtained by grouping exponential data into equal intervals (Kalbfleisch and Prentice 1980). I have done so, but the results are virtually identical to those obtained when a continuous distribution is assumed. For simplicity I report only the latter results. Estimates of the regression coefficients and their standard errors are shown under Constant Hazard in Table 2. Codings for the covariates are shown in the note to the table. To summarize these results, the hazard of residential mobility was lower among those whose first job was farming, those with more than a high school education, and those who were born in the North. Race, number of siblings, father's education, and parents' economic status had virtually no impact.

To evaluate the fit of the model, the residuals $u_i \exp(\beta z_i)$ described by Lawless (1982) were calculated, and the survival curve of the residuals was plotted. No departures from the exponential distribution were apparent. Correlations among the covariates and among the coefficient estimates were moderate to low.

3. A TIME-DEPENDENT MODEL

Although the assumption of a constant hazard may be a reasonable approximation in some settings, it is likely to be unduly restrictive in many others. Suppose that the model is relaxed by assuming that

$$\lambda\{t \mid N(t), z\} = \lambda(t; z), \tag{3.1}$$

which allows the hazard to depend on time and the covariates, but not on the previous event history. If we suppress dependence on the covariates, we have a time-dependent or nonhomogeneous Poisson process (Cox and Lewis 1966). Note that t is the length of time since the origin of the process, *not* the length of time since the previous event. Thus this model does not allow for what is commonly referred to as duration dependence.

The distribution of U, the backward recurrence time, is readily obtained. $\Pr(U > u)$ is equivalent to the probability that no events occur in the interval $[\tau - u, \tau]$. Ignoring dependence on z, from Cox and Lewis (1966) we have

$$\Pr(U > u) = \exp\left\{ - \int_{\tau-u}^{\tau} \lambda(x)\, dx \right\}, \tag{3.2}$$

which implies that the density for U is given by

$$f(u) = \lambda(\tau - u) \exp\left\{ - \int_{\tau-u}^{\tau} \lambda(x)\, dx \right\}. \tag{3.3}$$

Hence the hazard function for U, which is denoted by $\gamma(u)$, is given by

$$\gamma(u) = f(u)/\Pr(U > u) = \lambda(\tau - u). \tag{3.4}$$

We now introduce covariates by assuming that a proportional hazards model applies to $\lambda(\cdot)$; that is,

$$\lambda(t; z) = \lambda_0(t) \exp(\beta z), \tag{3.5}$$

where z is a vector of fixed covariates. If follows immediately that a proportional hazards model also applies to $\gamma(\cdot)$, that is,

$$\gamma(u; z) = \gamma_0(u) \exp(\beta z), \tag{3.6}$$

where $\gamma_0(u) = \lambda_0(\tau - u)$. If τ is constant across observations, then $\gamma_0(\cdot)$ is the same function across observations.

It is apparent, then, that the backward recurrence time U behaves just like an ordinary survival time arising from a proportional hazards model. Thus Cox's partial likelihood (1972a) should be an appropriate method for estimating β.

Table 2. Estimates for Regression Models Based on Backward Recurrence Times

	Constant Hazard			Time-Dependent Hazard			Age-Dependent Hazard			Duration-Dependent Hazard		
Covariate	β	SE	β/SE	β	SE	β/SE	β	SE	β/SE	β	SE	β/SE
FARM	−.512	.144	−3.56	−.524	.144	−3.65	−.348	.145	−2.40	−.259	.070	−3.72
ED	−.304	.080	−3.80	−.313	.080	−3.90	−.430	.080	−5.41	−.158	.037	−4.25
RACE	−.120	.087	−1.38	−.127	.087	−1.46	−.037	.087	− .42	−.051	.040	−1.29
SIBS	.016	.013	1.13	.016	.014	1.16	.020	.014	1.43	.006	.006	.88
ECON	.081	.073	1.10	.084	.073	1.15	.053	.074	.72	.030	.033	.93
FAED	.066	.093	.71	.067	.093	.71	−.064	.094	− .69	.020	.042	.48
NORTH	−.422	.146	−2.88	−.445	.147	−3.04	−.428	.147	−2.91	−.206	.066	−3.11
NC	−.174	.136	−1.28	−.189	.136	−1.38	−.218	.137	−1.59	−.088	.061	−1.45
SOUTH	−.206	.136	−1.52	−.220	.136	−1.62	−.211	.135	−1.56	−.119	.061	−.194

NOTE: Covariates are coded as follows: RACE = 1 if white, 0 otherwise; ED = 1 if more than high school education, 0 otherwise; SIBS is number of siblings; FARM = 1 if first occupation is farmer, 0 otherwise; ECON = 0 if parents were "poor," 1 otherwise; FAED = 1 if father had more than a grade school education, 0 otherwise; NORTH = 1 if respondent grew up in the North, 0 otherwise; NC = 1 if respondent grew up in the North Central Region, 0 otherwise; SOUTH = 1 if respondent grew up in the South, 0 otherwise. SE = standard error.

