ESTIMATION OF THE DURATION MODEL BY NONPARAMETRIC MAXIMUM LIKELIHOOD, MAXIMUM PENALIZED LIKELIHOOD, AND PROBABILITY SIMULATORS

Keun Huh and Robin C. Sickles*

Abstract—Failure to properly treat heterogeneity components in longitudinal analyses can result in an incorrect parameterization of the duration model. Estimation bias is not limited to duration dependence but also extends to the structural parameters. Our paper uses Monte Carlo methods to examine the finite sample behavior of three estimators for this problem: nonparametric maximum likelihood, maximum penalized likelihood, and the probability simulator. Our results on the estimators' finite sample behavior for this class of model add to limited experimental evidence. They highlight the estimators' computational feasibility and point to their relative strengths in empirical duration modeling.

I. Introduction

THE presence of unobservable individual effects in models of duration (survivor) hazards is problematic when the underlying hazard exhibits duration dependence (Lancaster and Nickell, 1980; Lancaster, 1979; Neyman and Scott, 1948). Theoretical treatments (Simar, 1976; Laird, 1978; Lindsay, 1983a, b; Heckman and Singer, 1984; Manton, Stallard, and Vaupel, 1986) have provided ways to control for unobserved heterogeneity using a mixture probability density. Two estimation approaches have distinguished themselves in the literature: the nonparametric approach of Heckman and Singer (1984) and the sufficient statistic method of Andersen (1970).

Heckman and Singer adopted a nonparametric method to identify an unobserved distribution from a mixed distribution assuming random effects, while Andersen used sufficient statistics to avoid the incidental parameters problem in assuming fixed effects. Andersen's application of conditional likelihood, which estimates structural parameters conditional on sufficient statistics for unobserved fixed effects, has had limited appeal due to the difficulty in finding the sufficient statistics for particular applications.

In this paper, we consider two additional estimators for the survivor model with heterogeneity: maximum penalized likelihood estimation and the simulated frequency method. Maximum penalized likelihood provides estimates of the mixed joint density while smoothing the influences of unobserved heterogeneity.[1] The probability simulator maximizes the log likelihood of simulated relative frequencies of duration times based on axioms that describe the data generating process. Discontinuities in the objective function are smoothed using kernel based procedures. The main objective of the paper is to examine the finite sample performance of these estimators using Monte Carlo analysis.

In the next section we outline the mass point method and its extension to the finite mixture model considered by Heckman and Singer. Section III presents the maximum penalized likelihood estimator and outlines algorithms that can be used to implement it in duration modeling.

[1] For the use of the maximum penalized likelihood estimator as well as the Heckman and Singer and maximum likelihood alternatives in studies of mortality and its economic determinants, see Behrman, Sickles, and Taubman (1988, 1990) and Behrman, Sickles, Taubman, and Yazbeck (1991). These studies extended the work of Sickles and Taubman (1986) in which individual heterogeneity was introduced into a joint model of leisure and morbidity and of Sickles (1989) which reviewed various treatments for heterogeneity in simultaneous limited dependent variable systems.

[683]

Section IV outlines how the probability simulator can be set up to handle mixture densities based on axioms that govern the behavior of the data generating process. This estimator was introduced by Lerman and Manski (1981) and first discussed in the survivor model context by Thompson, Atkinson, and Brown (1987). It has been recently analyzed in depth by Gourieroux and Monfort (1992), McFadden (1989) and Pakes and Pollard (1989). The nonparametric maximum likelihood and maximum penalized likelihood methods yield equivalent consistent estimators in large samples, while the simulated probability estimator, under certain assumptions about the number of simulated draws, is a multinomial maximum likelihood estimator when the underlying stochastic process has been correctly specified. However, finite sample evidence on these three estimators' comparative performance is quite limited for this class of model, due in part to the substantial nonlinearities intrinsic to the model and to the treatments for unobserved heterogeneity. We address these issues in section V, in which we present the data generation design and results from a series of Monte Carlo experiments that assess the relative performance of the nonparametric maximum likelihood estimator (NPMLE), the maximum penalized likelihood estimator (MPLE), and the simulated frequency method (SFM) for the duration model with heterogeneity. We use the Weibull as our experimental baseline hazard, in part because of its ubiquitous use in applied work and also because it is the only distribution that can generate both a proportional hazard and an accelerated time to failure model. Section VI concludes.[2]

II. Treatments of Unobserved Heterogeneity Based on Mass Point Methods

We begin with some preliminaries on the specification of the marginal likelihood for single-spell models when the density of duration times is mixed with unobserved heterogeneity. Let t_i be the absolutely continuous time of a completed spell, $\underline{X}_i$ be an m-vector of strictly exogenous and (possibly) time varying covariates, and let unobserved scalar heterogeneity be θ_i. Censored observations are given by $T_i = \min(t_i, t_c)$ and by $d_i = I(t_i < t_c)$, where t_c is the censored time of an incomplete spell and I is an indicator function: $d_i = 1$ if $t_i < t_c$ and $d_i = 0$ otherwise. The joint density of the mixed distribution is

$$f(t_i, \theta_i | \underline{X}_i) = g(t_i | \underline{X}_i, \theta_i) \mu(\theta_i). \tag{1}$$

Assuming independence over n individual spells, the marginal likelihood of duration times $f(t_i | \underline{X}_i)$ is

$$L = \prod_{i=1}^{n} \int g(t_i | \underline{X}_i, \theta_i) \, d\mu(\theta). \tag{2}$$

The likelihood function (2) is a typical form of the statistical mixture model. The problem of how to control for the unobserved mixing distribution $\mu(\theta)$ has been addressed by a number of authors (Lancaster, 1979; Lancaster and Nickell, 1980; Heckman and Singer, 1982, 1984). Standard approaches to the estimation of the parameters of (2) require the specification of a distribution on θ. However, if the density function $\mu(\theta)$ is specified, then estimation bias due to an incorrect parameterization of $\mu(\theta)$ is not limited to duration dependence effects, but extends to the structural parameters of included variables as well. Moreover, Heckman and Singer (1984) show that the problem of overparameterization can lead to the observational equivalence of two different sets of distributions.

The class of nonparametric estimators which can avoid the ad hoc specification of the mixing distribution $\mu(\theta)$ in (2) is the nonparametric maximum likelihood estimator (Robbins, 1964; Laird, 1978; Lindsay, 1983a, b; Heckman and Singer, 1982, 1984).

