Treatment of Hidden Heterogeneity in Event History Analysis

Nancy B. Tuma
Anatoli I. Yashin

July 1986
CP-86-20

INTERNATIONAL INSTITUTE FOR APPLIED SYSTEMS ANALYSIS
2361 Laxenburg, Austria

Treatment of Hidden Heterogeneity in Event History Analysis

Nancy B. Tuma, Anatoli I. Yashin***

1. INTRODUCTION

In recent years there has been a tremendous increase in analyses of temporal (or over time) data on individuals to study demographic events, such as births, marriages, residential changes, and deaths. Since many phenomena studied by demographers involve discrete changes in life conditions that can (in principle) occur at any moment in time, it is not surprising that attention has been drawn especially to modeling these phenomena as finite-state, continuous-time stochastic processes and then to estimating these models from event (or life) histories, which record the date of every change of state of individuals in some time period (for example, the dates of birth of all children ever born to women in some sample).

The popularity of this form of demographic analysis has been greatly enhanced by two factors. First, event history data pertaining to a wide range of demographic phenomena are becoming widely available. One of the most notable examples of such data are the fertility histories collected from samples of women in 44 countries during the last decade [1]. But there also exist data on marital histories, migration histories, job histories, and health histories. Second, new statistical methods designed particularly for event history analysis have been developed and incorporated into various computer software packages [2, 3, 4].

To date the methods that have been developed for event history analysis have focused mainly on attempts to relate the observed event histories to measured covariates thought to explain (or at least to predict) the occurrence and timing of demographic events in a person's life. These developments are all to the good. Yet further developments are needed. In particular, most of the statistical tools that have previously been developed for event history analysis have ignored one universal problem--

*Nancy B. Tuma, Department of Sociology, Stanford University, Stanford, CA 94305, USA
**Anatoli I. Yashin, Population Program, IIASA, A-2361 Laxenburg, Austria

namely, the life histories that we observe are influenced not only by the covariates that we measure but also by many things that we do not measure. That is, individuals are not only heterogeneous in ways we observe (i.e., measured covariates), but also in ways that we do not observe (i.e., random "nuisance" factors). These unobserved factors or disturbances influence a person's life history, and we are likely to draw erroneous conclusions if we ignore them.

A related argument in the favor of treatment of the unobserved heterogeneity is as follows. There are already results from many demographic, economic, sociological, and medical studies available in compact form (i.e., as models, descriptions, and other forms of knowledge). Since the studies have different goals, the variables or processes that were the focal point in one study may be nuisance variables or processes in other research. In theory one can combine data from several studies and develop a unified approach to analyzing the combined data. However, even if a huge data bank were created (which is unrealistic), some important causal variables or processes are still unlikely to be measured.

In this paper we describe some statistical tools for event history analysis when there is unobserved heterogeneity of particular types as well as measured covariates. For expository purposes we also discuss several potential applications of these statistical methods. Actual application of these methods to demographic data remains as a goal for future work.

2. PRELIMINARIES

Let ξ_t, $t \geq 0$ denote a finite-state stochastic process that makes a finite number of jumps for any $0 \leq t < \infty$. This process can also be represented in terms of a sequence of random times at which jumps occur, T_n, and random variables, Y_n, where $Y_n = \xi_{T_n}$. (We assume that the trajectories of ξ_t are right-continuous.) This process may be considered as a particular form of a multivariate point process as described in [5]. The realization of the process ξ_t for the i-th individual on time interval $[0,t]$ is just the event history for individual i from time 0 to time t. We will denote it by $\xi_{0t}^{(i)}$.

The simplest case of such a process is the discrete time Markov chain. If p_{jk}, $j,k = 1,\psi$ are the unknown transition probabilities of change from state j to state k, then the maximum likelihood ratio approach leads to the following formula for the estimator of $\hat{p}_{jk}(t)$ using information about transitions (time is discrete)

$$\hat{p}_{jk}(t) = \frac{n_{jk}(t)}{n_j(t)}$$

where $n_{jk}(t)$ is the number of an individual's transitions from j to k on the set of times $\{0,1,\ldots,t\}$ and $n_j(t)$ is the number of times when the individual occupies the state j on the set $\{0,1,\ldots,t\}$.