For the residential mobility data described previously, I estimated a proportional hazards model in which the baseline hazard was allowed to vary arbitrarily with calendar year. Covariates were the same as those used in the exponential regression analysis. Estimation was by partial likelihood, using the SAS supplemental procedure PHGLM (SAS Institute Inc. 1980). This program uses Breslow's (1974) approximation for tied data. The coefficients and standard errors shown under Time-Dependent Hazard in Table 2 are remarkably similar to those estimated under the assumption of a constant hazard. The fit of the model was evaluated by plotting the survival curve of the residuals, as suggested by Kalbfleisch and Prentice (1980). The fit appeared to be good, although Crowley and Storer (1983) have raised questions about the sensitivity of residuals from a proportional hazards model to departures from the model.

To this point, I have assumed that the point of interruption is the same for every observation in the sample. If the hazard is specified to be a function of calendar time and if the survey is conducted at approximately a single point in calendar time, then this is clearly an appropriate assumption. In some cases, however, different individuals in the sample may be questioned at quite different points in calendar time. In other cases, moreover, it may be more realistic to specify the hazard as a function of age since birth or some other natural starting point (Breslow et al. 1983; Farewell and Cox 1979). The hazard for a residence change, for example, is likely to vary more with age than with calendar time. If all individuals are interviewed at the same point in calendar time, the point of interruption will occur at different ages for different individuals.

When interruption times vary across observations, the previous argument for the use of the partial likelihood method is not directly applicable. Following Cox's original argument, the contribution to the partial likelihood for individual k with an event at time t_k is

$$\lambda(t_k;\, z_k) \Big/ \sum_{l \in R(t_k)} \lambda(t_l;\, z_l), \tag{3.7}$$

where $R(t)$ is the set of indices for individuals known to be at risk at time t. When τ is the same for all individuals, the size of $R(t)$ is an increasing function of t (i.e., a decreasing function of $u = \tau - t$). When τ varies across individuals, however, $R(t)$ consists of all j such that $t_j \leqslant t < \tau_j$. As a consequence, as t moves from the smallest t_j to the largest τ_j, $R(t)$ may both increase and decrease in size.

Although this creates no difficulty in theory, virtually all standard computer programs for partial likelihood estimation assume that the risk set never increases with time. What is needed is a program that will allow for "transient risk sets" (Mantel and Byar 1974), that is, for individuals to both leave and enter the risk set during the span of observation. Note that this problem is not peculiar to backward recurrence times. It will occur whenever one deals with repeatable events and a limited observation period that varies across individuals. This is because when events are repeatable, one cannot assume that no events occurred during periods when the individual was not under direct observation.

I modified a conventional partial likelihood program to allow for transient risk sets and used it to estimate a proportional hazards model for the residential mobility data. The baseline hazard was specified as an arbitrary function of respondent's age. Covariates were the same as those in the preceding analyses. The coefficient estimates under Age-Dependent Hazard in Table 2 are similar to those for the constant-hazard model, but not nearly so similar as the coefficients for the time-dependent model. The most noteworthy differences are the reduced coefficient for the farm indicator and the increased coefficient for the education indicator.

Fixed covariates have been assumed to this point. It is well known that the proportional hazards model and the partial likelihood method can be generalized to allow for time-dependent covariates, and this generalization would seem to be appropriate for backward recurrence times. There is one danger that demands careful consideration, however. Let us write the proportional hazards model as

$$\lambda[t;\, z(t)] = \lambda_0(t)e^{\beta z(t)}, \tag{3.8}$$

where $z(t)$ is a vector of possibly time-dependent covariates. In many situations, $z(t)$ may change in value as a consequence of an event at time t. If the event is a job change, for example, many characteristics of the job will change. With prospective data, it is usually understood that a causal interpretation requires that the time-dependent covariates realize their values prior (often just prior) to the occurrence of the event. This is still essential in the case of backward recurrence times, even though estimation usually proceeds as if time ran backwards from the point of interruption to the occurrence of the event. Thus in the case of residential mobility, covariates describing the residence must refer to the home vacated, not the home entered. Unfortunately, such information is rarely collected in retrospective surveys that produce backward recurrence times. For the residential mobility data, for instance, there is no information regarding characteristics of the previous residence.

