Estimation of dynamic models of recurring events with censored data

by

Tue Gorgens

and

Sanghyeok Lee

ANU Working Papers in Economics and Econometrics
655

21 November 2017

JEL: C33, C41, C51

ISBN: 0 86831 655 5

Estimation of dynamic models of recurring events with censored data

Sanghyeok Lee* Tue Gørgens†

2017-11-21

Abstract: In this paper we consider estimation of dynamic models of recurring events (event histories) in continuous time using censored data. We develop maximum simulated likelihood estimators where missing data are integrated out using Monte Carlo and importance sampling methods. We allow for random effects and integrate out the unobserved heterogeneity using a quadrature rule. In Monte Carlo experiments, we find that maximum simulated likelihood estimation is practically feasible and performs better than both listwise deletion and auxiliary modelling of initial conditions.

Keywords: Duration analysis; survival analysis; failure-time analysis; reliability analysis; event history analysis; hazard rates; data censoring; panel data; initial conditions; random effects; maximum simulated likelihood; Monte Carlo integration; importance sampling.

JEL classification codes: C33, C41, C51.

Acknowledgements: This research was supported in part by Australian Research Council Grant DP1096862.

*Address: Research School Economics, 25A Kingsley Street, The Australian National University, Acton ACT 2601, Australia. E-mail: sanghyeok.lee@anu.edu.au.

†Address: Research School Economics, 25A Kingsley Street, The Australian National University, Acton ACT 2601, Australia. E-mail: tue.gorgens@anu.edu.au.

1 Introduction

Data censoring is a pervasive problem in the analysis of the occurrence and timing of events. Often the observation process is such that some individuals are not under observation continuously during the time they are at risk, and therefore some events may be missing in the data available for analysis. For example, the observation period may begin and end at fixed calendar times and only events that occur within this window are available for analysis. The event histories are said to be left-censored or right-censored if events before the start or after the end of the observation period are missing, respectively. In some longitudinal surveys, participants provide information annually about events that have occurred in the previous year, and participants who skip an interview will have a gap in their recorded event histories.

In practice, event history models are estimated by the method of maximum likelihood (ML). Usually it is assumed that the observation process is independent of the event process (and the former is not modelled). In this case, it is straightforward to include right-censored event histories, and gaps can be handled by artificially right-censoring the histories at the start of the gap. If there are not too many gaps, the data loss may be acceptable. However, left-censoring remains a difficult problem in most applications. Since consistent estimates can be obtained from the non-left-censored histories, a common solution is simply to drop all left-censored histories from the analysis. For example, Doiron and Gørgens (2008) and Cockx and Picchio (2012, 2013) studied transitions between labour force states and avoided the left-censoring issue by focusing on young people who first entered the labour force during the observation period (so their initial labour market outcomes are observed). Similarly, Bhuller, Brinch, and Königs (2016) studied dynamic aspects of the receipt of welfare benefits, and selected a sample of individuals who turned 18 and thus became eligible for the first time during the study period. Dropping left-censored histories from the analysis comes at the cost of a smaller sample size. For example, by restricting their sample to school leavers Doiron and Gørgens (2008) used only one third of the total sample.

The problem of left-censoring in event history analysis is related to the well-known problem of initial conditions in discrete-time dynamic panel data models of binary responses or other limited dependent variables. In these models, the "structural" equation involves lagged dependent variables whose coefficients (or partial effects) are parameters of interest. The dilemma is that the structural equation cannot be evaluated for the initial observations since lagged information is not available, but conditioning on the initial observations leads to inconsistent estimates in the presence of unobserved heterogeneity. In the context of a first-order Markov model of binary responses, Heckman (1981) proposed to supplement the structural model with an approximate reduced-form model for the initial conditions, based on exogenous information available for the initial periods, a flexible specification of the influence of unobserved heterogeneity, and imposing no parameter restrictions across submodels. Alternatively, Wooldridge (2005) proposed to supplement a structural first-order Markov model with an auxiliary model of the distribution of unobserved heterogeneity in terms of the initial conditions and exogenous explanatory variables. Wooldridge's method has been applied and extended to higher-order models and discrete-time duration models for example by Stewart (2007) and Bhuller, Brinch, and Königs (2016). Heckman's method has been applied for example in continuous-time duration analysis by Gritz (1993) and in discrete-time duration analysis by Ham and LaLonde (1996), Cappellari, Dorsett, and Haile (2010), and Gørgens and Hyslop (2016). Skrondal and Rabe-Hesketh (2014) provided a recent comparison between these and related methods for estimating first-order Markov dynamic panel data models of binary responses.

In this paper we consider estimation of continuous-time dynamic event history models with censored data by maximising a simulated likelihood function using all available data. The likelihood function is specified in terms of observed and unobserved events, and unobserved events are then "integrated out" using Monte Carlo and importance sampling methods. We allow for unobserved heterogeneity in the form of so-called random effects and integrate out unobserved heterogeneity using a Gaussian quadrature rule. Our maximum simulated likelihood (MSL) estimator uses all available data and does not involve additional functional-form assumptions or additional ad hoc parameters. The method is applicable when the times during which individuals are at risk of experiencing events are known.[1] For simplicity, we focus on recurring events. This class of models covers a wide range of applications: purchases of specific goods or services, health events such as heart attacks or dental fillings, child births, time between earth quakes or geyser eruptions, etc.

The method of maximum simulated likelihood estimation has been successfully applied in other contexts. For example, Lerman and Manski (1981) were the first econometricians to consider the frequency simulator of (multinomial probit) choice probabilities. Keane (1994) studied MSL estimation of binary response models with serially correlated errors, with the multinomial probit model as the leading case. McCulloch (1997) considered latent class (mixture) models. Kamionka (1998) sketched a general framework for continuous-time transition models and provided some simulation results for estimating continuous-time time-homogeneous Markov processes using data measured on a discrete time scale. Keane and Sauer (2010) developed a method for estimating discrete-time dynamic panel data models with unobserved endogenous state variables. Their method assumed that the dependent variables are measured with error. Some authors have compared MSL estimation with estimation using the EM algorithm and found that the latter performed better. Brinch (2012) argued that the negative assessment of MSL estimation among some authors is at least partly due to suboptimal choices made in the implementation.

