Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study

Michael Baker and Angelo Melino*
Department of Economics
University of Toronto
150 St. George St.
Toronto, Ontario, Canada
M5S 3G7

June 14, 1999

Abstract

We examine the behaviour of the nonparametric maximum likelihood estimator (NPMLE) for a discrete duration model with unobserved heterogeneity and unknown duration dependence. We find that a nonparametric specification of either the duration dependence or unobserved heterogeneity, when the other feature of the hazard is known to be absent, leads to estimators that are well behaved even in modestly sized samples. In contrast, there is a large and systematic bias in the parameters of these components when both are specified nonparametrically, as well as a complementary bias in the coefficients on observed heterogeneity. Furthermore, these biases diminish very gradually as sample size increases. We find that a minor modification of the quasilikelihood that penalizes specifications with many points of support leads to a dramatic improvement.

*We gratefully acknowledge the research support of SSHRC. We thank David Green, John Ham, Jim Heckman, Tom Mroz and Gary Solon for helpful comments. Mike Campolieti provided invaluable research assistance.

1 Introduction

Hazard models of event durations find application in many areas of applied economics ranging from the analysis of unemployment spells to studies of fecundity. There is now widespread acknowledgment that inference in these models can be subject to specification error from a number of sources. One cause is unobserved heterogeneity arising from the omission of (possibly unobservable) variables that affect the hazard. A common antidote to this problem is to model the unobserved heterogeneity as individual-specific random effects. There is growing recognition, however, that inference can be sensitive to the assumed distribution of the unobserved heterogeneity. This issue is of particular importance when the hazard (the conditional probability of exit) exhibits duration dependence. As in many other areas of econometrics, nonparametric approaches have been proposed in the absence of any guidance from economic theory about functional form. In hazard models, the nonparametric approach has been applied with some success to both the specification of the baseline hazard (e.g., Ham and Rea 1987, Meyer 1990) and the distribution of unobserved individual effects (e.g., Nickell 1979, Heckman and Singer 1984).

Empirical researchers increasingly take heed of these results, but compromises are made. For example, semiparametric models, that is models made up of a mixture of parametric and nonparametric components, are often used. Typically, the hazard is characterized by a parametric form, but either the duration dependence or the unobserved heterogeneity distribution is modelled nonparametrically. While there is an obvious argument to be flexible in both dimensions, numerical difficulties often hinder the estimation of models with nonparametric specifications of both the baseline hazard and the unobserved heterogeneity (e.g., Meyer 1990, Baker and Rea 1998). In fact, largely in response to computational problems, the nonparametric estimators of either component are often only partially adopted. For example, the nonparametric maximum likelihood estimator (NPMLE) of the unobserved heterogeneity distribution proposed by Heckman and Singer (1984) in effect assumes that the unobserved random effects are drawn from a discrete distribution with unknown support and an unknown number of mass points. In many studies, the number of mass points is simply specified a priori as at most two, or the search for additional mass points is abandoned when numerical problems are encountered. Similar compromises are made in the specification of the duration dependence. In a common nonparametric

approach, the baseline hazard is specified as a step function. Therefore, the width of the steps must be chosen: narrow steps will allow for very flexible shapes for the duration dependence but may require enormous amounts of data for reliable estimation. In practice there appears to be a trade off between widening the step (and thus imposing more parametric structure on the duration dependence) and relaxing the specification of the unobserved effects (e.g., Narendranathan and Stewart 1993), so researchers must decide which of these two features to emphasize.

While particular choices may be necessitated by the realities of the data or machine arithmetic, there is no a priori reason to believe they are without consequence. In this paper, we attempt to provide some practical guidance to their effects for a class of discrete hazard models incorporating nonparametric specifications of unobserved heterogeneity and duration dependence.

We begin by outlining a computational strategy for the NPMLE with unobserved heterogeneity. This is a more familiar formulation of the strategy due to Lindsay (1983) that was adapted by Heckman and Singer (1984), and provides a systematic approach to determining the location and number of mass points for the heterogeneity distribution. We next provide Monte Carlo evidence on the behaviour of the NPMLE in a variety of environments. Both the data generating process (DGP) used to construct the sample data, and the quasi-likelihood (QL) specification used for estimation, vary across the experiments. The latter differ by the specification of the time functions used to model the true duration dependence. To anticipate our main results, we find that a nonparametric specification for either the duration dependence or unobserved heterogeneity, when the other feature of the hazard is known to be absent, leads to estimators that are well behaved even in modestly sized samples. However, the combination of a flexible specification for both duration dependence and unobserved heterogeneity leads to a very reliable and systematic bias. The estimated time functions are biased toward finding positive duration dependence and the NPMLE overestimates the dispersion of the random effects. Often both duration dependence and unobserved heterogeneity are of secondary interest, so it is important to emphasize that these faults lead to a large and significant bias for the parameters on observed heterogeneity, which we will refer to as β. It is well known that ignoring unobserved heterogeneity, even though it is independent of observed heterogeneity, biases β towards zero. It is perhaps not surprising that we find the exaggerated dispersion in the estimated heterogeneity distribution coincides

with a bias in β away from zero. Moreover, the bias is extremely large and goes away very slowly even as the sample grows to sizes that are enormous by today's standards. It appears that the poor sampling performance of the NPMLE of unobserved heterogeneity stems almost entirely from the fact that it finds too many "spurious" points of support. A minor modification of the QL to include a term that penalizes specifications with many points of support (for example, the Schwartz or Hannan-Quinn Information Criterion) seems like a natural solution, and was proposed in a similiar setting by Leroux (1992). We find that the penalized NPMLE leads to a dramatic improvement in the sampling properties of the various estimators and a much more reliable estimator for β.

2 A Computational Strategy

We next describe the algorithm for estimation of the NPMLE with unobserved heterogeneity. Aside from some minor details, the algorithm is due to Lindsay (1983) and can be found also in Heckman and Singer (1984). Nevertheless, we suspect that many applied researchers are unaware of this strategy, and can benefit from the more heuristic interpretation supplied here. Our suspicion is based in introspection (we developed the algorithm independently before finding it in the cited studies); the observation that estimation strategies adopted by many researchers in this area are essentially ad hoc; and, numerous conversations with researchers. Our presentation avoids the Gateaux differential used by Heckman and Singer and replaces it with the more familiar Kuhn-Tucker multiplier.

Assume that heterogeneity is indexed by a single parameter θ, which is known to lie in the interval $[\theta_L, \theta_H]$. Let $L_{ih} \equiv L_h(\alpha, \theta_i)$ denote the likelihood for individual h given the heterogeneity parameter takes on the value θ_i, where $\alpha \in \Re^p$ are other parameters. Suppose that the heterogeneity parameter is distributed as a discrete random variable with N_θ points of support. The loglikelihood for the individual is

$$\ln L_h \equiv \ln \left(\sum_{i=1}^{N_\theta} L_{ih} P_i \right) \tag{1}$$

where the probability weights satisfy $\sum P_i = 1$ and $P_i \geq 0$. The loglikelihood

for the sample is just

$$\ln L \equiv \sum_{h=1}^{N_h} \ln L_h. \tag{2}$$

The parameters of the heterogeneity distribution are N_θ (the number of points of support), along with $(\theta_i, P_i; i = 1, ..., N_\theta)$. We proceed assuming that these parameters are functionally independent of α. Given that N_θ is an integer, a sensible strategy is to pick a value for the number of points of support, say $\bar{N}_\theta$, estimate the remaining parameters and then check if the likelihood can be increased by adding an additional point of support to the distribution of θ.[1]

Suppose we have the MLE $\{\hat{\alpha}, \hat{\theta}_i, \hat{P}_i; i = 1, ..., \bar{N}_\theta\}$, conditional on the assumption that there are $\bar{N}_\theta$ points of support. Let $\bar{\theta}$ be a candidate for an additional point of support. Our approach is to consider the optimal values of $(P_1, ..., P_{\bar{N}_\theta+1})$, fixing $\alpha = \hat{\alpha}$, $\theta_i = \hat{\theta}_i$ for $i = 1, ..., N_\theta$ and $\theta_{\bar{N}_\theta+1} = \bar{\theta}$. If the optimal value of $P_{\bar{N}_\theta+1} = 0$, then adding a point of support at $\bar{\theta}$ can't increase the sample likelihood. If this is true for all $\bar{\theta} \in [\theta_L, \theta_H]$ then we have found the NPMLE. If instead the optimal value of $P_{\bar{N}_\theta+1} > 0$, then it is possible to increase the sample likelihood by adding a point of support at $\bar{\theta}$.

Formally we can write the problem as

$$max\{P_1, ..., P_{\bar{N}_\theta+1}\} \ln L = \sum_h \ln \left(\sum_{i=1}^{\bar{N}_\theta+1} L_{ih} P_i \right) + \lambda \left(\sum_{i=1}^{\bar{N}_\theta+1} P_i - 1 \right) + \sum_{j=1}^{\bar{N}_\theta+1} \mu_i P_i. \tag{3}$$

Notice that the objective function is concave in $\{P_i\}$ and the constraints are linear, so the optimum is characterized by the Kuhn-Tucker first order conditions

$$\frac{d \ln L}{d P_i} = \sum_h \left(\frac{L_{ih}}{\sum_j L_{jh} P_j} \right) + \lambda + \mu_i = 0, \tag{4}$$

$$\frac{d \ln L}{d \lambda} = \sum P_i - 1 = 0, \tag{5}$$

[1] While this is essentially the incremental strategy adopted in some studies in the area, in practice it can be problematic if the location of the additional mass point is "poorly" chosen. Our point in this section is to suggest a systematic approach to identifying the location of additional points of support. An alternative strategy would be to estimate the model using a grid of values for the θ_i over a "reasonable" range. See for example Lemieux and MacLeod (1995).

and

$$P_i \geq 0; \quad \mu_i \geq 0; \quad \mu_i P_i = 0. \tag{6}$$

It's not possible to solve for the optimal probabilities in closed form from the first order conditions, but we can use them to evaluate a particular candidate. First, multiply (4) by P_i, sum over i, and use equations (5) and (6) to obtain $\lambda = -N_h$. Substituting back into (4) we eliminate λ, and obtain

$$\mu_i = \sum_h \left(1 - \frac{L_{ih}}{\sum_j L_{jh} P_j}\right). \tag{7}$$

Consider the candidate $P_i = \hat{P}_i$ for $i = 1, ..., \bar{N}_\theta$, and $P_{\bar{N}_\theta+1} = 0$. Because the $\{\hat{P}_i\}$ maximize the likelihood for $\bar{N}_\theta$ points of support, it is easy to show that μ_i will be zero for $i = 1, ..., \bar{N}_\theta$ (the right hand side of (7) is just the gradient of the loglikelihood (2) with respect to these probabilities). If $\mu_{\bar{N}_\theta+1} \geq 0$ then the first order conditions are satisfied at the candidate probabilities so the likelihood can't be improved by adding a point of support at $\bar{\theta}$. If this is true for all $\bar{\theta} \in [\theta_L, \theta_H]$ then we have found the NPMLE. If, on the other hand, $\mu_{\bar{N}_\theta+1} < 0$ for some choice of $\bar{\theta}$, then the candidate probabilities violate the first order conditions and it is possible to increase the sample likelihood by adding a point of support at $\bar{\theta}$.

Heckman and Singer (1984) show that the (negative of the) Kuhn-Tucker multipliers given by (7) are Gateaux derivatives, and a necessary and sufficient condition for the NPMLE (using our notation) is $\mu_{\bar{N}_\theta+1} \geq 0$ for all $\bar{\theta} \in [\theta_L, \theta_H]$. This result suggests the following algorithm:

Step 0: Set $\bar{N}_\theta = 1$ and $P_1 = 1$. Choose initial values for α and θ_1.

Step 1: Given the current value of $\bar{N}_\theta$, maximize the likelihood over α and $(\theta_i, P_i; i = 1, ..., \bar{N}_\theta)$.

Step 2: Evaluate $\mu_{\bar{N}_\theta+1}$ for a grid of values of $\bar{\theta} \in [\theta_L, \theta_H]$.

a. If $\mu_{\bar{N}_\theta+1} \geq 0$ for all choices of $\bar{\theta}$ then **STOP**.

b. Else, set $\theta_{\bar{N}_\theta+1}$ to the value of $\bar{\theta}$ that yields the smallest value for $\mu_{\bar{N}_\theta+1}$.

Step 3: Solve (3) numerically to obtain new initial values for the probabilities. Increase the value of $\bar{N}_\theta$ by 1. Return to Step 1.

In our application, the QL is concave in the parameters (α, θ), so the initial values chosen in Step 0 can be arbitrary. The problem solved in Step 3 is

also very well behaved; we use the reduced gradient method described in Luenberger (1984, Ch. 11). With more than one point of support, Step 1 is the most time consuming and difficult due to the vagaries of nonlinear estimation. We used a modified Newton-Raphson algorithm, but the finite mixture literature often relies on the EM algorithm (see Lancaster (1990), ch. 8). Step 2 appears to work reasonably well in practice (but see section 6.3 for a suggested improvement).

