Estimating a Semi-Parametric Duration Model without Specifying Heterogeneity

Jerry A. Hausman and Tiemen M. Woutersen[‡]
MIT and Johns Hopkins University

Draft, August 2005

Abstract. This paper presents a new estimator for the mixed proportional hazard model that allows for a nonparametric baseline hazard and time-varying regressors. In particular, this paper allows for discrete measurement of the durations as happens often in practice. The integrated baseline hazard and all parameters are estimated at regular rate, $\sqrt{N}$, where N is the number of individuals. A hazard model is a natural framework for time-varying regressors if a flow or a transition probability depends on a regressor that changes with time since a hazard model avoids the curse of dimensionality that would arise from interacting the regressors at each point in time with one another.

Keywords: Mixed Proportional Hazard Model, Time-varying regressors, Heterogeneity.

1. Introduction

The estimation of duration models has been the subject of significant research in econometrics since the late 1970s. Since Lancaster (1979), it has been recognized that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity causes the estimated hazard rate to decrease more with the duration than the hazard rate of a randomly selected member of the population. Moreover, the estimated proportional effect of explanatory variables on the population hazard rate is smaller in absolute value than that on the hazard rate of the

*Comments are welcome, jhausman@mit.edu and woutersen@jhu.edu.

†We thank Su-Hsin Chang, Matthew Harding, and Marcel Voia for research assistance. We have received helpful comments from Bo Honoré, Moshe Buchinsky and seminar participants at Harvard-MIT, UCLA, Texas A&M, Rice University, Yale University, UC Santa Barbara, the University of Maryland, and the University of Virginia.

average population member and decreases with the duration. To account for unobserved heterogeneity Lancaster proposed a parametric Mixed Proportional Hazard (MPH) model, a generalization of Cox's (1972) Proportional Hazard model, that specifies the hazard rate as the product of a regression function that captures the effect of observed explanatory variables, a base-line hazard that captures variation in the hazard over the spell, and a random variable that accounts for the omitted heterogeneity.

Lancaster's MPH model was fully parametric, as opposed to Cox's semi-parametric approach, and from the outset questions were raised on the role of functional form and parametric assumptions in the distinction between unobserved heterogeneity and duration dependence[1]. This question was resolved by Elbers and Ridder (1982) who showed that the MPH model is semi-parametrically identified if there is minimal variation in the regression function. A single indicator variable in the regression function suffices to recover the regression function, the base-line hazard, and the distribution of the unobserved component, provided that this distribution does not depend on the explanatory variables. Semi-parametric identification means that semi-parametric estimation is feasible, and a number of semi-parametric estimators for the MPH model have been proposed that progressively relaxed the parametric restrictions.

Nielsen et al., (1992) showed that the Partial Likelihood estimator of Cox (1972) can be generalized to the MPH model with Gamma distributed unobserved heterogeneity. Their estimator is semi-parametric because it uses parametric specifications of the regression function and the distribution of the unobserved heterogeneity. The estimator requires numerical integration of the order of the sample size, which further limits its usefulness and makes it impractical for most situation in econometrics. Heckman and Singer (1984) considered the non-parametric maximum likelihood estimator of the MPH model with a parametric baseline hazard and regression function. Using results of Kiefer and Wolfowitz (1956), they approximate the unobserved heterogeneity with a discrete mixture. The rate of convergence and the asymptotic distribution of this estimator are not known. Another estimator that does not require the specification of the unobserved heterogeneity

[1] Heckman (1991) gives an overview of attempts to make this distinction in duration and dynamic panel data models.

distribution was suggested by Honoré (1990). This estimator assumes a Weibull baseline hazard and only uses very short durations to estimate the Weibull parameter.

Han and Hausman (1990) and Meyer (1990) proposes an estimator that assumes that the baseline hazard is piecewise-constant, to permit flexibility, and that the heterogeneity has a gamma distribution. We present simulations and a theoretical result that show that using a nonparametric estimator of the baseline hazard with gamma heterogeneity yields inconsistent estimates for all parameters and functions if the true mixing distribution is not a gamma, which limits the usefulness of the Han-Hausman-Meyer approach. Thus, we find it important to specify a model that does not require a parametric specification of the unobserved heterogeneity.

Horowitz (1999) was the first to propose an estimator that estimates both the baseline hazard and the distribution of the unobserved heterogeneity nonparametrically. His estimator is an adaptation of the semi-parametric estimator for a transformation model that he introduced in Horowitz (1996). In particular, if the regressors are constant over the duration then the MPH model has a transformation model representation with the logarithm of the integrated baseline hazard as the dependent variable and a random error that is equal to the logarithm of a log standard exponential minus the logarithm of a positive random variable. In the transformation model the regression coefficients are identified only up to scale. As shown by Ridder (1990) the scale parameter is identified in the MPH model if the unobserved heterogeneity has a finite mean. Horowitz (1999) suggests an estimator of the scale parameter that is similar to Honoré's (1990) estimator of the Weibull parameter and consistent if the finite mean assumption holds so that his approach allows estimation of the regression coefficients (not just up to scale). However, the Horowitz approach only permits estimation of the regression coefficients at a slow rate of convergence and it is not $N^{-1/2}$ consistent, where N is the sample size. In practice, there may be three obstacles for applying Horowitz (1999) MPH estimator. First, the durations need to be measured at a continuous scale in order to estimate the transformation model. This condition often does not hold in economic data, e.g. unemployment duration data as discussed in Han and Hausman (1990). Second, like the transformation model, the

MPH estimator does not allow for time-varying regressors. Finally, the estimator relies on arbitrarily short durations to estimate the scale parameter and, therefore, converges slowly. Thus, the regression coefficient estimates, which are often of primary interest, are often not estimated very precisely[2].

In this paper, we derive a new estimator for the mixed proportional hazard model (with heterogeneity) that allows for a nonparametric baseline hazard and time-varying regressors. No parametric specification of the heterogeneity distribution nor nonparametric estimation of the heterogeneity distribution is necessary. Intuitively, we condition out the heterogeneity distribution, which makes it unnecessary to estimate it. Thus, we eliminate the problems that arise with the Lancaster (1979) approach to MPH models. In our new model the baseline hazard rate is nonparametric and the estimator of the integrated baseline hazard rate converges at the regular rate, $N^{-1/2}$, where N is the sample size. This convergence rate is the same rate as for a duration model without heterogeneity. The regressor parameters also converge at the regular rate. A nice feature of the new estimator is that it allows the durations to be measured on a finite set of points. Such discrete measurement of durations is important in economics; for example, unemployment is often measured in weeks. In the case of discrete duration measurements, the estimator of the integrated baseline hazard only converges at this set of points, as would be expected.

It may be argued that the bias in the estimates of the regression coefficients is small, if the estimates of the MPH model indicate that there is no significant unobserved heterogeneity. The problem with this argument is that estimates of the heterogeneity distribution are usually not very accurate. Given the results in Horowitz (1999) this finding should not come as a surprise. The simulation results in Baker and Melino (2000) show that it is empirically difficult to find evidence of unobserved heterogeneity, in particular if one chooses a flexible parametric representation of the baseline hazard. However, Han-Hausman (1990) and applications of their approach have found significant heterogeneity using a flexible approach to the baseline hazard. Bijwaard and Ridder (2002) find that the

[2]It should be noted that the slower than $N^{-1/2}$ convergence of Horowitz (1999) estimator is a property of the model. Hahn (1994) and Ishwaran (1996a) show that no estimator can converge at rate $N^{-1/2}$ under the assumptions of Horowitz (1999). Horowitz (1999) assumes that the first three moments of the heterogeneity distribution exist and Ishwaran (1996b) shows that the fastest possible rate of convergence is $N^{-2/5}$ for that case and Horowitz' (1999) estimator converges arbitrarily close to that rate.

bias in the regression parameters is largely independent of the specification of the baseline hazard. Hence, failure to find significant unobserved heterogeneity should not lead to the conclusion that the bias due to correlation of the regressors and the unobservables that affect the hazard is small.

Because it is empirically difficult to recover the distribution of the unobserved heterogeneity, estimators that rely on estimation of this distribution may be unreliable. Therefore, we avoid estimating the unobserved heterogeneity distribution. Nevertheless, we can identify and estimate the regression parameters and the integrated baseline hazard. We find the removal of the requirement to estimate the heterogeneity distribution a major advantage of our approach[3]. Our estimator is related to the estimator by Han (1987). Han derives an estimator, up to scale, of the regression coefficients. However, Han's estimator cannot handle time-varying regressors and we estimate the regression coefficients when time-varying regressors are present as well as the scale of the regression coefficients. In particular, by estimating the regression coefficients up to scale, each regression coefficient can be interpreted as the elasticity of the hazard with respect to its regressor. Similarly, Chen's (2002) estimator of the transformation model cannot handle time-varying regressions and only gives the transformation function up to scale. While Horowitz's (1999) estimator is not subject to the limitation of estimating the regression coefficients up to scale only, it converges slowly and it is not $N^{-1/2}$ consistent which makes standard inferences techniques inapplicable unless N is very large.

A hazard model is a natural framework for time-varying regressors if a flow or a transition probability depends on a regressor that changes with time since a hazard model avoids the curse of dimensionality that would arise from interacting the regressors at each point in time with one another. A nonconstructive identification proof for the duration model with time-varying regressors can be produced using techniques similar to Honoré (1993b) and Honoré (1993a) gives such a proof. In particular, Honoré (1993a) does not assume that the mean of the heterogeneity distribution is finite[4]. Ridder and Woutersen (2003) argue that it is precisely the finite mean assumption that makes the identification

[3] An unconditional approach is also used, in another context, by Heckman (1978) who develops unconditional tests to distinguish true and spurious state dependence.

[4] nor does Honoré (1993a) assume a tail condition as in Heckman and Singer (1985).

of Elbers and Ridder (1982) 'weak' in the sense that the model of Elbers and Ridder (1982) cannot be estimated at rate $N^{-1/2}$. As in Honoré (1993a), we do not need the finite mean assumption which gives an intuitive explanation why we can estimate the model at rate $N^{-1/2}$.

This paper is organized as follows. Section 2 discusses the mixed proportional hazard model (with heterogeneity) and presents our estimator. Section 3 shows that our estimator converges at the regular rate and is asymptotically normally distributed. Section 4 adjust the objective function for the case of an endogenous regressor. Section 5 shows that misspecifying the heterogeneity yields inconsistent estimates, even if the baseline hazard is nonparametric. Section 6 presents an empirical example and section 7 concludes.