The use of time-dependent covariates also makes it possible to let the hazard depend on both calendar time and age. The simplest approach, which can be done with standard programs, is to specify the hazard as an arbitrary function of calendar time and let age enter as a time-dependent covariate. An alternative approach is to use age to define time-dependent strata (Kalbfleisch and Prentice 1980; Breslow et al. 1983).

4. MODELS WITH DURATION DEPENDENCE

In modeling repeated events, it is common to assume that the hazard depends on the length of time since the immediately preceding event (Prentice et al. 1981; Gail et al. 1980; Flinn and Heckman 1982). In the notation used here, this can be specified as

$$\lambda\{t \mid N(t)\} = \lambda(t - t_{n(t)}). \tag{4.1}$$

This defines an ordinary renewal process so that interval lengths are iid random variables. If we allow for dependence on covariates, we have a modulated renewal process (Cox 1972b). Suppressing such possible dependence, let $F(\cdot)$ be the common distribution function for completed intervals. For renewal processes, it is well known that the backward recurrence time has a distribution function $G(\cdot)$ that differs from F (except in the special case in which F is the exponential distribution function).

In general, G will depend on τ, the length of time between the origin of the process and the point of interruption. For typical survey data, however, the origin of the process may not be known, at least not with much certainty. In many cases, it may be reasonable to assume that the point of interruption is relatively far from the origin. Then one can employ the well-known limiting distribution of the backward recurrence time (Karlin and Taylor 1975). The limiting density is given by

$$g(u) = \mu^{-1}[1 - F(u)], \qquad (4.2)$$

where μ is the mean length of completed intervals. The distribution function is thus

$$G(u) = \mu^{-1} \int_0^u [1 - F(x)]\, dx. \qquad (4.3)$$

Using this result, one can in principle transform models for F into models for G. The integral in (4.3) may be difficult to evaluate, however, except in a few special cases. One such case is the Weibull distribution

$$F(t) = 1 - \exp\{-(\lambda t)^\alpha\}, \qquad (4.4)$$

where α is a shape parameter and λ is a scale parameter. We then have

$$g(u) = [\lambda/\Gamma(1 + \alpha^{-1})] \exp\{-(\lambda u)^\alpha\} \qquad (4.5)$$

and

$$G(u) = I[\alpha^{-1}, (\lambda u)^\alpha], \qquad (4.6)$$

where I is the incomplete gamma function (as defined by Lawless 1982, p. 512). This distribution is a member of the family of generalized gamma distributions defined by Stacy (1962), which also includes the Weibull and two-parameter gamma distributions as special cases.

This result is convenient because Farewell and Prentice (1977) used the generalized gamma distribution as the basis for a regression model for censored data. If U is the backward recurrence time and z is a vector of fixed covariates, their model can be written as

$$\log U = -\beta z + v/\alpha, \qquad (4.7)$$

where v has the density $\exp(kv - e^v)/\Gamma(k)$. If we impose the restriction $k = 1/\alpha$, the distribution of U is G in (4.6) with $\lambda = \exp(\beta z)$.

I used the procedure proposed by Farewell and Prentice (1977) and Lawless (1982) to obtain maximum likelihood (ML) estimates of α and β for the residential mobility data, using the same covariates as in the earlier analyses. As noted in Section 2, a discrete-time model would actually be more appropriate for these data, but the results would probably be so similar that it would not be worth the substantial additional effort necessary to develop such a model.

The estimate for α was 3.06, implying that the underlying Weibull distribution has an increasing hazard; that is, the risk of moving increases with time since the last move. Note that $\alpha = 1$ corresponds to the exponential model estimated earlier. A likelihood ratio test of the exponential versus Weibull models yields a chi-square of 384 with 1 df. Estimates for β are presented in Table 2 under Duration-Dependent Hazard. Qualitatively, the results are similar to those obtained via exponential regression and partial likelihood; the ratios of the estimates to their standard errors are virtually identical. Both the estimates and their standard errors are somewhat smaller than those in the earlier analyses, however.