The algorithm guarantees that the likelihood will increase each time a point of support is added. Heckman and Singer (1984) describe a strategy for choosing $[\theta_L, \theta_H]$ in a systematic way that must include any additional point of support that can increase the QL. We were unable to implement their suggestion. Instead, we chose this interval implicitly so that the probability of exit in the first period for all types was restricted to the set $[\varepsilon, 1-\varepsilon]$. Initially, we set ε to 10^{-5}, but this value was decreased adaptively if the estimated points of support approached a boundary.

In practice, we make a few minor modifications to this algorithm to deal with computational problems. In Step 1, we reparameterize the probabilities using a logistic transformation to keep them in the unit simplex so that we can use familiar and well tested algorithms for unconstrained maximization. Unfortunately, this means that probabilities can be made small but never set exactly to zero. So, before evaluating the Gateaux derivative we take estimates from Step 1 and use the same code as in Step 3 to maximize the likelihood over the $\bar{N}_\theta$ probabilities. This small detour allows us to drop points of support with probabilities that should have been set to zero and it lessens the possibility that in Step 2 we will add a point of support more or less on top of one that has already been included[2]. Finally, if adding a point of support in Step 3 leads us to drop one of the existing points, we do not increase $\bar{N}_\theta$. This situation can arise because we are not guaranteed to obtain a global optimum in Step 1.

[2]It is not numerically possible to achieve a gradient of the log-likelihood that is exactly zero, but we found that Step 2 can be sensitive to even very small errors in computing the optimal probabilities. After a good deal of experimenting, we concluded (at least in our implementation) that inserting an extra step to 'polish' the estimated probabilities had no noticeable impact on the parameter estimates but it improved the numerical robustness of our code. In those cases where a point of support was dropped or we achieved even a modest increase in the loglikelihood, we returned immediately to Step 1.

3 Related Research

Although duration data used in economic applications are always reported in discrete units (eg., days or weeks) and are therefore integer valued, the econometric literature is dominated by continuous-time duration models in which the durations can take on all values on the positive real line. In particular, many theoretical and applied studies adopt the mixed proportional hazard model (MPH). The leading special case of the MPH has an instanteous hazard rate of the form

$$\lambda_{ht} = \exp(X_h\beta + f(t) + \theta_h) \tag{8}$$

where the three components in (8) represent observed heterogeneity, duration dependence, and unobserved heterogeneity, respectively. θ_h is assumed to be an i.i.d. draw from some unknown distribution that must be estimated along with β and the parameters of the baseline hazard $\exp(f(t))$.

Several analytical results are known for the MPH model. Invoking fairly general assumptions, Elbers and Ridder (1982) showed that the parameters of the MPH model are identified (but see Ridder 1990 and Ishwaran 1996), and Heckman and Singer(1984) proved that the NPMLE is consistent. The asymptotic distribution of the NPMLE for this model is not yet known, but Hahn (1994) and Ishwaran (1996) show that, at least for a leading special case, it is not $\sqrt{n}$ consistent.[3]

Some Monte Carlo evidence of the sampling distribution of the NPMLE for the MPH model is also available. Heckman and Singer (1984) reported the results from roughly a dozen artificial samples. They considered a rich variety of heterogeneity distributions but restricted attention to a tightly parameterized model of duration dependence (i.e. Weibull models). They found that the NPMLE reproduced the structural parameters fairly well but provided an unreliable estimate of the mixing distribution. Ridder (1987) showed that if the durations are uncensored and the baseline hazard is known then the ML estimates of β are insensitive to misspecification of the mixing distribution. Using a one-parameter model for $f(t)$, he found that estimation of the baseline hazard did not much affect this conclusion. Ridder did find, however, that very heavy censoring (about 80%) could generate a bias in β as

[3]Bearse et al (1996) show how to construct a semiparametric estimator using kernel methods that is generically $\sqrt{n}$ consistent and asymptotically normal. Campolieti (1997) uses a Dirichlet process prior to construct an estimator that is effectively nonparametric but amenable to standard Bayesian inferential techniques in finite samples.

large as 15%. More recently, Huh and Sickles (1994) compare the NPMLE with alternatives that smooth the estimated distribution function for the Weibull proportional hazard model. They also restrict attention to a one-parameter model for duration dependence. They find that the NPMLE works reasonably well although there appears to be a small sample advantage to their alternative estimators.

The discrete duration model that we investigate in this paper has been used by many authors. We suspect that this model should lead to estimates with properties that are very similiar to those of the MPH. We know of no analytical results for this model, however, and comparisons of our results with Monte Carlo results for the MPH model can be viewed at best as suggestive.

Closely related to our discrete duration model is the literature on binary choice. Because it is an index model, the results of Coslett (1983) can be used to show that the regression coefficients in our discrete duration model are identified at least up to scale, even if we allow the exit probabilites to vary freely with duration. Cameron and Heckman (1998) provide identification results for a general class of discrete duration models. They show that identification is enhanced if the index varies with duration. If the index is constant, as is the case in our model, then identification is delicate and requires some additional structure on the transition probabilities. Because it has a discrete factor structure, their Theorem 4 shows that our model is identified, at least if we restrict attention to finite mixture distributions. We have been unable to extend their results to the general case.[4]

Finally, there is a large statistical literature on the estimation of mixing distributions (see Lindsay and Lesperance 1995 for a recent survey). While relevant and suggestive, much of this literature is difficult to adapt to our setting because it excludes the case where there are other parameters of interest. In econometric applications, the main interest is often in the parameters α that are implicit in (1). Of course, it is possible that the MLE estimates of α may be efficient and asymptotically normal even though these properties do not extend to the estimated mixing distribution (Van der Vaart 1996).

[4]We can show that if the duration dependence function is known to be constant for the first two periods, then the model is identified for general mixtures.

4 Sample Design

We investigated the properties of the NPMLE in a variety of settings. These settings varied with (a) the true data generating process (DGP) for the data; (b) the quasilikelihood (QL) used for estimation; and (c) the number of sample observations used in estimation.

4.1 DGPs

In calibrating our DGP, we tried to choose true hazards that resembled those typically observed in data on unemployment spells, measured in weeks. Ham and Rea (1987) report empirical hazards of .171, .170, .098 and .066 for the first four weeks of UI insured unemployment spells in Canada, at a time when most spells were insured. Sider (1985) estimates that the continuation probability out of the first month of unemployment ranged from .41 to .59 in the U.S. over the period 1969 to 1982. This implies (constant) weekly hazards over the first month of unemployment ranging from 0.12 to 0.19[5]. These estimates provide some room for choice. As a benchmark, we tried to generate hazards with the probability of exiting in the first period equal to about .15, with about half the sample exiting by the fourth week, and that were declining with duration.

We assumed that the probability that a given observation h in our sample survived from time t-1 to time t, hereafter referred to as the *continuation function*[6], was of the logistic form, that is

$$S_{ht} = \frac{\exp(z_{ht})}{1 + \exp(z_{ht})} \tag{9}$$

where the index is of the form

$$z_{ht} = X_h\beta + f(t) + \theta_h \tag{10}$$

The three terms represent observed heterogeneity, duration dependence, and unobserved heterogeneity, respectively. We fixed the values of X across all our simulations, but we considered 3 cases for the duration dependence and

[5]These estimates are constructed assuming 4.3 weeks per month. Meyer (1990) reports that the empirical hazards for the first four weeks of insured spells in the US are .082, .066, .056, and .061. for the period 1978-1983.

[6]The discrete time hazard is one minus the continuation probability.

3 cases for the unobserved heterogeneity. For each of these nine DGPs, we generated 100 random samples of size 5000 and stored them[7]. These became the data for our NPMLE calculations. For computational reasons, we censored observations that lasted more than forty periods; that is, we decided to record durations of greater than 40 as incomplete durations that lasted at least 40 periods. Figure 1 shows the average values of the hazards obtained from each of our nine DGPs. We now discuss in turn how we chose each of the components of (10)

4.1.1 Observed Heterogeneity

Because of computational costs, we only considered the case where observed heterogeneity is summarized by a scalar. To ease interpretation and without loss of generality, we set $\beta = 1$ in all our simulations. Many authors have warned that various properties of estimators can be hidden in a design that considers only symmetric distributions for X. Nonetheless, it is also useful to maintain comparability with previous studies, so we followed Heckman and Singer (1984), Ridder (1987), and Huh and Sickles (1994) and assumed that $X \sim NID(0, \sigma_x^2)$. Setting $E(X) = 0$ is a useful normalization that clarifies the interpretation of the remaining parameters. The choice of σ_x^2 is less obvious. The value of σ_x^2 determines the relative importance of observable heterogeneity and is a key determinant in practice of not only how accurately we can estimate β but whether or not we can distinguish duration dependence from unobserved heterogeneity. One way to choose this parameter is to try to match the R^2 from a regression of the log of duration on X to values typically observed. On this basis, we chose $\sigma_x^2 = 0.25$. This put the average R^2 in our DGP with no duration dependence and no unobserved heterogeneity at about .08. Moreover, this value for σ_x^2 more or less kept the average R^2 in all of our DGPs described below in the range .05-.10[8].

4.1.2 Duration Dependence

We considered three forms for the true duration dependence: none, negative duration dependence, and positive duration dependence. Setting $f(t) \equiv 0$

[7]All the random numbers were generated using IMSL routines DRNUNF and DRNGAM. We used the multiplier 950706376 and shuffling.

[8]Using the data from Baker and Rea (1998), the R^2 for log duration of weekly unemployment spells is about .05.

allows us to gauge the efficiency loss that comes from trying to allow for duration dependence when none is present. The case where, other things equal, the hazard is declining with duration is called negative duration dependence. Of course, a declining hazard is equivalent to a rising continuation function, so that in our parameterization negative duration dependence is associated with an *increasing* path for $f(t)$. We model negative duration dependence by setting

$$f(t) = 1 - \exp(\frac{1-t}{5}) \tag{11}$$

Notice that $f(1) = 0$. This normalization was chosen to facilitate the interpretation of the heterogeneity parameter θ. The time function rises smoothly and asymptotes to the value 1. The case of negative duration dependence seems to be the most empirically relevant. For completeness, we also considered the case of positive duration dependence where the time function $f(t)$ was set equal to the negative of the right hand side of (11).

4.1.3 Unobserved Heterogeneity

We assumed that the unobserved heterogeneity parameter θ_h was a random draw from a given distribution[9]. We considered three cases: a degenerate distribution (that is, no unobserved heterogeneity); a discrete distribution with two points of support; and a (translated) gamma distribution. In the case of no unobserved heterogeneity, we simply set $\theta_h = 1.8$. This value was chosen so that the unconditional probability of exiting at time t=1 was about .15. For the discrete distribution, we assumed that each of the two points was equally likely. We set the mean to 1.8 and the variance to unity. This uniquely determined the points of support at $\theta = 0.8$ and $\theta = 2.8$. The gamma distribution is often used to model heterogeneity in continuous time hazard models. In contrast to the previous choices, it has a density and the distribution isn't symmetric about the mean. We first tried to draw gamma random variables from a distribution with mean 1.8 and variance 1, matching the same two moments as in our discrete heterogeneity distribution case. However, we were unhappy with the resulting hazard shapes. After some

[9]None of the estimators that we consider in this paper deal with the possibility that the unobserved heterogeneity may be correlated with X, but it is easy to see what the consequences of such correlation would be in one special case. Suppose we model unobserved heterogeneity as $\widetilde{\theta} = \theta + \lambda X$, where θ is a random effect. This model can be mapped into our setup by redefining the coefficient on X as $\widetilde{\beta} = \beta + \lambda$.

experimentation, we first drew θ_h^1 from a gamma distribution with mean .5 and variance 1, and then constructed $\theta_h = \theta_h^1 + 1.3$[10].

4.2 Quasilikelihoods

We took as the QL the likelihood formed from (9) and (10) but with the true duration dependence $f(t)$ proxied by $\phi(t)$ where the assumed duration dependence $\phi(t)$ took on one of three forms : none, a cubic polynomial in duration, and a "nonparametric" step function.

By setting the duration dependence in the QL to zero when that is the right thing to do, we get some idea of the value of knowing the true duration dependence for estimating the distribution of unobserved heterogeneity. Similarly, by ignoring duration dependence in the QL when the true DGP has duration dependence, we get some idea of how well the duration dependence in the true DGP can be "mopped up" by a flexible parameterization of the unobserved heterogeneity.

Low order polynomials are often used to model duration dependence. Notice that neither the negative nor the positive duration dependence in our true DGP is due to a cubic polynomial, so we can get some idea of whether or not small errors in the parameterization of the duration dependence have large consequences for parameter estimation. For numerical reasons, we found it useful to normalize the time polynomial, so in the cubic case we set

$$\phi(t) = \alpha_1 \frac{(t-1)}{10} + \alpha_2 \frac{(t-1)^2}{100} + \alpha_3 \frac{(t-1)^3}{1000} \tag{12}$$

Notice that we normalize $\phi(1) = 0$, to facilitate comparison between the estimated and true heterogeneity parameters.