2. Mixed Proportional Hazard Model

Lancaster (1979) introduced the mixed proportional hazard model in which the hazard is a function of a regressor X, unobserved heterogeneity v, and a function of time $\lambda(t)$,

$$\theta(t \mid X, v) = ve^{X\beta_0}\lambda(t). \tag{1}$$

The function $\lambda(t)$ is often referred to as the baseline hazard. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t \mid X)$ to be decreasing with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, v, and genuine duration dependence, $\lambda(t)$, see Lancaster (1990) and Van den Berg (2001) for overviews. Lancaster (1979) uses functional form assumptions on $\lambda(t)$ and distributional assumptions on v to identify the model. Examples by Lancaster and Nickell (1980) and Heckman and Singer (1984), however, show the sensitivity to these functional form and distributional assumptions. We avoid theses functional form and distributional assumptions and consider the mixed proportional hazard model with time-varying regressors,

$$\theta(t|x(t), v) = ve^{x(t)\beta_0}\lambda(t) \tag{2}$$

where $x(t)$ is a set of regressors that can vary with time, v denotes the heterogeneity and is independent of $x(t)$ and $\lambda(t)$ denotes the baseline hazard. We also use $x(t)$ to denote the sequence of the regressors $x(s)$ for $s = 0$ to $s = t$. The mixed proportional hazard

model of equation (2) implies the following survival probabilities,

$$\begin{aligned} P(T \geq t|x(t), v) = \bar{F}(t|x(t), v) = \exp(-v \int_0^t e^{x(s)\beta_0} \lambda(s) ds) \text{ and} \\ P(T \geq t|x(t)) = E_v\{\bar{F}(t|x(t), v)\} = E_v\{\exp(-v \int_0^t e^{x(s)\beta_0} \lambda(s) ds)\}. \end{aligned} \tag{3}$$

In applied work, duration are measured discretely and to fix ideas we assume that the duration are measured on a weekly scale. We also assume that the regressors could only change at the beginning of the week. Let the regressor x_{i1} denote the vector of regressors of individual i during week 1, x_{i2} the regressors of individual i during week two etc. We now can write equation (3) as follows,

$$P(T \geq t|x(t)) = E_v\{\bar{F}(t|x(t), v)\} = E_v\{\exp(-v \sum_{s=1}^{t} e^{x_s\beta_0 + \delta_{0,s}})\},$$

where t is a natural number, $\delta_{0,s} = \ln\{\int_{s-1}^{s} \lambda(s) ds\}$ and we normalize $\delta_{0,1} = 0$. This specification is similar to Han-Hausman (1990) who specify $\delta_{0,s}$ in a similar manner, but who specify and estimate v parametrically, a requirement we remove in this paper.

Kendall (1938) proposes a statistic for rank correlation. If we are interested in the rank correlation between T and the index $X\beta$, then Kendall's (1938) rank correlation has the following form,

$$Q(\beta) = \frac{1}{N(N-1)} \sum_i \sum_j 1\{T_i > T_j\} 1\{X_i\beta > X_j\beta\}.$$

Han (1987) proposes an estimator that maximizes $Q(\beta)$, the rank correlation between T and the index $X\beta$. Under certain assumptions, including that the T only depends on X through the index $X\beta$, maximizing $Q(\beta)$ yields an estimate for β up to scale, excluding the intercept which cannot be estimated.[5].

However, Kendall's (1938) rank correlation cannot be used for the case of time-varying regressors since it is unclear which regressor one should use. We therefore propose the following modification of the rank correlation. In particular, in our model, the expectation does depend on an index, although it has a more complicated form. Define $Z_i(l; \beta, \delta) =$

[5] For this reason, Han (1987) estimates $\beta/||\beta||$; alternatively, a nonzero coefficient of a regressor could be normalized to be one in absolute value, e.g. $|\beta_1| = 1$.

$\sum_{s=1}^{l} e^{X_{is}\beta+\delta_s}$. We propose minimizing the following objective function,

$$Q(\beta,\delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)\}. \tag{4}$$

Thus, $Z_i(l;\beta,\delta)$ is the index *during* the l^{th} period. The intuition for this objective function is the following. We are comparing two different individuals as does Han's objective function. However, we are now also taking account of the outcome in each period through the parameters for the integrated hazard function, δ. The probability that individual i survives period l is larger than the probability that individual j survives period k if and only if $Z_i(l;\beta_0,\delta_0) < Z_j(k;\beta_0,\delta_0)$. Vice versa if $Z_i(l;\beta_0,\delta_0) > Z_j(k;\beta_0,\delta_0)$. Thus, we use the outcomes for individuals i and j together with these probabilities to yield an objective function that permits identification of the parameters β_0 and δ_0, without the restriction of only up to scale as in the Han approach. Consider the expectation of the objective function,

$$E\{Q(\beta,\delta)\} = \frac{\sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K}}{N(N-1)} E[\{e^{-vZ_i(l;\beta_0,\delta_0)} - e^{-vZ_j(k;\beta_0,\delta_0)}\} \cdot 1\{Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)\}].$$

This expectation of the objective function is minimized at the true value of the parameters. To see this, suppose that $Z_i(l;\beta_0,\delta_0) > Z_j(k;\beta_0,\delta_0)$ so that $e^{-vZ_i(l;\beta_0,\delta_0)} < e^{-vZ_j(k;\beta_0,\delta_0)}$. Thus, $\{\beta,\delta\} = \{\beta_0,\delta_0\}$ minimizes $[E_v\{e^{-vZ,_i(l;\beta_0,\delta_0)} - e^{-vZ_j(k;\beta_0,\delta_0)}\} 1\{Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)\} | Z]$ for each set $\{i,j,k,l\}$ and therefore the expectation of the sum.[6] Note that our approach focuses on the probability than an individual i survives period l (measured from time 0) which permits a convenient treatment of the heterogeneity in comparison with the "traditional" approach that focuses on the hazard function. By only using comparisons measured from time $t = 0$ we are able to "condition out" the heterogeneity distribution. The more traditional hazard approach considers the probability of survival conditional on individual i surviving up to period l which requires an explicit treatment of the heterogeneity distribution.

The definition of $Q(\beta,\delta)$ that is given above contains a double sum so that the number of computational operations for calculating $Q(\beta,\delta)$ is N^2 (note that L and K are fixed).

[6] In the appendix 1 we show that the true value *uniquely* minimizes the expectation of the objective function.

In order to reduce the number of computational operations to be of the order $N \ln N$, we use the rank operator. In particular, let $d_r = 1\{T \geq r\}$ for the vector T of length N. Let d be constructed by stacking the vectors d_r vertically for all $r = 1, ..., K$. Now both d and Z are of dimension $NK \times 1$. If a regressor is continuously distributed conditional on the other regressors, then we can re-write $Q(\beta, \delta)$ using these vectors and the rank function,

$$Q(\beta, \delta) = \frac{1}{N(N-1)} \sum_{j=1}^{NK} d(j)[2 \cdot Rank(Z(j)) - NK].$$

The computational burden to calculate[7] $Q(\beta, \delta)$ is proportional to $N \ln(N)$.

Note that we have identification of β rather than identification only up to an unknown scale coefficient, which is the usual outcome of most previous approaches to the problem. Also, note that by focussing on survival from the beginning of the sample, we have eliminated the requirement to specify the heterogeneity distribution since no survival bias (dynamic sample selection) occurs in our sample comparisons. Our identification is somewhat similar to the nonconstructive identification result of Elbers and Ridder (1982) in the sense that we also assume a continuously distributed regressor. However, our identification results differs in two important ways. First, our identification proof is constructive in the sense that it suggests an estimator. Second, our identification result does *not* rely on an iterative procedure. An iterative procedure typically precludes $\mathrm{N}^{1/2}$ consistency[8].

3. Large Sample Properties

In this section, we derive the large sample properties of our estimator. Suppose that $x^k = \{x_1, ..., x^k\}$, $x_1, ..., x_k$ are scalars. We say that the density of the regressor is positive around x^k if at least one element of x^k is continuously distributed and the density is positive around all continuously distributed elements of x^k. We assume that we observe $\{T_i, x_i\}$ where T_i is a natural number and $T_i \in [0, K]$, $K > 1$. For example, we observe unemployment duration, which is measured in weeks, and want to estimate the integrated baseline hazard at the end of each week. We assume the following.

[7] Suppose we have an ordered vector of length $N-1$; calculating the rank of a new, N^{th} observation is $\ln(N)$. We can see this by observing that having $2(N-1)$ elements to begin with would require us to compare the '*new*' observation to the median of the $2(N-1)$ elements; we are then back to comparing the new element to $N-1$ observation. Thus, the extra cost is $\ln(N)$. The summation then yields the rate $N \ln(N)$.

[8] Indeed, Hahn (1994) shows that the identification result of Elbers and Ridder (1982) holds for singular information matrices, so that no $\sqrt{N}$ estimator exists.

ASSUMPTION 1 (TIME-VARYING REGRESSORS): *Let (i)* $\{T, v, x\}$ *be a random sample,* $x = \{x_1, ..., x^K\}$, $x_1, ..., x_K$ *are scalars,* $\{T, x\}$ be observed and $K \geq 2$; *(ii)* v *and* x *are independent, (iii)* $Pr(T \geq l|x) = E_v \exp\{-v \sum_{s=1}^{l} e^{x_s\beta+\delta_s}\}$ *for* $l = 1, ..., K$; *(iv)* δ_1 *is normalized to be zero, let* $\{\beta, \delta\} \in \Theta$, *which is compact; (v) let* G *be a* K *by* K *matrix and let the element* G_{lk} *be equal to one if* $\exists$ *a pair* $\{x^l, x^k\} \in \mathbb{R}^{2K}$ *such that* $Pr(T \geq l|x^l) = Pr(T \geq k|x^k)$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x^l *or* x^k *and let* G_{lk} *be zero otherwise; let the matrix* G *represent a connected graph; (vi) either (a)* $x^r = x_c$ if $x_1 = x_2 = ... = x_r = x_c$ (thus, $x^1 = x_c$), $Pr(x^K = x_c) > 0$ *for some* $x_c \in \mathbb{R}$, let G^* *be a* K *by* K *matrix and let the element* G^*_{lk} *be equal to one if* $\exists$ *a pair* $\{x_c, x^k\} \in \mathbb{R}^{K+1}$ *such that* $Pr(T \geq l|x_c) = Pr(T \geq k|x^k)$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x_c *or* x^k *and let* G^*_{lk} *be zero otherwise; let the matrix* G^* *represent a connected graph; or (b)* $x_{l,1}|x_{l,2}, x_{l,3}, ..., x_{l,K}, x_{l-1}, x_{l-2}, ...$ *is continuously distributed for all* l, *and* $Pr(T \geq l|x^l) = Pr(T \geq k|x^k)$ *where* $x_{l,1}$ *is in the interior of the support of* $x_{l,1}|x_{l,2}, x_{l,3}, ..., x_{l,K}, x_{l-1}, x_{l-2}, ...$ *for all* k; *(vii)* $\exists$ *a pair* $\{x_1, x'_1, x'_2\} \in \mathbb{R}^3$, $x'_1 \neq x'_2$, *such that* $0 < \Pr(T \geq 1|x_1, x_2) = Pr(T \geq 2|x'_1, x'_2) < 1$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x_1 *or* $\{x'_1, x'_2\}$.

Conditions (i)-(vi(a)) ensure identification up to scale and condition (i)-(vii) ensures complete identification[9]. Condition (iii) is satisfied if the data generating process is the mixed proportional hazard model of equation (2) with exogenous regressors that can change at the beginning of each period. Condition (vi) assumes that either (a) the regressor are constant with positive probability or (b) a regressor is continuously distributed conditional on the other regressors and earlier realizations of the regressors. The substantial restriction of condition (vii) is that one of the regressors varies with time; the theorem below still holds if condition (vii) holds after relabelling the periods. For example, one can label week 1 through 8 as period 1 so that condition (vi) holds while the other condition hold before or after relabelling.