Nonparametric characterization of the marginal density $f(t_i | \underline{X}_i)$ takes the form

$$f(t_i | \underline{X}_i) = \sum_{j=1}^{k} g(t_i | \underline{X}_i, \theta_j) p_j, \tag{3}$$

where $\sum p_j = 1$, $p_j \geq 0$, $j = 1, \ldots, k$, k is the number of points of supports, p_j is probability mass point and θ_j, $j = 1, \ldots, k$, is a locator of p_j such that $p_j = \text{prob}(\theta = \theta_j)$. Under random sam-

---

[2] The focus of our research is on estimators of compound processes. However, a number of authors have pointed out that focusing attention on the mixing heterogeneity distribution at the expense of a richly parameterized baseline duration distribution may have serious specification error consequences. (Behrman, Sickles and Taubman, 1989; Han and Hausman, 1990; Kiefer, 1988; Newman and McCulloch, 1984; Ridder, 1986). The trade-offs between these sources of possible misspecification is a fertile research topic not addressed in our paper.

pling, the log likelihood for duration times is

$$\ln L = \sum_{i=1}^{N} \ln \sum_{j=1}^{k} g(t_i | \underline{X}_i, \theta_j) p_j. \qquad (4)$$

Lindsay (1983a) provides necessary and sufficient conditions for the existence and uniqueness of solutions to the maximization of (4) based on a geometric interpretation of the likelihood set. Some care must be given in estimating the parameters of the marginal density, e.g., in setting the range of θ over which $g(t_i | \underline{X}_i, \theta_i)$ should have its support. For the Weibull conditional hazard model used by Heckman and Singer (1984) and on which we base our experimental design, this range is established as follows. The conditional hazard is $h(t_i | \underline{X}_i, \theta_i) = \exp(\gamma \ln t_i) \exp(\underline{X}_i \beta + \theta_i)$. Let $t_i^* = \int_0^{t_i} s^\gamma \exp(\underline{X}_i \beta) \, ds$ and $\exp(\theta_i) = v_i^*$. The conditional survival function in terms of t_i^* is

$$S(t_i^* | \underline{X}_i, \theta_i) = \exp(-t_i^* v_i^*), \qquad (5)$$

while the conditional density function of t_i^* is

$$f(t_i^* | \underline{X}_i, \theta_i) = \begin{bmatrix} v_i^* \exp(-t_i^* v_i^*), & \text{if } d_i = 1 \\ \exp(-t_i^* v_i^*), & \text{if } d_i = 0. \end{bmatrix} \qquad (6)$$

This implies that $v_i^* = 1$ if the ith observation is censored in which case $\theta_i = 0$ and $\theta_{\min} = 0$. On the other hand, for an uncensored observation, the maximum occurs at $dg(t_i^* | \underline{X}_i, \theta_i)/dv_i^* = 0$ and

$$\exp(-t_i^* v_i^*) - v_i^* t_i^* \exp(-t_i^* v_i^*) = 0. \qquad (7)$$

By solving (7) we have $v_i^* = 1/t_i^*$ where t_i^* is positive and bounded. Thus,

$$\theta_i = -\ln\left[1 \Big/ \left(\int_0^t s^\gamma \exp(\underline{X}_i \underline{\beta}) \, ds \right) \right], \qquad (8)$$

where $\theta_i > 0$ if $\int_0^t s^\gamma \exp(\underline{X}_i \beta) \, ds > 1$, $\theta_i \leq 0$ if $\int_0^t s^\gamma \exp(\underline{X}_i \beta) \, ds \leq 1$. θ_i is thus the Cox-Snell residual. Choosing the largest and smallest value of v_i^* from the uncensored observations gives $\theta_{\max}$ and $\theta_{\min}$. As the data become more censored identification of the tail distribution becomes more problematic since a relatively small number of mass points around clustered observations may not be able to trace out a possibly long-tailed heterogeneity distribution.

Having established an interval for θ, the EM algorithm (Dempster et al., 1977) provides a nat-

ural method for solving the likelihood equations. Application to the heterogeneity model is achieved by treating the sequence of unobservables $\{\theta_i\}$ as missing data.[3] The Heckman-Singer estimator is consistent for mixing distributions characterized by a finite number of points of support. As a practical matter the number of these must be small enough for their identification to be empirically feasible.

III. Maximum Penalized Likelihood Estimation (MPLE)

Maximum penalized likelihood estimation (MPLE) was introduced by Good and Gaskins (1971) and developed by de Montricher, Tapia and Thompson (1975), and Silverman (1982). They consider the piecewise smooth estimation of an unknown density function by adding a term to the likelihood which penalizes unsmooth estimates. The general form of a penalized log likelihood under random sampling is

$$\log L = \sum_{i=1}^{n} \log f(x_i) - \alpha R\{f(x)\}, \qquad (9)$$

where $f(x)$ is an unknown density, $R\{f(x)\} < \infty$, R is a functional, and α is the smoothing parameter. The choice of α controls the balance between smoothness and goodness-of-fit, while the choice of penalty functional R determines the type of behavior in the estimated density considered undesirable.[4] In this section we consider the application of maximum penalized likelihood estimation to the hazard function with a mixing het-

---

[3] Heckman and Singer suggested the steps to find a global maximum. For the first step, start to maximize with one point of support $(k = 1)$ with initial value for $\delta^{(m)}$. Let $\delta^{(m+1)}$ denote the estimated parameter. Divide the interval $[\theta_{\min}^{(1)}, \theta_{\max}^{(1)}]$ into a representative mass of points of support $(k = 1)$ and find the points which have the Gateaux derivative of the log likelihood function, $D(\theta, \mu_1) > 0$, for all $\theta \in \Theta$. If there is no point showing the positive Gateaux derivative, a global solution has been found. If $D(\theta, \mu_1) > 0$, add more points of support and divide the interval. Proceed to the subsequent step until there is termination by the criterion $D(\theta, \mu) \leq 0$, for all $\theta \in \Theta$.

[4] For example, if R is defined as the norm of the first derivative, then a penalty functional R will smooth the slope of the density $f(x)$ which is semi-discontinuous. If R uses the norm of the second derivative, the curvature will be smoothed as well. Therefore, this smooth estimator is an application of the spline function.

erogeneity distribution defined in (1).[5] The penalty functional in this problem has both a classical and a Bayesian motivation.