Because our DGPs yield discrete survivor data taking values from 1 to 40, we can easily specify a nonparametric model for the duration dependence by using a step function, namely

$$\phi(t) = \sum_{\tau=2}^{40} \phi_\tau D_{t\tau} \tag{13}$$

[10]Our designs set $\mathrm{Var}(\theta)/\mathrm{Var}(\beta X)$ to 0 or 4 depending on whether or not we have unobserved heterogeneity in the true DGP. In Section 6, we report some results where this ratio is 1. Our results do not appear sensitive to the value of this ratio, but we had no idea what a sensible value would be. We would like to thank a referee for bringing to our attention that Lancaster (1979) using a continuous- time model reports a ratio of 0.5.

where $D_{t\tau}$ is a dummy variable that takes on the value 1 if $t = \tau$ and zero otherwise, and the ϕ_τ are coefficients. However, the specification (13) introduces 39 new parameters and would be computationally very demanding, especially for a Monte Carlo study. As mentioned previously, researchers often compromise on the fully nonparametric specification by restricting some of the coefficients in the step function to be equal. We adopt this strategy and use a step function where the coefficients ϕ_τ are unrestricted for $\tau = 2, .., 20$, but the remaining coefficients are grouped so that $\phi(t)$ is constant over the intervals of length four given by $\tau = 21 - 24, 25 - 28, .., 37 - 40$. This specification for duration dependence is still very flexible, but reduces the number of parameters in $\phi(t)$ from 39 to a slightly more manageable 24.

4.3 Sample Size

We complete our design by investigating the effect of sample size on the NPMLE. We consider 3 values for the number of observations in the sample: 500, 1000 and 5000. The experiments involving a sample size of N are constructed using the first N observations of the 5000 generated from the true DGP. This means, for example, that in the experiments using a sample of size 1000 we simply add 500 observations to those used in the experiments involving a sample size of 500. This facilitates comparisons across the different sample sizes and reduces sampling error.

In all, therefore, we have 81 experiments in our sample design constructed from 9 DGPs, 3 specifications for the QL, and 3 sample sizes.

5 Monte Carlo Results

We organize the results in four steps. First we describe what happens in a single sample to both the parameters and to the Gateaux derivative as we add points of support using, respectively, table 1 and figure 2. In the second subsection, we look at the sampling distribution of the coefficient on X (the source of observable heterogeneity). Here the results are summarized in table 2. The third subsection summarizes the results for the 'nuisance' parameters, that is, the fitted time function used to model duration dependence and the estimated heterogeneity distribution. The last section looks beyond the parameter estimates and compares the true and fitted hazard functions.

5.1 A detailed look at a single sample

The example is based on a DGP that incorporates negative duration dependence and a discrete distribution of unobserved heterogeneity with two points of support. In the QL, the duration dependence is approximated by a cubic. In table 1, for each iteration of the search algorithm presented in section 2, we report: estimates of the parameter on the observable heterogeneity β; the parameters of the cubic in duration (α_1-α_3); and, the points of support of the distribution of unobserved heterogeneity along with their associated probabilities (θ_i and P_i).

The results using a sample of size 500 are reported in the first panel. On the first iteration we obtain an estimate of β which is far below its true value of 1.0. The cubic in duration exhibits negative duration dependence over the relevant range, but it lies everywhere above the true $f(t)$ given by (11). Note that at this point we effectively ignore unobserved heterogeneity and observe the usual biases. The parameter on observable heterogeneity is biased towards 0, and we overestimate the degree of negative duration dependence in the hazard. Moreover, the magnitude of these biases is large. The Gateaux derivative (recall that this is the *negative* of the KT multiplier of Section 2) is evaluated over a grid of candidate θ's ranging from -11.51 to 11.51. The profile of the derivative across the grid for this iteration is reported as the solid line in panel A of figure 2. We see here a result which is quite common across the samples. The maximum is achieved at a corner solution: the algorithm indicates that the next point of support should be entered at the lower boundary of the grid (-11.51). This result is reported in the second last column of table 1. Finally, in the last column of the table we report the value of the probability associated with the new point of support as calculated in Step 3 of the algorithm.

In the second iteration there are marginal changes in the estimates of β and the α_i, and the estimate of the new point of support ends up at a more extreme value than the initial guess of -11.51. The profile of the Gateaux derivative (panel A of figure 2) is now sharply lower in the negative range of the grid (we have suppressed its value at $\hat{\theta}_2$ (-25.560) to maintain a reasonable scale in the figure). The new maximum of the Gateaux derivative is located at -0.588. Note that in this iteration the QL is "correctly specified " in the sense that it incorporates a distribution of heterogeneity with two points of support. Nevertheless, the probability associated with the second point is very small and the estimate of β is still too small.

In contrast, in the third iteration the estimate of β is almost equal to 1.0. Also, the weight in the distribution of unobserved heterogeneity is nearly equally distributed across two points of support (θ_1 and θ_3) that are very close to the true values, while the second point continues to have small probability. Finally, the cubic exhibits less negative duration dependence than in earlier iterations. It now matches the true $f(t)$ fairly closely up to about $t = 8$ (about the first three-quarters of the sample) before diverging. Note that the Gateaux derivative displays a similar profile to the previous iteration and the suggested choice for the next point of support is again just less than 0.

In the fourth iteration the point of support at -25.56 is dropped, and the weight in the heterogeneity distribution is re-distributed almost equally across the remaining 3 points. This mis-specification in the QL has clear effects on the estimates of the other parameters: the estimate of β has "moved beyond" 1.0, and the cubic now shows signs of positive duration dependence. Similar trends are observed in the fifth and final iteration. The MLE of β is more than 170 percent of its true value, and there is greater evidence of positive duration dependence in the $\hat{\alpha}_i$. This latter finding is not surprising given the well known result that unobserved heterogeneity can lead to spurious inference of negative duration dependence. Here, as the distribution of unobserved heterogeneity is over parameterized, the resulting "excess" negative correlation between the hazard and time is offset by positive duration dependence in the cubic.

In the second and third panels of table 1 and panels B and C of figure 2, we document the iterations in samples of 1000 and 5000 using the same DGP and QL. Very similar patterns are apparent. First, the maximum of the Gateaux derivative is often obtained at a corner solution. Second, β is estimated with reasonable precision on the iteration for which the QL is correctly specified.[11] Note also the rough congruence in the estimates of the α_i across samples on this iteration. Third, the NPMLE leads to an over-parameterization of the distribution of unobserved heterogeneity. Fourth, the MLE of the cubic displays relative positive duration dependence to offset this mis-specification.

The clear message of these examples is that in the absence of other information, maximizing the likelihood function leads to over-parameterization

[11]The iteration for which the QL is correctly specified is number 3 for sample sizes 500 and 1000, and number 2 for sample size 5000.

of the unobserved heterogeneity, and important biases in the estimates of other parameters of the model. This tendency is not attenuated to any large degree as sample size grows over the range typically encountered in longitudinal data sets. In the next section we provide evidence that this conclusion remains true across a wide variety of DGPs and QLs.

5.2 Sampling distribution of β

In table 2 we report the mean and standard deviation for the estimates of β in the 81 experimental settings[12]. For each of our nine DGPs, we took the 100 simulated samples and estimated β using each of the three QLs and the three sample sizes. To show how the estimates vary as we add points of support to the distribution of the unobserved heterogeneity, we report the statistics at 4 values for $\hat{N}_\theta$: $\hat{N}_\theta = 1$; $\hat{N}_\theta = 2$; $\hat{N}_\theta = \hat{N}_\theta^{MLE}$ and $\hat{N}_\theta = \hat{N}_\theta^{HQ}$ (defined below).

The results when $\hat{N}_\theta = 1$ allows us to judge the consequences of ignoring unobserved heterogeneity whether or not that is the right thing to do. Many researchers fix the value N_θ to 2, a priori. The results when $\hat{N}_\theta = 2$ demonstrate the performance of this "parametric" specification. The third column contains the descriptive statistics for the estimated β when $\hat{N}_\theta$ is chosen according to the nonparametric MLE described in section 2. Our fourth column is based on estimating $\hat{N}_\theta$ by the Hannan-Quinn Information Criterion (HQIC). Information criterion are typically of the form

$$\ln L - c\,p \tag{14}$$

where p is the number of parameters in the model and c is a penalty function. The Schwarz or Bayesian Information Criterion (BIC) sets $c = \ln(N_h)/2$. The HQIC sets $c = \ln(\ln(N_h))$. The BIC was originally proposed for a choosing the explanatory variables in regression setting and the HQIC for determining the order of an autoregressive model[13]. Leroux (1992) proposes a penalized MLE based on (14) for pure mixture models and provides conditions under

[12]Our calculations were done on a Pentium 133 PC. We used a Watcom Fortran compiler. For Step 1, we reparameterized the probabilities via a logistic transformation to constrain them to the unit simplex. We then used repeated calls to the DFP algorithm from GQOPT to guarantee convergence. The calculations in Table 2 took about 5 months of CPU time.

[13]Sin and White (1996) survey the properties of various information measures in a wide variety of settings.

which it will lead to a consistent estimator of the mixing distribution. Because his conditions do not apply directly to our setting, we propose using the HQIC as an ad hoc rule whose properties deserve further exploration[14].

The results in table 2 lead to the following conclusions about the estimates of β:

First, a nonparametric specification for either duration dependence or unobserved heterogeneity, when the other feature is known to be absent, leads to estimates that are well behaved for all sample sizes considered.

Second, mis-specification matters. Although it is difficult to distinguish unobserved heterogeneity from duration dependence, it is not sufficient to model one of these features in a flexible way while ignoring the other as suggested by some authors (e.g. Meyer 1990, Ridder 1987). This practice can lead to significant biases.

Third, the combination of a flexible specification for both duration dependence and the distribution of unobserved heterogeneity leads to a large and systematic bias for the estimated β that declines very slowly with sample size. It appears that the poor sampling performance stems almost entirely from the fact that the NPMLE finds too many "spurious" points of support and overestimates the dispersion of the unobserved heterogeneity while compensating for this mis-specification through the estimate of the duration dependence. The excessive dispersion causes the hazard to decline too sharply with time. A flexible specification for $\phi(t)$ in the QL allows the NPMLE to offset the effects of excessive dispersion of the unobserved heterogeneity on the hazard with spurious positive duration dependence. Notice that this interaction and the resulting problems do not arise when there is no duration dependence in either the DGP or the QL. Excessive dispersion of the unobserved heterogeneity also leads to a large bias in the estimated β. It is well known that ignoring heterogeneity biases the estimate of β towards zero. We find that the NPMLE leads to an estimated β that is biased away from zero. Moreover, this bias is so large that, for the sample sizes considered, researchers would be better off ignoring the unobserved heterogeneity altogether.

Our fourth finding is that the problems associated with the NPMLE are greatest when the true DGP displays negative duration dependence. Unfor-

[14]Using the BIC in place of the HQIC sometimes leads to the selection of models with fewer points of support, but the statistics reported in Table 2 would be virtually identical if we used the BIC rather than the HQIC.

tunately, this is probably the case of greatest interest to applied researchers.

Somewhat ironically, the computational difficulties faced by applied researchers have served to avoid the large biases associated with the NPMLE. Numerical difficulties due to machine arithmetic, ad hoc algorithms for choosing additional points of support that fail to guarantee a strict increase in the likelihood, and the general difficulty in optimizing over all the parameters in Step 2 of our computational algorithm, all tend to reduce the number of points of support found for the heterogeneity distribution. A more defensible approach to eliminating "spurious" heterogeneity is to use an information criterion. Our last finding is that choosing $\hat{N}_\theta$ to maximize the HQIC leads to a dramatic improvement over the NPMLE. This is perhaps not surprising when the true DGP is discrete with only one or two points of support. But note that we find qualitatively similar results when the true DGP has Gamma heterogeneity. In many cases, we do almost as well in large samples as we could if we actually knew the true value of N_θ. However, for sample sizes of 500 or 1000, we see (at least with some DGPs) that the penalized MLE displays a slight negative bias when the QL has either a cubic or a step function. The HQIC in these situations is conservative in that better parameter estimates would have been obtained if the penalty on adding points of support was reduced.

In figure 3, we plot a kernel density estimate of the sampling distribution for $\widehat{\beta}$ obtained using the HQIC, for a few cases. We also plot a normal density standardized to have the same mean and variance. To conserve space, we restrict attention to the case where the DGP has no duration dependence, the QL has a cubic polynomial and the sample size is 5000. Because we only have 100 draws for each experiment, these graphs provide at best a rough guide to the unknown large sample distribution of the estimated β. Nonetheless, it appears that in these three cases, a normal approximation to the sampling distribution seems to work reasonably well. Unfortunately, in other cases we find that a normal approximation is much less reliable. In smaller samples, the $\widehat{\beta}$ obtained using the HQIC often has bimodal distribution. The second mode appears to decline with sample size, but in several of our designs, the second mode is still noticeable even with a sample size of 5000. The corresponding plots for the NPMLE estimator of β (not shown) reveal that although it is very badly biased in many cases, the sampling distribution appears to be fairly well approximated by a normal density in large samples for all the DGP/QL pairs in our design.