Note that although the Weibull regression model satisfies the proportional hazards assumption, the distribution of the backward recurrence time obtained from a Weibull renewal process does *not* satisfy the proportional hazards assumption. I suspect that this pattern holds more generally—that proportional hazards models for completed intervals imply hazard functions for backward recurrence times that are not proportional.

5. DEPENDENCE ON THE NUMBER OF PREVIOUS EVENTS

All of the models considered so far have been simplified in one key respect: they do not allow the hazard $\lambda\{t \mid N(t), z\}$ to depend on $n(t)$, the number of events that have already occurred by time t. This restriction is likely to be violated for human and animal subjects, since most events are positively or negatively reinforcing to some degree.

It is easy to formulate models that allow for such dependence, but it is extremely difficult to estimate them with backward recurrence times. This is obvious when the only datum is the backward recurrence time itself. But even if we are given the additional knowledge of the number of prior events, the estimation problem is formidable. Consider, for example, the relatively simple case of

$$\lambda\{t \mid N(t), z\} = \lambda_{n(t)+1}, \qquad (5.1)$$

which implies that intervals between events are independent and that the jth interval is exponentially distributed with parameter λ_j. Suppose that we have complete knowledge of the event history in $[0, \tau]$, and we wish to estimate a particular λ_k. The ML estimate of λ_k is easily obtained by considering only those individuals who experience at least $k - 1$ events and then applying the standard ML estimator for a possibly censored exponential variate to the kth interval. Censoring occurs only if the kth interval extends beyond τ.

It is tempting to take the same approach when we are given only the time of the most recent event and the number of prior events. That is, for a particular λ_k, we could restrict the sample to those who experienced exactly $k - 1$ events in the interval $[0, \tau]$ and treat the backward recurrence time as being exponentially distributed with parameter λ_k. Sheps et al. (1970), however, showed that this procedure may be highly misleading. For the model under consideration, their results imply that the distribution of the backward recurrence time conditional on the number of previous events is *not* exponential. In fact, the conditional density is given by

$$h(u) = [e^{-\lambda_k u} g_{k-1}(\tau - u)]/[G_{k-1}(\tau) - G_k(\tau)], \qquad (5.2)$$

where G_k and g_k are, respectively, the distribution function and the density function for the convolution of the first k intervals. Hence the distribution of the backward recurrence time depends on λ_j $(j = 1, \ldots, k)$ and not on λ_k alone. Estimation of λ_k thus appears to be quite intractable.

Let us specialize further by considering the case of $\lambda_j = \lambda$ for all j, which is just the Poisson process discussed in Section 2. Suppose that we did not know that $\lambda_j = \lambda$, however, and we tried to estimate the λ_j's by the procedure just described. That is, we stratified the sample by number of previous events and, within each stratum, used ML for an exponential variate. Now we know already that the marginal distribution of the backward recurrence time in this case is exponential with parameter λ. Nevertheless, the conditional density, given that j events already occurred, is given by

$$h(u) = j(\tau - u)^{j-1}/\tau^j, \tag{5.3}$$

which depends on the number of prior events but not on λ. Hence in this case, it is impossible to estimate λ by stratifying the sample. Furthermore, the hazard function associated with this density is $j/(\tau - u)$. Thus one might be led to conclude that the underlying hazard is increasing with the number of events when in fact it is constant. Intuitively, the reason for this anomaly is that the backward recurrence time is merely τ minus the sum of the intervals up to the last observed event. When the number of intervals (events) is large, we expect the backward recurrence time to be small.

A second example is provided by the "linear growth with immigration" process (Karlin and Taylor 1975), which has

$$\lambda\{t \mid N(t), z\} = a(z) + \theta n(t), \tag{5.4}$$

where a is some nonnegative function of the covariates and $\theta > 0$. It can be shown that the distribution function of U, conditional on both $n(\tau)$ and z, is given by

$$\Pr[U < u \mid n(\tau) = x, z] = 1 - \left(\frac{e^{\theta(\tau-u)} - 1}{e^{\theta\tau} - 1}\right)^x. \tag{5.5}$$

Thus conditioning on the number of prior events eliminates any dependence on the covariates.

In sum, it appears that models that allow the hazard to vary as a function of the number of previous events are not amenable to estimation with backward recurrence times. The attempt to do so can produce wholly misleading results.