The corresponding optimization problem using the log penalized likelihood function and a particular choice of the penalty function R, suggested by de Montricher, et al. (1975)[6] is

$$\log L = \sum_{i=1}^{n} \log f(t_i, \theta_i | \underline{X}_i) - \alpha \| f^{(2)} \|^2, \quad (10)$$

where the penalty function is the square of the norm of the second derivative with respect to the observed variables and can vary according to differing assumptions about the correlation structure between unobserved heterogeneity and the observed variables. The norm of the second derivative weighted by the smoothing parameter(s) determines the roughness of the densities while estimating the best fitting structural relationship by maximum likelihood. The density of unobserved heterogeneity $\mu(\theta_i)$ can be approximated by the penalty term nonparametrically, e.g., by splines, if roughness, or lack of smoothness in terms of the metric of the penalty term R, is due to unobserved heterogeneity. Since the smoothing parameter(s) is a hyper-prior which sets the degree of acceptable roughness and thus the weight of the penalty term in (10), it can be interpreted as forcing a prior density on heterogeneity, giving the maximum penalized likelihood estimator a Bayesian interpretation.[7] As a practi-

cal matter, the effects of heterogeneity can either be ignored, in which case the penalty term serves the purpose of smoothing out the misspecified density of duration times, or if a mass point method is used to integrate out the density, the penalty term can be viewed as a way to smooth out bumps in the finite support discrete histogram used to characterize the marginal density of heterogeneity. In the Monte Carlo experiments that follow in section V, we utilize this second option by combining the Heckman-Singer mass point estimator for the first term on the right-hand-side of (10) with the penalty function intended to smooth out bumps in the heterogeneity distribution. These bumps may be induced by the discrete nature of the histogram used to characterize the distribution of unobservables and may become more problematic as censoring rates increase. As mentioned at the end of section II, a large percentage of clustered observations makes it difficult for the finite support histogram to identify the heterogeneity distribution in finite samples.

Proofs of the existence of the MPLE when there is unobserved individual specific and time-specific heterogeneity are available from the authors on request. Although there has been very little work on the asymptotic properties of MPLE, it is clear that for our Case 1 problem considered in the Monte Carlo experiments below, MPLE is consistent as $\alpha / \sqrt{n} \to 0$ for bounded α, if the mixing distribution can be characterized by a finite number of points of support. The reason is that the NPMLE and the MPLE converge to the same function for large N since the penalty term becomes negligible as estimates of unobserved heterogeneity become smoother.[8]

IV. The Probability Simulator

Monte Carlo approaches to probability calculations are well known in the area of computer

---

[5] The problems we pursue here are quite different from previous applications which were limited to univariate modeling under the assumption that the functional form for the density is unknown. For example, Bartoszynski et al. (1981) applied this method to estimate Cox's (1982) proportional hazard function. However, their analysis concerned a smoothed pointwise estimate of an unknown hazard distribution which is characterized as a dirac delta function. Other writers who were concerned with the parametric curve-fitting problem discussed the possibility that the method could smooth out the random error component in the linear least squares regression model (Kimeldorf and Wahba, 1970a, b; Anselone and Laurent, 1968).

[6] Good and Gaskins (1971), the first authors who applied this method, based the penalty function on the first derivative $R(f) = \int (f')^{1/2}$. Since then, a number of alternative penalty functions have been introduced by de Montricher et al. (1975) and Silverman (1982).

[7] A referee has pointed out that the Bayesian interpretation of the MPLE has much appeal in terms of both the alternative perspective it brings to the modeling of unobserved heterogeneity and in terms of its potential to handle mixing distributions which are not independent of the covariates. While the importance of this estimator as an alternative to e.g., Chamberlain's estimator, may be great, a study of this issue is left to future research.

[8] For our numerical experiments, we have used the discrete minimization routine (ZXCGR) in the IMSL library since a step function approximation was employed in the interval (x_{min}, x_{max}) divided by $(m - 1)$. The time interval was divided by a natural time unit for Case 2 and 3. When the subjective choice method for the hyper-prior α was introduced, we increased or decreased α until there was no significant pattern in $\{\theta_j\}$ and/or $\{\theta_i\}$ since if roughness in the function is not smoothed out, estimates of θ will exhibit wide fluctuations.

simulation and in econometrics (McFadden, 1989; Pakes and Pollard, 1989). Simulation methods have much potential and are gaining wide use in empirical studies. Early applications were based on frequency or density simulation (Lerman and Manski, 1981; Diggle and Gratton, 1984). Recent uses of simulation estimators to the multinomial probit can be found in McFadden and Ruud (1992), Hajivassiliou and Ruud (1993), and Geweke et al. (1994).

To outline the estimator and the intuition behind it for the duration model with heterogeneity, suppose that we wish to estimate duration dependence (γ) in the absence of covariates. Individual failure times are recorded by $t = (t_1 \leq t_2 \leq \cdots \leq t_N)$. Divide the time axis into k bins, the lth of which contains n_l observations. Based on a particular value of γ, say γ_0, the axiomatic data generating process is used to generate simulated failure times $s = (s_1 \leq s_2 \leq \cdots \leq s_M)$, where $M \gg N$. Let m_{kl} be the number of simulated observations which fall into the lth bin. The simulated bin probabilities $\hat{P}_{kl}(\gamma_0) = m_{kl}/M$ should approximate the probability that sample observations fall in the same bin, $P_l = n_l/N$ for values of γ_0 close to truth. The simulation mechanism used here is the cumulative distribution function $F(t)$. A random number u_i, $i = 1, \ldots, N$, is drawn from the uniform distribution and a simulated time to failure (s_i) is generated by inverting $F(t)$.

This probability simulator has been used in survival modeling by Atkinson et al. (1988) and Thompson et al. (1987). They suggest three possible criterion functions (under the equal binning scheme) which can be used to minimize distance between $\hat{P}_{kl}(\gamma_0)$ and P_l. The first two criteria to assess the deviation of the simulated probabilities from the actual probabilities are defined by maximizing multinomial likelihoods

$$\text{Max } S_1(\gamma_0) = \sum_{l=1}^{k} n_j \ln \hat{P}_{kl}(\gamma_0)$$

or

$$\text{Max } S_2(\gamma_0) = \sum_{l=1}^{k} \ln \hat{P}_{kl}(\gamma_0)$$

depending on the binning scheme. The third criteria is the "Neyman-Modified χ^2" (Cressie and Reed, 1988) which is a minimum chi-squared estimator based on Pearson's function and re-

mains unchanged when, e.g., two cells are combined into a single cell.[9] Goodness-of-fit over the k bins is defined as

$$S_3(\gamma_0) = \sum_{l=1}^{k} \frac{\left(\hat{P}_{kl}(\gamma_0) - P_l\right)^2}{P_l}, \qquad (11)$$

and is minimized when $\hat{P}_{kl}(\hat{\gamma}) = P_l$, $l = 1, \ldots, k$, or when either S_1 or S_2 are maximized when the equal binning scheme we adopt is used.