Although there is no asymptotic theory to justify it, practitioners rou-

tinely compute and report standard errors for the NPMLE based on inverting the Hessian. Table 3 gives the average standard errors for $\widehat{\beta}$ so obtained at $\hat{N}_\theta = 1$, $\hat{N}_\theta = 2$, $\hat{N}_\theta = \hat{N}_\theta^{MLE}$, and $\hat{N}_\theta = \hat{N}_\theta^{HQ}$ for the same DGPs and QL used to construct figure 3, but using a smaller sample size of 1000. Comparing to Table 2, we see that these standard errors tend to underestimate the sampling standard deviations and these biases are particularly large for both the NPMLE and the penalized MLE.

5.3 Sampling Distribution of Estimated Duration Dependence and Heterogeneity Distribution

The difficulty in estimating β is reflected in the sample estimates of the duration dependence function and the distribution of unobserved heterogeneity. To save space, we restrict attention to two representative experiments. The sample size in each case is 1000 observations. The DGP for both cases is identical, but the specifaction of duration dependence in the QL varies. For each experiment, we computed the difference between the average of the fitted value of $\phi(t)$ and the true time function $f(t)$. The results are plotted in figure 4. These plots illustrate a general result: biases in the estimated β and the biases in the estimated duration dependence function are closely related. Estimates of β below 1 are usually accompanied by excessively negative duration dependence. Conversely, estimates of β above 1 are usually accompanied by positive bias in the estimated duration dependence.

For example, in panel A if figure 4 the true DGP displays negative duration dependence and a two-point heterogeneity distribution, and the QL has a cubic time polynomial. The average estimated value of $\phi(t)$ obtained with the penalized MLE overestimates $f(t)$. Note from table 2 that the average value of β in this case is 0.885. In contrast, using the NPMLE leads to an estimated value of $\phi(t)$ that is on average below the true value. Recall that in our parameterization, negative values of $\phi(t)$ correspond to a hazard that is *increasing* with duration, so the NPMLE is biased toward positive duration dependence. This bias in the time function matches up with a bias in β; the average estimated value here is 1.247 (table 2).

The bias in the estimated duration dependence can have important implications. For example, as noted by Heckman and Singer (1984), many search models of unemployment predict positive duration dependence in the unemployment hazard because an individual's reservation wage declines the

longer she is unemployed. The negative duration dependence found in many studies of unemployment data therefore implies a rising reservation wage. Heckman and Singer argue that mis-specification of unobserved heterogeneity may lie behind this counterintuitive result. They re-examine a sample of unemployment spells in which many previous studies had found evidence of negative duration dependence. Although ignoring the unobserved heterogeneity or trying to capture it with a tightly parameterized heterogeneity distribution leads to estimates of negative duration dependence, they find that the NPMLE indicates positive duration dependence in these data. Our results suggest that this reversal in the estimated duration dependence may be due to the small sample bias associated with the NPMLE rather than the relaxation of an incorrect specification. More generally, our results suggest that finding positive duration dependence is even more likely if we combine the NPMLE with a flexible specification of the baseline hazard.

The distribution of unobserved heterogeneity is usually treated as a nuisance parameter and is rarely of direct interest. In part, this reflects the belief that this distribution cannot be estimated with much precision. We find that the NPMLE is in fact a poor estimator. It often leads to an estimated distribution that is incorrectly centered and tends to put to much probability on extreme values for θ. In contrast, use of the HQIC leads to a much more reliable estimator. In figure 5, we plot the average estimated distribution function of unobserved heterogeneity obtained via the HQIC for the two cases described above. The agreement between the true and average fitted cumulative distribution function is remarkably good, and it clearly improves with sample size.

5.4 Predicted Hazards

Although the separate components of the hazard are difficult to estimate and can be subject to large biases, particularly if we overparameterize, maximizing the QL virtually guarantees a close fit between the observed and the predicted hazards. Let $\lambda_t(X)$ denote the exit hazard at time t for an individual with observable characteristics X. The predicted hazard is constructed as a weighted sum of the hazards $\lambda_{it}(X)$ for an individual of type i. The weights $w_{it-1}(X)$ give the probability that an individual with observable characteristics X who has survived to time $t-1$ is of type i. With discrete

types, we have

$$\lambda_t(X) = \sum_i \lambda_{it}(X)\, w_{it-1}(X)$$

In figure 6, we plot the difference between the true hazard and the average of the predicted hazards for three values of X (the observed heterogeneity). The predicted hazards are formed using the NPMLE estimates. The three values of X chosen are its mean, and its mean plus/minus two standard deviations. For convenience, we restrict ourselves to the same DGP/QL pairs as in the previous section. We see that even though the separate components of the true hazard may be poorly estimated, the various biases are almost always combined in a way such that the difference between the true and predicted hazards is very small. The exceptions occur in a setting where an individual with characteristic X is unlikely to survive in the sample for very long (e.g., X=-1). In such a case, the observed hazard at long durations is uninformative and cannot impose agreement between the true and predicted hazards.

Many questions of interest to researchers depend on the structural parameters only through the predicted hazard. Notable examples are the expected duration of a spell given the vector of observables X_1, or the expected change in duration if the observables are varied from X_1 to X_2. For such questions, the difficulty in estimating the structural parameters is largely irrelevant if the values of X_1 and X_2 are 'well-represented' in the sample, and we can ignore censoring.

In practice, the predicted hazard often leads to a defective distribution for duration (that is, it gives a non-zero probability to the event that the spell never ends), so the expected duration is not well defined. Let $1_{t<T^*}$ denote the indicator variable that equals 1 if the duration t is less than T^*, and zero otherwise. It is convenient to summarize the fitted hazard via the estimates for the two quantities $E(1_{t<T^*}t|X)$ and $(1-E(1_{t<T^*}|X))$ at various values for T^*. As T^* increases to ∞, $E(1_{t<T^*}t|X)$ converges to $E(t|X)$ when the latter exists, but it is also well behaved for most defective distributions. $(1 - E(1_{t<T^*}|X))$ gives the probability that a spell doesn't end before time T^*.

Using the parameter estimates from the first panel of Table 1 (that is, those obtained with 500 observations), we computed $E(1_{t<T^*}t|X)$ and $(1 - E(1_{t<T^*}|X))$, at various values of X. The results are reported in Table 4. Setting $T^* = 41$ matches the censoring point in our sample, so the estimated quantities depend only on the predicted hazard for periods and regressors that are observed. Although the parameter values change dramat-

ically as we go from iteration 1 to 5 (among other things, the coefficient on observed heterogeneity increases from .73 to 1.74), the values for the two summaries of the predicted hazard duration barely change. Researchers may also be interested, however, in using the structural parameters to explore out of sample experments. For example, a researcher may want to know the impact of a change in X on the expected duration for complete spells. With censoring, the sample data are only partially informative and the predicted hazard must be extrapolated beyond the observed range of the data. As a consequence, bias in the estimated structural parameters can have important consequences. Ham et al (1998) report that the predicted expected duration for complete spells can be very sensitive to the number of points in the estimated heterogeneity distribution[15]. Although we find that extrapolating the cubic polynomial beyond the range of the data never works well, we obtain a similiar sensitivity to the number of points of support using our example in Table 4. In our case, the predicted distribution for duration is defective at 1 point support (so the mean duration is infinite), but duration has a finite mean when we use the estimates from the NPMLE.

6 Some Extensions

This section contains some extensions that we have investigated but not as intensively as in the main Monte Carlo.

6.1 Very Large Samples

The results reported in Table 2 show that estimates of β display large biases when we combine the NPMLE of the heterogeneity distribution with a flexible duration dependence specification in the QL. However, the bias appears to decline with sample size. In order to investigate more fully the effect of sample size, we combined our 100 samples of 5,000 observations into 5 samples each containing 100,000 observations. To reduce computational costs, we considered only the cases where the true DGP displays no duration

[15]Using data from the Slovak Republic, Ham et al (1998) estimate the impact of an additional week's entitlement to unemployment insurance on the expected duration of an unemployment spell. They report that the predicted duration rises from 0.4 to 1.1 to 1.9 weeks as the number of points of support in the fitted heterogeneity distribution increases from 1 to 2 to 3, respectively.

dependence. Note that this implies that all the QL specifications contain the true DGP as a special case. The results of the nine experiments are reported in Table 5.

In those cases where the heterogeneity distribution has either one or two points of support, the HQIC leads to exactly the same estimates as we obtained by imposing a priori the true number of points of support. The biases are negligible and the standard errors decline roughly in line with the square-root of sample size. When the unobserved heterogeneity is drawn from a gamma distribution, use of the HQIC leads to estimates that are very close to the true value. If anything, however, the estimator is a bit conservative when the QL contains either a cubic or step function in that better estimates would have been obtained if the penalty on the number of parameters was slightly smaller.

The NPMLE appears to be converging to the true value $\beta = 1$. However, with a very flexible duration dependence specification in the QL, the rate of convergence appears to be extremely slow and we still see nontrivial sampling error even with such large samples.

6.2 Increasing the Variance of Observable Heterogeneity

An increase in the variance of observed heterogeneity should reduce the sampling variance of $\widehat{\beta}$ and may help to distinguish the relative contributions of duration dependence and unobserved heterogeneity. To investigate the consequences of increasing σ_x^2, we conducted the following experiment. It was not feasible to consider all nine combinations of duration dependence and unobserved heterogeneity, so we restricted attention to a DGP with negative duration dependence and a discrete heterogeneity distribution with two points of support. We increased the variance of X, σ_x^2, from 0.25 to 1.00. To maintain comparability with the results from Section 5, we used exactly the same draws from the heterogeneity distribution and the values of X were simply multiplied by two. Once again, we used the specified DGP to generate 100 samples of 5,000 observations. The increase in σ_x^2 had some minor effects on the distribution of observed duration: the average duration of a spell increased by about 0.6 to 14.86, and the fraction of censored observations increased marginally from about 19% to 21%. However, the increase in σ_x^2 did have a large effect on the relative importance of observed heterogeneity:

the average R^2 from a regression of log-duration on X across the 100 samples jumped from .071 to .228.

Table 6 summarizes the results. We see that the increase in σ_x^2 results in a marginal improvement in the sampling distribution of the MLE and penalized MLE but does not alter the substantive conclusions reached in Section 5.

6.3 A Computational Alternative

Using the value of θ that maximizes the Gateaux derivative in Step 2b of our algorithm, as suggested by Heckman and Singer, often leads to a corner solution; That is, the value of θ so chosen is either extremely large or extremely small. This choice has potentially important consequences for both computational efficiency and inference. For example, an extreme point of support may be added on the second or third iteration of our algorithm only to spend many CPU cycles bringing it back to something in the 'middle' of our chosen range. In other cases, the algorithm finds a local optimum at an extreme value of θ but assigns it a negligible probability, so that the quasilikelihood barely increases. Because the HQ criterion is sensitive to the order in which points of support are added, adding a point with negligible probability on the second iteration of our algorithm can lead the HQ criterion to settle for too little heterogeneity. Both of these issues are nicely illustrated by our example in Section 5.1.

Is there a better way to choose the candidate for the next point of support in Step 2b of our algorithm? The Heckman and Singer suggestion is based on choosing θ to maximize the slope of a linear approximation to the quasi log-likelihood. An alternative strategy that we have developed is to choose θ to maximize a quadratic approximation. More precisely, we can re-write slightly the problem faced in (3) as

$$max\tau \ln L = \sum_h \ln \left((1-\tau) \sum_{i=1}^{\bar{N}_\theta} L_{ih} P_i + \tau L_{\bar{N}_\theta + 1, h} \right) \tag{15}$$

where we set $\{\alpha, \theta_i, P_i; i = 1, ..\bar{N}_\theta\}$ equal to the MLE estimates conditional on having $\bar{N}_\theta$ points of support, and where $\overline{\theta}$ is the candidate for the next point of support. Suppose we choose τ to maximize a second order Taylor series approximation to the right hand side of (15). Then, in Step 2b of our

algorithm, we choose the candidate $\overline{\theta}$ that yields the highest value for

$$\frac{\sum_h (\frac{L_{\bar{N}_\theta+1,h}}{\sum_{i=1}^{\bar{N}_\theta} L_{ih} P_i} - 1)}{\sqrt{\sum_h (\frac{L_{\bar{N}_\theta+1,h}}{\sum_{i=1}^{\bar{N}_\theta} L_{ih} P_i} - 1)^2}} \tag{16}$$

Notice that the numerator in (16) is the Gateaux derivative. The same terms appear in both numerator and denominator of (16), so choosing the new point of support is only marginally more difficult than in the original Heckman and Singer (1984) approach. Also, this alternative rule shares the desirable property that the selected value of θ is guaranteed to increase the likelihood function.