If the available information about population movement consists of the transition records for I independent identical individuals, then the estimator $\hat{p}_{jk}(t)$ has the form

$$p_{jk}(t) = \frac{\sum_{i=1}^{I} n_{jk}^{(i)}(t)}{\sum_{i=1}^{I} n_j^{(i)}(t)}$$

where the index i is related to the i-th particular individual. Notice that

$$\sum_{i=1}^{I} n_j^{(i)}(t) = \sum_{u=0}^{t} s_j(u)$$

where $s_j(u)$ is the number of the individuals who occupy state j at time u.

More realistic than a discrete-time model is a continuous-time model, represented by a finite-state continuous-time Markov process. Denoting the unknown constant transition intensities by r_{jk}, $j,k = 1,\ldots,\psi$, one can easily obtain the maximum likelihood estimators $\hat{r}_{jk}(t)$ when sample paths of individuals are observed:

$$r_{jk}(t) = \frac{n_{jk}(t)}{\tau_j(t)} \quad , \quad j = 1,2,\ldots,\psi; \quad t \geq 0$$

where $n_{jk}(t)$ is defined as above and $\tau_j(t)$ is the time spent in state j during time interval $[0,t]$.

In reality the situation is even more complicated. Transition intensities usually change over time and are subject to various impacts. Some of the influential variables are observed or measured; others are not. All of these circumstances lead to the following general statement of the problem.

3. STATEMENT OF THE PROBLEM

Assume that the random variable z, random process ξ_t, and an additional random variable $\mathbf{x}$ are given on probability space (Ω,H,P). We assume that variable $\mathbf{x}$ and random process ξ_t are observed and that random variable z is unobserved. We also assume that the joint probability distribution of ξ_t and z depends on the vector of measured variables $\mathbf{x}$. As before, ξ_t denotes a finite-state continuous-time process that

makes a finite number of jumps for any $0 \leq t < \infty$.

We let $\mathbf{x}_i$ denote the value of the observed vector for individual i. Finally, we assume that a vector of unknown parameters α, which are the same for all individuals in the population, indicate how the observed $\mathbf{x}$ influences the joint probability distribution of z and ξ_t. When z, α, and $\mathbf{x}$ are known, the evolution of the process ξ_t can be described by the transition intensities $r_{jk}(t,z,\alpha,\mathbf{x})$, $j,k = 1,\ldots,\psi$ where ψ is the size of the state space for the process ξ_t.

The goal is to estimate the unknown parameters α using data on the event histories for I individuals, $\xi_{0t}^{(i)}$, $i = 1,\ldots,I$.

4. STRATEGY

One of the most popular ways of estimating parameters in a model is the method of maximum likelihood. The functional form of the likelihood ratio for a multivariate point process is well known [5]. More traditional in sociological applications is the notion of the likelihood function. Denoting this function for individual i by $L_i(\xi_{0t})$ and omitting index i from ξ_{0t} for simplicity, we have:

$$L_i(\xi_{0t}) = \prod_{T_n \leq t} r_{\xi_{T_n-0}\xi_{T_n}}(T_n - 0, z, \alpha, x)\, e^{-\int_0^t r_{\xi_{T_n-0}\xi_{T_n}}(u,z,\alpha,x)du}$$

where T_n, $n = 1,2,\ldots$, denotes the times of jumps in the history ξ_{0t} (i.e., the times when events occurred), and $T_n - 0$ denotes an instant before the nth jump. This form of the likelihood cannot be used directly because it depends on the unobserved variables z, as well as on the measured variables $\mathbf{x}$. It is necessary, therefore, to find a way of representing the probabilistic characteristics of the process ξ_t that does not depend on z. The following theorem, which can be proved using the results for predictable compensators in martingale theory [6], implies that such a form exists.

Theorem. *The process ξ_t may be represented in terms of the initial distribution $p_j(0) = p\{\xi_0 = j\}$, $j = 1,\ldots,\psi$, and transition intensities $r_{jk}(t,\alpha,\mathbf{x})$, $j,k = 1,\ldots,\psi$, that have the form*

$$\bar{r}_{jk}(t,\alpha,\mathbf{x}) = E[r_{jk}(t,z,\alpha,\mathbf{x}) \mid \xi_{0t},\alpha,\mathbf{x}] \qquad (1)$$

where E denotes the operator of mathematical expectation with respect to z and ξ_{0t} denotes the history of the process from time 0.