The previous discussion was based on a data generating process lacking covariates. Suppose that for individual i the Poisson parameter h is now specified as the Weibull hazard in section II. The problem is to estimate the parameters $\delta = (\beta, \gamma)$ by either maximizing S_1 or S_2 or minimizing S_3. Due to the presence of heterogeneity, it is possible that h is not monotone in $\underline{X}$ and t if $\underline{X}$ is time-varying. The nonmonotonicity prevents us from using conventional equal probability binning methods. The equal binning assumption is essential to establish the asymptotic properties in the Thompson et al. (1987) algorithm as well as being a necessary condition for minimizing (11) without defining optimization problems for the choice of bin width and the number of simulated observations for each bin. It is particularly important in dampening the variability of (11) with respect to $\hat{\gamma}$ in the neighborhood of the true γ. To address these points we need a robust variance reduction procedure for each bin probability ($\hat{P}_{kl}$, $l = 1, \ldots, k$) for incremental changes in estimates of δ. We also require a procedure that avoids empty bins, since in this case the criterion function in (11) becomes uninformative if $P_j = 0$.

Solutions to these problems can be illustrated by considering a two-dimensional duration model containing a single covariate. Let $t = \{t_i(X_i), i = 1, \ldots, N\}$ be failure time data conditional on exogenous variable X_i and let k_1 and k_2 be the number of bins dividing the time axis and the covariate axis, respectively. Again let M be the number of simulated observations and let the frequency with which simulated durations and values for the exogenous covariate fall in the $(l_1, l_2)^{\text{th}}$ bin be m_{l_1, l_2}, $l_1 = 1, \ldots, k_1$, $l_2 = 1, \ldots, k_2$. If a particular δ_0 is close to truth, then

[9] We thank an anonymous referee for pointing out the proper lineage of this criteria.

the simulated bin probability

$$\hat{P}_{l_1 l_2}(\delta_0) = \frac{m_{l_1 l_2}}{M} \qquad (12)$$

should approximate the corresponding portion of data (time and a covariate) in the same bin,

$$P_{l_1 l_2} = \frac{n_{l_1 l_2}}{N}. \qquad (13)$$

For large M and N, (12) should converge to (13) as δ_0 converges to the true δ.

A minor modification in the criterion function is necessary since the presence of empty bins will make Pearson's goodness-of-fit uninformative. To prevent this, the modified Pearson goodness of fit becomes

$$S_m(\delta) = \begin{cases} \sum_{l_1=1}^{k_1} \sum_{l_2=1}^{k_2} \frac{\left(\hat{P}_{l_1 l_2}(\delta) - P_{l_1 l_2}\right)^2}{P_{l_1 l_2}}, \\ \qquad \text{if } \hat{P}_{l_1 l_2}(\delta), P_{l_1 l_2} \neq 0 \\ \sum_{l_1=1}^{k_1} \sum_{l_2=1}^{k_2} \frac{\left(\hat{P}_{l_1 l_2}(\delta) - P_{l_1 l_2}\right)^2}{\hat{P}_{l_1 l_2}(\delta)}, \\ \qquad \text{if } \hat{P}_{l_1 l_2}(\delta) \neq 0, P_{l_1 l_2} = 0 \\ 0, \quad \text{otherwise.} \end{cases} \qquad (14)$$

The modified chi-squared criterion substitutes observed probabilities with simulated probabilities when the observed probabilities for certain bins are zero. The criterion is minimized when $P_{l_1 l_2} = \hat{P}_{l_1 l_2}(\delta)$, $l_1 = 1, \ldots, k_1$, $l_2 = 1, \ldots, k_2$.

It is well known (c.f. McFadden, 1989) that discontinuities in the simulated objective function S_1, S_2, or S_3 can lead to serious numerical breakdowns in most gradient optimization methods. We utilize a method introduced by Scott (1979, 1985, 1992) and referred to as the average-shifted histogram (ASH) method which is well-suited to smooth our binned frequencies of duration times. The histogram method minimizes the integrated mean square error (IMSE) of the bin probabilities. Minimization of the IMSE yields an optimal bin width, $b_i^* = 3.5 s_i n^{-(1/(2+d))}$, where i denotes each axis in the multidimensional (d) space, s_i is the standard deviation, n is the number of observations. The formal algorithms are outlined in Scott (1992).[10] We use the biweight (quartic) ker-

[10] We thank Professor David Scott for making these algorithms available to us.

nel to weight the bin counts. Other smoothing techniques for the simulated frequency method, maximum simulated likelihood method, and simulated method of moments are discussed in McFadden (1989), Stern (1992), Hajivassiliou and Ruud (1993), and Geweke et al. (1994).

Consistency properties of the simulated probability estimator as $M/N^{1/2}$ goes to infinity are discussed in Lee (1992) and McFadden and Ruud (1992). McFadden (1989) and Pakes and Pollard (1989) consider alternative simulation estimators that are moment-based and prove consistency when the number of simulations is finite while Brown (1994) examines a class of efficient residual-based simulation estimators. While the moment based approaches to estimation by simulation are appealing in that their limiting distributions do not depend on the number of simulations, construction of the multinomial likelihood requires that the number of simulations be large as well, since in this case the simulated probabilities converge to the sample probabilities. Part of the appeal of the exercise in constructing the simulated probabilities of duration times is to assess the practicality and performance of this particular simulated probability estimator using supercomputing technologies which allow for the large number of simulations needed to assure proper asymptotics.

V. Monte Carlo Results

A. Design of Experiments and Data Generation

Although the asymptotic distribution theory for the three estimators introduced above has been addressed in the literature, their finite sample properties have not been worked out. The bulk of empirical work which utilizes the duration model with heterogeneity has used samples of less than 1000, and it is unclear what inferential properties exist for these alternative estimators in the realistic setting in which asymptotic appeals cannot be justified. Owing to the highly nonlinear nature of the duration model with heterogeneity and of the estimators proposed to estimate such models, Monte Carlo experience is limited. Thus it would be informative to examine the finite sample behavior of estimators for this class of model. We carry out a highly capital intensive set of Monte Carlo experiments to highlight both the computational feasibility of these alternative estimators

and to indicate their strengths and shortcomings using sample sizes that are frequently encountered in empirical duration modeling. We examine the feasibility of implementing these alternatives and assess their comparative performance under different forms of heterogeneity, rates of censoring, sizes of samples, and parameterizations of covariates and duration dependence terms.