The MSL approach has both advantages and disadvantages over the alternatives. As mentioned above, dropping left-censored histories from the analysis (listwise deletion) makes for easy ML estimation but can be very costly in terms of sample size. Specifying auxiliary models for either the distribution of the initial conditions in terms of unobserved heterogeneity (Heckman, 1981) or the distribution of unobserved heterogeneity in terms of initial conditions (Wooldridge, 2005) also allow for standard ML estimation, but specification error potentially affects the bias and consistency of the estimates and the

[1] In a study of transitions into and out of female headship, Moffitt and Rendall (1995) were able to integrate out unobserved events analytically because the distribution of missing data was discrete in their model.

additional parameters lead to a loss of degrees of freedom. The MSL approach is expected to have higher efficiency, because the full data set can be used, and because no auxiliary parameters are involved. By increasing the number of simulations, MSL estimates can be made arbitrarily close to the exact ML estimates, and hence MSL estimates can be asymptotically efficient under standard conditions.

A potential disadvantage of the MSL approach is computational difficulties. First, numerical integration in high dimensions is known to be difficult, whether by quadrature rules or Monte Carlo methods. In practice, limits on computing capacity may restrict the level of accuracy that can be achieved within reasonable time. Second, when the integration is carried out using Monte Carlo methods the simulated likelihood function is discontinuous, which causes trouble for standard maximisation algorithms such as Newton's method. However, importance sampling methods can be used to smooth the simulated likelihood function (see e.g. Gouriéroux and Monfort, 1991).

The present paper contributes to the literature by showing how MSL estimation can be applied in the context of dynamic models of recurring events in continuous time with censored data. We provide Monte Carlo evidence to show that MSL estimation is practically feasible, and we confirm that MSL estimation can provide substantial efficiency gains over listwise deletion and Heckman's approximate reduced-form modelling.

The paper is organised as follows. Section 2 sets up the notation and discusses maximum likelihood estimation. Section 3 presents the results of our Monte Carlo experiments. Section 4 concludes.

2 Maximum likelihood estimation

2.1 The likelihood function

When analysing censored data, it is necessary to distinguish between the underlying event process and the observation process. For example, the statistics literature talks about time at risk and time under observation. Let time be partitioned into j_i periods, $(c_{ij-1}, c_{ij}]$ for $j = 1, 2, \ldots, j_i$, such that c_{i0} is the time individual i becomes at risk, c_{ij_i}

is the last time individual i is both at risk and under observation, and the individual is alternatingly either under observation or not during each period. Thus, individuals are either under observation in all odd periods or in all even periods. Analysis time is defined by normalising $c_{i0} = 0$.

The interaction between the event process and the observation process necessitates notation which number event times within observation periods. Hence, let k_{ij} denote the number of (observed or unobserved) events during individual i's period j, and let b_{ijk} for $k = 1, \dots, k_{ij}$ denote event times within period j. For convenience, define the vector $\boldsymbol{b}_{ij} = (b_{ijk_{ij}}, \dots, b_{ij1})$; if $k_{ij} = 0$ then $\boldsymbol{b}_{ij}$ denotes a zero-dimensional vector. To simplify certain expressions, define also b_{ij0} by setting $b_{ij0} = c_{ij-1}$. We assume that the event process and the observation process are independent. We postpone the discussion of observed and unobserved heterogeneity until later.

In general, the likelihood of an event at any given time may depend on the history of events prior to that time. Let $s_i(t)$ denote all individual i's history at time t. That is, $s_i(t)$ includes all event times until and including t, the fact that no events occurred between the most recent event and time t, and the observation period boundaries. Let $h(t|s(t'), \theta)$ for $t > t'$ denote the conditional hazard function for events evaluated at time t given the history until time t', $s(t')$, where θ is the unknown parameter vector to be estimated. Also let $H(t|s(t'), \theta)$ for $t > t'$ denote the associated value of the cumulative hazard function from time t' until time t. That is, H is defined by $H(t|s(t'), \theta) = \int_{t'}^{t} h(y|s(t'), \theta)\, dy$. Furthermore, let $f(t|s(t'), \theta)$ denote the conditional event density at t given the history $s(t')$, and let F denote the corresponding cumulative distribution function. Then we have that

$$f(t|s(t'), \theta) = h(t|s(t'), \theta) \exp\bigl(-H(t|s(t'), \theta)\bigr), \quad t > t'. \tag{1}$$

Here the exponential term on the right-hand side captures the non-occurrence of events during $(t', t]$. Finally, let g_j be the conditional joint density of events during period j given previous events. Using $\boldsymbol{b}_j$ without subscript i to denote a generic vector of event times in period j and using k_j for the corresponding number of events, we have

$$g_j(\boldsymbol{b}_j|\boldsymbol{b}_{j-1},\ldots,\boldsymbol{b}_1,\theta) = \left(\prod_{k=1}^{k_j} f(b_{jk}|s(b_{jk-1}),\theta)\right)\exp\bigl(-H(c_j|s(b_{jk_j}),\theta)\bigr). \tag{2}$$

The exponential term on the right-hand side represents the fact that no events occurred during $(b_{jk_j}, c_j]$ if $k_j > 0$ or during $(c_{j-1}, c_j]$ if $k_j = 0$. (Recall that we have defined $b_{j0} = c_{j-1}$.) By convention the product of the sequence on the right-hand side of (2) is defined to be 1 if $k_j = 0$ (and $\boldsymbol{b}_j$ is zero-dimensional).

The likelihood contribution for individual i in terms of observed and unobserved terms (i.e. the complete-data likelihood contribution, apart from right-censoring) is[2]

$$L_i^\star(\theta) = \prod_{j=1}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1},\ldots,\boldsymbol{b}_{i1},\theta). \tag{3}$$

The full complete-data likelihood function is defined as the product of $L_i^\star(\theta)$ over i.

The complete-data likelihood function cannot be evaluated when the data are not complete. Simply omitting terms that involve missing data in (3) and maximising the computable part of the likelihood function generally does not yield a consistent estimator of θ. This is because the resulting truncated sample may not be representative of the population (see e.g. Moffitt and Rendall, 1995).