6. ALTERNATING PROCESSES

Suppose that individuals may alternate between two states—for example, married or unmarried, employed or unemployed, on welfare or not on welfare. Suppose further that the data consist of the current state at some time τ and the length of time since the last change of states. Thus a survey might ask unemployed persons how long they have been out of work and employed persons how long it has been since they were last unemployed. The objective is to estimate the hazard to each state, which may vary with time or as a function of covariates.

Models for two-state processes can be difficult to estimate with this kind of data. Nevertheless, there is one simple model that can be estimated easily by using conventional methods. I assume that the data arise from a two-state birth and death process (Hoel et al. 1972) such that the hazard for moving from state 1 to state 2 is a constant λ, and for moving from state 2 to state 1, a constant μ. For the moment, I assume no dependence on covariates. Under this model, completed intervals in state 1 are exponentially distributed with parameter λ and completed intervals in state 2 are exponentially distributed with parameter μ. For an individual who is in state 1 at time τ and has been in that state since time t, the likelihood of the data is the product of three factors: (a) the probability of being in state 2 just prior to t, (b) the hazard for leaving state 2, and (c) the probability of staying in state 1 from t to τ. Using results in Hoel et al. (1972), the likelihood for an individual who was in state 1 at time 0 is therefore

$$(\lambda - \lambda e^{-(\lambda+\mu)t})/(\lambda + \mu) \times \mu \times e^{-\lambda(\tau-t)}. \tag{6.1}$$

For an individual who was in state 2 at time 0, the likelihood is

$$(\lambda + \mu e^{-(\lambda+\mu)t})/(\lambda + \mu) \times \mu \times e^{-\lambda(\tau-t)}. \tag{6.2}$$

These expressions have limited value because (a) we may not know how long it has been since time 0 and (b) we may not know the state at time 0. It is expedient then to assume that τ and t are both far from the origin, which implies that $\exp[-(\lambda + \mu)t]$ will be near zero. Regardless of initial state, the approximate likelihood for individuals who moved to state 1 at time t and remained there until time τ is therefore

$$\lambda\mu e^{-\lambda(\tau-t)}/(\lambda + \mu). \tag{6.3}$$

By a similar argument, the likelihood for an individual who, at time τ, has been in state 2 since time t will have a limiting value (as t and τ get large) of

$$\lambda\mu e^{-\mu(\tau-t)}/(\lambda + \mu). \tag{6.4}$$

For a sample of n independent individuals, the joint likelihood is thus

$$[\lambda\mu/(\lambda + \mu)]^n\, e^{-\lambda T_1 - \mu T_2}, \tag{6.5}$$

where T_1 and T_2 are the total amounts of time that individuals are known to be in each of the two states. Maximum likelihood estimators for λ and μ are readily found to be

$$\hat{\lambda} = n/(T_1 + \sqrt{T_1T_2}) \quad \text{and} \quad \hat{\mu} = n/(T_2 + \sqrt{T_1T_2}). \tag{6.6}$$

Although these estimators are easily computed, it is instructive to consider estimators conditional on the state occupied at τ. The limiting probability of being in state 1 is $\mu/(\lambda + \mu)$, and the limiting probability of being in state 2 is $\lambda/(\mu + \lambda)$. It follows that given a state of 1 at time τ, the conditional likelihood for the last state change being at time t is $\lambda \exp\{-\lambda(\tau - t)\}$ and, similarly, the conditional likelihood for a person in state 2 at τ is $\mu \exp\{-\mu(\tau - t)\}$. In short, the backward recurrence times have conditional distributions that are exponential with a parameter corresponding to the distribution of completed intervals in that state. Hence standard estimators for exponential distributions may be applied. Specifically, if n_1 and n_2 are the number of individuals currently in state 1 and state 2 (and there is no censoring on the left), the conditional estimators are

$$\hat{\lambda} = n_1/T_1 \quad \text{and} \quad \hat{\mu} = n_2/T_2, \tag{6.7}$$

where T_1 and T_2 are as defined for (6.5). Concretely, one could estimate the hazard for becoming unemployed merely from knowledge of the length of employment among those currently employed.

The conditional estimators are necessarily less efficient than the unconditional estimators because they discard information contributed by the proportion of observations in each state at time τ. On the other hand, they can be much more easily generalized. If there is censoring on the left, for example, conventional estimators for a censored exponential variate are appropriate. Unconditional estimators for censored data are obtainable, but they are much more complicated than the conditional estimators. More important, if we introduce covariates by letting $\lambda = \exp\{\beta z\}$ and $\mu = \exp\{\gamma z\}$, the conditional estimators are equivalent to the standard estimators for regression analysis of an exponential variate, which we discussed in Section 2. The only thing new is that the sample is split into two groups depending on current state.