[9] Matrices with only zeros and ones can be represented by graphs; a connected graph means that, informally speaking, you can 'travel' from one point to any other point but not necessarily directly. Condition (v) is considerably weaker than a condition that a regressor has a positive conditional density on the whole real line.

Theorem 1:

Let assumption 1 hold. Let δ be contained in a compact subset D of $\mathbb{R}^K$ and normalize $\delta_1 = 0$. Let $\{\hat{\beta}, \hat{\delta}\} = \underset{\beta,\delta}{argmin}\ Q(\beta, \delta)$ where

$$Q(\beta,\delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{K} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l) < Z_j(k)\}.$$

Then

$$\{\hat{\beta}, \hat{\delta}\} \underset{p}{\rightarrow} \{\beta, \delta\}.$$

Suppose that the regressor is a vector instead of a scalar. The easiest way to prove identification for that case is by noting that one can identify the regressor up to scale using only observations of the first period. In particular, the parameter vector could be estimated up to scale using the maximum rank correlation estimator (MRC). Rank correlation was introduced by Kendall (1938) and Han (1987) proposed the MRC estimator. Han (1987) assumes that one regressor has infinite support conditional on all other regressors. We weaken this assumption. In particular, if all regressors are distributed continuously, then we only require one regressor is continuously distributed conditional on the others without any support restrictions. In order to estimate β up to scale, we assume the following.

ASSUMPTION 2: *Let (i) β be contained in a compact subset $\widetilde{B}$ of $\mathbb{R}^q$ (ii) $Pr(T \geq 1|x_1) = G(x_1\beta, v)$* where $G(.,.)$ is a strictly monotonic decreasing function in its first argument; *(iii) $\{T, v, x\}$ be a random sample (iv) let $x_1 = \{x_{1,1}, \tilde{x}_1\}$, let $\tilde{S}$ denote that support or a subset of the support of $\tilde{x}_1$, let $\tilde{S}_1$ denote the interior of the support of the continuously distributed $x_{1,1}$ conditional on $\tilde{x}_1$ and let, for all $\tilde{x}_1 \in \tilde{S}_1$, there be an $x_{i1,1} \in \tilde{S}_1$ such that $0 < \Pr(T \geq 1|x_{1,1}, \tilde{x}_1) = p < 1$ for some p; (v) let S denote the support of x conditional on $\tilde{x} \in \tilde{S}_1$ and assume that this support of x, S, is not contained in any proper linear subspace of $\mathbb{R}^q$, (v) $\beta_1 \neq 0$, and (vi) v and x are independent.*

Proposition 1

Let assumption 2 hold. Let $\{\hat{\kappa}\} = \underset{\kappa:|\kappa_1|=1}{\operatorname{argmin}} Q(\kappa)$ *where*

$$Q(\kappa) = \frac{1}{N(N-1)} \sum_i \sum_j [1\{T_i \geq 1\} - 1\{T_j \geq 1\}] 1\{Z_i(1) < Z_j(1)\}.$$

Then

$$\hat{\kappa} \underset{p}{\to} \beta/|\beta_1|.$$

This proposition states that β can be estimated up to scale under weaker support conditions than presented by Han (1987). In particular, if all regressors are distributed continuously, then we only require that $x_{1,1}$ is continuously distributed on a small interval without assuming that it has support over the whole real line[10]. If regressor has a discrete distribution and the support of the continuously distributed variables is small, then we can first condition on the regressor with the discrete distribution and identify the whole model using theorem 1 and proposition 1. We then can try to identify the coefficient on the regressor using the objective function of equation (4). Alternatively, we can check whether this objective function empirically identifies the parameters. Suppose that none of the regressors varies with time (most likely, this would be due to quality of the data) and that we want to estimate β up to scale. We can then use the objective function of equation (4). Besides the mild support condition, this objective function can also handle known censoring points that depend on the regressors while Han's (1987) objective function cannot handle such censoring.

Choosing $G(x_1\beta) = E_v \exp(-ve^{x_1\beta})$ in proposition 1 and combining the theorem 1 and proposition 1 yields a consistency result for $\{\hat{\beta}, \hat{\delta}_2, ..., \hat{\delta}_K\}$. Thus, instead of estimation of β up to scale, the objective function $Q(\beta, \delta)$ permits estimation of the β, including the scale.

Theorem 2 (Consistency):

Let assumption 1-2 hold. Let $Pr(T \geq l|x) = E_v \exp(-\sum_{s=1}^{s=l} ve^{x_s\beta+\delta_s})$. *Let* δ *be contained*

[10] To see this, consider choosing $\tilde{S}$ such that $x_1^*\beta_{1,0} < \tilde{x}\tilde{\beta}_0 < x_1^{**}\beta_{1,0}$ for some x_1^* and $x_1^{**} \in \tilde{S}_1$ and note that there must exist such an x_1^* and x_1^{**} since $\tilde{S}_1$ contains an interval.

in a compact subset D of $\mathbb{R}^K$ and normalize $\delta_1 = 0$. Then

$$\{\hat{\beta}, \hat{\delta}\} \underset{p}{\to} \{\beta, \delta\}.$$

3.1 Asymptotic Distribution

In this subsection, we derive the asymptotic distribution of our estimator. As before, we use the following objective function, where $\theta = \{\beta, \delta\}$,

$$Q_N(\theta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l) < Z_j(k)\}.$$

In the appendix, we show that

$$Q_N(\theta) = \frac{\sum_i}{N} \sum_{l=1}^{L} 1\{T_i \geq l\} K[1 - 2\hat{F}_Z\{Z_i(l)\}]. \tag{5}$$

where $\hat{F}_Z\{Z_i(l)\} = \frac{\sum_j}{N-1} \frac{\sum_{k=1}^{K}}{K} 1\{Z_j(k) < Z_i(l)\}$. Note that $E[\hat{F}_Z\{Z_i(l)\}|Z_i(l)] = F_Z\{Z_i(l)\}$ where F_Z is the cumulative distribution function of $Z_i(l)$ for $l = 1, ..., K$ and $i = 1, ..., N$. Let $Q_0(\theta)$ be twice continuously differentiable at θ_0 with respect to θ and let H denote the second derivative divided by the constant K and evaluated at θ_0, i.e.

$$H = \frac{1}{K} \nabla_{\theta\theta} Q_0(\theta_0).$$

We assume the following.

Assumption 3 (Interior): *Let $\theta_0 = (\beta, \delta) \in$ Interior(Θ), where Θ is compact.*

Let $f_Z\{Z_i(l)\}$ denote the density of $Z_i(l)$.

Assumption 4: *Let (i) the second derivative H be nonsingular; (ii) let $f_Z(z)$ be differentiable and let $|f_Z(z)\frac{\partial Z}{\partial \theta}| < M$ for all θ, $|\frac{df_Z\{z\}}{dz}| < M$ for all z and for some $M < \infty$.*

Assumption 4 is a standard regularity condition and supports an argument based on a Taylor expansion[11].

[11] We cannot immediately apply Sherman (1993) since he requires that $Q_N(\theta_0) - Q_0(\theta_0) = O_p(N^{-1})$, an assumption that is violated for our objective function. Therefore, we apply Newey (1991) and Newey and McFadden (1994, lemma 2.8 and section 7).

Theorem 3 (Asymptotic Normality)

Let assumption 1-4 hold. Then

$$\sqrt{N}\{\hat{\theta}-\theta\} \underset{d}{\rightarrow} N(0, H^{-1}\Omega H^{-1})$$

where $\Omega = E[D_N(\theta_0)D_N(\theta_0)']$ and

$$D_N(\theta) = -2\frac{\sum_i}{\sqrt{N}}[\sum_{l=1}^{L} 1\{T_i \geq l\} f_Z\{Z_i(l)\}\frac{\partial Z_i(l)}{\partial \theta} - E[\sum_{l=1}^{L} 1\{T_i \geq l\} f_Z\{Z_i(l)\}\frac{\partial Z_i(l)}{\partial \theta}]].$$

The function $D_N(\theta)$ is an 'approximate derivative' and an 'influence function' in the terminology of Newey and McFadden (1994). It allows to view the asymptotic behavior of an estimator as an average, multiplied by $\sqrt{N}$. Moreover, bootstrapping an asymptotically normally distributed estimator that can be represented by an influence function yields a consistent variance-covariance matrix and consistent confidence intervals, see Horowitz (2001, theorem 2.2)[12]. In the application, we bootstrap the estimator.

The matrix $\Omega = E[D_N(\theta_0)D_N(\theta_0)']$ can be estimated using a sample analogue where $f_Z\{Z_i(l)\}$ can be estimated using a second order kernel that omits observation i. In order to estimate H let e_i denote the ith unit vector, ε_N a small positive constant that depends on the sample size, and $\hat{H}$ the matrix with i, jth element

$$\hat{H}_{ij} = \frac{1}{4\varepsilon_N^2}[\hat{Q}(\hat{\theta}+e_i\varepsilon_N+e_j\varepsilon_N)-\hat{Q}(\hat{\theta}-e_i\varepsilon_N+e_j\varepsilon_N)-\hat{Q}(\hat{\theta}+e_i\varepsilon_N-e_j\varepsilon_N)+\hat{Q}(\hat{\theta}-e_i\varepsilon_N-e_j\varepsilon_N)].$$

Lemma 1 (Newey and McFadden, Estimating H)

Let the conditions of theorem 3 be satisfied. Let $\varepsilon_N \rightarrow 0$ and $\varepsilon_N\sqrt{N} \rightarrow \infty$. Then $\hat{H} \underset{p}{\rightarrow} H$.

Theorem 3 requires the regressors to be exogenous. Sometimes a regressor can qualify as an exogenous regressor, even if its value depend on survival up to a certain point. For example, a treatment that is randomly assigned with probability p_h to individuals who survived h periods may appear to be endogenous since it depends on survival. However, in this duration framework, we can relabel the treatment as if it is given at the beginning of the spell with probability p_h and consider the randomly assigned treatment

[12] Horowitz (2001, theorem 2.2) averages $g_n(X_i)$.

exogenous[13]. In the next section, we consider endogenous regressors, such as randomly assigned treatment with partial compliance.

Our estimates of $\{\delta_1, ..., \delta_K\}$ imply an estimate for the the integrated hazard. In particular, suppose that we measure survival at $\{0, 1, ..., K\}$, e.g. weekly unemployment data, then

$$\widehat{\Lambda(t)} = \sum_{s=1}^{s=t} \exp(\hat{\delta}_s) \text{ where } t \in \{0, 1, ..., K\}.$$

We define the average hazard on the interval $[a, b)$ to be the value λ for which $\int_a^b \lambda(s)ds = \Lambda(b) - \Lambda(a)$. This gives an expression for the average hazard,

$$\widehat{\lambda(s)} = \exp(\hat{\delta}_t) \text{ for } t - 1 < s < t.$$

If the duration are measured on a fine grid, then one could also approximate the hazard by numerically differentiating the integrated hazard $\widehat{\Lambda(t)}$. Thus, we can estimate the integrated hazard rate at each point and also approximate the hazard rate at each point. This differs considerably from Chen (2002), who only estimates the logarithm of the integrated hazard up to a unknown scalar, so that we do not know whether the hazard is increasing or decreasing.