To get the likelihood contribution of the observed events, the unobserved events must be integrated out. For an individual who is under observation during odd-numbered periods (so j_i is odd), the incomplete-data likelihood contribution is[3]

$$\begin{aligned} L_i(\theta) = \iint\cdots\int\Biggl(&\prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{j-1},\boldsymbol{b}_{ij-2},\ldots,\boldsymbol{b}_2,\boldsymbol{b}_{i1},\theta)\Biggr) \\ &\times\Biggl(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_j|\boldsymbol{b}_{ij-1},\boldsymbol{b}_{j-2},\ldots,\boldsymbol{b}_2,\boldsymbol{b}_{i1},\theta)\Biggr)\,d\boldsymbol{b}_{j_i-1}\ldots d\boldsymbol{b}_4\,d\boldsymbol{b}_2 \\ = \mathsf{E}_{\boldsymbol{B}_{i2}^\theta}\Biggl[\cdots\mathsf{E}_{\boldsymbol{B}_{ij_i-1}^\theta}\Biggl[&\prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{B}_{ij-1}^\theta,\boldsymbol{b}_{ij-2},\ldots,\boldsymbol{B}_{i2}^\theta,\boldsymbol{b}_{i1},\theta) \end{aligned} \tag{4}$$

[2]This ignores the likelihood contribution of the entry and exit times, c_{ij-1} and c_{ij}, which leads to valid inference under the maintained assumption that these are independent of the event times. To focus on computational aspects we assume θ is identified and do not further discuss this issue.

[3]Admittedly the notation is sloppy here, since the dimension of the terms integrated out are random, and the limits of the definite integrals are omitted. The notation could be made formally correct by conditioning on and summing over the possible dimensions of the vectors.

$$\left| \boldsymbol{B}^{\theta}_{ij_i-2} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}^{\theta}_{i2}, \boldsymbol{B}^{\theta}_{i1} = \boldsymbol{b}_{i1} \right] \cdots \left| \boldsymbol{B}^{\theta}_{i1} = \boldsymbol{b}_{i1} \right],$$

where $\boldsymbol{B}^{\theta}_{ij}$ denotes a random vector of potential event times for individual i in period j, whose conditional probability density function given prior history is given in (2), taking individual i's realised observation period endpoints $c_{i0}, \ldots, c_{j_i}$ as given. The superscript θ serves as a reminder that this distribution is governed by the θ at which the likelihood contribution is evaluated, not the so-called true value behind the realised events $\boldsymbol{b}_{ij}$.

Similarly, for an individual who is under observation during even-numbered periods (so j_i is even), the incomplete-data likelihood contribution is

$$\begin{aligned} L_i(\theta) &= \iint \cdots \int \left(\prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_j | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \right) \\ &\qquad \times \left(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{j-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \right) d\boldsymbol{b}_{j_i-1} \ldots d\boldsymbol{b}_3 \, d\boldsymbol{b}_1 \\ &= \mathsf{E}_{\boldsymbol{B}^{\theta}_{i1}} \left[\cdots \mathsf{E}_{\boldsymbol{B}^{\theta}_{ij_i-1}} \left[\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{B}^{\theta}_{ij-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}^{\theta}_{i1}, \theta) \right.\right. \\ &\qquad\qquad \left.\left. \left| \boldsymbol{B}^{\theta}_{ij_i-2} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}^{\theta}_{i2} = \boldsymbol{b}_{i2}, \boldsymbol{B}^{\theta}_{i1} \right] \cdots \right]. \end{aligned} \tag{5}$$

Note the outermost expectation is unconditional here, since there is no history prior to period 1.

The full incomplete-data likelihood function is defined as the product of $L_i(\theta)$ over i. Since this is the exact likelihood function for the observed data, the maximiser is a consistent and asymptotically efficient estimator of θ. However, computing this function is hampered by the fact that in general the integrals (expectations) cannot be solved analytically. In typical model specifications, the event density function depends non-linearly on previous events, and the integrals are not separable.

2.2 Monte Carlo integration

Our proposal is to use Monte Carlo simulation to integrate out the unobserved terms. For each individual we draw R independent pseudo-histories for periods with missing information. For a given value of θ, we then approximate the likelihood function by averaging over the R pseudo-histories. That is, for an individual who is under observation during odd-numbered periods, we compute

$$L_i^R(\theta) = \frac{1}{R}\sum_{r=1}^{R} \prod_{j=1:j\,\mathrm{odd}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1}^r, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}^r, \boldsymbol{b}_{i1}, \theta), \tag{6}$$

and for and individual who is under observation during even-numbered periods, we compute

$$L_i^R(\theta) = \frac{1}{R}\sum_{r=1}^{R} \prod_{j=1:j\,\mathrm{even}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1}^r, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta), \tag{7}$$

where for each $r = 1, \ldots, R$ and $j = 1, \ldots, j_i$ the $\boldsymbol{b}_{ij}^r$ are sequences of simulated event times specific to individual i's period j, compatible with the individual's observed and simulated event history, and compatible with the density evaluated at θ. That is, each $\boldsymbol{b}_{ij}^r$ is drawn from the conditional distribution g_j given in (2), with simulated prior event times replacing actual when the latter are unobserved, and using the θ at which the likelihood function is evaluated. (For simplicity, the dependence of $\boldsymbol{b}_{ij}^r$ on θ is suppressed in the notation.) Let k_{ij}^r denote the dimension of $\boldsymbol{b}_{ij}^r$. Standard arguments (the law of large numbers) imply that L_i^R converges to L_i pointwise as R diverges to infinity.

The dynamic nature of the density function g_j means that the simulation must be done sequentially. Recall that f denotes the conditional density of events, and F is the corresponding cumulative distribution function. For common parametric specifications of the hazard function, f, F and F^{-1} are easily evaluated using closed-form formulae. Pseudo-histories can therefore be created using the inversion method.

Suppose first that $(c_{i0}, c_{i1}]$ is a period where individual i is not under observation. To simulate a first event time for this individual, we draw a pseudo-random number u_{i11}^r from the uniform distribution and then compute a candidate event time by $b_{i11}^r = F^{-1}(u_{i11}^r|s_i(c_{i0}), \theta)$. If $b_{i11}^r > c_{i1}$, we decide that no events happened during $(c_{i0}, c_{i1}]$ and set $k_{i1}^r = 0$. If $b_{i11}^r \leq c_{i1}$, we keep b_{i11}^r and draw a second candidate event time. In general, having drawn $b_{i1k-1}^r, \ldots, b_{i11}^r$ with $b_{i1k-1}^r \leq c_{i1}$, we draw a candidate for the kth event time by $b_{i1k}^r = F^{-1}(u_{i1k}^r|s_i^r(b_{i1k-1}^r), \theta)$ where u_{i1k}^r is another (independent) draw from the uniform distribution and $s_i^r(b_{i1k-1}^r)$ includes the simulated previous events b_{i1k-1}^r, $\ldots$,

b^r_{i11}. If $b^r_{i1k} > c_{i1}$, the rth pseudo-history for period j is complete with $k^r_{i1} = k - 1$ and $\boldsymbol{b}^r_{ij} = (b^r_{i1k^r_{ij}}, \ldots, b^r_{i11})$. If $b^r_{i1k} \leq c_{i1}$, we increment k and consider the next candidate event time.