The usefulness of this approach depends, of course, on the plausibility of the simplifying assumptions, principally that (a) the hazards are constant over time and (b) the recorded state changes are far enough from the origin that the limiting distribution of states is applicable. The assumption of constant hazards, though unlikely to hold exactly in any situation, may be a reasonable approximation in many circumstances. It is not easily relaxed, however, nor is it easy to verify with the data. If one knows the origin time and state, it *is* possible to relax the assumption of the limiting distribution by using (6.1), (6.2), and similar expressions for other state combinations. Only in rare cases is this likely to be worth the effort, however. Somewhat more helpful is the fact that the data may provide some evidence for or against this assumption. Use of the limiting distribution requires that $\exp\{-(\lambda + \mu)t\}$ be negligible. Having obtained estimates of λ and μ, one can get an internal check by computing $\exp\{-(\hat{\lambda} + \hat{\mu})t\}$. Furthermore, other data sources may suggest the approximate size of λ and μ to gauge the appropriateness of this assumption.

7. FORWARD RECURRENCE TIMES

Many of the results in the previous sections also apply to forward recurrence times, that is, the length of time between some point of interruption (which is fixed or independent of the occurrence of events) and the first subsequent event. Such data are much less common than backward recurrence times, however, which is why I have focused on the latter. It *is* common to begin observing the occurrence of events at some point in an ongoing process, but usually observation is discontinued at some fixed point in time rather than at the occurrence of an event. For such data, many of the issues dealt with earlier do not arise.

For completeness, I shall briefly discuss some of the similarities and differences in the analysis of forward and backward recurrence times. Results in Sections 2–4 apply to forward recurrence times with virtually no modification. It is well known, for example, that under the constant-hazard model discussed in Section 2, both the forward and backward recurrence times are exponentially distributed with the same parameter. Similarly, for the renewal process discussed in Section 4, both the forward and backward recurrence times have the same limiting distribution, which differs from the common distribution of completed intervals. For the time-dependent model of Section 3, the construction of the likelihood for forward recurrence times is similar to that for backward recurrence times, and the same estimation procedures may be employed. Note, however, that the problems associated with the use of time-varying covariates do not arise with forward recurrence times.

For the sample of male heads of households examined in earlier sections, it was possible to obtain forward recurrence times for residential mobility. Respondents were interviewed annually beginning in 1968, and in each year they were asked if they had moved since the previous interview. Using data through 1972, I constructed the number of years between the 1968 interview and the first subsequent residence change, with moves after the 1972 interview treated as censored. I then estimated the same models as for the backward recurrence times by using the same estimation procedures and the same covariates. Although this was done for the sake of comparing the forward and backward results, many more covariates could have been included in the forward analysis, since all of the variables measured in the 1968 survey were possible candidates. Results are shown in Table 3. The similarities are striking. As in the backward analysis, the qualitative conclusions are not very sensitive to the choice of model. For the most part those conclusions are the same as in the backward analysis. Mobility rates are lower among farmers, the more educated, and those born in the North. We do not find any effects of number of siblings, father's education, and parents' economic status. The one noteworthy difference is that the forward analysis clearly indicates that blacks were more mobile than whites, whereas the backward analysis found no significant difference.

In Section 5, I observed that models in which the hazard at time t depends on the number of events prior to t are very difficult to estimate with backward recurrence times. These difficulties arise because there is a purely artifactual dependence of the distribution of the backward recurrence time on the number of prior events. No such difficulty arises with forward recurrence times, however. If data are available on the number of events prior to the survey, models incorporating dependence