Equation (1) means that the functional form of $\bar{r}_{jk}(t,\alpha,\mathbf{x})$ depends on the initial distribution of the unobserved random variable z and the functional form of $r_{jk}(t,z,\alpha,\mathbf{x})$. Below we consider several special cases.

Special Case I. z is discrete and time invariant for every individual i.

It is sometimes reasonable to assume that the random variable z has a finite number of values M, $\{z_m\}$, $m = 1,\ldots,M$, with known *a priori* probabilities in the population, $q_1,q_2,\ldots,q_m$. In this instance equation (1) simplifies to

$$r_{jk}(t,\alpha,\mathbf{x}) = \sum_{m=1}^{M} r_{jk}(t,z_m,\alpha,\mathbf{x})\pi_m(t,\alpha,\mathbf{x}) \tag{2}$$

where $\pi_m(t,\alpha,\mathbf{x})$ is the conditional probability that $z = z_m$ given ξ_{0t}, $\mathbf{x}$ and satisfies the system of nonlinear stochastic equations

$$\pi_m(t,\alpha,\mathbf{x}) = q_m - \int_0^t [r_{\xi_u\xi_u}(u,z_m,\alpha,\mathbf{x}) - \bar{r}_{\xi_u\xi_u}(u,\alpha,\mathbf{x})]\pi_m(u,\alpha,\mathbf{x})du$$

$$+ \sum_{T_n \le t} \pi_m(T_n - 0,\alpha,\mathbf{x})\left[\frac{r_{\xi_{T_n}-0\xi_{T_n}}(T_n - 0,z_m,\alpha,\mathbf{x})}{\bar{r}_{\xi_{T_n}-0\xi_{T_n}}(T_n - 0,\alpha,\mathbf{x})} - 1\right] . \tag{3}$$

Although this system of nonlinear equations can be solved analytically, in general the solution will appear very complicated.

To clarify this approach, we apply it to a concrete problem—a youth's entry into the labor force for the first time. Since we concentrate on the *first* event in a person's work history (i.e., the first job) and do not consider what kind of job the youth obtains, our application is a particularly simple case.

We assume that there are measurements on many personal attributes of a youth related to the speed with which he or she enters the labor force, for example, gender, parents' educational levels and income, ethnicity, and grades in school. Moreover, we assume that prior research and theory gives us confidence that the relationship between the rate of entering the first job and these attributes $\mathbf{x}$ is as follows:

$$h(t,z,\alpha,\mathbf{x}) = z\,\mu(t,\alpha,\mathbf{x}) \tag{4}$$

where $\mu(t,\alpha,\mathbf{x})$ has a known form, but z is unobserved for every individual. In this hypothetical application, z might describe, for example, a youth's relative opportunities to work in a particular geographical place, which depends on the place's industrial

structure, unemployment rate and the extent of opportunities for educational advancement. Although these place-specific variables certainly affect the rate at which youths enter jobs, in many studies these variables are not measured. We also may not even know where youths in the sample live.[1] However, the data analyst often knows that respondents were selected from M different geographical regions in proportions $q_1,\ldots,q_M$, with $\sum_{m=1}^{M} q_m = 1$. One might assume that the many unobserved characteristics of region m affecting a youth's rate of finding a first job raise or lower $\mu(t,\alpha,\mathbf{x})$ by some unknown multiplicative factor z_m.

For this example equation (2) becomes

$$\bar{h}(t,\alpha,\mathbf{x}) = \sum_{m=1}^{M} h(t,z_m,\alpha,\mathbf{x})\pi_m(t,\alpha,\mathbf{x}) \tag{5}$$

$$= \mu(t,\alpha,\mathbf{x}) \sum_{m=1}^{M} z_m \pi_m(t,\alpha,\mathbf{x})$$

where for convenience we also assume that

$$\sum_{m=1}^{M} z_m q_m = 1 \quad . \tag{6}$$

This last assumption actually just normalizes the z's. We can also write equation (3) for this special case. It is

$$\pi_m(t,\alpha,\mathbf{x}) = q_m - \int_0^t [h(u,z_m,\alpha,\mathbf{x}) - \bar{h}(u,\alpha,\mathbf{x})]\,\pi_m(u,\alpha,\mathbf{x})du$$

$$= q_m - \int_0^t \mu(u,\alpha,\mathbf{x})\left[z_m - \sum_{l=1}^{M} z_l \pi_l(t,\alpha,\mathbf{x})\right]\pi_m(u,\alpha,\mathbf{x})du \quad . \tag{7}$$