The simulation procedure is similar for other periods where an individual is not under observation. The only difference is that the history includes the observed event times during prior periods where the individual is under observation as well as simulated event times during prior periods where the individual is not under observation. For example, if individual i is under observation during $(c_{i0}, c_{i1}]$ but not during $(c_{i1}, c_{i2}]$, then $s^r_i(b^r_{i2k-1})$ includes the simulated events b^r_{i2k-1}, ..., b^r_{i21} as well as the observed events $\boldsymbol{b}_{i1}$.

As pointed out by several authors (see e.g. Stern, 1997; Brinch, 2012), it is essential for successful numerical maximisation to use the same underlying draws from the uniform distribution in all the evaluations of the likelihood function (including computation of numerical derivatives).

The full incomplete-data simulated likelihood function is defined as the product of $L^R_i(\theta)$ over i. Maximising the simulated likelihood function yields a consistent and asymptotically efficient estimator under standard conditions provided $\sqrt{N}/R \to 0$ as $N \to \infty$ where N is the number of individuals in the sample (Gouriéroux and Monfort, 1991).

2.3 Importance sampling

The simulated likelihood contributions described above are not everywhere continuous. Discontinuities occur when a small change in θ leads to a switch in the decision of whether to retain or discard a candidate event time (b^r_{ijk}). These discontinuities mean that standard maximisation methods for differentiable functions such as Newton's method may not work well.

Since the magnitude of the discontinuities are of order $1/R$, one approach to numerical maximisation of the likelihood function is to use a standard derivative-based method with R very large, and increase R whenever a discontinuity is causing problems. Another approach is to use a non-gradient method such as simplex algorithm (see e.g. Keane and Sauer, 2010). These approaches will generally lead to convergence, but are expected to be slow.

An appealing method is to smooth the likelihood contributions using importance sampling techniques. In the present context, an importance sampling distribution for b_{ij} can be any given conditional distribution of events during period j given previous events. For concreteness, we choose g_j evaluated at some fixed value θ^*. For an individual who is under observation during even-numbered periods (the odd-numbered case is similar), the incomplete-data likelihood contribution can be written as

$$\begin{aligned}
L_i(\theta) = \iint \cdots \int \Bigg(&\prod_{j=1:j\,\text{odd}}^{j_i} g_j(\boldsymbol{b}_j|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \\
&\qquad\qquad \times \frac{g_j(\boldsymbol{b}_j|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta^*)}{g_j(\boldsymbol{b}_j|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{j-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta^*)} \Bigg) \\
&\times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{j-1}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_1, \theta) \Bigg)\, d\boldsymbol{b}_{j_i-1} \ldots d\boldsymbol{b}_3\, d\boldsymbol{b}_1 \\
= \mathsf{E}_{\boldsymbol{B}_{i1}^{\theta^*}} \Bigg[\cdots \mathsf{E}_{\boldsymbol{B}_{ij_i-1}^{\theta^*}} \Bigg[\Bigg(&\prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{B}_{ij}^{\theta^*}|\boldsymbol{b}_{ij-1}, \boldsymbol{B}_{ij-2}^{\theta^*}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta)}{g_j(\boldsymbol{B}_{ij}^{\theta^*}|\boldsymbol{b}_{ij-1}, \boldsymbol{B}_{ij-2}^{\theta^*}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta^*)} \Bigg) \\
&\times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{B}_{ij-1}^{\theta^*}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*}, \theta) \Bigg) \\
&\qquad\qquad \Bigg| \, \boldsymbol{B}_{ij_i-2}^{\theta^*} = \boldsymbol{b}_{ij_i-2}, \ldots, \boldsymbol{B}_{i2}^{\theta^*} = \boldsymbol{b}_{i2}, \boldsymbol{B}_{i1}^{\theta^*} \Bigg] \cdots \Bigg].
\end{aligned} \tag{8}$$

The corresponding simulated likelihood contribution is

$$\begin{aligned}
L_i^R(\theta) = \frac{1}{R} \sum_{r=1}^{R} \Bigg(&\prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{b}_{ij}^r|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^r, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta)}{g_j(\boldsymbol{b}_{ij}^r|\boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^r, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta^*)} \Bigg) \\
&\qquad \times \Bigg(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij}|\boldsymbol{b}_{ij-1}^r, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^r, \theta) \Bigg),
\end{aligned} \tag{9}$$

where $\boldsymbol{b}_{ij}^r$ for $r = 1, \ldots, R$ and $j = 1, \ldots, j_i$ are drawn from the importance sampling distribution $g_j(\cdot|\cdot, \theta^*)$ instead of the "correct" distribution $g_j(\cdot|\cdot, \theta)$. The principle underpinning importance sampling is that the "error" can be fixed by reweighting using the ratio of correct density over the importance sampling density.

One of the advantages of the importance sampling approach is that the simulated event times do not depend on the value of θ at which the likelihood contribution is evaluated, and hence the simulated likelihood function is continuous and differentiable. A potential drawback is that a very large R may be needed in order to achieve a good approximation to the likelihood function. Keane and Sauer (2010) suggest that it may be advantageous to scale the importance sampling weights to sum to 1 over r.

2.4 Covariates

So far we have ignored covariates, in order to focus on missing event times. In practice, covariates can be time-invariant or time-varying. Incorporating covariates is straightforward when the covariate paths are completely observed. Usually covariates with incompletely observed paths can also be incorporated, using an extended simulation procedure. For example, in some cases the observation process is such that time-varying covariates are missing during the same periods when the event times are not observed. These covariates can be incorporated by specifying an auxiliary model for their paths, and using this model to integrate out the missing parts of the covariate paths.[4]

2.5 Unobserved heterogeneity

Allowing for individual-specific time-invariant effects is standard in the literature. These effects capture correlation across event times ("frailty" in the statistics literature). It is well-known that omitting individual-specific time-invariant effects can lead to a bias towards negative duration dependence (see e.g. Elbers and Ridder, 1982; Heckman and Singer, 1984a). The effects are usually assumed to be independent of covariates ("random effects" in the econometrics literature). The distribution of the random effects is specified either as discrete (following Heckman and Singer, 1984b) or as continuous such as a normal distribution with mean 0.