Table 3. Estimates for Regression Models Based on Forward Recurrence Times

	Constant Hazard			*Time-Dependent Hazard*			*Age-Dependent Hazard*			*Duration-Dependent Hazard*		
Covariate	$\hat{\beta}$	*SE*	$\hat{\beta}/SE$	$\hat{\beta}$	*SE*	$\hat{\beta}/SE$	$\hat{\beta}$	*SE*	$\hat{\beta}/SE$	$\hat{\beta}$	*SE*	$\hat{\beta}/SE$
FARM	−.517	.166	−3.12	−.507	.166	−3.06	−.386	.167	−2.31	−.388	.123	−3.15
ED	−.195	.093	−2.10	−.190	.093	−2.06	−.260	.093	−2.80	−.150	.068	−2.20
RACE	−.262	.099	−2.64	−.255	.099	−2.57	−.219	.100	−2.19	−.195	.073	−2.68
SIBS	.002	.016	.14	.003	.016	.16	.004	.016	.25	.001	.012	.08
ECON	.004	.085	.05	.006	.085	.08	−.047	.085	− .55	−.002	.062	− .03
FAED	.091	.108	.84	.089	.108	.83	.055	.109	.50	.073	.079	.93
NORTH	−.561	.166	−3.38	−.544	.166	−3.28	−.544	.166	−3.28	−.402	.121	−3.34
NC	−.399	.155	−2.58	−.386	.154	−2.50	−.431	.155	−2.78	−.285	.112	−2.55
SOUTH	−.317	.153	−2.07	−.304	.153	1.99	−.309	.152	−2.03	−.220	.111	−1.99

NOTE: See note to Table 2.

on that number may be estimated in a straightforward fashion, either by stratifying the sample or by inclusion of the number of previous events as a covariate.

For the alternating process discussed in Section 6, estimation is somewhat less restrictive with forward recurrence times. In particular, although the conditional ML estimators are the same as those given in Section 6, their derivation does not require the assumption that the limiting distribution of states has been reached. Furthermore, contrary to the case with backward recurrence times, it is straightforward to estimate models that allow for a time-dependent hazard. In fact, estimation conditional on the state at time τ (the point of interruption) is identical to that described in Section 3.

8. DISCUSSION

I have shown that data on backward recurrence times can often by analyzed by using standard methods of survival analysis, treating time as though it ran backwards from the time of the survey to the time of the most recent event. This only makes sense if events are repeatable, however, which demands that the occurrence of events be modeled as some sort of stochastic process. An important limitation is that models in which the hazard depends on the number of previous events must be ruled out. And though certain models with duration dependence are estimable, the estimation procedure is considerably more complicated than that used in most survival analysis.

Given the frequency with which such data are collected, I believe that the methods described here are a useful addition to the methodology of survival analysis. Nevertheless, considering the limitations just mentioned, it is reasonable to ask whether such data should continue to be collected or whether more effort should be expended in getting complete event histories. Certainly more complete data are always desirable when it is practical to collect them. On the other hand, eliciting accurate data on long histories of events can be a difficult and costly process. For many kinds of events, moreover, the ability of the respondent to accurately recall the event history is questionable. In a recent study of contraceptive usage, for example, respondents were frequently inconsistent in constructing histories of their contraceptive use and failure over the previous 12 months (Furstenberg et al. 1983). Similarly, Cannell et al. (1981) found that respondents made many errors in reporting hospitalizations in the previous year. In such cases, the limitations imposed by focusing only on the most recent event may well be outweighed by the greater accuracy of recall.

For those who would collect such data, however, there is one important recommendation. Every effort should be made to determine the values of any covariates prior to the occurrence of the most recent event. This is rarely done, but the failure to do so usually means that only a small fraction of variables coded are available for inclusion as covariates.

[Received May 1983. Revised April 1984.]

REFERENCES

Breslow, N. E. (1974), "Covariance Analysis of Censored Survival Data," *Biometrics*, 30, 89–99.

Breslow, N. E., Lubin, J. H., Marek, P., and Langholtz, B. (1983), "Multiplicative Models and Cohort Analysis," *Journal of the American Statistical Association*, 78, 1–12.

Cannell, Charles F., Miller, Peter V., and Oksenberg, Lois (1981), "Research on Interviewing Techniques," in *Sociological Methodology 1981*, ed. Samuel Leinhardt, San Francisco: Jossey-Bass, pp. 389–437.

Cox, D. R. (1962), *Renewal Theory*, London: Methuen.

——— (1972a), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society*, Ser. B, 34, 187–202.

——— (1972b), "The Statistical Analysis of Dependencies in Point Processes," in *Stochastic Point Processes*, ed. P. A. W. Lewis, New York: John Wiley, pp. 55–66.

Cox, D. R., and Lewis, P. A. W. (1966), *The Statistical Analysis of Series of Events*, London: Methuen.

Crowley, J., and Storer, B. E. (1983), Comment on "A Reanalysis of the Stanford Heart Transplant Data," *Journal of the American Statistical Association*, 78, 277–281.