We found that our second order method for choosing θ was less likely to wander off to a corner. For instance, in the three samples reported in table 1, maximizing the Gateaux derivative lead us four times to a corner value of -11.513. In contrast, maximizing (16) only lead us to the corner once. Although it saved some CPU cycles, maximizing (16) did not affect the parameter estimates. We reached exactly the same MLE conditional on the number of points of support as reported in table 1[16].

To obtain a better idea of the consequences of choosing new points of support to maximize (16) rather than the Gateaux derivative, we applied this rule to some of our experiments. To keep the computational costs down, we considered only the case where the true DGP has two points of support and negative duration dependence. Further, we restricted attention to the case where the QL has a cubic polynomial. Compared to the results reported in Table 2, we found that maximizing (16) rather than the Gateaux derivative had some computational advantages but lead to virtually identical parameter estimates. The savings in CPU time varied with sample size, but they were consistently positive and averaged almost 10% across the three designs. The two rules for choosing additional points of support lead to identical NPMLE estimates of β in all cases, and to the same estimates conditional on a given number of points of support in the vast majority of cases. In the few cases where different local optima where reached conditional on the number of points of support, neither rule lead to consistently higher values for the quasilikelihood.

[16]The only difference is that in these examples our suggestion for choosing the new candidate never lead us to add a point of support on one interation only to drop it on the next.

7 Conclusion

Our Monte Carlo results demonstrate that recent improvements in computing power, coupled with some care in designing the algorithm, make it computationally feasible to combine the NPMLE estimator of the unobserved heterogeneity distribution with a very flexible specification for duration dependence. However, our results also show that this estimation strategy has poor sampling properties.

We find that a nonparametric specification for either duration dependence or unobserved heterogeneity, when the other feature of the hazard is known to be absent, leads to estimators that are well behaved even in modestly sized samples. However, the combination of a flexible specification for both duration dependence and unobserved heterogeneity leads to very reliable and systematic biases in each of the components of the estimated hazard. Applied researchers often sacrifice efficiency by adding extra parameters to safeguard against mis-specification. Our results suggest that this strategy is particularly questionable in this setting. Adding superfluous parameters not only sacrifices efficiency, it also introduces a potentially very large bias, even in very large samples. With a flexible specification for duration dependence, the NPMLE is biased towards finding an excessively dispersed distribution of unobserved heterogeneity. The fit to the empirical hazard is maintained by compensating with a positive bias to the estimated duration dependence and a bias to the coefficient on observed heterogeneity away from zero. In fact, we found (almost without fail) that the estimates of $\phi(t)$ and β moved in the directions consistent with these biases *each* time we added a point of support to our estimated heterogeneity distribution. On the other hand, ignoring unobserved heterogeneity leads to a negative bias in estimated duration dependence and biases the coefficient on observed heterogeneity towards zero.

Given the biases in the estimates of $\phi(t)$ and β associated with the NPMLE and the behaviour of these estimates as we add points of support, a minor modification of the quasilikelihood to include a term that penalizes specifications with many points of support seems like a natural solution. We find that using the Hannan-Quinn Information Criterion to choose the number of points of support leads to a dramatic improvement in the sampling properties of the estimated components of the hazard and, in particular, for a much more reliable estimator for β.

References

[1] Baker, M., and S. Rea, (1998), "Employment Spells and Unemployment Insurance Eligibility Requirements", *Review of Economics and Statistics,* 80, 80-94.

[2] Bearse, P., J. Canals and P. Rilstone, (1996), "Semiparametric Maximum Likelihood Estimation of Duration Models", unpublished paper, York University.

[3] Cosslett, Stephen R., (1983), "Distribution-Free Maximum Likelihood Estimator of the Binary Choice Model", *Econometrica*, 51, 765-782.

[4] Cameron, S.V. and J.J. Heckman, (1998), "Life Cycle Schooling and and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males", *Journal of Political Economy*, 106, 262-333.

[5] Campolieti, M., (1997), "Bayesian Estimation of Discrete Duration Models", PhD Thesis, University of Toronto

[6] 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.

[7] Gunderson M., and A. Melino, (1990), "The Effects of Public Policy on Strike Duration" *Journal of Labor Economics*, 8, 295-316.

[8] Hahn, J., (1994) "The Efficiency Bound of the Mixed Proportional Hazard Model", *Review of Economic Studies*, 61, 607-629.

[9] Ham, J.C., and S.A. Rea, Jr., (1987), "Unemployment Insurance and Male Unemployment in Canada", *Journal of Labor Economics*, 5, 325-353.

[10] Ham, J.C., J. Svejnar and K. Terrell, (1998) "Unemployment, the Social Safety Net and Efficiency in Transition: Evidence from Micro Data on Czech and Slovak Men", *American Economic Review*, 88, 1117-1142.

[11] Huh, K. and R.C. Sickles, (1994), "Estimation of the Duration Model by Non-Parametric Maximum Likelihood, Maximum Penalized Likelihood, and Probability Simulators", *The Review of Economics and Statistics*, 76, 683-694.

[12] Heckman, J., and B. Singer, (1984), "A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data", *Econometrica*, 52, 271-320.

[13] Ishwaran, H., (1996) "Uniform Rates of Estimation in the Semiparametric Weibull Mixture Model", *The Annals of Statistics*, 24, 1572-1585.

[14] Lancaster, T. (1979), "Econometric Methods for the Duration of Unemployment", *Econometrica*, 47, 939-956.

[15] Lancaster, T., (1990), **The Econometric Analysis of Transition Data**, Cambridge University Press.

[16] Lemieux, T. and W.B. MacLeod, (1995), "State Dependence and Unemployment Insurance", Ottawa: Human Resources and Development Canada.

[17] Leroux, B. G., (1992), "Consistent Estimation of a Mixing Distribution", *The Annals of Statistics*, 20, 1350-1360.

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

[19] Lindsay, B. G. and M. L. Lesperance, (1995), "A review of semiparametric mixture models", *Journal of Statistical Planning and Inference*, 47, 29-39.

[20] Luenberger, D. G., (1984), **Linear and Nonlinear Programming** (second edition), Addison-Wesley .

[21] Meyer, B.D., (1990), "Unemployment Insurance and Unemployment Spells", *Econometrica*, 58, 757-782.

[22] Narendranathan, W., and M.B. Stewart, (1993), "How Does the Benefit Effect Vary as Unemployment Spells Lengthen?", *Journal of Applied Econometrics*, 8, 361-381.

[23] Nickell, S. (1979), "Estimating the Probability of Leaving Unemployment", *Econometrica*, 47, 1249-1266.

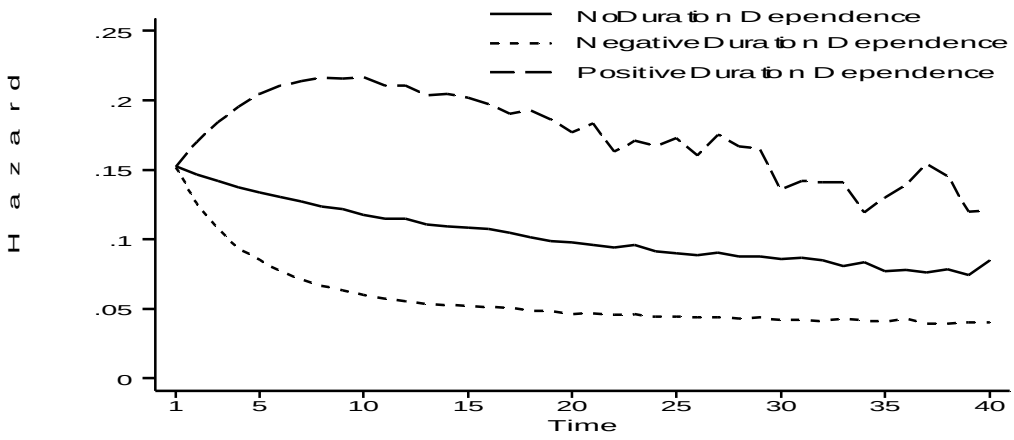
[24] Ridder, G., (1987), "The Sensitivity of Duration Models to Misspecified Unobserved Heterogeneity and Duration Dependence", (Working paper, University of Groningen).

[25] Ridder, G., (1990), "The Non-Parametric Identification of Generalized Accelerated Failure-Time Models", *Review of Economic Studies*, 57, 167-182.

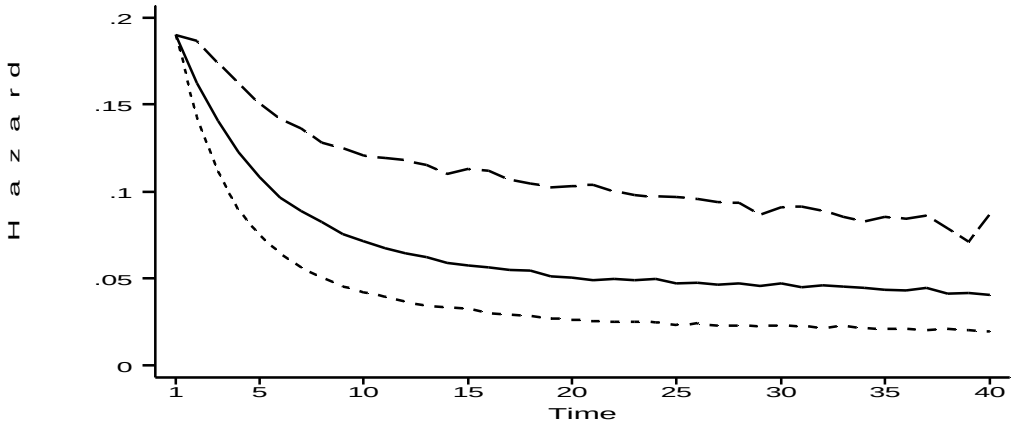
[26] Sider, H., (1985) "Unemployment Duration and Incidence: 1968-82", *American Economic Review*, 75, 461-472.

[27] Sin, C-Y., and H. White, (1996), "Information Criteria for Selecting Possibly Misspecified Parametric Models", Journal of Econometrics, 71, 207-225.

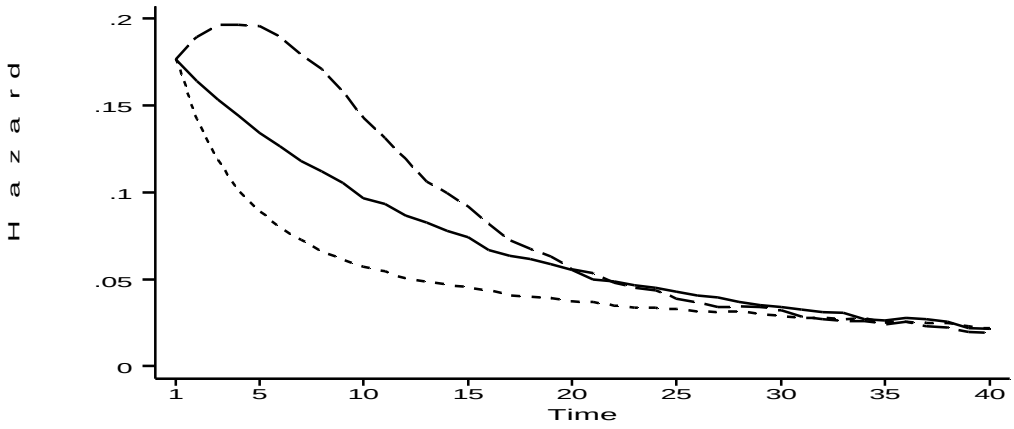
[28] Van der Vaart, A., (1996), "Efficient Maximum Likelihood Estimation in Semiparametric Mixture Models", *The Annals of Statistics*, 24, 862-878.



A) Empirical Hazard for DGP with No Heterogeneity



B) Empirical Hazard for DGP with Two-Point Heterogeneity



C) Empirical Hazard for DGP with Gamma Heterogeneity

Figure 1: The Empirical Hazards for the Different Specifications of the Data Generating Process (DGP).