Let v_i denote the realised unobserved random effect for individual i, and consider the complete-data likelihood function given in (3). Including and integrating out the random effects gives

$$L_i^{\star}(\theta) = \int_{-\infty}^{\infty} \left(\prod_{j=1}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{ij-1}, \ldots, \boldsymbol{b}_{i1}, v, \theta) \right) dZ(v), \tag{10}$$

[4]See e.g. Keane and Sauer (2010) for a similar approach in a discrete-time setting.

where Z denotes the cumulative distribution function of v_i, and implicitly θ has been augmented to include unknown parameters of the distribution of v_i. For simplicity, we also reuse the symbols g_j, f, h, and H to denote the corresponding functions which depend on the random effect. The modification required to include a random effect is similar in the other likelihood contributions given above.

In practice, if Z is continuous then the integration is carried out using Gaussian quadrature. While straightforward, this increases the computational burden somewhat. For example, with Q evaluation points $v_1, \ldots, v_Q$ and weights $w_i, \ldots, w_Q$, the simulated likelihood contribution in (9) becomes

$$L_i^R(\theta) = \sum_{q=1}^{Q} w_q \frac{1}{R} \sum_{r=1}^{R} \left(\prod_{j=1:j\,\text{odd}}^{j_i} \frac{g_j(\boldsymbol{b}_{ij}^{qr} | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^{qr}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta)}{g_j(\boldsymbol{b}_{ij}^{qr} | \boldsymbol{b}_{ij-1}, \boldsymbol{b}_{ij-2}^{qr}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta^*)} \right) \times \left(\prod_{j=1:j\,\text{even}}^{j_i} g_j(\boldsymbol{b}_{ij} | \boldsymbol{b}_{ij-1}^{qr}, \boldsymbol{b}_{ij-2}, \ldots, \boldsymbol{b}_{i2}, \boldsymbol{b}_{i1}^{qr}, v_q, \theta) \right). \tag{11}$$

Note that the same underlying random draws from the uniform distribution can be used for each q, but the simulated event times, and even the number of compatible simulated event times, k_{ij}^{qr}, will be different.

2.6 Estimation based on Heckman's method

The likelihood contribution for individual i's period j given in (2) is made up of subcontributions representing each of the events, and a term representing the final right-censored period when no events occurred. In general, the hazard function at any given time may depend on the entire previous history of events. However, in many applications it can be assumed that the hazard function depends only on recent history. For example, the hazard rate for an event occurring at time t may depend only on whether or not an event occurred (or the number of events that occurred) in the period $(t-\tau, t)$ for some fixed τ. In applications where the influence of history is limited, missing data may affect only some and not all of the event subcontributions. If so, then the terms in the likelihood function that do not depend on missing data are "computable", and it may be feasible to handle the "uncomputable" parts by adapting the ideas of Heckman (1981).

To describe how Heckman's idea can be applied in the present context, define d_{ijk} to be 1 if $h(b_{ijk}|s_i(b_{ijk-1}), v_i, \theta)$ is computable, and define d_{ijk} to be 0 otherwise. Define also $d_{ijk_{ij}+1}$ so that $\exp\bigl(-H(c_{ij}|s_i(b_{ijk_{ij}}), v_i, \theta)\bigr)$ is computable if and only if $d_{ijk_{ij}+1} = 1$.

It is helpful to begin with a simple two-period observation process, so suppose individual i is under observation in period 2 but not in period 1. By definition, the computable terms are those that do not depend on the unobserved events in period 1. Since they don't depend on period 1 events, they can be factored out of the integral in the incomplete-data likelihood contribution for individual i. Allowing for unobserved heterogeneity, we have from (5) that

$$\begin{aligned} L_i(\theta) &= \int_{-\infty}^{\infty} \biggl\{ \int g_2(\boldsymbol{b}_{i2}|\boldsymbol{b}_1, v, \theta) g_1(\boldsymbol{b}_1|v, \theta)\, d\boldsymbol{b}_1 \biggr\}\, dZ(v) \\ &= \int_{-\infty}^{\infty} \biggl\{ \biggl[\int \biggl(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}), v, \theta)^{1-d_{i2k}} \biggr) \\ &\qquad\qquad \times \exp\bigl(-H(c_{i2}|s_i(b_{i2k_{i2}}), v, \theta)\bigr)^{1-d_{i2k+1}} g_1(\boldsymbol{b}_1|v, \theta)\, d\boldsymbol{b}_1 \biggr] \\ &\qquad\qquad \times \biggl(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}), v, \theta)^{d_{i2k}} \biggr) \\ &\qquad\qquad \times \exp\bigl(-H(c_{i2}|s_i(b_{i2k_{i2}}), v, \theta)\bigr)^{d_{i2k+1}} \biggr\}\, dZ(v). \end{aligned} \tag{12}$$

The integral with respect to $\boldsymbol{b}_1$ is uncomputable, because the necessary history is not observed. Heckman's idea was to approximate this using a reduced-form density that is based on as much predetermined information as is available, incorporates unobserved heterogeneity, and uses a flexible parametric specification. How much information is available depends on the details of how the hazard rate depends on previous history.

Let $h^{\dagger}(t|s(t'), v, \xi)$ for $t > t'$ be an approximate conditional hazard function evaluated at time t given the event history until time t'. For simplicity, we do not introduce new notation for the observed history itself. The principle is that $h^{\dagger}$ is parameterised so that it depend only on the part of $s(t')$ that is observed at time t'. Hence, $h^{\dagger}(t|s(t'), v, \xi)$ is computable even though $s(t')$ is not fully observed.[5] Let $H^{\dagger}$ denote the corresponding

[5]In practice, flexible specifications with different parameters may be used depending on the amount of history available at time t'.

cumulative hazard function from time t' to time t, and define $f^\dagger$ by

$$f^\dagger(t|s(t'),v,\xi) = h^\dagger(t|s(t'),v,\xi)\exp\big(-H^\dagger(t|s(t'),v,\xi)\big), \quad t > t'. \tag{13}$$

Then the hope is that given θ for some ξ we have that

$$\begin{aligned}\int\Bigg(\prod_{k=1}^{k_{i2}} f(b_{i2k}|s_i(b_{i2k-1}),v,\theta)^{1-d_{ijk}}\Bigg)&\\ \times\exp\big(-H(c_{i2}|s_i(b_{i2k_{i2}}),v,\theta)\big)^{1-d_{ijk+1}} g_1(\boldsymbol{b}_1|v,\theta)\,d\boldsymbol{b}_1&\\ \approx\Bigg(\prod_{k=1}^{k_{i2}} f^\dagger(b_{i2k}|s_i(b_{i2k-1}),v,\xi)^{1-d_{ijk}}\Bigg)\exp\big(-H^\dagger(c_{i2}|s_i(b_{i2k_{i2}}),v,\xi)\big)^{1-d_{ijk+1}}.&\end{aligned} \tag{14}$$

Substituting the approximation into (12) gives an approximate likelihood contribution as a function of (θ,ξ).