4. An Endogenous Regressor

The last section dealt with exogenous regressors. However, some regressors are endogenous in the sense that the regressor depends on the unobserved heterogeneity. This situation occurs often in panel data and the genesis of the problem and an approach to a solution to the problem are discussed in e.g. Mundlak (1961), Hausman and Wise (1979) and Hausman and Taylor (1981). For example, in the National Supported Work Demonstration[14] data, long term unemployed individuals are randomly offered training but some choose not to participate. Thus, there is a partial compliance problem and the treatment indicator can depend on unobserved heterogeneity. See also Heckman, LaLonde, and Smith (1999). The duration model of this paper gives a natural framework to handle survival

[13] In particular, individuals that do not survive up to period h will be assigned treatment with probability p_h; an alternative is to use a weighting function that gives the weiths p_h and $(1 - p_h)$ to both possible outcomes.

[14] Ham and LaLonde (1996) discuss this data

selection and time-varying regressors. As discussed in the last section, the model can handle survivor selection without using instrumental variables. However, some treatment are not just endogenous in the sense of dynamic or survival selection but are also endogenous in the sense that the treatment still depends on the unobserved heterogeneity v, even after conditioning on survival. Let $R \in \{0,1\}$ denote the treatment assignment and let $X \in \{0,1\}$ denote actual treatment. Let R be randomly assigned among the individuals that are unemployed at time h. Suppose that an individual can refuse treatment, that is, we can observe $R = 1$ and $X = 0$ for a particular individual. The refusal of treatment, or equivalently, the choice of participating, can potentially depend on the unobserved heterogeneity v or on the observed regressors. If the probability of X depends on v, the distribution $p(v|X = 1)$ is different from $p(v|X = 0)$. In particular, let X be a function of R, v and other exogenous regressor and random noise. Since the distribution of v depends on X, we have, in general,

$$E_v\{e^{-vZ_i(l)}|Z_i(l), X = 0\} \neq E_v\{e^{-vZ_j(l)}|Z_j(l) = Z_i(l), X = 1\}.$$

Therefore, $E_v\{e^{-vZ_i(l)}|Z_i(l), X\}$ may not be decreasing in $Z_i(l)$. Therefore, we need to adjust the objective function $Q(\beta, \delta)$ that was introduced above,

$$Q(\beta,\delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)\}. \tag{6}$$

One can view the indicators $1\{T_i \geq l\}$ and $1\{T_j \geq k\}$ as estimators of survival functions. In order to deal with the self selection into treatment, we replace these indicators by other estimates of survivor functions. In particular, we replace $1\{T_i \geq l\}$ and $1\{T_j \geq k\}$ by survivor functions that have the same unobserved heterogeneity distribution so that we do not have to explicitly model the distribution of the heterogeneity. Suppose that individuals are treated at the beginning of period h. In order to avoid survival bias, we condition on survival up to h and also on the index at $h-1$, $Z(h-1)$. Let R denote the treatment intention[15], X the actual treatment and $R, X \in \{0,1\}$. For now, we assume that $R = 0$ implies $X = 0$ and that $P(X = 1|R = 1) > P(X = 1|R = 0)$. Suppose that there are individuals that could have different values of R or X but that have identical values of

[15] The support of any instrument can be reduced to two points.

exogenous regressors. For example, they became unemployed at the same time and also have the same exogenous regressors but can have different values of R or X. Let these groups be denoted by $G_1, ..., G_M$ and let κ be an estimate of the odds ratio, $\kappa = \frac{1-\hat{p}}{\hat{p}}$ where $\hat{p} = \frac{\sum_i 1\{T_i \geq h\} 1\{R_i = 1\}}{\sum_i 1\{T_i \geq h\}}$. For each group, we construct the following distribution functions,

$$\hat{F}_{g,11} = \frac{\sum_{i \in g} \sum_{l=h}^{K} 1\{T_i \geq l\} 1\{X_i = 1\}}{\sum_{i \in g} \sum_{l=h}^{K} 1\{T_i \geq h\} 1\{X_i = 1\}} \text{ and}$$

$$\begin{aligned} \hat{F}_{g,00} &= \frac{\sum_{i \in g} \sum_{l=h}^{K} [1\{T_i \geq l\} 1\{R_i = 0\} 1\{X_i = 0\} - \kappa \cdot 1\{T_i \geq l\} 1\{R_i = 1\} 1\{X_i = 0\}]}{\sum_{i \in g} \sum_{l=h}^{K} [1\{T_i \geq h\} 1\{R_i = 0\} 1\{X_i = 0\} - \kappa \cdot 1\{T_i \geq h\} 1\{R_i = 1\} 1\{X_i = 0\}]} \\ &= \frac{\sum_{i \in g} \sum_{l=h}^{K} 1\{T_i \geq l\} 1\{X_i = 0\} [1\{R_i = 0\} - \kappa \cdot 1\{R_i = 1\}]}{\sum_{i \in g} \sum_{l=h}^{K} 1\{T_i \geq h\} 1\{X_i = 0\} [1\{R_i = 0\} - \kappa \cdot 1\{R_i = 1\}]}. \end{aligned}$$

We use these functions instead of $1\{T_i \geq l\}$ and $1\{T_j \geq k\}$ in equation (6). Define $Z_{g,11}(l)$ as the index of the the individuals of group g with $R = X = 1$. Similarly, $Z_{g,00}(l)$ is the index of group g for which $R = X = 0$. Define

$$Q_1^*(\beta, \delta) = \frac{\sum_g}{M} \sum_{l=h}^{K} \sum_{k=h}^{K} [\hat{F}_{g,11}(l) - \hat{F}_{g,00}(k)] 1\{Z_{g,11}(l) < Z_{g,00}(k)\}.$$

For the periods $1, 2, ..., (h-1)$ we can use the same statistic[16] as in earlier sections,

$$Q_2^*(\beta, \delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{h-1} \sum_{k=1}^{h-1} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l; \beta, \delta) < Z_j(k; \beta, \delta)\}.$$

Moreover, since $R = 0$ implies $X = 0$ we can also define R^* as in the last section. In particular, Let $R^* = R$ for all individuals who have been assigned treatment and for all others we assign $R^* = 0$ with probability $1 - \hat{p}$, i.e. $P(R^* = 0|R = 0) = 1$ and $P(R^* = 0|T_i < h) = 1 - \hat{p}$. We then define

$$Q_3^*(\beta, \delta) = \frac{\sum_i}{N} 1(R_i^* = 0) \frac{\sum_j}{N-1} 1(R_j^* = 0) \sum_{l=1}^{K} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l; \beta, \delta) < Z_j(k; \beta, \delta)\}.$$

We then minimize the following objective function,

$$Q^*(\beta, \delta) = w_1 Q_1^*(\beta, \delta) + w_2 Q_2^*(\beta, \delta) + (1 - w_1 - w_2) Q_3^*(\beta, \delta)$$

where $0 < w_1 < 1$ and $0 < w_2 < 1$. Let the random assignment or instrumental variable assumptions mentioned above hold and let the specification assumptions of theorem 2

[16] If we calculate the rank of the indices Z_{ik} for $k = 1, ..., (h-1)$ and $i = 1, ..., N$, then we can use $Q(\beta, \delta) = \frac{1}{N(N-1)} \sum_{j=1}^{N(h-1)} d(j)[2 \cdot Rank(Z(j)) - N(h-1)]$.

hold so that $Q_1^*(\beta, \delta)$ has a maximum at the true value of the parameters. If M, the number of groups, is finite, than $Q_1^*(\beta, \delta)$ does not point identify any parameters. This complicates the choice of the weights. Suppose that the limit of $Q_1^*(\beta, \delta)$, $Q_2^*(\beta, \delta)$ and $Q_3^*(\beta, \delta)$, together, identify the parameters but that $Q_1^*(\beta, \delta)$ does not play a role in local asymptotics. Let the assumptions of theorem 2 and 3 hold for the objective functions $Q_2^*(\beta, \delta)$ and $Q_3^*(\beta, \delta)$. In that case, consistency and asymptotic normality follows from theorem 2 and 3 for any $0 < w_1 < 1$ and $0 < w_2 < 1$. In particular, after choosing $w_1 = w_2 = \frac{1}{3}$ to get an initial estimate, one can calculate the asymptotic variance-covariance matrix using theorem 2 and lemma 1 using $Q_2^*(\beta, \delta)$ and $Q_3^*(\beta, \delta)$. The ratio $w_2/(1 - w_1 - w_2)$ can then be chosen to minimize a function of the asymptotic variance. In a second step, one can minimize $Q^*(\beta, \delta)$ using the estimate the ratio estimates for the weights [17]. For finite M, one can use $Q_1^*(\beta, \delta)$ without using $Q_2^*(\beta, \delta)$ or $Q_3^*(\beta, \delta)$ to derive bounds but this is beyond the scope of the paper.

The objective function $Q_1^*(\beta, \delta)$ can be interpreted as conditioning on both survival up to the end of period h as well as $Z(h)$ which removes possible dependence between treatment assignment and the unobserved heterogeneity term. This data generating process resembles the data of Ham and LaLonde (1996); see also Heckman, LaLonde, and Smith (1999). We can extend the analysis in a straightforward manner to the situation of noncompliance in both treatment and control individuals, so that $R = 1$ and $X = 0$ for a particular individual and $R = 0$ and $X = 1$ for another individual. However, since the latter situation is relatively unlikely to occur in practice, we leave the details as an exercise.

5. Gamma Mixing Distribution

Han and Hausman (1990) and Meyer (1990) use a flexible baseline hazard and model the unobserved heterogeneity as a gamma distribution. In this section we discuss the sensitivity of the estimators of the MPH model to misspecification of the mixing distribution. In particular, misspecifying the heterogeneity yields inconsistent estimators and having a flexible integrated baseline hazard $\Lambda(t)$ does not compensate for a failure to control for

[17] An easier but computationally more intensive way is to determine the asymptotic variance using bootstrap and to try several values for $w_1 \in [0, 1]$, $w_2 \in [0, 1]$.

heterogeneity. We illustrate this using two examples.