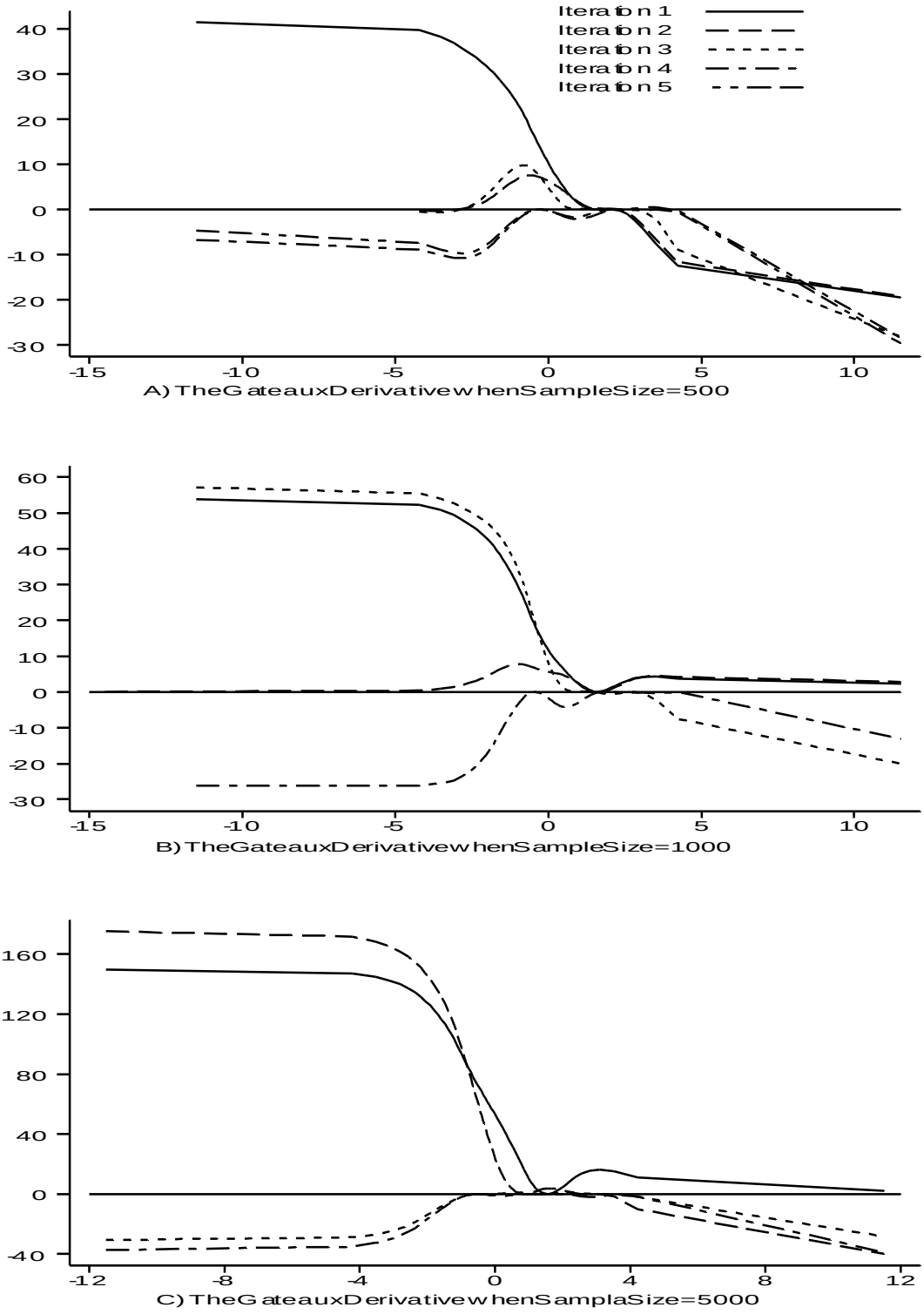
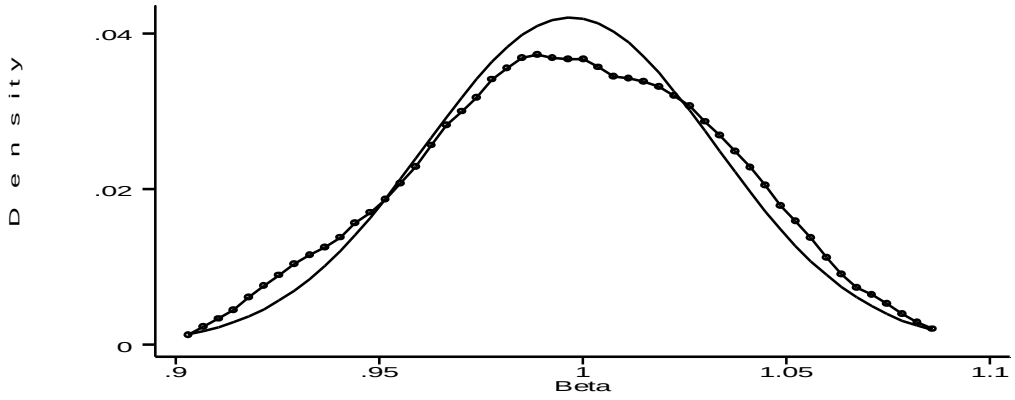
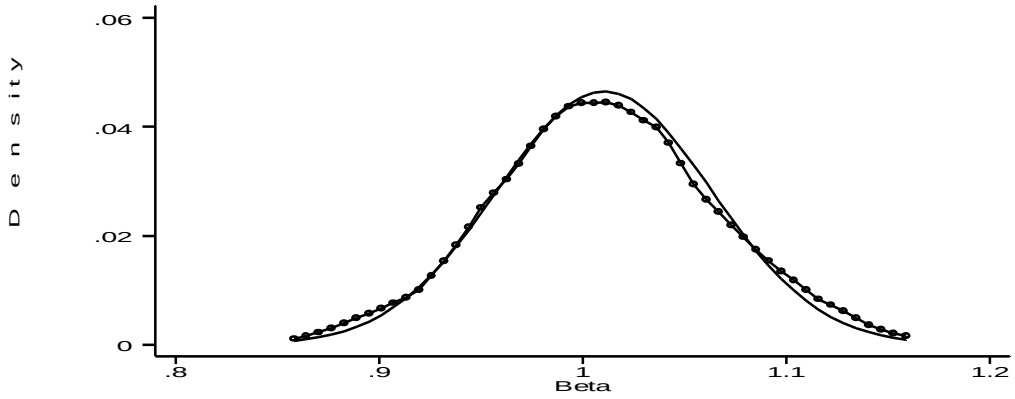


Figure 2: The Profile of the Gateaux Derivative.

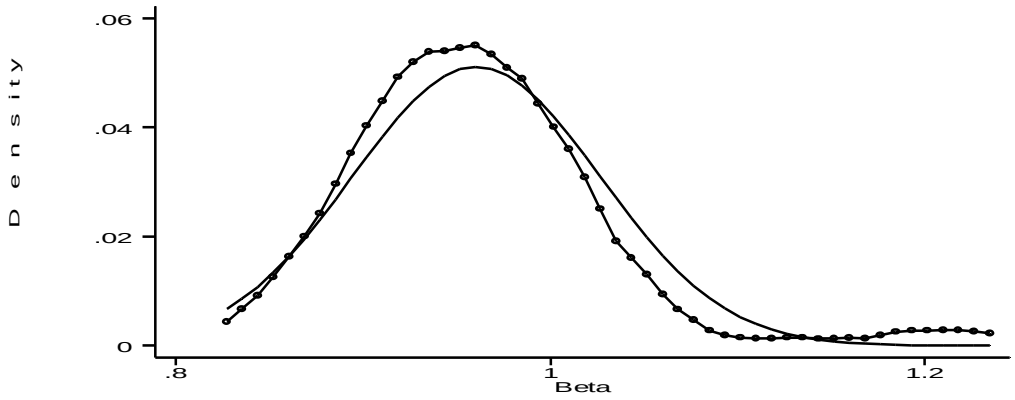
Notes: The derivative is evaluated along the grid $q \in [-11.51, 11.51]$. The reported results are for a data generating process incorporating negative duration dependence and unobserved heterogeneity following a discrete distribution with two points of support. In the quasi-likelihood, duration dependence is approximated by a cubic.



A) DGP: NoHeterogeneity, NoDuration Dependence. QL: Cubic



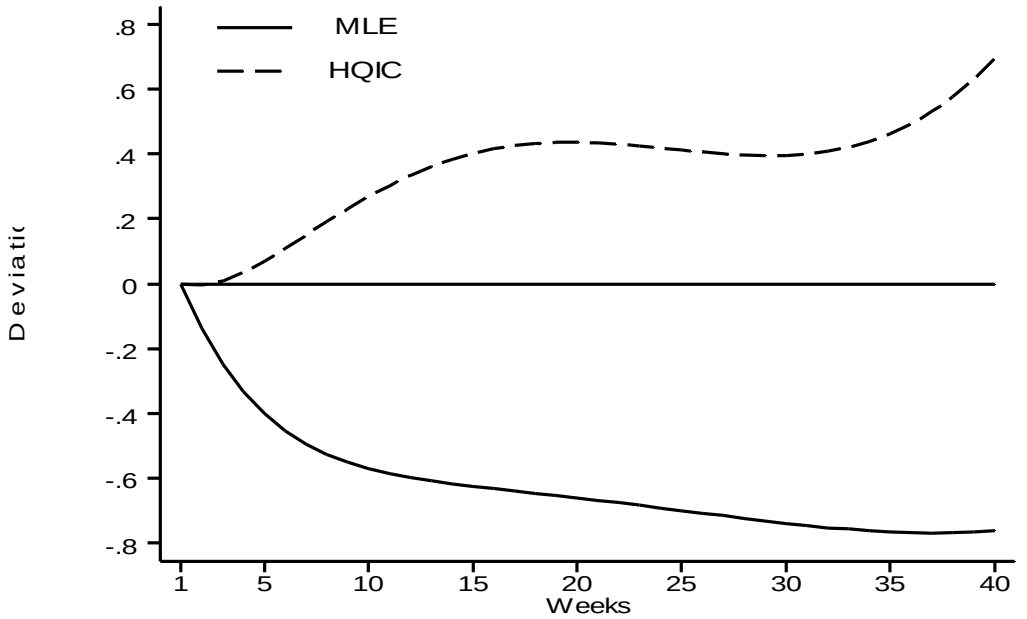
B) DGP: Two-Point Heterogeneity, NoDuration Dependence. QL: Cubic



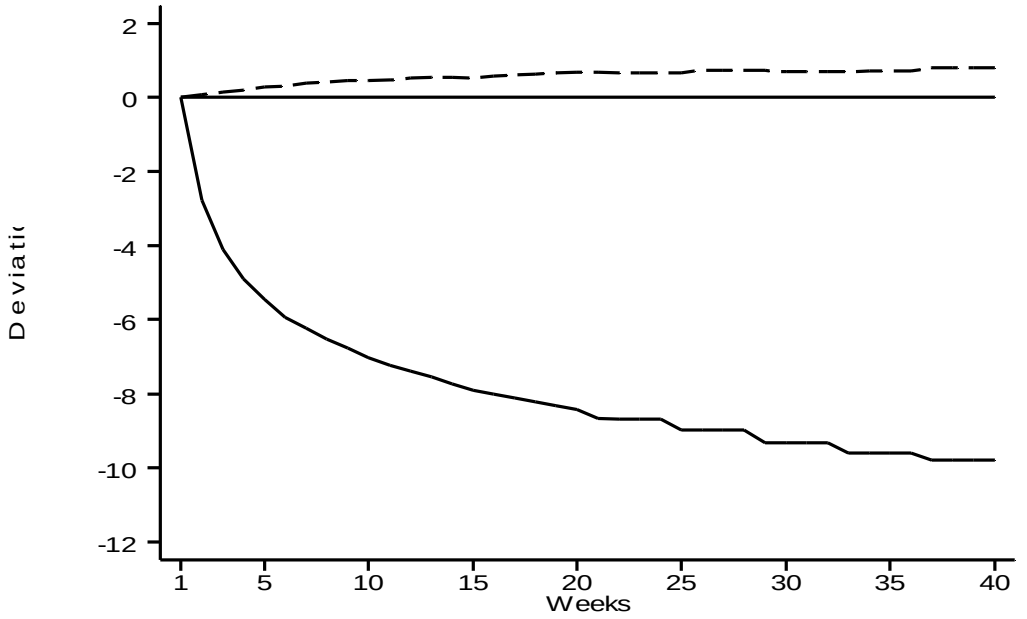
C) DGP: GammaHeterogeneity, NoDuration Dependence. QL: Cubic

Figure 3: Kernel Density Estimates of the Sampling Distribution of $\hat{b}$.

Notes: DGP is data generating process. QL is quasi likelihood. Sample size is 5000. The underlying estimates of b are obtained using the Hannan-Quinn Information Criterion. The solid lines are for a normal density standardized to have the same mean and variance.