In the general multi-period case, the approximate likelihood contribution for an individual who is under observation during even-numbered periods (the odd-numbered case is similar) is

$$\begin{aligned}L_i^\dagger(\theta,\xi) = \int_{-\infty}^{\infty}\Bigg\{\prod_{j=1:j\text{ even}}^{j_i}\Bigg(\prod_{k=1}^{k_{ij}} f(b_{ijk}|s_i(b_{ijk-1}),v_i,\theta)^{d_{ijk}}&\\ \times f^\dagger(b_{ijk}|s_i(b_{ijk-1}),v_i,\xi)^{1-d_{ijk}}\Bigg)\exp\big(-H(c_{ij}|s_i(b_{ijk_{ij}}),v_i,\theta)\big)^{d_{ijk_{ij}+1}}&\\ \times\exp\big(-H^\dagger(c_{ij}|s_i(b_{ijk_{ij}}),v_i,\xi)\big)^{1-d_{ijk_{ij}+1}}\Bigg\}\,dZ(v).&\end{aligned} \tag{15}$$

Maximising the corresponding full likelihood function yields a consistent estimator of θ, provided the approximate reduced-form model is in fact correctly specified. Generally the hope is that the approximation is good enough that the magnitude of the inconsistency is acceptable.

3 Monte Carlo experiments

To investigate the performance of the MSL approach, we carried out a small set of Monte Carlo experiments. The designs feature mixed proportional hazards with a Weibull baseline hazard function, a single time-invariant covariate, x_i, and a continuous random effect,

v_i. The covariate and the random effect are realisations from a standard normal distribution.

Separate models are specified for the first event and for subsequent events. Current duration dependence is captured in the baseline hazards. After the first event, the hazard rates also depend on whether an event occurred or not during a recent period of fixed length (i.e. a moving window). Specifically, the hazard function for the first event is

$$h_1(t|s(0), x, v, \theta) = \alpha_1 t^{\alpha_1 - 1} \exp(x\beta_1 + \mu_1 + v\delta_1), \quad t > 0. \tag{16}$$

With t' representing the most recent event time before t, the hazard function for subsequent events is

$$h_2(t|s(t'), x, v, \theta) = \alpha_2 t^{\alpha_2 - 1} \exp(1(t < t' + \tau)\gamma + x\beta_2 + \mu_2 + v\delta_2), \quad t > t', \tag{17}$$

where $\theta = (\alpha_1, \beta_1, \mu_1, \delta_1, \gamma, \alpha_2, \beta_2, \mu_2, \delta_2)'$ and τ is a constant that varies across experiments. We normalise $\delta_1 > 0$ and $\delta_2 > 0$. The parameters used in the data-generating processes are fixed at $\alpha_1 = 1$, $\beta_1 = 0.2$, $\mu_1 = -0.5$, $\gamma = 0.5$, $\alpha_2 = 1$, $\beta_2 = 0.2$, and $\mu_2 = -0.5$, while either $\delta_1 = 0$, $\delta_2 = 0$ or $\delta_1 = 1$, $\delta_2 = 1$ as indicated in the tables.

Note that baseline time does not reset after an event in these designs. Alternatively, the baseline hazard rate can be specified in terms of $t - t'$. More flexible models can be obtained by specifying separate hazard functions for second events, third events, etc. Less flexible models can be obtained by assuming $\alpha_1 = \alpha_2$, $\beta_1 = \beta_2$, $\mu_1 = \mu_2$, and $\delta_1 = \delta_2$. In this case, the model effectively consists of a single hazard specification since (16) is simply (17) with $\gamma = 0$. Such a specification was adopted for example by Keane and Sauer (2010). Our designs satisfy these restrictions, but we do not impose them in the estimation.

The observation process mimics a sampling procedure where analysis time is age and data are collected from the population stock over a fixed calendar period. Specifically, half the sample are observed over the age range $(0, 1]$ while the other half is observed over $(1, 2]$. That is, the former is right-censored at time 1 (and not left-censored), while the latter is left-censored at time 1 and right-censored at time 2. The number of non-left-censored individuals in the samples is $N_1 = 250$ and while the number of left-censored individuals is either $N_2 = 250$ or $N_2 = 500$ as indicated in the tables.

Across all designs, about half of the individuals in a sample do not have any events during their observation period. For those who do have observed events, the mean time until the first event is about 0.38. Since $\alpha_1 = 1$ and $\alpha_2 = 1$ imply memoryless exponential hazard functions, these statistics apply to both the left-censored and the non-left-censored.

We compute several estimators to compare the MSL approach with simple estimators that may be considered in practice. Estimator ISU is an MSL estimator which uses importance sampling techniques without scaling of the weights, while estimator ISN has the weights normalised to sum to one. For simplicity, we use the true data-degenerating process as the importance sampling distribution, and we set $R = 100$.

Estimator NLC uses only individuals with non-left-censored data; that is, half the sample in the experiments with $N_2 = 250$ and a third of the sample when $N_2 = 500$.

Estimator HKM uses the approximate reduced-form idea of Heckman (1981) to handle the left-censoring problem. For the designs considered here, the only uncomputable term in the likelihood contribution for the left-censored individuals concerns the first observed event in period 2, b_{i21}, if this happens to happen within the period $(1, 1 + \tau]$. This is because $1(b_{i21} < b_{i1k_{i1}} + \tau)$ cannot be computed as the time of the last event in period 1, $b_{i1k_{i1}}$, is unknown, while $1(b_{i21} < b_{i1k_{i1}} + \tau) = 0$ can be inferred if $b_{i21} > c_{i1} + \tau$ for $c_{i1} = 1$. Since no useful information is available in $s(1)$, we specify the auxiliary hazard function for b_{i21} as

$$h_3(t|s(1), x, v, \theta) = \alpha_3 t^{\alpha_3 - 1} \exp(x\beta_3 + \mu_3 + v\delta_3), \quad t > 0. \tag{18}$$

The literature on dynamic panel data models usually does not distinguish between the start of the event process and the start of the observation period, although these are associated with conceptually distinct problems: at the start of the event process lags cannot exist so logically a different structural equation is required, whereas at the start of the observation period lags may exist so a method for dealing with missing data is required. Here we maintain the distinction between left-censoring and genuine first events. That is, our HKM implementation estimates the parameters of all three hazard functions.