We consider the Weibull proportional hazard model discussed in section II and assume that $\underline{X}_i = (x_{1i}, x_{2i})$, $\beta = (\beta_1, \beta_2)$ and θ_i is unobserved heterogeneity. Heterogeneity is assumed to be distributed identically and independently. Artificial samples (of size n) are generated by first drawing uniform random variables u_i, $i = 1, \ldots, n$, in the interval $[0, 1]$ and then generating θ_i according to the implicit function, μ, where $\theta_i = \mu^{-1}(u_i)$ and where μ^{-1} is the inverse of the cumulative distribution function. The two exogenous variables $\underline{X}_i = (x_{1i}, x_{2i})$, $i = 1, \ldots, n$, are drawn from a standard normal distribution. Another uniform random number in the interval $[0, 1]$ is drawn for the survival function $S_i = [1 - F(\cdot)]$, $i = 1, \ldots, n$ and we then solve for the implied duration t_i, $i = 1, \ldots, n$, from the survival function with given values of parameters, $\underline{\beta}$ and γ. Thus,

$$t_i = \exp\left[\{\ln(-\ln S_i) + \ln(\gamma + 1) - (\beta_1 x_{1i} + \beta_2 x_{2i} + \theta_i)\} \frac{1}{\gamma + 1} \right]. \quad (15)$$

Different distributions for θ_i are used to compare the performances of the different estimators. The standard normal is the unimodal contamination. A mixture of univariate normals and the multinomial distribution is used for the multimodal heterogeneity. Experiments are also conducted on samples with right-censoring rates of 15% and 20%. Samples of 100, 500 and 1000 are used. These are in the range of sample sizes used for the bulk of empirical duration studies.[11]

The method employed to choose the smoothing parameter a of MPLE is the subjective choice method (Bartoszynski et al., 1981). We attempted cross-validation but it became too computationally burdensome even for data sets of size 100. Searching for a maximum requires no less than the number of function evaluations times the number of observations times the number of function evaluations with new a. In a typical case, about 14700 iterations were needed for data sets of size 100. The adaptation of cross-validation methods to our model merits further investigation.

The simulated frequency method (SFM) simulation algorithm can be adapted easily for more complicated models. We set the number of simulated failure times at $N = 10,000, 50,000, 100,000$ for sample sizes $n = 100, 500, 1000$. The number of simulated data sets (M) for constructing the bootstrap standard deviations is set at 30.

B. Comparisons among Different Estimators

Typical outcomes of our Monte Carlo experiments based on 100 replications are shown in tables 1–3.[12] These results are suggestive of some discrepancies among the different estimators in different cases but also suggest substantial comparability between them when the underlying stochastic process is not too complicated and has been correctly modeled. Moreover, they point to the relatively good performance of all three estimators when sample sizes of 1000 are available. Table 1 presents results based on the three estimators as we vary sample size and the censoring rates. Heterogeneity is drawn from a standardized normal distribution. Experiments with NPMLE were begun using 2 points of support to identify the heterogeneity distribution. Since the standard normal distribution has a mass point at 0.0, one point of support locates at -3.0 with cumulative probability 0.0 and the other point is set to locate at 0.0 with the expected cumulative probability 0.5. Adding two more points of sup-

---

[11] Computing algorithms were developed in Fortran77. In addition to the computing source codes, STEPIT of Chandler (1969) is employed as the maximization method for SFM. STEPIT is useful for the SFM procedure because when the steps oscillate it detects the fashion of zigzags and shortcuts the optimizations. The minimization routine ZXCGR in the IMSL library was used for MPLE. We also use the computer code of CTM documented by Yi et al. (1987) for Heckman

and Singer's NPMLE, which is based on the EM algorithm of Dempster, et al. (1977). Approximate run time for the SPE was 65 c.p.u. minutes on an ES9000 using 1000 observations, while the NPMLE was 15 and the MPLE was approximately 25.

[12] Additional experimental results are available from the authors on request.

TABLE 1.—COMPARISON OF ESTIMATORS WITH DIFFERENT CENSORING RATES AND STANDARD NORMAL HETEROGENEITY
(true parameters are $\gamma = \beta_1 = \beta_2 = 1.0$)

		No Censoring			15% Right-Censored			20% Right-Censored		
		NPMLE	MPLE	SPE	NPMLE	MPLE	SPE	NPMLE	MPLE	SPE
	$\hat{\gamma}$	0.749 (0.011)	0.661 (0.017)	0.756 (0.012)	0.622 (0.018)	0.632 (0.018)	0.803 (0.008)	0.621 (0.013)	0.612 (0.019)	0.770 (0.031)
$n = 100$	$\hat{\beta}_1$	0.799 (0.012)	0.781 (0.015)	1.182 (0.010)	0.689 (0.013)	0.766 (0.017)	0.819 (0.010)	0.686 (0.015)	0.762 (0.016)	0.869 (0.008)
	$\hat{\beta}_2$	0.760 (0.007)	0.721 (0.009)	0.852 (0.012)	0.667 (0.011)	0.707 (0.010)	0.867 (0.007)	0.661 (0.016)	0.699 (0.016)	0.856 (0.009)
	$\hat{\gamma}$	0.970 (0.004)	0.933 (0.005)	1.017 (0.005)	0.904 (0.007)	0.901 (0.012)	1.091 (0.008)	0.844 (0.007)	0.836 (0.013)	1.155 (0.008)
$n = 500$	$\hat{\beta}_1$	0.973 (0.007)	0.983 (0.010)	1.058 (0.010)	0.812 (0.007)	0.913 (0.011)	1.021 (0.010)	0.732 (0.007)	0.856 (0.012)	1.092 (0.004)
	$\hat{\beta}_2$	0.960 (0.009)	0.950 (0.009)	1.082 (0.011)	0.851 (0.012)	0.902 (0.015)	0.985 (0.007)	0.722 (0.013)	0.859 (0.017)	1.085 (0.006)
	$\hat{\gamma}$	1.054 (0.004)	0.936 (0.006)	0.991 (0.004)	0.916 (0.003)	0.921 (0.005)	0.989 (0.003)	0.887 (0.005)	0.865 (0.006)	0.931 (0.003)
$n = 1000$	$\hat{\beta}_1$	0.975 (0.008)	0.961 (0.008)	1.011 (0.009)	0.799 (0.004)	0.927 (0.009)	1.023 (0.005)	0.769 (0.005)	0.907 (0.008)	1.118 (0.004)
	$\hat{\beta}_2$	0.970 (0.004)	0.969 (0.005)	0.987 (0.004)	0.815 (0.005)	0.933 (0.007)	1.011 (0.007)	0.809 (0.007)	0.913 (0.008)	1.027 (0.005)

Note: Average standard errors are in parentheses below estimates averaged over the 100 replications for each experiment.