Since $\sum_{m=1}^{M} \pi_m(t,\alpha,\mathbf{x}) = 1$, this system of equations can be solved explicitly. The result (see [7]) is

$$\pi_m(t,\alpha,\mathbf{x}) = \frac{q_m \exp[-z_m H(t,\alpha,\mathbf{x})]}{\sum_{l=1}^{M} q_l \exp[-z_l H(t,\alpha,\mathbf{x})]} \tag{8}$$

---

[1] In the U.S., for example, detailed information on place of residence is often withheld to protect the identity of respondents to a survey.

where $H(t,\alpha,\mathbf{x}) = \int_0^t \mu(u,\alpha,\mathbf{x})du$. Notice that equation (8) is just a generalization of the usual logistic equation. Together with formula (5) for $\bar{h}(t,\alpha,\mathbf{x})$, this equation lets one write expressions for $\bar{S}(t,\alpha,\mathbf{x}) = Pr[T \geq t \mid \alpha,\mathbf{x}]$ and $\bar{f}(t,\alpha,\mathbf{x}) = \frac{\partial}{\partial t} Pr(T \leq t \mid \alpha,x)$ that can be used to write a likelihood function

$$\bar{L} = \prod_{i=1}^{I} \bar{h}(T_i,\alpha,x_i)\, e^{-\int_0^{T_i} \bar{h}(u,\alpha,x_i)du} \quad .$$

where the bar over L denotes the likelihood for the sample of I individuals with the transition rates given by function $\bar{h}(u,\alpha,x)$. This likelihood function can then be maximized with respect to α.

Assume that $h(t,z,\alpha,x)$ can be represented in a multiplicative form

$$h(t,z,\alpha,x) = \lambda(t)e^{\alpha_0 z + \alpha_1 x}$$

where α_0 and α_1 are unknown constants. This implies

$$\bar{h}(t,\alpha,x) = \lambda(t)\gamma(t,\alpha_0,\alpha_1)e^{\alpha_1 x} \quad .$$

where

$$\gamma(t,\alpha_0,\alpha_1) = \sum_{m=1}^{M} e^{\alpha_0 z_m}\pi_m(t,\alpha_0,\alpha_1,x)$$

One can see that the presence of unobservables in a traditional Cox regression scheme [8] creates a dependence of the underlying hazard rate [which equals $\lambda(t)\gamma(t,\alpha_0,\alpha_1)$] on unknown parameters.

Special Case II. z is discrete but can jump from time to time.

Sometimes z has a finite number of values M, $\{z_m\}$, $m = 1,\ldots,M$, with known probabilities in the population at time 0 (as in Case I), but an individual's value of z may jump from one value to another according to a finite-state jump process described by the transition intensities

$$\lambda_{lm}(t,\beta,\mathbf{x}), \quad l,m = 1,\ldots,M \quad . \tag{9}$$

In this case equation (2) still holds; however, equation (3) does not. Instead, $\pi_m(t,\alpha,\mathbf{x})$ are the solutions of the more complex system of nonlinear equations:

$$\pi_m(t,\alpha,\mathbf{x}) = q_m + \int_0^t \sum_{l=1}^{M} \lambda_{lm}(u,\beta,\mathbf{x})\pi_l(u,\alpha,\mathbf{x})du$$

$$- \int_0^t [r_{\xi_u \xi_u}(u, z_m, \alpha, \mathbf{x}) - r_{\xi_u \xi_u}(u, \alpha, \mathbf{x})] \pi_m(u, \alpha, \mathbf{x}) du$$

$$+ \sum_{T_n \le t} \pi_m(T_n - 0, \alpha, \mathbf{x}) \left[\frac{r_{\xi_{T_n - 0} \xi_{T_n}}(T_n - 0, z_m, \alpha, \mathbf{x})}{r_{\xi_{T_n - 0} \xi_{T_n}}(T_n - 0, \alpha, \mathbf{x})} - 1 \right] \tag{10}$$

for $m = 1,\ldots,M$. Unfortunately the exact analytical solution of this system of equations is unknown. However, in principle, knowledge of the form of the equations permits them to be solved numerically.