There are 1000 samples in each experiment. In designs with random effects, unobserved heterogeneity is integrated out using Gauss-Hermite quadrature with $Q = 10$ evaluation points.[6]

Table 1 shows root mean square errors (RMSE) for the four estimators for designs without random effects. The likelihood function is separable in the parameters pertaining to the first and subsequent events, respectively. Consequently, the NLC and HKM estimates for the parameters of the first hazard function are identical. The RMSEs for the IS estimates are slightly lower. For the second hazard functions, the HKM estimates improve dramatically on the NLC estimates. This is because the usable sample is twice as large, and the HKM involve only a few more parameters. The RMSEs for the IS estimates are lower again, especially for γ and μ_2.

The value of τ does not affect the first hazard function, but the higher τ, the more history data are needed to estimate the second hazard function. The problem of missing data therefore becomes more severe and higher RMSEs are expected. This is confirmed in table 1. The results for the first hazard function do not change, because the same data are used. For the second hazard function, the RMSEs for $\ln \alpha_2$ and β_2 also remain roughly constant, while the RMSEs for γ and μ_2 increase. The increase occurs because the number of individuals with no recent events becomes small when τ is large, and hence it becomes difficult to estimate μ_2 accurately.[7] Since individuals who have recent events identify the sum $\gamma + \mu_2$, the uncertainty in the estimates of μ_2 are mirrored in the estimates of γ. However, the HKM estimator is better than the NLC estimator, since it uses much more sample, and the two IS estimators are better than the HKM estimator, since they use the sample efficiently.

Table 2 shows results for designs with random effects. Looking first at the case where $\tau = 0.3$ and $N_2 = 250$, the patterns are similar to those without random effects. The HKM estimator improves on the NLC estimator and the IS estimators perform better

[6] The results omit a few samples (max 3 per experiment) where the estimation procedure did not converge.

[7] In the extreme, if these individuals experience no further events, the estimated hazard should be zero, which means $\hat{\mu}_2 = -\infty$.

than the HKM estimator. Estimation of distributions of random effects is notoriously difficult, so it is not surprising to find much higher RMSEs for $\ln \delta_1$ and $\ln \delta_2$.

As τ increases, the results for the first-event parameters and for $\ln \alpha_2$ and β_2 do not change much. Similar to the designs without random effects, estimation of γ and μ_2 becomes more difficult when τ is large, so the RMSEs for those parameters increase for all estimators. The increase is very large for the NLC and HKM estimators but only modest for the IS estimators, so the efficiency gain of the latter becomes more substantial. The patterns for the RMSEs of $\ln \delta_1$ and $\ln \delta_2$ are complex and not entirely intuitive. For example, the RMSEs for the NLC estimator of $\ln \delta_2$ tend to increase with τ, but decrease for the HKM estimator. Presumably this is because the "practical identification" of these parameters is weak, so small approximation errors in the likelihood function can have large effects of the estimates.

When the number of left-censored individuals is increased from $N_2 = 250$ to $N_2 = 500$, the results for the first-event parameters hardly change, while there is some improvement for the parameters relating to the second hazard function. This is particularly true for the difficult parameters $\ln \delta_1$ and $\ln \delta_2$, and to a lesser extent for γ and μ_2.

To conclude, it is clear that there are potentially large efficiency gains in using MSL estimation over methods based on listwise deletion or Heckman's approximate reduced-form modelling of initial conditions. The gains are particularly high for parameters that are difficult to estimate. The fact that the results for the ISU and ISN estimators are not identical reveal a disadvantage of MSL estimation; namely, that numerical integration inevitably involves some approximation error. As a practical guide, we suggest computing several MSL estimates, using different importance sampling distributions with and without scaling of the weights. If the estimates are too different, then the values of R and Q can be increased until all estimates agree.

4 Concluding remarks

This paper considers ML estimation of dynamic models of recurring events in continuous time using censored data. We propose to deal with censoring by integrating out missing data from the likelihood function using Monte Carlo simulation and importance sampling techniques. We compare MSL estimation with estimators that either ignore left-censored individuals (listwise deletion) or deal with censoring using ad hoc modifications to the likelihood function (Heckman's method). The results show that there can be substantial efficiency gains in maximising the full simulated likelihood function.

We assume that the censoring and the event processes are independent, and we focus on settings where time origins and covariate paths are known. We anticipate that these assumptions can be relaxed, at the costs of further computational complications. Given the encouraging results for models of recurring events, it is also likely that similar efficiency gains are available for example in multi-state transition models.

References

Bhuller, M., C. N. Brinch, and S. Königs (2016). Time aggregation and state dependence in welfare receipt. *Economic Journal*. Forthcoming.

Brinch, C. N. (2012). Efficient simulated maximum likelihood estimation through explicitly parameter dependent importance sampling. *Computational Statistics 27*, 13–28.

Cappellari, L., R. Dorsett, and G. Haile (2010). State dependence and unobserved heterogeneity in the employment transitions of the over-50s. *Empirical Economics 38*(3), 523–554.

Cockx, B. and M. Picchio (2012). Are short-lived jobs stepping stones to long-lasting jobs? *Oxford Bulletin of Economics and Statistics 74*(5), 646–675.

Cockx, B. and M. Picchio (2013). Scarring effects of remaining unemployed for long-term unemployed school-leavers. *Journal of the Royal Statistical Society Series A 176*(4), 951–980.

Doiron, D. and T. Gørgens (2008). State dependence in youth labor market experiences, and the evaluation of policy interventions. *Journal of Econometrics 145*, 81–97.

Elbers, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazards model. *Review of Economic Studies 49*, 402–411.

Gørgens, T. and D. Hyslop (2016). The specification of dynamic discrete-time two-state panel data models. ANU Working Papers in Economics and Econometrics #631, Australian National University.

Gouriéroux, C. and A. Monfort (1991). Simulated based inference in models with heterogeneity. *Annales d'Economie et de Statistique 20/21*, 69–107.