Example 1:

Suppose we estimate the following hazard model, $\theta(t|v,x) = \phi^x \lambda(t)$. The function $\lambda(t)$ is nonparametric and one could (incorrectly) think that the flexibility of this function 'compensates' for the lack of unobserved heterogeneity. This model implies that the following survivor function, $P(T \geq t|x) = \bar{F}(t \mid x) = \exp(-\phi^x \Lambda(t))$. Suppose we observe $\bar{F}(t \mid x)$ for $x = 0, 1$ and all $t \geq 0$. We define $\bar{F}_0(t) = \bar{F}(t \mid x = 0)$ and estimate $\Lambda(t)$,

$$\hat{\Lambda}(t) = -\ln \bar{F}(t \mid x = 0)$$

For a given $\hat{\Lambda}(t) = -\ln \bar{F}(t \mid x = 0)$, the MLE of ϕ can be derived (see appendix) and it can be shown that

$$\operatorname*{plim}_{N\to\infty} \hat{\phi} = \frac{-1}{E[\ln\{\bar{F}_0(T)\}|x = 1]}.$$

where $\bar{F}_0$ is the survival function for $x = 0$. Suppose that $v \sim \text{Gamma}(\alpha, \alpha)$, so that $\bar{F}_0(t) = \left(1 + \frac{\Lambda(t)}{\alpha}\right)^{-\alpha}$ and $-\ln F_0(t) = \alpha \ln\left(1 + \frac{\Lambda(t)}{\alpha}\right)$. Note that $\phi^x \Lambda(T) = \frac{Z}{v}$ where Z has an exponential distribution with mean one. This yields

$$\operatorname*{plim}_{N\to\infty} \hat{\phi} = \frac{1}{E[\alpha \ln\{1 + Z/(\phi v \alpha\}]}.$$

where $v \sim$Gamma(α, α). Note that ϕ only appears in the denominator of an argument of a logarithmic function. This does not bode well for consistency. Using $N = 10,000$ we find the following,

True ϕ	True α	plim $\hat{\phi}$
$\phi = 2$	$\alpha = 1$	$\hat{\phi} = 1.46$
$\phi = 2$	$\alpha = 2$	$\hat{\phi} = 1.09$
$\phi = 10$	$\alpha = 1$	$\hat{\phi} = 4.04$
$\phi = 10$	$\alpha = 2$	$\hat{\phi} = 3.20$

■

Example 2:

Suppose we estimate the following hazard model, $\theta(t|v,x) = ve^{x\beta}\lambda(t)$ where v has a gamma distribution. The function $\lambda(t)$ is nonparametric and this time one could (incorrectly) think that the flexibility of this function 'compensates' for the restrictive assumption that v has a gamma distribution.

Suppose the data is generated by a hazard model, $\theta(t|v,x) = ve^{x}\lambda(t)$ where $p(v) = e^{c-v}$, $v \geq c$ and $c \geq 0$. Thus, v is an exponential stochast to which the nonnegative number c is added and the true value of β equals one.

Consider estimating this model under the assumption of gamma heterogeneity. Without loss of generality, we can write the integrated baseline hazard as follows,

$$\Lambda(t) = H(t)^{d}$$

where $H(t)$ is unrestricted and $d > 0$. Horowitz (1996) and Chen (2002) show how to estimate $H(t)$ at rate $\sqrt{N}$. Suppose that the conditions of Horowitz (1996) or Chen (2002) are fulfilled and that one first estimates $H(t)$ using one of these methods. Estimating d is then like estimating a Weibull model. In the appendix, we show that the inconsistency of β does not depend on the distribution of the regressors. Using $N = 10,000$, we found the following,

c	β	γ_v	δ_v	$\beta; \gamma_v = 2, \delta_v = 1$
0	1	1	1	1
0.1	1.11	1.12	0.96	1.06
0.2	1.15	1.23	0.89	1.09
0.3	1.16	1.30	0.84	1.12
0.5	1.17	1.42	0.76	1.14
1	1.21	1.75	0.54	1.21
2	1.30	1.87	0.33	1.27

For $c = 0$, correct specification, all parameters can be consistently estimated; the last column gives estimation results for β for $\gamma_v = 2$ and $\delta_v = 1$. The simulation results show that the inconsistencies increases with c.

■

Note that the asymptotic bias in the examples above does not depend on the shape of the hazard. The following lemma gives a reason for the asymptotic bias.

Lemma 2: Let $\theta(t \mid v,x) = ve^{x\beta}\lambda(t)$ where $v \perp x$. Let $v - c \mid T \geq 0 \sim \text{Gamma}(\gamma_v, \delta_v)$. If $c = 0$, then $\bar{F}(t|x)$ decreases at a polynomial rate. If $c > 0$, then $\bar{F}(t|x)$ decreases at an exponential rate.

The lemma states that the survivor probability as a function of time decreases at a polynomial rate if the unobserved heterogeneity distribution is a gamma distribution but that

the survivor probability decreases at an exponential rate if the unobserved heterogeneity distribution is a shifted gamma distribution. As the examples show, misspecification of the heterogeneity distribution cannot, in general, be corrected by a flexible baseline hazard. The estimator presented in this paper does not rely on specifying or estimating the heterogeneity distribution which explains its better performance in terms of asymptotic bias and consistency.

6. Empirical Results

We estimate our new duration model on a sample of 15,491 males who received unemployment benefits beginning in 1998 in a data set called the Study of Unemployment Insurance Exhaustees public use data. The study was designed to examine the characteristics, labor market experiences, unemployment insurance (UI) program experiences, and reemployment service receipt of UI recipients.[18]

The study sample consists of UI recipients in 25 states who began their benefit year in 1998 and received at least one UI payment, and is designed to be nationally representative of UI exhaustees and non-exhaustees. The data description is:

> "The data come from the UI administrative records of the 25 sample states and telephone interviews conducted with a subsample of these UI recipients. Telephone interviews were conducted in English and Spanish between July 2000 and February 2001 using a two-stage process. For the first 16 weeks, all 25 participating states used mail, phone, and database methods to locate sample members, who were then asked to complete the survey. The second stage, conducted in 10 of the sample states, added field staff to help locate non-responding sample members. The administrative data include the individual's age, race, sex, weekly benefit amount, first and last payment date, the state where benefits were collected, and whether benefits were exhausted." (op. cit.)

The survey data contain individual level information about labor market and other activities from the time the person entered the UI system through the time of the interview. However, we limit our econometric study to the first 25 weeks of unemployment

[18] The following description follows from http://www.upjohninst.org/erdc/uie/datasumm.html which has further details of the sample design and results.

due to the recognized change in behavior in week 26 when UI benefits cease for a significant part of the sample, see e.g. Han-Hausman (1990). The data include information about the individual's pre-UI job, other income or assistance received, and demographic information.

We use two indicator variables, race and age over 50 in our index specification. We also use the replacement rate which is the weekly benefit amount divided by the UI recipient's base period earnings. Lastly, we use the state unemployment rate of the state from which the individual received UI benefits during the period in which the individual filed for benefits. This variable changes over time. Table 1 gives the means and standard deviations for the variables we use in our empirical specification:

Table 1 here

We first estimate the unknown parameters of the model using the gamma heterogeneity specification of Han-Hausman (1990) and Meyer (1990) (HHM). This specification allows for a piecewise constant baseline hazard, which does not restrict the specification since unemployment duration is recorded on a weekly basis. However, it does impose a gamma heterogeneity distribution on the specification which can lead to inconsistent estimates as we discussed above. We estimate the model using a gradient method and report the HHM estimates and bootstrap standard errors in Table 2.

Table 2 here

The estimates of the parameters, as reported in table 2, should not depend on how many weeks of data we use (6, 13 or 24 weeks). However, the coefficients differ significantly. We find significant evidence of heterogeneity in the two larger samples, while in the 6 period sample we do not estimate significant heterogeneity. We also find the expected negative estimates for all of the coefficients with the state unemployment rate a significant factor in affecting the probability of exiting unemployment. When comparing

the estimates of the β_i across the 3 samples, the scaling changes depending on the variance of the estimate gamma distribution. Thus, the ratios of the coefficients should be compared. The ratios of the coefficients across samples remain similar with the results for the 13 period and 24 period very close to each other.

We now turn to estimate of the new duration specification, which does not require estimation of a heterogeneity distribution using the same samples as above. Optimization of the objective function can now create a problem because of its lack of smoothness. Usual Newton-type gradient methods or conjugate gradient (simplex) methods do not work in this situation. To date we have found that generalized pattern search algorithms perform best.[19] We use the pattern search routine from Matlab to estimate the parameters. See the Appendix for further details of our computational approach. The basic idea is to begin with the gamma heterogeneity estimates and to construct a "bounding box" around each parameter estimates of 3 standard deviations. We then find new estimates and increase the bounding box until we do not find an increase in the objective function. The routine converges relatively rapidly. We estimate standard errors using a bootstrap approach. In Table 3 we give the estimates of the new duration model. We also check our pattern search results using a genetic optimization approach that is also discussed in the appendix. The genetic optimization approach has the advantage of not depending on initial values. However, it has the disadvantage of taking much longer to solve so it cannot be used feasibly to bootstrap the results to estimate the standard errors. However, the results of the pattern search algorithm and the genetic optimization algorithm are very similar as we describe in the appendix.

Table 3 here

Again we find that all of the estimated coefficients have the expected negative signs. The coefficients are also estimated with a high degree of statistical precision, although this

[19] Further research would be helpful here. We have also used gradient algorithms on a smoothed objective function to obtain initial estimates and then employed Nelder-Mead routines to find the optima. However, the pattern search algorithms appear to work best. See e.g. Audet and Dennis (2003) for a recent survey of pattern search algorithms.

finding may be a result of our large sample size of 15,491 individuals. We again find that the ratio of coefficients remains relatively stable across the three different samples with the exception of the replacement rate which becomes increasingly larger with respect to the state unemployment rate as the sample length increases. The change in the estimated coefficient for the replacement rate for the 24 week sample appears to arise because most recipients' unemployment insurance terminates after 26 weeks. Han-Hausman (1990) found a significant change in behavior at week 26. As individuals start to approach week 26 the size of the replacement rate has a diminished effect on their behavior as they foresee the end of their unemployment benefits beginning to draw near.

In Figures 1 and 2 we plot the survival curves for the 13 week and 24 week gamma heterogeneity estimates and for the estimates from the new model. We fit the survival curves using a second order local polynomial estimator which takes account of the standard deviations of the estimated period coefficients in Table 2 and 3.[20] The estimated local polynomial survival curves fit the data well for all specifications.

Figure 1 here

Figure 2 here

We find that the results of the new model gives extremely similar results for the 6 period data and the 13 period data. Indeed, a Hausman (1978) specification test on the slope coefficients is 0.42 with 4 degrees of freedom. Thus, we find that the new model is not sensitive to the number of periods used to estimate the model. For the 24 period model we find the coefficients again very close to the other results except for the coefficient of the replacement rate. A Hausman test now rejects the equality of the slope coefficients with a value of 234.3, based essentially on the change in the replacement rate coefficient. However, since most individuals' unemployment benefits run out in the 26th week, the

[20] We explain our approach in more detail in the appendix.

change in the estimated coefficient is likely because of unmodeled dynamics at the point of benefit exhaustion. Lastly, if we test the ratios of the gamma heterogeneity model versus the new duration model we do not reject the ratios are the same for 6 periods with a test value of 3.5; we marginally reject equality of coefficient ratios for 13 periods with a test value of 6.2; and we do reject equality of coefficient ratios for 24 weeks with a test value of 12.4. Thus, the new duration model does find differences from the previous gamma heterogeneity model. The new duration model also has the advantage that the absolute value of the estimated coefficients is not sensitive to the length of the data period, while the gamma heterogeneity model does not have this property.

The main difference we find between the results of the gamma heterogeneity survival curves and the semi-parametric survival curves is that the gamma heterogeneity survival curves are initially steeper. Thus, the gamma heterogeneity results predict a higher probability of exiting unemployment in the early periods than do the semi-parametric results. However, again the differences are not substantial. We reject equality of the survival curves due to the extremely small standard errors we estimate given our very large sample.