TABLE 2.—COMPARISON OF ESTIMATORS USING UNCENSORED DATA
AND NON-UNIMODAL HETEROGENEITY DISTRIBUTIONS
(true parameters are $\gamma = \beta_1 = \beta_2 = 1.0$)

	Estimate	Bimodal Heterogeneity[a]		Multimodal Heterogeneity[b]	
		NPMLE	MPLE	NPMLE	MPLE
	$\hat{\gamma}$	0.788 (0.007)	0.757 (0.009)	0.828 (0.005)	0.812 (0.007)
$n = 100$	$\hat{\beta}_1$	0.854 (0.007)	0.843 (0.011)	0.847 (0.009)	0.839 (0.012)
	$\hat{\beta}_2$	0.853 (0.006)	0.838 (0.007)	0.836 (0.008)	0.842 (0.011)
	$\hat{\gamma}$	0.953 (0.002)	0.922 (0.003)	0.985 (0.002)	0.992 (0.002)
$n = 500$	$\hat{\beta}_1$	0.929 (0.005)	0.921 (0.005)	0.985 (0.003)	0.980 (0.003)
	$\hat{\beta}_2$	0.921 (0.004)	0.921 (0.006)	0.969 (0.002)	0.964 (0.003)
	$\hat{\gamma}$	0.954 (0.003)	0.954 (0.004)	0.986 (0.002)	0.994 (0.024)
$n = 1000$	$\hat{\beta}_1$	0.975 (0.002)	0.970 (0.004)	0.988 (0.002)	0.985 (0.002)
	$\hat{\beta}_2$	0.977 (0.003)	0.975 (0.006)	0.972 (0.002)	0.971 (0.002)

Note: Average standard errors are in parentheses below estimates averaged over the 100 replications for each experiment.
[a] $d\mu(\theta) = p(2\pi\sigma_1^2)^{-1/2} \exp\{-\theta^2/2\sigma_1^2\} + (1 - p)(2\pi\sigma_2^2)^{-1/2} \exp\{-\theta^2/2\sigma_2^2\} d\theta$ with $p = 0.5$ and $\sigma_1^2 = \sigma_2^2 = 1.0$.
[b] $d\mu(\theta_i) = p_i$, $i = 1, \ldots, 7$, with $p_1 = p_3 = p_5 = p_7 = 0.1$, $p_2 = p_4 = p_6 = 0.2$.

TABLE 3.—COMPARISON OF ESTIMATORS USING UNCENSORED DATA AND ALTERNATIVE
PARAMETERIZATIONS OF THE DURATION MODEL
(standard normal heterogeneity)

Sample Size	Parameter Estimate	$\gamma = 2, \beta_1 = \beta_2 = 1$		$\gamma = 3, \beta_1 = 1, \beta_2 = 2$	
		NPMLE	MPLE	NPMLE	MPLE
	$\hat{\gamma}$	1.558 (0.014)	1.412 (0.013)	2.502 (0.015)	2.213 (0.028)
$n = 100$	$\hat{\beta}_1$	0.810 (0.010)	0.783 (0.012)	0.734 (0.011)	0.728 (0.018)
	$\hat{\beta}_2$	0.801 (0.012)	0.771 (0.013)	1.635 (0.010)	1.576 (0.021)
	$\hat{\gamma}$	1.940 (0.002)	1.872 (0.005)	2.976 (0.004)	2.774 (0.005)
$n = 500$	$\hat{\beta}_1$	0.967 (0.002)	0.955 (0.004)	0.991 (0.003)	0.973 (0.002)
	$\hat{\beta}_2$	0.975 (0.003)	0.948 (0.006)	1.928 (0.003)	1.829 (0.006)
	$\hat{\gamma}$	1.953 (0.002)	1.883 (0.003)	2.987 (0.001)	2.778 (0.004)
$n = 1000$	$\hat{\beta}_1$	0.971 (0.002)	0.962 (0.003)	0.992 (0.001)	0.982 (0.025)
	$\hat{\beta}_2$	0.976 (0.003)	0.956 (0.002)	1.935 (0.002)	1.838 (0.006)

Note: Average standard errors are in parentheses below estimates averaged over the 100 replications for each experiment.

port reduced bias in parameter estimates substantially $[(\theta, \mu(\theta)) = \{(.012, .276), (.232, .343), (1.22, .760)\}]$. We adopted the strategy of giving an additional point of support to the heterogeneity distribution until no directional directives show positive values and no improvement in the likelihood value is shown (Heckman and Singer, 1984). The simulated probability estimator (SPE) is based on stochastic axioms which are consistent with the data generating process, absent parameter values for the duration model. Samples of size of 1000, 500, and 100 require seven, six, and four bins for each dimension, respectively. For MPLE, the squared norm of the second derivative of the hazard function with respect to $\underline{X}$, $\|h(\underline{X})\|^2$, was used as the penalty function.[13] We used ten bins to calculate derivatives in the interval $[\underline{X}_{min}, \underline{X}_{max}]$.