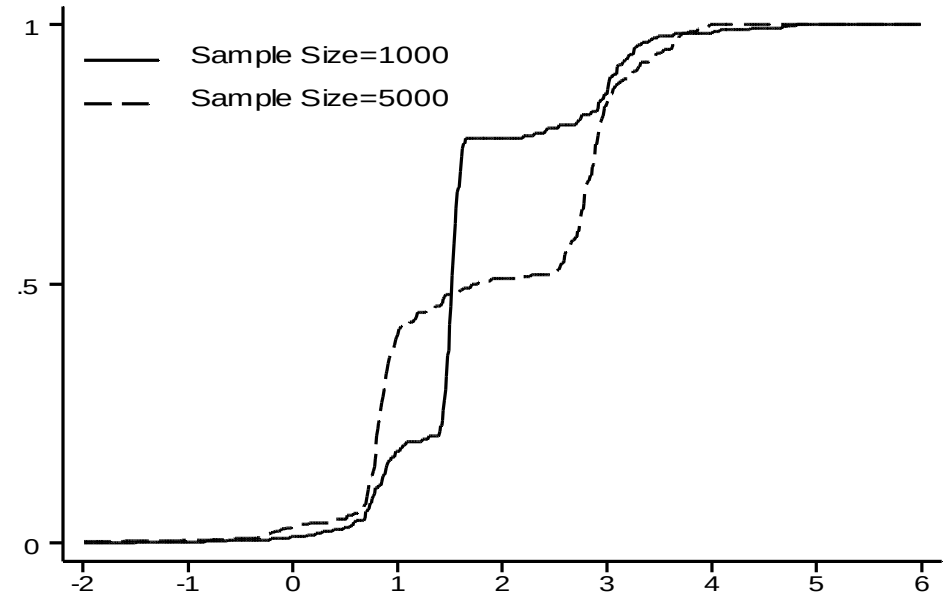


A) DGP: Two-Point Heterogeneity, Negative Duration Dependence
QL: Cubic



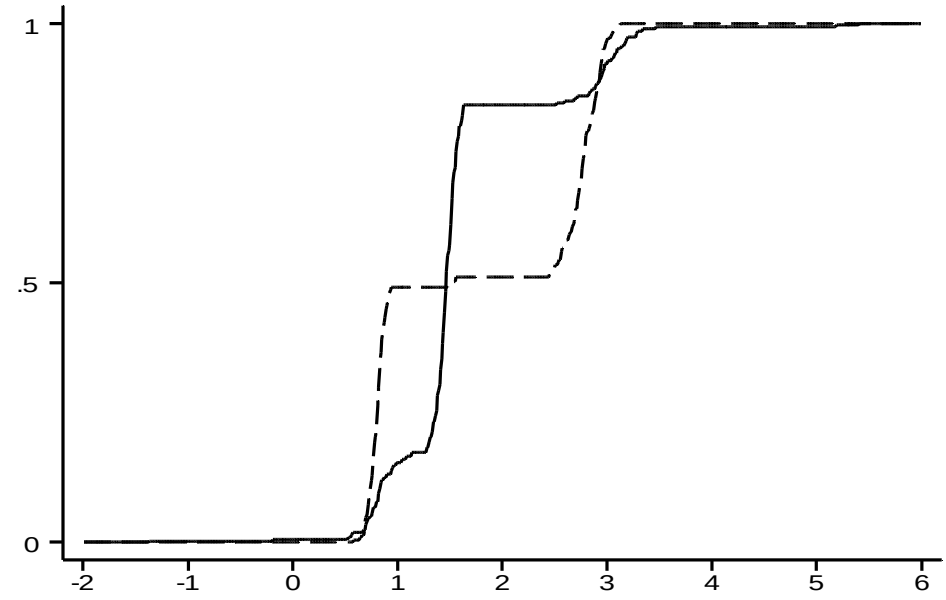
B) DGP: Two-Point Heterogeneity, Negative Duration Dependence
QL: Step Function

Figure 4: Estimated Duration Dependence for Selected Specifications
Notes: DGP is data generating process. QL is quasi likelihood. Sample size is 1000. The reported profiles are the deviations (by week) of the estimated duration dependence from the true duration dependence. MLE are maximum likelihood estimates. HQIC are the estimates that result from using the Hannan-Quinn Information Criterion.



A) DGP:Two-Point Heterogeneity, Negative Duration Dependence

QL: Cubic

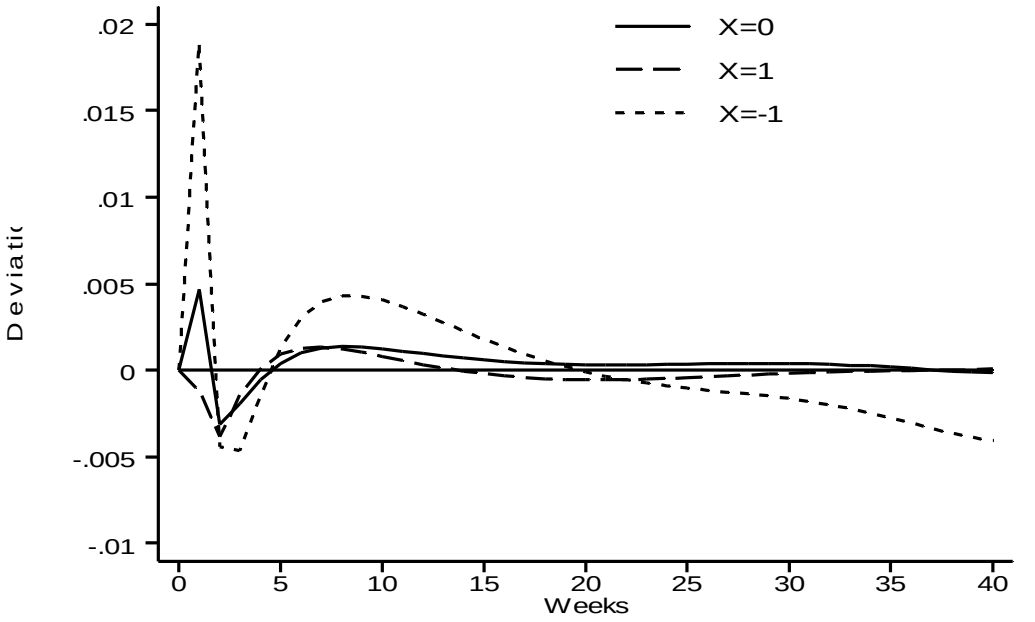


B) DGP: Two-Point Heterogeneity, Negative Duration Dependence

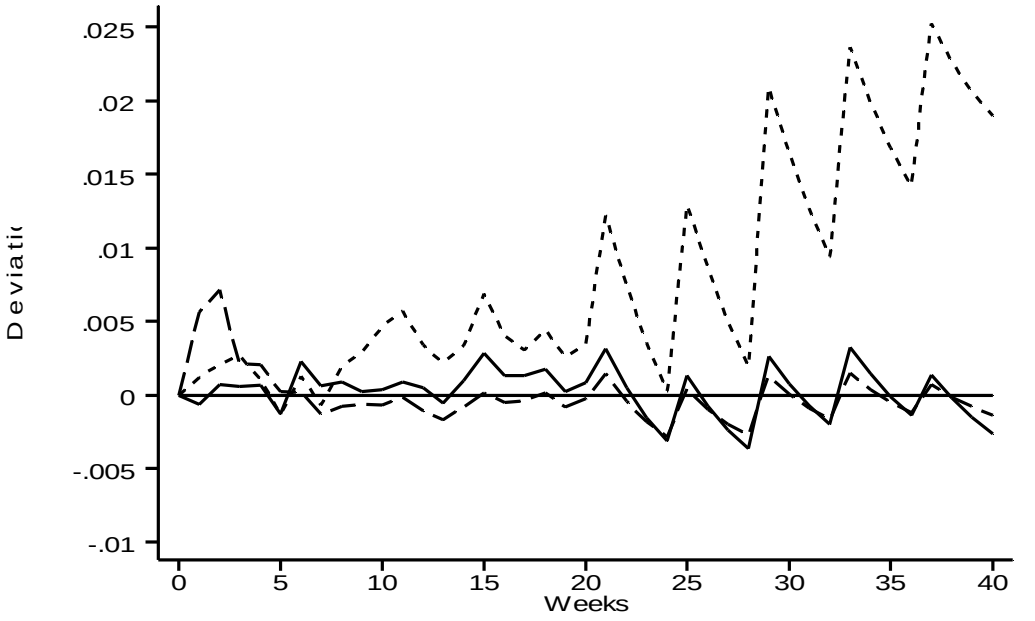
QL: Step Function

Figure 5: The Average Estimated Cumulative Distribution Function of Unobserved Heterogeneity for Selected Specifications.

Notes: DGP is data generating process. QL is quasi likelihood. Sample size is 1000. The reported distribution functions are constructed from estimates using the Hannan-Quinn Information Criterion.



A) DGP: Two-Point Heterogeneity, Negative Duration Dependence
QL: Cubic



B) DGP: Two-Point Heterogeneity, Negative Duration Dependence
QL: Step Function

Figure 6: Predicted Hazards for Selected Specifications.
Notes: DGP is data generating process. QL is quasi likelihood. X is the value specified for the observable heterogeneity. Sample size is 1000. The underlying estimates are obtained using non-parametric maximum likelihood estimation. The reported profiles are the deviations (by week) of the predicted hazard from the true hazard.

Table 1: Estimation Results by Iteration of the Search Algorithm. DGP: Duration Dependence = Negative; Unobserved Heterogeneity = Discrete, Two Points of Support. QL: Duration Dependence = Cubic.

	β	α_1	α_2	α_3	θ_1	P_1	θ_2	P_2	θ_3	P_3	θ_4	P_4	$LogL$	$\overline{\theta}_i$	$\overline{P}_i$
Sample Size=500															
1	0.731	2.603	-1.197	0.181	1.565	1.000							-1418.22	-11.513	0.017
2	0.740	2.224	-0.965	0.145	1.706	0.961	-25.560	0.039					-1417.39	-0.588	0.034
3	1.079	1.282	-0.626	0.098	2.788	0.528	-25.560	0.032	0.944	0.439			-1415.61	-0.847	0.067
4*	1.615	-0.879	0.347	-0.045	4.692	0.338	-0.261	0.285	2.328	0.376			-1412.57	3.106	0.004
5	1.737	-1.179	0.471	-0.062	5.115	0.285	-0.294	0.287	2.263	0.309	3.506	0.118	-1412.52		
Sample Size=1000															
1	0.665	2.486	-1.049	0.150	1.570	1.000							-2811.36	-11.513	0.011
2	0.671	2.244	-0.900	0.125	1.658	0.975	-11.513	0.025					-2810.68	-0.987	0.036
3*	0.976	1.602	-0.752	0.113	2.785	0.464	0.980	0.536					-2808.03	-11.513	0.013
4	1.333	-0.228	0.132	-0.021	4.178	0.352	1.862	0.441	-0.442	0.207			-2804.76		
Sample Size=5000															
1	0.670	2.558	-1.040	0.139	1.549	1.000							-14067.84	-11.513	0.006
2	1.066	1.276	-0.563	0.077	2.935	0.493	0.852	0.507					-14042.82	-11.513	0.008
3	1.216	0.507	-0.190	0.020	3.543	0.418	1.364	0.434	-0.292	0.147			-14038.13	2.197	0.002
4	1.269	0.340	-0.124	0.011	3.767	0.380	0.980	0.288	-0.530	0.111	1.926	0.221	-14037.78		

Notes: DGP = data generating process, QL = quasi-likelihood. β is the parameter on the observed heterogeneity, the α_i are the parameters of the cubic in duration and the θ_i and P_i are the parameters of the distribution of unobserved heterogeneity. $LogL$ is the value of the loglikelihood. $\overline{\theta}_i$ is the position suggested by the search algorithm for the new point of support on the next iteration. $\overline{P}_i$ is the suggested value for the probability associated with the new point of support. * indicates that a point of support was dropped in this iteration.

Table 2: A Summary of Estimates of β Across DGP's and Quasi Likelihoods

Sample Size		500				1000				5000			
DGP	Quasi Likelihood	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC
DD: none Het.: none	DD: none	1.000 (0.108)	1.007* (0.108)	1.004 (0.110)	1.000 (0.108)	0.992 (0.068)	0.993* (0.071)	0.996 (0.069)	0.993 (0.068)	0.998 (0.033)	0.998* (0.033)	0.999 (0.033)	0.998 (0.033)
	DD: cubic	1.001 (0.117)	1.123* (0.181)	1.394 (0.505	1.011 (0.140)	0.992 (0.076)	1.079* (0.117)	1.177 (0.219)	1.001 (0.106)	0.997 (0.036)	1.033* (0.053)	1.075 (0.076)	0.997 (0.036)
	DD: step	1.004 (0.118)	1.149 (0.181)	4.023 (1.813)	1.012 (0.132)	0.993 (0.076)	1.079 (0.125)	3.289 (1.295)	0.996 (0.084)	0.997 (0.036)	1.039 (0.055)	2.499 (0.879)	0.997 (0.036)
DD: none Het.: discrete	DD: none	0.905 (0.118)	1.015 (0.132)	1.028 (0.132)	1.015 (0.132)	0.896 (0.083)	1.004 (0.091)	1.012 (0.091)	1.004 (0.091)	0.901 (0.039)	1.005 (0.042)	1.010 (0.043)	1.006 (0.042)
	DD: cubic	0.701 (0.086)	0.959* (0.226)	1.372 (0.594)	0.827 (0.234)	0.692 (0.060)	0.967* (0.173)	1.148 (0.225)	0.875 (0.228)	0.698 (0.029)	0.990 (0.077)	1.062 (0.065)	1.011 (0.053)
	DD: step	0.702 (0.087)	1.003 (0.206)	3.436 (1.559)	0.843 (0.247)	0.697 (0.060)	0.995 (0.161)	2.914 (0.875)	0.901 (0.238)	0.698 (0.029)	0.998 (0.068)	2.268 (0.845)	1.010 (0.051)
DD: none Het.: Gamma	DD: none	0.776 (0.150)	0.981 (0.132)	1.003 (0.133)	0.984 (0.132)	0.774 (0.115)	0.981 (0.099)	0.998 (0.100)	0.983 (0.098)	0.766 (0.051)	0.980 (0.043)	0.998 (0.044)	0.993 (0.045)
	DD: cubic	0.651 (0.113)	0.943 (0.158)	1.430 (0.644)	0.944 (0.198)	0.649 (0.087)	0.942 (0.111)	1.164 (0.225)	0.952 (0.128)	0.642 (0.036)	0.940 (0.046)	1.060 (0.084)	0.960 (0.066)
	DD: step	0.652 (0.113)	0.948 (0.154)	3.917 (1.461)	0.946 (0.181)	0.650) (0.087)	0.943 (0.111)	3.491 (1.310)	0.954 (0.119)	0.643 (0.036)	0.940 (0.046)	2.631 (0.896)	0.966 (0.073)
DD: negative Het.: none	DD: none	1.191 (0.120)	1.260 (0.135)	1.310 (0.142)	1.265 (0.137)	1.198 (0.081)	1.273 (0.096)	1.316 (0.099)	1.284 (0.101)	1.194 (0.040)	1.261 (0.044)	1.303 (0.047)	1.290 (0.051)
	DD: cubic	1.011 (0.101)	1.140* (0.155)	1.523 (0.483)	1.042 (0.209)	1.014 (0.067)	1.126* (0.132)	1.339 (0.241)	1.045 (0.123)	1.005 (0.035)	1.057 (0.057)	1.200 (0.094)	1.030 (0.080)
	DD: step	1.012 (0.102)	1.160 (0.152)	4.212 (1.473)	1.027 (0.131)	1.015 (0.068)	1.144 (0.134)	3.913 (1.203)	1.026 (0.087)	1.004 (0.034)	1.049 (0.050)	2.608 (1.031)	1.004 (0.034)