Gritz, R. M. (1993). The impact of training on the frequency and duration of employment. *Journal of Econometrics 57*(1–3), 21–51.

Ham, J. C. and R. J. LaLonde (1996). The effect of sample selection and initial conditions in duration models: Evidence from experimental data on training. *Econometrica 64*(1), 175–205.

Heckman, J. J. (1981). The incidental parameter problem and the problem of initial conditions in estimating a discrete time–discrete data stochastic process. In C. F. Manski and D. McFadden (Eds.), *Structural Analysis of Discrete Data with Econometric Applications*, pp. 179–195. Cambridge, Massachusetts: MIT Press.

Heckman, J. J. and B. Singer (1984a). Econometric duration analysis. *Journal of Econometrics 24*, 63–132.

Heckman, J. J. and B. Singer (1984b). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica 52*, 271–320.

Kamionka, T. (1998). Simulated maximum likelihood estimation in transition models. *Econometrics Journal 1*(1), C129–C153.

Keane, M. P. (1994). A computationally practical simulation estimator for panel data. *Econometrica 62*(1), 95–116.

Keane, M. P. and R. M. Sauer (2010). A computationally practical simulation estimation algorithm for dynamic panel data models with unobserved endogenous state variables. *International Economic Review 51*(4), 925–958.

Lerman, S. R. and C. F. Manski (1981). 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*, pp. 303–319. Cambridge, Massachusetts: MIT Press.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*(437), 162–170.

Moffitt, R. A. and M. S. Rendall (1995). Cohort trends in the lifetime distribution of female family headship in the United States, 1968–1985. *Demography 32*(3), 407–424.

Skrondal, A. and S. Rabe-Hesketh (2014). Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity. *Journal of the Royal Statistical Society Series C 63*(2), 211–237.

Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature 35*(4), 2006–2039.

Stewart, M. B. (2007). The interrelated dynamics of unemployment and low-wage employment. *Journal of Applied Econometrics 22*, 511–531.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of Applied Econometrics 20*, 39–54.

Table 1: RMSE for designs without random effects

Parameter	NLC	HKM	ISU	ISN
$\tau = 0.3$				
$\ln \alpha_1$	0.083	0.083	0.073	0.076
β_1	0.113	0.113	0.098	0.099
μ_1	0.096	0.096	0.088	0.088
γ	0.334	0.224	0.206	0.206
$\ln \alpha_2$	0.281	0.149	0.148	0.149
β_2	0.168	0.101	0.092	0.092
μ_2	0.338	0.252	0.238	0.239
$\tau = 0.5$				
$\ln \alpha_1$	0.083	0.083	0.073	0.077
β_1	0.113	0.113	0.098	0.098
μ_1	0.096	0.096	0.088	0.088
γ	0.467	0.299	0.246	0.247
$\ln \alpha_2$	0.265	0.144	0.143	0.145
β_2	0.157	0.092	0.086	0.086
μ_2	0.482	0.329	0.285	0.287
$\tau = 0.7$				
$\ln \alpha_1$	0.083	0.083	0.075	0.079
β_1	0.113	0.113	0.098	0.098
μ_1	0.097	0.096	0.089	0.089
γ	4.273	1.661	0.348	0.349
$\ln \alpha_2$	0.256	0.140	0.139	0.140
β_2	0.151	0.091	0.084	0.084
μ_2	4.267	1.665	0.380	0.382

NLC: estimation using non-left-censored individuals; HKM: estimation using Heckman's approach; ISU: importance sampling unnormalised; ISN: importance sampling normalised. Results for the parameters in the HKM auxiliary equation not shown. See text for DGP and implementation of estimators.

Table 2: RMSE for designs with random effects

		$N_2 = 250$			$N_2 = 500$		
Parameter	NLC	HKM	ISU	ISN	HKM	ISU	ISN
$\tau = 0.3$							
$\ln \alpha_1$	0.139	0.132	0.126	0.123	0.131	0.128	0.120
β_1	0.150	0.145	0.126	0.128	0.146	0.116	0.113
μ_1	0.152	0.147	0.126	0.123	0.145	0.121	0.117
$\ln \delta_1$	1.248	1.120	0.662	0.662	1.376	0.517	0.617
γ	0.282	0.193	0.177	0.175	0.164	0.142	0.138
$\ln \alpha_2$	0.135	0.112	0.107	0.110	0.117	0.112	0.104
β_2	0.153	0.137	0.090	0.114	0.107	0.074	0.086
μ_2	0.340	0.267	0.225	0.242	0.249	0.204	0.211
$\ln \delta_2$	0.336	1.593	0.304	0.344	0.532	0.284	0.316
$\tau = 0.5$							
$\ln \alpha_1$	0.128	0.125	0.121	0.117	0.123	0.123	0.114
β_1	0.147	0.143	0.125	0.125	0.144	0.116	0.113
μ_1	0.151	0.146	0.126	0.124	0.143	0.118	0.116
$\ln \delta_1$	1.379	1.000	0.490	0.491	0.939	0.491	0.452
γ	0.439	0.300	0.261	0.256	0.240	0.221	0.213
$\ln \alpha_2$	0.124	0.131	0.105	0.105	0.115	0.111	0.104
β_2	0.147	0.164	0.090	0.107	0.124	0.073	0.089
μ_2	0.491	0.356	0.284	0.289	0.300	0.259	0.251
$\ln \delta_2$	0.339	0.794	0.300	0.338	0.518	0.284	0.317
$\tau = 0.7$							
$\ln \alpha_1$	0.163	0.121	0.119	0.115	0.120	0.121	0.111
β_1	0.169	0.144	0.124	0.125	0.144	0.112	0.110
μ_1	0.160	0.143	0.127	0.126	0.141	0.119	0.117
$\ln \delta_1$	6.663	1.101	1.004	0.949	0.747	0.484	0.451
γ	4.369	1.077	0.440	0.429	0.712	0.370	0.352
$\ln \alpha_2$	0.133	0.101	0.104	0.103	0.120	0.109	0.102
β_2	0.193	0.096	0.089	0.103	0.117	0.072	0.092
μ_2	4.359	1.094	0.444	0.433	0.727	0.391	0.364
$\ln \delta_2$	0.584	0.338	0.304	0.335	1.538	0.288	0.322

NLC: estimation using non-left-censored individuals; HKM: estimation using Heckman's approach; ISU: importance sampling unnormalised; ISN: importance sampling normalised. Results for the parameters in the HKM auxiliary equation not shown. See text for DGP and implementation of estimators.