7. Conclusion

Since Lancaster (1979), it has been recognized that it is important to account for unobserved heterogeneity in models for duration data. Failure to account for unobserved heterogeneity makes the estimated hazard rate decreases more with the duration than the hazard rate of a randomly selected member of the population. In this paper, we derive a new estimator for the mixed proportional hazard model that allows for a nonparametric baseline hazard and time-varying regressors. By using time varying regressors we are able to estimate the regression coefficients, instead of estimates only up to scale as in some of the previous literature. We also do not require explicit estimation of the heterogeneity distribution in estimating the baseline hazard and regression coefficients. The baseline hazard rate is nonparametric and the estimator of the integrated baseline hazard rate converges at the regular rate, $N^{-1/2}$, where N is the sample size. This is the same rate as for a duration model without heterogeneity. The regressor parameters also converge at

the regular rate. A nice feature of the new estimator is that it allows the durations to be measured on a finite set of points. Such discrete measurement of durations is important in economics; for example, unemployment is often measured in weeks. In that case, the estimator of the integrated baseline hazard only converges at this set of points.

Appendix 1: Proof of Theorem 1

Proof of **Theorem 1**:

We first establish identification and then show that the estimator converges in probability.

Identification:

Let assumption 1 (i)-(vi(a)) and (vii) hold so that a regressor can stay constant over time with positive probability. To simplify the proof we first consider a two period model. Without loss of generality, let $\beta_0 > 0$ (if $\beta_0 < 0$, multiply x by -1). Consider the following reparametrization, $\delta_2 = \ln(e^{\beta c} - 1)$ for some $c > 0$. The same reasoning as in the main text yields that the true values $\{\beta_0, c_0\}$ yield *a* minimum of the expectation of the objective function for any set $\{i, j, k, l\}$ and for any regressor. We now argue that c yields a *unique* minimum. The expectation of the contributions of a subset of the observations that compares realizations of the first and second period, $i \neq j$, of the objective function have the following form,

$$\begin{aligned} & E[\{e^{-vZ_i(l=2;\beta_0,\delta_0)} - e^{-vZ_j(k=1;\beta_0,\delta_0)}\} \cdot 1\{Z_i(l=2;\beta,\delta) < Z_j(k=1;\beta,\delta)\} | x_i, x_j, x_{i1} = x_{i2}] \\ = \; & E([\exp\{-v(e^{x_{i1}\beta} + e^{x_{i1}\beta+\delta})\} - \exp\{-ve^{x_{j1}\beta}\}] \cdot 1\{e^{x_{i1}\beta} + e^{x_{i1}\beta+\delta} < e^{x_{j1}\beta}\}] | x_i, x_j, x_{i1} = x_{i2}). \end{aligned}$$

Using $\delta_2 = \ln(e^{c\beta} - 1)$ for some $c > 0$ yields $e^{x_{i1}\beta} + e^{x_{i1}\beta+\delta} = e^{x_{i1}\beta+c\beta}$. Thus,

$$\begin{aligned} & E[\{e^{-vZ_i(l=2;\beta_0,\delta_0)} - e^{-vZ_j(k=1;\beta_0,\delta_0)}\} \cdot 1\{Z_i(l=2;\beta,\delta) < Z_j(k=1;\beta,\delta)\} | x_i, x_j, x_{i1} = x_{i2}] \\ = \; & E([\exp\{-v(e^{x_{i1}\beta+c_0\beta})\} - \exp\{-ve^{x_{j1}\beta}\}] \cdot 1\{e^{x_{i1}\beta+c\beta} < e^{x_{j1}\beta}\}] x_i, x_j, x_{i1} = x_{i2}) \\ = \; & E([\exp\{-v(e^{x_{i1}\beta+c_0\beta})\} - \exp\{-ve^{x_{j1}\beta}\}] \cdot 1\{c - (x_{j1} - x_{i1}) < 0\}] | x_i, x_j, x_{i1} = x_{i2}) \qquad (7) \\ = \; & E([\exp\{-v(e^{x_{i1}\beta+c_0\beta})\} - \exp\{-ve^{x_{j1}\beta}\}] \cdot 1\{c - x_{ij} < 0\}] | x_i, x_j, x_{i1} = x_{i2}). \end{aligned}$$

where $x_{ij} = x_{j1} - x_{i1}$. Next note that under assumption 1, $\{c_0 - x_{ij}\}$ and $e^{x_{i1}\beta+c_0\beta} - e^{x_{j1}\beta}$ have both support around zero since $\exists$ a pair $\{x_a, x_{b,1}, x_{b,2}\} \in \mathbb{R}^3$, $x_{b,1} = x_{b,2}$, such that $Pr(T \geq 1 | x_{a,1}, x_{a,2}) = Pr(T \geq 2 | x_{b,1}, x_{b,2})$ (equivalently, $x_a = c_0 + x_b$) where the density of the regressor is positive in an arbitrarily small neighborhood around $x_{a,1}$ or $\{x_{b,1}, x_{b,2}\}$. Thus, we can find neighborhoods B in $\mathbb{R}$ such that $dF_{x_{ij}}(B) > 0$, and for each $x_{ij} \in B$ we

have $1\{c_0 - x_{ij} < 0\} \neq 1\{c - x_{ij} < 0\}$. This implies that, for $c \neq c_0$, $i \neq j$ and $x_{i1} = x_{i2}$,

$$\begin{aligned} & E\{Q(\beta^*, c_0)\} - E\{Q(\beta^*, c)\} \\ \geq\ & E[\{e^{-vZ_i(l=2;\beta^*,c_0)} - e^{-vZ_j(k=1;\beta^*,c_0)}\}\cdot \\ \cdot[1\{Z_i(l\ =\ & 2;\beta^*,c_0) < Z_j(k=1;\beta^*,c_0)\} - 1\{Z_i(l=2;\beta^*,c) < Z_j(k=1;\beta,c)\}] \\ \geq\ & E\{([\exp\{-v(e^{x_{i1}\beta+c_0\beta})\} - \exp\{-ve^{x_{j1}\beta}\}]\cdot \\ \cdot[1\{c_0 - x_{ij}\ <\ & 0\} - 1\{c - x_{ij} < 0\}])|x_{ij} \in B\}P(x_{ij} \in B) > 0. \end{aligned}$$

The last equation implies that c_0 is identified.

In order to show identification of β, define

$$\begin{aligned} H_{ij}(\beta, c) &= e^{x_{i1}\beta} + e^{x_{i2}\beta+\delta_2} - e^{x_{j1}\beta} \\ &= e^{x_{i1}\beta} + e^{x_{i2}\beta+c\beta} - e^{x_{i2}\beta} - e^{x_{j1}\beta} \end{aligned} \tag{8}$$

using $\delta_2 = \ln(e^{c\beta} - 1)$. Define

$$H^*_{ij}(\beta, c) = 1 + e^{(x_{i2}-x_{i1}+c)\beta} - e^{(x_{i2}-x_{i1})\beta} - e^{(x_{j1}-x_{i1})\beta}.$$

Differentiating with respect to β gives

$$\frac{\partial H^*_{ij}(\beta, c)}{\partial \beta} = (x_{i2} - x_{i1} + c)e^{(x_{i2}-x_{i1}+c)\beta} - (x_{i2} - x_{i1})e^{(x_{i2}-x_{i1})\beta} - (x_{j1} - x_{i1})e^{(x_{j1}-x_{i1})\beta}.$$

Let $P(T_i \geq 2|x_i) \geq P(T_j \geq 1|x_j)$ so that $E[\exp\{-v(e^{x_{i1}\beta_0} + e^{x_{i2}\beta_0+\beta_0 c_0} - e^{x_{i2}\beta_0})\}|x_i] \geq E[\exp\{-v(e^{x_{j1}\beta_0})\}|x_j]$. This implies that $H_{ij}(\beta_0, c_0) = e^{x_{i1}\beta_0} + e^{x_{i2}\beta_0+\beta_0 c_0} - e^{x_{i2}\beta_0} - e^{x_{j1}\beta_0} \leq 0$ and that $H^*_{ij}(\beta_0, c_0) = 1 + e^{(x_{i2}-x_{i1}+c_0)\beta_0} - e^{(x_{i2}-x_{i1})\beta_0} - e^{(x_{j1}-x_{i1})\beta_0} \leq 0$. Suppose that $x_{i2} - x_{i1} < 0$ so that $1 - e^{(x_{i2}-x_{i1})\beta_0} > 0$ for any value of $\beta_0 > 0$. This implies that $e^{(x_{i2}-x_{i1}+c_0)\beta_0} < e^{(x_{j1}-x_{i1})\beta_0}$ so that $(x_{i2} - x_{i1} + c_0) < (x_{j1} - x_{i1})$. This implies that $\frac{\partial H^*_{ij}(\beta,c_0)}{\partial \beta} < 0$ for all $\beta > 0$ so that $H^*_{ij}(\beta, c_0) < H^*_{ij}(\beta_0, c_0)$ if $\beta > \beta_0$ and $H^*_{ij}(\beta, c_0) > H^*_{ij}(\beta_0, c_0)$ if $\beta < \beta_0$. In particular, given the assumption 1 (v), for those values of the regressors for which $P(T_i \geq 2|x_i, x_{i1} > x_{i2}) \geq P(T_j \geq 1|x_j)$ and $x_{i2} - x_{i1} < 0$, the conditional expectations of the contributions to the objective functions,

$$\{P(T_i \geq 2|x_i, x_{i1} > x_{i2}) - P(T_j \geq 1|x_j)\} * 1\{H^*_{ij}(\beta, c_0) < 0\}$$

are maximized for any value of β for which $\beta \geq \beta_0$.

Now consider $P(T_i \geq 2|x_i, x_{i1} > x_{i2}) \leq P(T_j \geq 1|x_j)$ then $E[\exp\{-v(e^{x_{i1}\beta_0}+e^{x_{i2}\beta_0+\beta_0 c_0}-e^{x_{i1}\beta_0})\}] \leq E[\exp\{-v(e^{x_{j1}\beta_0})\}]$. This implies that $H_{ij}(\beta_0, c_0) = e^{x_{i1}\beta_0} + e^{x_{i2}\beta_0+\beta_0 c_0} - e^{x_{i2}\beta_0} - e^{x_{j1}\beta_0} \geq 0$ and that $H^{**}(\beta_0, c_0) = e^{(x_{i1}-x_{i2})\beta_0} + e^{\beta_0 c_0} - 1 - e^{(x_{j1}-x_{i2})} \geq 0$. Again, suppose that $x_{i2} - x_{i1} < 0$ so that $e^{(x_{i1}-x_{i2})\beta_0} - 1 > 0$ for any value of $\beta_0 > 0$. This implies that $e^{c_0\beta_0} > e^{(x_{j1}-x_{i2})\beta_0}$ so that $c_0 > (x_{j1} - x_{i2})$. This implies that $\frac{\partial H^{**}(\beta, c_0)}{\partial \beta} > 0$. Similar reasoning as above implies that the conditional expectations of the contributions to the objective functions,

$$\{P(T_i \geq 2|x_i, x_{i1} > x_{i2}) - P(T_j \geq 1|x_j)\} * 1\{H^*(\beta, c_0) < 0\}$$

are maximized for any value of β for which $\beta \leq \beta_0$. Thus, β_0 is identified if $x_{i2} - x_{i1} < 0$. A similar reasoning applies if $x_{i2} - x_{i1} < 0$ so that β_0 is identified under the assumptions. Identification of $\{\beta, \delta\}$ is equivalent to identification of $\{\beta, c\}$. Now consider a model with multiple periods. Consider the following reparametrization, $\rho_k = \ln\{\sum_{s=1}^{k} \exp(\delta_s)\}$ for all k. For those individuals whose regressors do not change, we have

$$\begin{aligned} Z_i(k, \beta, \delta) &= \sum_{s=1}^{k} \exp(x_i\beta + \delta_s) = \exp(x_i\beta) \sum_{s=1}^{k} \exp(\delta_s) \\ &= \exp(x_i\beta) \exp[\ln\{\sum_{s=1}^{k} \exp(\delta_s)\}] = \exp(x_i\beta + \rho_k) = Z_i(k, \beta, \rho_k). \end{aligned}$$

Thus, for a subset of the data, we have a single index and assumption 1 identifies these single index parameters up to scale, $\{\beta/|\beta|, \rho_2/|\beta|, \rho_2/|\beta|, ...\}$, using a simplified version of the proof of proposition 1 below. In particular, note that $\exp(x_i\beta + \rho_k) = \exp\{|\beta|(x_i\beta/|\beta| + \rho_k/|\beta|)\}$ and that $\rho_k/|\beta|$ is like a dummy of a particular time period. Then, note that G^*_{1r} must be nonzero for some r since G is a connected graph. This identified ρ_r. Next note that an element of $\{G^*_{1s}, G^*_{rs}\}$ is nonzero for some s since G^* is a connected graph and so on. Thus, only β remains to be identified and we identify it using the same reasoning as for a two scaler period.