Notes: DGP = data generating process. DD = Duration Dependence; Het. = unobserved Heterogeneity. The reported statistics are the mean and standard deviation of β across 100 samples on the indicated iteration of the algorithm. Initial = on the first iteration; Two Points = on the iteration in which the distribution of unobserved heterogeneity is specified as having two points of support; MLE = the Maximum Likelihood estimate; HQIC = the estimate using the Hannon-Quinn Information Criterion. In the column titled Two Points, means denoted with an "*" are calculated over less than 100 samples, as the search algorithm stopped before reaching the the second point of support.

Table 2 (cont.)

Sample Size		500				1000				5000			
DGP	Quasi Likelihood	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC
DD: negative Het.: discrete	DD: none	0.897 (0.152)	1.138 (0.232)	1.325 (0.240)	1.223 (0.269)	0.890 (0.105)	1.128 (0.162)	1.310 (0.175)	1.250 (0.181)	0.867 (0.048)	1.107 (0.070)	1.269 (0.077)	1.249 (0.077)
	DD: cubic	0.678 (0.110)	0.897* (0.246)	1.409 (0.523)	0.830 (0.337)	0.676 (0.077)	0.929 (0.184)	1.247 (0.264)	0.885 (0.283)	0.662 (0.036)	0.901 (0.163)	1.152 (0.109)	1.054 (0.108)
	DD: step	0.677 (0.110)	0.995 (0.223)	4.123 (1.744)	0.813 (0.340)	0.674 (0.077)	0.980 (0.162)	3.670 (1.247)	0.829 (0.264)	0.660 (0.036)	0.987 (0.084)	2.768 (0.823)	0.994 (0.076)
DD: negative Het.: Gamma	DD: none	0.912 (0.153)	1.155 (0.205)	1.318 (0.177)	1.278 (0.183)	0.925 (0.116)	1.168 (0.160)	1.323 (0.130)	1.296 (0.128)	0.920 (0.052)	1.143 (0.077)	1.308 (0.057)	1.293 (0.058)
	DD: cubic	0.721 (0.114)	1.001 (0.151)	1.411 (0.386)	0.958 (0.234)	0.727 (0.087)	0.995 (0.128)	1.258 (0.230)	1.011 (0.174)	0.723 (0.038)	0.977 (0.051)	1.166 (0.099)	1.028 (0.104)
	DD: step	0.721 (0.115)	1.004 (0.150)	4.253 (1.817)	0.951 (0.222)	0.727 (0.087)	0.997 (0.120)	3.992 (1.174)	0.983 (0.154)	0.722 (0.038)	0.973 (0.051)	2.898 (1.155)	0.975 (0.053)
DD: positive Het.: none	DD: none	0.803 (0.074)		0.803 (0.074)	0.803 (0.074)	0.795 (0.055)		0.795 (0.055)	0.795 (0.055)	0.795 (0.026)		0.795 (0.026)	0.795 (0.026)
	DD: cubic	1.008 (0.101)	1.147* (0.156)	1.954 (1.183)	1.024 (0.154)	0.995 (0.076)	1.099* (0.121)	1.407 (0.783)	1.003 (0.096)	0.994 (0.036)	1.048* (0.059)	1.061 (0.086)	0.994 (0.036)
	DD: step	1.012 (0.103)	1.172 (0.162)	4.003 (1.981)	1.039 (0.177)	0.998 (0.076)	1.118 (0.131)	3.529 (1.541)	1.009 (0.102)	0.994 (0.036)	1.048 (0.063)	2.851 (1.258)	0.994 (0.036)
DD:positive Het.: discrete	DD: none	0.796 (0.110)	0.806 (0.113)	0.808 (0.113)	0.806 (0.113)	0.795 (0.075)	0.807 (0.077)	0.808 (0.077)	0.808 (0.077)	0.794 (0.031)	0.806 (0.032)	0.807 (0.032)	0.806 (0.032)
	DD: cubic	0.712 (0.098)	0.946* (0.205)	1.299 (0.659)	0.800 (0.214)	0.711 (0.067)	0.953* (0.171)	1.104 (0.256)	0.863 (0.208)	0.707 (0.029)	0.969* (0.097)	1.008 (0.092)	0.978 (0.088)
	DD: step	0.715 (0.098)	0.974 (0.205)	2.910 (1.297)	0.835 (0.259)	0.712 (0.067)	0.958 (0.165)	2.607 (0.816)	0.893 (0.245)	0.708 (0.029)	0.966 (0.097)	2.317 (0.855)	0.988 (0.067)
DD: positive Het.: Gamma	DD: none	0.667 (0.132)	0.794 (0.105)	0.800 (0.107)	0.794 (0.105)	0.670 (0.103)	0.796 (0.076)	0.801 (0.077)	0.796 (0.076)	0.664 (0.043)	0.789 (0.033)	0.792 (0.034)	0.789 (0.033)
	DD: cubic	0.627 (0.101)	0.908 (0.146)	1.413 (0.562)	0.935 (0.173)	0.629 (0.077)	0.920 (0.130)	1.188 (0.263)	0.959 (0.123)	0.623 (0.031)	0.919 (0.040)	1.042 (0.070)	0.991 (0.056)
	DD: step	0.632 (0.102)	0.916 (0.136)	3.709 (1.550)	0.952 (0.240)	0.632 (0.077)	0.911 (0.102)	3.200 (1.025)	0.937 (0.125)	0.625 (0.031)	0.909 (0.040)	2.555 (0.927)	0.972 (0.074)

Table 3: Estimated values of $E(1_{t<T^*}t|x)$ and $(1-E(1_{t<T^*}t|x))$

$T^*=41$	$E(1_{t<T^*}t\|x)$				$(1-E(1_{t<T^*}\|x))$			
	IT=1	IT=3	IT=5	True	IT=1	IT=3	IT=5	True
x=1	6.46	7.02	7.24	7.26	0.41	0.39	0.39	0.37
x=0	7.17	7.02	6.94	6.83	0.16	0.17	0.17	0.18
x=-1	5.13	5.27	5.54	5.56	0.03	0.02	0.01	0.03
$T^*=\infty$	$E(1_{t<T^*}t\|x)$				$(1-E(1_{t<T^*}\|x))$			
	IT=1	IT=3	IT=5	True	IT=1	IT=3	IT=5	True
x=1	6.92	7.88	31.14	64.13	0.40	0.38	0.00	0.00
x=0	7.54	7.75	16.51	21.88	0.15	0.15	0.00	0.00
x=-1	5.25	5.49	6.16	7.33	0.02	0.01	0.00	0.00

Notes: Estimates are obtained using the parameter values from the first panel of Table 1 (500 observations). IT=iteration. The statistics in the columns labeled `True' are constructed using the true parameter estimates.

Table 4: A Summary of Estimates of the Standard Errors of β

Sample Size		1000			
DGP	Quasi Likelihood	Initial	Two Points	MLE	HQIC
DD: none Het.: none	DD: cubic	0.074	0.114*	0.172	0.074
DD: none Het.: discrete	DD: cubic	0.071	0.110*	0.189	0.091
DD: none Het.: Gamma	DD: cubic	0.070	0.094	0.180	0.095

Notes: DGP = data generating process. DD = Duration Dependence; Het. = unobserved Heterogeneity. The reported statistics are the mean of the estimated standard errors of β across 1000 samples on the indicated iteration of the algorithm. Initial = on the first iteration; Two Points = on the iteration in which the distribution of unobserved heterogeneity is specified as having two points of support; MLE = the Maximum Likelihood estimate; HQIC = the estimate using the Hannon-Quinn Information Criterion. In the column titled Two Points, means denoted with an "*" are calculated over less than 100 samples, as the search algorithm stopped before reaching the the second point of support.

Table 5: A Summary of Estimates of β in Large Samples

Sample Size		100000			
DGP	Quasi Likelihood	Initial	Two Points	MLE	HQIC
DD: none Het.: none	DD: none	0.997 (0.008)	0.997* (0.009)	0.998 (0.008)	0.997 (0.008)
	DD: cubic	0.996 (0.009)	0.999 (0.010)	1.008 (0.012)	0.996 (0.009)
	DD: step	0.996 (0.009)	0.999 (0.010)	1.079 (0.098)	0.996 (0.009)
DD: none Het.: discrete	DD: none	0.901 (0.008)	1.005 (0.008)	1.006 (0.008)	1.005 (0.008)
	DD: cubic	0.698 (0.007)	1.004 (0.010)	1.011 (0.011)	1.004 (0.010)
	DD: step	0.697 (0.007)	1.005 (0.010)	1.128 (0.110)	1.005 (0.010)
DD: none Het.: Gamma	DD: none	0.765 (0.016)	0.979 (0.010)	0.996 (0.009)	0.996 (0.009)
	DD: cubic	0.642 (0.012)	0.938 (0.010)	1.000 (0.013)	0.989 (0.012)
	DD: step	0.642 (0.012)	0.938 (0.010)	1.169 (0.150)	0.989 (0.012)

Notes: DGP = data generating process. DD = Duration Dependence; Het. = unobserved Heterogeneity. The reported statistics are the mean and standard deviation of β across 5 samples on the indicated iteration of the algorithm. Initial = on the first iteration; Two Points = on the iteration in which the distribution of unobserved heterogeneity is specified as having two points of support; MLE = the Maximum Likelihood estimate; HQIC = the estimate using the Hannon-Quinn Information Criterion. In the column titled Two Points, means denoted with an "*" are calculated over less than 5 samples, as the search algorithm stopped before reaching the the second point of support.

Table 6: Estimates of β When the Variance of Observable Heterogeneity is Increased

Sample Size		500				1000				5000			
DGP	Quasi Likelihood	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC	Initial	Two Points	MLE	HQIC
DD: negative Het.: discrete	DD: none	0.906 (0.071)	1.205 (0.098)	1.294 (0.109)	1.246 (0.110)	0.902 (0.054)	1.190 (0.071)	1.282 (0.073)	1.255 (0.077)	0.914 (0.027)	1.192 (0.037)	1.281 (0.037)	1.267 (0.038)
	DD: cubic	0.687 (0.055)	0.962 (0.170)	1.302 (0.284)	0.977 (0.223)	0.683 (0.039)	0.960 (0.147)	1.204 (0.146)	1.013 (0.137)	0.690 (0.019)	0.984 (0.109)	1.122 (0.068)	1.033 (0.057)
	DD: step	0.685 (0.056)	0.994 (0.146)	3.551 (1.376)	0.919 (0.212)	0.681 (0.039)	0.984 (0.108)	3.295 (1.341)	0.972 (0.178)	0.687 (0.019)	0.996 (0.059)	2.130 (1.015)	1.001 (0.050)

Notes: DGP = data generating process. DD = Duration Dependence; Het. = unobserved Heterogeneity. The reported statistics are the mean and standard deviation of β across 100 samples on the indicated iteration of the algorithm. Initial = on the first iteration; Two Points = on the iteration in which the distribution of unobserved heterogeneity is specified as having two points of support; MLE = the Maximum Likelihood estimate; HQIC = the estimate using the Hannon-Quinn Information Criterion. For these results, the variance of observable heterogeneity (X) is set to 1.00 (it is set to 0.25 for the results in table 1-3).