For the selection of population parameters considered in table 1 ($\gamma = \beta_1 = \beta_2 = 1$), average point estimates display a pervasive pattern of

downward bias for both NPMLE and for MPLE, although the two estimators and their standard errors are quite comparable. This downward bias in the duration dependence term is expected when heterogeneity is ignored and the data exhibit positive duration dependence, but the source of the downward bias in the covariate effects as well is unclear. What is clear is that the bias is substantially removed when uncensored samples of 1000 are considered. Both the MPLE and the NPMLE show over a five-fold reduction in bias in moving from samples of 100 to samples of 1000 with samples of 1000 indicating a bias ranging between -6.4% (γ, MPLE) and 5.4% (γ, NPMLE). Standard errors for NPMLE, MPLE and SPE are also quite comparable. Results for the SPE are quite encouraging. Even at censoring rates of 20%, the simulated probability estimator tracks the underlying data generating process quite well, often offering a ten-fold bias reduction vis-à-vis NPMLE and MPLE. The relative performance of the simulated probability estimator when the underlying data generating process was correctly specified did not appear to suffer as we modified the parameterization and the form of heterogeneity. With censoring rates of 15%

[13] Details concerning the theoretical motivation for using the second derivative with respect to $\underline{X}$ can be found in an original appendix to this paper which no longer accompanies it. Interested readers can write the authors for the discussion.

NPMLE and MPLE continue to underestimate the true parameter values. However, the degree of underestimation for the structural parameters is greater with NPMLE than MPLE. MPLE benefits substantially from smoothing by the penalty term as censoring increases the number of observations in the tails of the distribution, making it more difficult for the discrete histogram used in NPMLE to identify and estimate the underlying mixture.

The next set of results in table 2 considers the comparative performance of NPMLE and MPLE and the robustness of the probability simulator which ignores a multimodal contaminating mixture. We consider two rather benign forms of modality. The first is one in which heterogeneity is a mixture of independent standardized normal variates and the second is one in which heterogeneity is discrete, taking on seven values with three modes (see footnote 4 in table 2). MPLE and NPMLE performed well when actual heterogeneity is not unimodal. However, NPMLE has mass points at (.274, .783, .823) for the bimodal distribution and showed all directional derivatives to be negative. For the multimodal distribution, 4 points of support appeared adequate. These results, as well as those with the unimodal distribution, are consistent with a common finding that the mass point method employed by NPMLE has difficulty reflecting the true distribution of heterogeneity and that the choice of optimal supporting points remains an empirical problem requiring further research. These results are broadly consistent with the findings of table 1 in that, for most experiments, NPMLE has the smaller bias and smaller standard errors than MPLE when no censoring occurs, regardless of the modality of the underlying heterogeneity. Moreover, the downward bias in the duration dependence term is essentially removed by NPMLE and MPLE for samples of 1000 when heterogeneity is discrete with a small number of points of support (7). This is to be expected since that is just the representation of heterogeneity used by MPLE and NPMLE.

Table 3 provides comparisons of the NPMLE and MPLE when several alternative parameterizations of the duration model are used as well as the robustness of the probability simulator under these alternative parameterizations when heterogeneity is ignored. A similar pattern of downward bias in the duration parameter and covariates is

found as well as an overall marginal preference of the NPMLE over MPLE when no censoring exists. These same experiments were repeated as many as 100 times using different seed values. The Monte Carlo results are quite stable and support the outcomes of typical experiments shown in tables 1–3.

VI. Conclusions

This paper has outlined and studied several methods well-suited to measure mixture distributions. Until the work of Heckman and Singer (1984), few in the field of econometrics had paid attention to semi-nonparametric estimation methods for the identification of mixed unobservables. We have proposed two additional estimators which also address the existence of an unobserved mixing distribution in the sample density. Maximum penalized likelihood estimation smooths out roughness while maximizing goodness of fit. We have provided proofs of the existence and uniqueness of this estimator for the duration model with general forms of unobserved heterogeneity. The simulated probability estimator is based on axioms which detail the data generating process. Our Monte Carlo results suggest that the nonparametric maximum likelihood estimator has some disadvantages in terms of finite sample behavior relative to the maximum penalized likelihood estimator when censoring exists but that the two estimators often are quite comparable. The simulated probability estimator performs quite well and provides relatively robust estimates in samples of 100 to 500 even when the data generating process is misspecified to ignore the heterogeneity present in the data. Finally, although the maximum penalized likelihood estimator performs well in most cases, the choice of smoothing parameters is an open question when alternative norms are used for the penalty function.

REFERENCES

Andersen, E. B., "Asymptotic Properties of Conditional Maximum Likelihood Estimators," *Journal of Royal Statistical Society, Series B* 32 (1970), 283–301.
Anselone, P. M., and P. J. Laurent, "A General Method for the Construction of Interpolating or Smoothing Spline Functions," *Numerische Mathematik* 12 (1968), 66–88.
Atkinson, E. Neeley, Robert Bartoszynski, Barry Brown, and James R. Thompson, "Simulation Techniques for Parameter Estimation in Tumor-related Stochastic Pro-

cesses," in *Proceedings of the 1983 Computer Simulation Conference* (New York: North-Holland, 1983), 754–757.

Atkinson, E. Neeley, Barry Brown, and James Thompson, "SIMEST and SIMDAT: Differences and Convergences," in *Proceedings of the 20th Symposium on the Interface: Computing Science and Statistics* (1988).

Bartoszynski, Robert B., Barry Brown, Raymond C. McBride, and James R. Thompson, "Some Nonparametric Techniques for Estimating the Intensity Function of a Cancer Related Nonstationary Poisson Process," *The Annals of Statistics* 9 (1981), 1050–1060.

Behrman, Jere, Robin C. Sickles, and Paul Taubman, "Age Specific Death Rates," in E. Lazear and R. Ricardo-Campbell (eds.), *Issues in Contemporary Retirement* (Stanford: Hoover Institution Press, 1988), 162–190.

———, "Survivor Functions with Covariates: Sensitivity to Sample Length, Functional Form, and Unobserved Fraility," *Demography* 27 (1990), 267–284.

Behrman, Jere, Robin C. Sickles, Paul Taubman and A. Yazbeck, "Black-White Mortality Inequalities," *Journal of Econometrics* 50 (1991), 183–204.

Brown, Bryan W., "Simulation-Based Semiparametric Estimation and Prediction in Nonlinear Systems," Rice University, mimeo (1994).

Chandler, J. P., "STEPIT," *Behavioral Science* 14 (1969), 81.

Cox, David R., "Regression Models and Life-Tables (with discussion)," *Journal of the Royal Statistical Society, Series B* 34 (1972), 187–202.

Cressie, Noel A. C., and Timothy R. C. Read, *Goodness of Fit Statistics for Discrete Multivariate Models* (New York: Springer Verlag, 1988).

Dempster, A. P., N. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Series B* 39 (1977), 1–38.

Diggle, Peter J., and Richard J. Gratton, "Monte Carlo Methods of Inference for Implicit Statistical Models," *Journal of the Royal Statistical Society, Series B* 46 (1984), 193–227.