Now suppose that assumption 1 (i)-(v), (vi(b)) and (vii) hold. We first consider identification in the two period model. By assumption 1 (vii), we have

$$Pr(T \geq 1|x_{a,1}, x_{a,2}) = Pr(T \geq 2|x_{b,1}, x_{b,2})$$

for some x_a, x_b. This implies that

$$E([\exp\{-v(\exp(x_{a,1}\beta) + \exp(x_{a,2}\beta + \delta))\}]) - \exp\{-v(\exp(x_{b,1}\beta))\}] = 0$$

thus,

$$E([\exp\{-v(\exp(x_{a,1}\beta) \cdot (1 + \exp(\{x_{a,2} - x_{a,1}\}\beta + \delta))\}]) - \exp\{-v(\exp(x_{b,1}\beta))\}] = 0$$

Define $\exp(\delta) = \exp\{(x_{a,1} - x_{a,2})\beta\} \cdot \{\exp(c^*\beta) - 1\}$ for some $c^* > 0$. This yields

$$E([\exp\{-v(\exp(x_{a,1}\beta + c^*))\}]) - \exp\{-v(\exp(x_{b,1}\beta))\}] = 0.$$

This yields the same expected contribution to the objective function as equation (7) so that $\{\beta, c^*\}$ are identified. Next, consider a model with multiple periods. Note that, by assumption 1 (vi(b)), $x_{l,1}|x_{l,2}, x_{l,3}, ..., x_{l,K}, x_{l-1}, x_{l-2}, ...$ is continuously distributed for all t, and $Pr(T \geq l|x^l) = Pr(T \geq k|x^k)$ where $x_{l,1}$ is in the interior of the support of $x_{l,1}|x_{l,2}, x_{l,3}, ..., x_{l,K}, x_{l-1}, x_{l-2}, ...$ for all k. Thus, $x_{l,1}$ is conditionally continuously distributed on an interval for all l so that all parameters are identified at least up to scale. Finally, β can be identified using the same reasoning as for a two scaler period.

Convergence in probability:

Define

$$\begin{aligned} Q_0(\beta, \delta) &= E\{Q_N(\beta, \delta)\} \\ &= E[E\{Q_N(\beta, \delta)|Z\}] \\ &= E[\frac{\sum_i}{N} \sum_{l=1}^{L} E_v\{e^{-vZ_i(l)}|Z_i(l)\} \sum_{k=1}^{K} [2 * F_Z(Z_i(l)) - 1]] \end{aligned}$$

where F_Z is the cdf of $Z_i(l)$ for $l = 1, ..., K$ and $i = 1, ..., N$. The function $Q_0(\beta, \delta)$ is continuous and minimized at the true value of the parameters. The function $Q(\beta, \delta)$ is stochastically equicontinuous and the conditions of Newey and McFadden (1994, lemma 2.8) are satisfied so that $Q(\beta, \delta)$ converges uniformly to $EQ(\beta, \delta)$. Moreover, Θ is assumed to be compact and the data are i.i.d., so that consistency follows from Newey and McFadden (1994, theorem 2.1). Note that these arguments do not require that there is unobserved heterogeneity; they still hold if all individuals have the same value of v.

Appendix 2: Proof of Proposition 1

Proof of **Proposition 1:**

Identification up to scale: Let W^* denote the random variable x_1 for which $\tilde{x}_1 \in \tilde{S}_1$. Let W be the difference between two realizations of W^*. The support of W^*, and therefore of W, is not contained in any proper linear subspace of $\mathbb{R}^q$. This implies that $E\{WW'\}$ is positive definite (e.g. see Newey and McFadden (1994, page 2125)). Therefore, for $\beta^* \neq \beta$, $W'(\beta - \beta^*) \neq 0$ on a set with positive probability so that $W'\beta \neq W'\beta^*$ on a set with positive probability. We need that $1\{W'\beta < 0\} \neq 1\{W'\beta^* < 0\}$ on a set with positive probability. To see that this is the case, note that the first component of W is the difference between two independent and continuously distributed random variables so that the first component of W is also continuously distributed. Next, let $W = \{W_1, \tilde{W}\}$ and note that the support of $\tilde{W}$ is not contained in a linear subspace of $\mathbb{R}^{q-1}$. Moreover, condition (iv) implies that $W'\beta$ is continuously distributed around zero so that β is identified up to scale.

Estimation up to scale: The probability limit of $EQ(\kappa)$ is uniquely maximized at $\kappa = \beta/|\beta_1|$. All conditions of Newey and McFadden (1994, theorem 2.1 and lemma 2.8) are satisfied and consistency follows.

Appendix 3: Proof of Theorem 2

Proof of **Theorem 2**: Note that one can consistently estimate the regressors up to scale by proposition 1 and that $Q(\beta, \delta)$ incorporates the objective function of proposition 1. Consider replacing $x_i\beta = x_i\kappa|\beta_1|$ by $\{x_i\hat{\kappa} \cdot |\beta_1|\}$ in the objective function and note that $x_i\hat{\kappa}$ then plays the role of x_i in theorem 1; note that an additional error term converges to zero in probability and consistency follows from theorem 1 and Newey and McFadden (1994, theorem 2.1 and lemma 2.8).

Appendix 4: Proof of Theorem 3: Asymptotic Normality

$$\begin{aligned}
Q_N(\theta) &= \frac{1}{N(N-1)}\sum_i\sum_j\sum_{l=1}^{L}\sum_{k=1}^{K}[1\{T_i \geq l\} - 1\{T_j \geq k\}]1\{Z_i(l) < Z_j(k)\} \\
&= \frac{1}{N(N-1)}\sum_i\sum_j\sum_{l=1}^{L}[1\{T_i \geq l\}\sum_{k=1}^{K}1\{Z_i(l) < Z_j(k)\} \\
&- \frac{1}{N(N-1)}\sum_i\sum_j\sum_{l=1}^{K}\sum_{k=1}^{L}1\{T_j \geq k\}1\{Z_i(l) < Z_j(k)\} \\
&= \frac{\sum_i}{N}\sum_{l=1}^{L}1\{T_i \geq l\}\frac{\sum_j}{N-1}\sum_{k=1}^{K}[1\{Z_i(l) < Z_j(k)\} - 1\{Z_i(l) > Z_j(k)\}] \\
&= \frac{\sum_i}{N}\sum_{l=1}^{L}1\{T_i \geq l\}\frac{\sum_j}{N-1}\sum_{k=1}^{K}[1 - 2 * 1\{Z_j(k) < Z_i(l)\}]
\end{aligned}$$

with probability one. Thus,

$$Q_N(\theta) = \frac{\sum_i}{N}\sum_{l=1}^{L}1\{T_i \geq l\}K[1 - 2\hat{F}_Z\{Z_i(l)\}].$$

Proof of theorem 2:

We proof theorem 3 by applying Newey (1991) and Newey and McFadden (1994, lemma 2.8 and section 7) and we follow their notation as much as possible.

$$\begin{aligned}
D_N &= -2\frac{\sum_i}{\sqrt{N}}[1\{T_i \geq l\} - E(1\{T_i \geq l\}|X_i)]\sum_{l=1}^{L}f_Z\{Z_i(l)\}\frac{\partial Z_i(l)}{\partial\theta} \\
-2[E(1\{T_i &\geq l\}|X_i)]\sum_{l=1}^{L}f_Z\{Z_i(l)\}\frac{\partial Z_i(l)}{\partial\theta} - E[\frac{\sum_i}{N}\sum_{l=1}^{L}1\{T_i \geq l\}f_Z\{Z_i(l)\}\frac{\partial Z_i(l)}{\partial\theta}]].
\end{aligned}$$

The assumption $|f_Z(z)\frac{\partial Z}{\partial\theta}| < M$, the random sample assumption of assumption 1 and the Lindeberg-Levy central limit theorem implies that $\sqrt{N}D_N(\theta)$ converges to a normal distribution with variance-covariance $\Omega = E[D_N(\theta_0)D_N(\theta_0)']$.

Note that

$$Q_N(\theta) - Q_N(\theta_0) = 2K\frac{\sum_i}{N}\sum_{l=1}^{L}1\{T_i \geq l\}[\hat{F}(Z_{0,i}(l)) - \hat{F}(Z_i(l))]$$

$$Q_0(\theta) - Q_0(\theta_0) = 2K * E_X[\frac{\sum_i}{N}\sum_{l=1}^{L}E\{1(T_i \geq l)|X_i\}[F_Z\{Z_{0,i}(l)\} - F_Z\{Z_i(l)\}]]$$

where the subscript zero denotes that $Z_{0,i}(l)$ is a function of θ_0, the true value. Let $1 - G(w)$ denote the cumulative distribution function of the logistic distribution, $G(w) = \frac{1}{1+\exp(w)}$, and let $G'(w) = -\frac{\exp(w)}{\{1+\exp(w)\}^2}$. Note that $G(u/h) - 1(u > 0)$ decreases exponentially in $1/h$ for all $u \neq 0$.