Elandt-Johnson, Regina C., and Johnson, Norman L. (1980), *Survival Models and Data Analysis*, New York: John Wiley.

Farewell, V. T., and Cox, D. R. (1979), "A Note on Multiple Time Scales in Life Testing," *Applied Statistics*, 28, 73–75.

Farewell, Vern T., and Prentice, Ross L. (1977), "A Study of Distributional Shape in Life Testing," *Technometrics*, 19, 69–75.

Feller, W. (1971), *An Introduction to Probability Theory and Its Applications* (Vol. 2), New York: John Wiley.

Flinn, C. J., and Heckman, J. J. (1982), "Models for the Analysis of Labor Force Dynamics," in *Advances in Econometrics*, eds. G. Rhodes and R. Basmann, New Haven, CT: JAI Press, pp. 35–95.

Furstenberg, F. F., Jr., Shea, J., Allison, P. D., Herceg-Baron, R., and Webb, D. (1983), "Contraceptive Continuation Among Adolescents Attending Family Planning Clinics," *Family Planning Perspectives*, 15, 211–217.

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.

Ginsberg, R. B. (1979), "Tests of Stochastic Models of Timing in Mobility Histories: Comparison of Information Derived From Different Observation Plans," *Environment and Planning A*, 11, 1387–1404.

Glasser, M. (1967), "Exponential Survival With Covariance," *Journal of the American Statistical Association*, 62, 561–568.

Hastings, Donald W., and Robinson, Walter W. (1975), "Open and Closed Birth Intervals for Once-Married Spouse-Present White Women," *Demography*, 12, 455–466.

Hoel, P. G., Port, S. C., and Stone, C. J. (1972), *Introduction to Stochastic Processes*, Boston: Houghton Mifflin.

Kalbfleisch, John D., and Prentice, Ross L. (1980), *The Statistical Analysis of Failure Time Data*, New York: John Wiley.

Karlin, Samuel, and Taylor, Howard M. (1975), *A First Course in Stochastic Processes* (2nd ed.), New York: Academic Press.

Lawless, J. F. (1982), *Statistical Models and Methods for Lifetime Data*, New York: John Wiley.

Leridon, H. (1969), "Some Comments on an Article by K. Srinivasan," *Population Studies*, 23, 102.

Mantel, N., and Byar, D. P. (1974), "Evaluation of Response-Time Data Involving Transient States: An Illustration Using Heart Transplant Data," *Journal of the American Statistical Association*, 69, 81–86.

Prentice, R. L., Williams, B. J., and Peterson, A. V. (1981), "On the Regression Analysis of Multivariate Failure Data," *Biometrika*, 68, 373–374.

SAS Institute Inc. (1980), *SAS Supplemental Library User's Guide, 1980 Edition*, Cary, NC: Author.

Sheps, Mindel C., Menken, Jane A., Ridley, Jeanne Clare, and Lingner, Joan W. (1970), "Truncation Effect in Closed and Open Birth Interval Data," *Journal of the American Statistical Association*, 65, 678–693.

Sørensen, Aage B. (1977), "Estimating Rates From Retrospective Questions," in *Sociological Methodology 1977*, ed. David R. Heise, San Francisco: Jossey-Bass, pp. 209–223.

Srinivasan, K. (1966), "The 'Open Birth Interval' as an Index of Fertility," *Journal of Family Welfare*, 13, 40–44.

——— (1968), "A Set of Analytical Models for the Study of Open Birth Intervals," *Demography*, 5, 34–44.

——— (1970), "Findings and Implications of a Correlation Analysis of the Closed and the Open Birth Intervals," *Demography*, 7, 401–410.

Stacy, E. W. (1962), "A Generalization of the Gamma Distribution," *Annals of Mathematical Statistics*, 33, 1187–1192.

Survey Research Center (1972), *A Panel Study of Income Dynamics: Study Design, Procedures, Available Data* (Vol. 1), Ann Arbor, MI: Institute for Social Research.

Venkatacharya, Kilambi (1972), "Some Problems in the Use of Open Birth Intervals as Indicators of Fertility Change," *Population Studies*, 26, 495–505.

Zippin, C., and Armitage, P. (1966), "Use of Concomitant Variables and Incomplete Survival Information in the Estimation of an Exponential Survival Parameters," *Biometrics*, 22, 665–672.