Geweke, John, Michael Keane, and David Runkle, "Alternative Computational Approaches to Inference in the Multinomial Probit Model," this REVIEW 76 (Nov. 1994).

Good, I. J., and R. A. Gaskins, "Nonparametric Roughness Penalties for Probability Densities," *Biometrika* 58 (1971), 255–277.

Gourieroux, Christian, and Alain Monfort, "Simulation Based Inference in Models with Heterogeneity," *Journal of Econometrics* 52 (1992), 159–199.

Hajivassiliou, Vassilis A., and Paul Ruud, "Classical Estimation Methods for LDV Models Using Simulation," Yale University, mimeo (1993).

Han, Aaron, and Jerry Hausman, "Semiparametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics* 5 (1990), 1–28.

Hasselblad, Victor, "Estimation of Parameters for a Mixture of Normal Distributions," *Technometrics* 8 (1966), 431–444.

Heckman, James, and Burton Singer, "The Identification Problem in Econometric Models for Duration Data," in W. Hildenbrand (ed.), *Advances in Econometrics, Proceedings of the World Meetings of the Econometric Society, 1980* (Cambridge: Cambridge University Press, 1982), 39–77.

———, "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data," *Econometrica* 52 (1984), 271–320.

Kiefer, Nicholas M., "Economic Duration Data and Hazard Functions," *Journal of Economic Literature* 26 (1988), 646–679.

Kiefer, J. M., and J. Wolfowitz, "Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters," *Annals of Mathematical Statistics* 27 (1956), 887–906.

Kimeldorf, George S., and Grace Wahba, "A Correspondence between Bayesian Estimation on Stochastic Processes and Smoothing by Splines," *Annals of Mathematical Statistics* 41 (1970a), 495–502.

———, "Spline Functions and Stochastic Process," *Sankhya(A)* 32 (1970b), 173–180.

Laird, Nan, "Nonparametric Maximum Likelihood Estimation of a Mixing Distribution," *Journal of American Statistical Association* 73 (1978), 805–811.

Lancaster, Tony, "Econometric Methods for the Duration of Unemployment," *Econometrica* 47 (1979), 939–956.

Lancaster, Tony, and Stephen Nickell, "The Analysis of Reemployment Probabilities for the Unemployed," *Journal of the Royal Statistical Society, Series A* 143 (1980), 141–165.

Lee, Lung-Fei, "Asymptotic Bias in Maximum Simulated Likelihood Estimation of Discrete Choice Models," University of Michigan, mimeo (1992).

Lerman, Steven R., and Charles F. Manski, "On the Use of Simulated Frequencies to Approximate Choice Probabilities," in C. F. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications* (Cambridge, MA: MIT Press, 1981), 305–319.

Lindsay, B. G., "The Geometry of Mixture Likelihoods: A General Theory," *The Annals of Statistics* 11 (1983a), 86–94.

———, "The Geometry of Mixture Likelihoods Part II: The Exponential Family," *The Annals of Statistics* 11 (1983b), 783–792.

Manton, Kenneth G., Eric Stallard, and James W. Vaupel, "Alternative Models for the Heterogeneity of Mortality Risks Among the Aged," *Journal of the American Statistical Society* 81 (1986), 635–644.

McFadden, Daniel, "A Method of Simulated Moments for Estimation of Discrete Response Models without Numerical Integration," *Econometrica* 57 (1989), 995–1026.

McFadden, Daniel, and Paul Ruud, "Estimation by Simulation," this REVIEW 76 (Nov. 1994).

de Montricher, Gilbert F., Richard A. Tapia, and James R. Thompson, "Nonparametric Maximum Likelihood Estimation of Probability Densities by Penalty Function Methods," *The Annals of Statistics* 3 (1975), 1329–1348.

Neyman, J., and Charles E. McCulloch, "A Hazard Rate Approach to the Timing of Births," *Econometrica* 52 (1984), 939–961.

Newman, John, and Elizabeth L. Scott, "Consistent Estimates Based on Partially Consistent Observations," *Econometrica* 16 (1948), 1–32.

Pakes, Ariel, and David Pollard, "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57 (1989), 1027–1058.

Ridder, Geert, "The Sensitivity of Duration Models to Misspecified Unobserved Heterogeneity and Duration Dependence," University of Amsterdam, mimeo (1986).

Robbins, Herbert, "The Empirical Bayes Approach to Statistical Decision Problems," *Annals of Mathematical Statistics* 35 (1964), 1–20.

Scott, David W., "On Optimal and Data Based Histograms," *Biometrika* 66 (1979), 605–610.

———, "Average Shifted Histograms: Effective Nonparametric Density Estimators in Several Dimensions," *Annals of Statistics* 13 (1985), 1024–1040.

———, *Multivariate Density Estimation* (New York: John Wiley & Sons, 1992).

Sickles, Robin C., "An Analysis of Simultaneous Limited Dependent Variable Models and Some Nonstandard Cases," in R. Mariano (ed.), *Advances in Statistical Analysis and Statistical Computing, Theory and Applications*, Vol. 2 (Greenwich: JAI Press, Inc., 1989), 85–122.

Sickles, Robin C., and Paul Taubman, "An Analysis of the Health and Retirement Status of the Elderly," *Econometrica* 54 (1986), 1339–1356.

Silverman, Bernard W., "On the Estimation of a Probability Density Function by the Maximum Penalized Likelihood Method," *The Annals of Statistics* 10 (1982), 795–810.

Simar, Leopold, "Maximum Likelihood Estimation of a Compound Poisson Process," *The Annals of Statistics* 4 (1976), 1200–1209.

Stern, Steven, "A Method for Smoothing Simulated Moments of Discrete Probabilities in Multinomial Probit Models," *Econometrica* 60 (1992), 943–952.

Thompson, James R., *Empirical Model Building* (New York: John Wiley & Sons, 1989), 114–131.

Thompson, James R., E. Neeley Atkinson, and Barry Brown, "SIMEST: An Algorithm for Simulation Based Estimation of Parameters Characterizing a Stochastic Process," in J. R. Thompson and B. Brown (eds.), *Cancer Modeling* (New York: Marcel Dekker, 1987), 387–415.

Yi, Kei-Mu, Bo Honore, and J. Walker, *CTM: A Program for the Estimation and Testing of Continuous Time Multi-Spell Models, User's Manual, Program Version 50* (ERC/NORC and University of Chicago, 1987).