A slight generalization of our earlier example illustrates this case. Suppose that we again are studying entry into the labor force by youths and observe personal attributes $\mathbf{x}$ of each youth in a sample but do not know in which region a youth is living. Before we assumed that a youth's region cannot change, which is a reasonable approximation if the length of our observation period is short. However, youths are often the most geographically mobile segment of a society. If the observation period is not short, it would be reasonable to assume that youths migrate from one region to another. In this situation z_m is not fixed for a given youth but can change in discrete jumps. This situation provides an example of Special Case II.

It is both customary and usually fairly plausible to assume that the histories of individuals in a sample are statistically independent. Denoting an individual's history by $L_i(\xi_{0t}^{(i)})$ and taking into account the result of the theorem given in (1), we have

$$L_i(\xi_{0t}^{(i)}) = \prod_{T_n^i \le t} \bar{r}_{\xi_{T_n^i - 0}^{(i)} \xi_{T_n^i}^{(i)}}(T_n^i - 0, \alpha, \mathbf{x})\, e^{-\int_0^t \bar{r}_{\xi_u^{(i)} \xi_u^{(i)}}(u, \alpha, \mathbf{x}) du}$$

where T_n^i are the jump times of the histories $\xi_{0t}^{(i)}$. Then, the likelihood for a sample of I individuals has the form

$$\bar{L} = \prod_{i=1}^{I} L_i(\xi_{0t}^{(i)}, \alpha, \mathbf{x}_i) \quad .$$

To maximize this function, the functional form of $r_{jk}(t, i, \alpha, \mathbf{x})$ should be specified. The presence of $\pi_m(t)$, $m = 1,\ldots M$, given by (10) in the formula for $r_{jk}(t, \alpha, \mathbf{x})$ predetermines to some extent the functional form of $r_{jk}(t, \alpha, \mathbf{x})$. Note that the hazard $\bar{\mu}(t, \alpha, \mathbf{x})$ does not factor into a product of time-dependent and covariate-dependent parts. Moreover, the unknown parameters have become inextricably intertwined with the dynamics of the proportions $\pi_m(t)$, $m = 1,\ldots,M$. This means that the traditional Cox model [8] is not applicable. Maximization of the likelihood must occur under con-

straints (10), which need to be specified for every individual in the sample.

5. CONCLUSION

Note that this approach can be developed also for the case with discontinuous cumulative transition rates. In addition, it is sometimes more realistic to assume that there is an observed random process X_t instead of a random vector of variables $\mathbf{x}$. In this case the process X_t is also a part of the individual's history. Its trajectories can be continuous or piecewise continuous [9].

References

1. J. Berent, E.F. Jones, and M.K. Siddiqui, "Basic Characteristics, Sample Designs and Questionnaires," *Comparative Studies: ECE Analyses of WFS Surveys in Europe and USA*(18 (June)) (1982).
2. N.B. Tuma, *Invoking RATE*, SRI International, Menlo Park, California (1979).
3. R.J. Baker and J.A. Nelder, *GLIM System, Release 3*, Oxford Numerical Algorithms Group (1978).
4. R.B. Avery and V.J. Hotz, *HotzTran User's Manual*, Unpublished manuscript, 1985.
5. J. Jacod, "Multivariate Point Processes: Predictable Projection, Radon-Nicodim Derivatives, Representation of Martingales," *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Gebiete* **31**, pp.235-253 (1975).
6. J. Jacod, *Calcules Stochastique et Probleme de Martingales. Lecture Notes in Mathematics, Vol. 714*, Springer-Verlag, Heidelberg (1979).
7. A. Yashin, *Unknown Hazards: How to Handle Them Best. WP-84-XX*, International Institute for Applied Systems Analysis, Laxenburg, Austria (Forthcoming).
8. D.R. Cox, "Regression Models and Life Tables," *Journal of Royal Statistical Society* **B 34**, pp.187-220 (1972).
9. A.I. Yashin, K.G. Manton, and J.W. Vaupel, "Mortality and Aging in a Heterogeneous Population: A Stochastic Process Model with Observed and Unobserved Variables," *Theoretical Population Biology* **27**(2), pp.154-175 (1985).