Let $\tilde{F}(.)$ denote the smoothed $\hat{F}(.)$,

$$\tilde{F}(Z_i(l)) = \frac{\sum_i}{N-1}\sum_{k=1}^{K} G\{\frac{Z_i(l) - Z_j(k)}{h}\}. \tag{9}$$

With probability one, $Z_i(l) - Z_j(k) \neq 0$. Consider u and u_0 and let $\Delta = u - u_0$.

$$G(u/h) = G(u_0/h + \Delta/h) = \frac{1}{1+\exp(u_0/h + \Delta/h)}.$$

Thus,

$$\begin{aligned} G(u/h) - G(u_0/h) &= \frac{1}{1+\exp(u_0/h+\Delta/h)} - \frac{1}{1+\exp(u_0/h)} \\ &= \frac{\exp(u_0/h) - \exp(u_0/h+\Delta/h)}{\{1+\exp(u_0/h)\}\{1+\exp(u_0/h+\Delta/h)\}} \\ &= \frac{\exp(u_0/h)}{\{1+\exp(u_0/h)\}}\frac{1-\exp(\Delta/h)}{\{1+\exp(u_0/h+\Delta/h)\}} \end{aligned}$$

Thus, for $\Delta \to_p 0$ for $N \to \infty$ and $h \propto N^{\delta}$, $\delta < 0$, we have $\{\frac{\sqrt{N}}{|\Delta|}[G(u/h) - G(u_0/h)]\} \underset{p}{\to} 0$. Define

$$q_N(\theta) - q_N(\theta_0) = 2\frac{\sum_i}{\sqrt{N}}\sum_{l=1}^{L} 1\{T_i \geq l\}\sum_{k=1}^{K}\{\tilde{F}(Z_{0,i}(l)) - \tilde{F}(Z_i(l))\}. \tag{10}$$

The above reasoning implies that $\{Q_N(\theta) - Q_N(\theta_0)\}/K$ is closely approximated by $q_N(\theta) - q_N(\theta_0)$. In particular,

$$\sup_{\theta\in\Theta} |\frac{\sqrt{N}}{||\theta-\theta_0||}[\frac{Q_N(\theta) - Q_N(\theta_0)}{K} - \{q_N(\theta) - q_N(\theta_0)\}]| \underset{p}{\to} 0$$

where the uniform convergence follows from Newey (1991). Define $q_0(\theta) - q_0(\theta_0) = E\{q_N(\theta) - q_N(\theta_0)\}$, and define

$$r_N(\theta) = q_N(\theta) - q_N(\theta_0) - \{q_0(\theta) - q_0(\theta_0)\}$$

Note that $r_N(\theta)$ is continuously differentiable. A Taylor approximation around $\theta = \theta_0$ yields

$$r_N(\theta) = \{\frac{\partial q_N(\theta)}{\partial\theta}|_{\theta=\bar{\theta}} - \frac{\partial q_0(\theta)}{\partial\theta}|_{\theta=\bar{\theta}}\}(\theta - \theta_0)$$

for some intermediate value $\bar{\theta} \in [\theta, \theta_0]$. For $h \propto N^{-1/5}$,

$$\begin{aligned}
r_N(\theta) &= \{\frac{\partial q_N(\theta)}{\partial \theta}|_{\theta=\bar{\theta}} - \frac{\partial q_0(\theta)}{\partial \theta}|_{\theta=\bar{\theta}}\}(\theta - \theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L} 1\{T_i \geq l\}\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta - \theta_0) \\
&- 2E[\frac{\sum_i}{N}\sum_{l=1}^{L} 1\{T_i \geq l\}\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta - \theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L}[1\{T_i \geq l\} - E(1\{T_i \geq l\}|X)]\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta - \theta_0) \\
&- 2\frac{\sum_i}{N}\sum_{l=1}^{L}[E(1\{T_i \geq l\}|X)]\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\} \\
&- E[E(1\{T_i \geq l\}|X)\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}]]|_{\theta=\bar{\theta}}(\theta - \theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L}[1\{T_i \geq l\} - E(1\{T_i \geq l\}|X)]\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta - \theta_0) \\
&- 2\frac{\sum_i}{N}\sum_{l=1}^{L}[E(1\{T_i \geq l\}|X)\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\} \\
&- E[E(1\{T_i \geq l\}|X)\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\}]]|_{\theta=\bar{\theta}}(\theta - \theta_0) + o_p(\frac{||\theta - \theta_0||}{\sqrt{N}}) \\
&= -\frac{D_N(\theta_0)}{\sqrt{N}}(\theta - \theta_0) + o_p(\frac{||\theta - \theta_0||}{\sqrt{N}}). \qquad (11)
\end{aligned}$$

The continuous differentiability of $r_N(\theta)$ with respect to θ implies that this convergence is uniform. Thus, $[Q_N(\theta) - Q_N(\theta_0) - \{Q_0(\theta) - Q_0(\theta_0)\}]/K$ can be approximated by r_N and the continuously differentiable r_N can be approximated by D_N. Define

$$R_N(\theta) = \sqrt{N}[Q_N(\theta) - Q_N(\theta_0) - K \cdot D_N(\theta - \theta_0) - \{Q_0(\theta) - Q_0(\theta_0)\}].$$

The above reasoning implies that, for any $\delta_N \to 0$, $\sup_{||\theta-\theta_N||\leq\delta_N}|R_N(\theta)/[1+\sqrt{N}||\theta - \theta_N||]| \underset{p}{\to} 0$. Thus, assumption (v) of Newey and McFadden (1994, theorem 7.1) is satisfied. Q.E.D.

Proof of **Lemma 1**

All conditions of Newey and McFadden theorem 7.4 are satisfied and the result follows.

Appendix: Examples

Example 1:

Assumption: Let (i) $\theta(t|v,x) = \phi^x \lambda(t)$ so that $\bar{F}(t \mid x) = \exp(-\phi^x \Lambda(t))$ (ii) $\bar{F}(t \mid x)$ be observed for $x = 0, 1$ and all $t \geq 0$.

We first estimate the integrated baseline hazard, $\hat{\Lambda}(t) = -\ln \bar{F}(t \mid x = 0) = \bar{F}_0(t)$. Assumption (i) implies the following density, $f(t \mid x = 1) = \phi\lambda(t)e^{-\phi\Lambda(t)}$. Suppose that $\lambda(t)$ and $\Lambda(t)$ are known, then the log likelihood and its derivative have the following form,

$$\begin{aligned} L(\phi) &= \ln \phi^x + \ln \lambda(t) - \phi^x \Lambda(t) \\ \frac{\partial L(\phi)}{\partial \phi} &= \frac{x}{\phi} - x\Lambda(t) \Rightarrow \\ \operatorname*{plim}_{N\to\infty} \hat{\phi}_{MLE} &= 1/E\{\Lambda(T)|x=1\} = -1/E[\ln\{\bar{F}_0(T)\}|x=1]. \end{aligned}$$

Let $v \sim$ Gamma(α, α), so that $\bar{F}_0(t) = \left(1 + \frac{\Lambda(t)}{\alpha}\right)^{-\alpha}$ and $-\ln F_0(t) = \alpha \ln\left(1 + \frac{\Lambda(t)}{\alpha}\right)$. Note that $\phi^x \Lambda(T) = \frac{Z}{v}$ where Z has an exponential distribution with mean one. This yields

$$\operatorname*{plim}_{N\to\infty} \hat{\phi} = \frac{1}{E[\alpha \ln\{1 + Z/(\phi v \alpha\}]}.$$

where $v \sim$Gamma(α, α).

Example 2:

After transforming the dependent variable using the transformation model of Horowitz (1996) we define $W = H(T)$. Note that $H(T)^{|\beta|}$ is distributed as an exponential random variable so that W is distributed as a Weibull random variable with parameter $|\beta|$. As in the example, let $\beta > 0$. Consider the Weibull model with a Gamma mixing distribution,

$$\begin{aligned} \theta\left(w_i \mid v, x_i\right) &= v e^{x_i\beta} \alpha w_i^{\alpha-1} \\ v &\sim \text{Gamma}\left(\gamma_v, \delta_v\right) \\ \bar{F}(w_i|x_i) &= E v e^{-v e^{x_i\beta} t_i^\alpha} = \frac{1}{\left(1 + \frac{e^{x_i\beta} w_i^\alpha}{\delta_v}\right)^{\gamma_v}} \\ f(w_i|x_i) &= \frac{\alpha\gamma_v e^{x_i\beta} w_i^\alpha}{\delta_v} \frac{1}{\left(1 + \frac{e^{x_i\beta} w_i^\alpha}{\delta_v}\right)^{\gamma_v+1}} \\ L^i(\alpha, \beta, \gamma_v, \delta_v) &= \ln\alpha + \ln\gamma_v + x_i\beta + \alpha \ln W_i - \ln\delta_v - (\gamma_v + 1)\ln\left(1 + \frac{e^{x_i\beta} W_i^\alpha}{\delta_v}\right) \end{aligned}$$

Imposing the restriction $\alpha = \beta$ and using

$$e^{x_i\beta}W_i^\beta = (e^{x_i\beta}W_i^\beta)^{\beta/\beta} = (Z_i)^{\beta/\beta}$$

where Z_i is distributed as an exponential stochast with mean one gives

$$L^i(\beta, \gamma_v, \delta_v) = \ln\beta + \ln\gamma_v + (\beta/\beta)\ln Z_i - \ln\delta_v - (\gamma_v + 1)\ln\left(1 + \frac{(Z_i)^{\beta/\beta}}{\delta_v}\right).$$

This likelihood does not depend on the regressor[21] x, which implies that the probability limit of β does not depend on the distribution of x.

Appendix: Computational Issues

by Matthew Harding, Jerry Hausman, and Tiemen M. Woutersen

We estimate the parameter vector (β, δ) from the following objective function which corresponds to a mass of indicator functions:

$$Q(\beta, \delta) = \sum_{i=1}^{n}\sum_{l=1}^{L} 1\{T_i \geq l\} \sum_{j=1}^{n}\sum_{k=1}^{K}[1\{Z_i(l) < Z_j(k)\} - 1\{Z_i(l) > Z_j(k)\}]. \tag{12}$$

Optimization of this objective function using iterated sums is not feasible since for the specification with 24 periods it takes approximately 15 minutes to evaluate one such objective function in Matlab. Note however that for all individuals i which pass the criterion $T_i \geq l$ the objective function evaluates the difference between the number of individuals with an index less than the index of individual i and the number of individuals with an index greater than the index of individual i. This information is also contained in the ranking of individual's indices and thus can be more efficiently extracted using the Rank function. This suggest that an efficient implementation of this optimization will be similar to that of Chen (2002).

We can define $d_k = 1\{T \geq k\}$ for the vector T of dimension $N \times 1$. Let d be constructed by stacking the vectors d_k vertically for all $k = 1, ..., K$. Similarly let Z be constructed by stacking the vectors $Z(k)$ for all $k = 1, ..., K$. Now both d and Z are of dimension $NK \times 1$. We can now rewrite $Q(\beta, \delta)$ using these vectors and the Rank function:

[21] The same reasoning holds for a negative β_0 (since the sign can be determined using Han (1987) and for a multivariate regressor (since this can be reduced to a scalar by estimating the regression coefficient up to scale using Han (1987).

$$Q(\beta,\delta) = \frac{1}{N(N-1)} \sum_{i=1}^{Nk} d(i) \left[2 \cdot Rank(Z(i)) - NK\right]. \tag{13}$$

This simpler yet numerically identical representation[22] will be more efficient to evaluate numerically because (i) it has only one summation sign and (ii) computation of the rank function requires sorting for which highly efficient algorithms are available. Indeed it now takes less than one second to estimate one such objective function for the specification with 24 periods.

Models with non-smooth objective functions in the parameters have been traditionally estimated using the Nelder-Mead simplex method (see, e.g. Abrevaya, 1999; Cavanah and Sherman, 1998). In this particular example the large number of local optima makes the Nelder-Mead method computationally unstable. The Nelder-Mead algorithm fails to converge or takes unreasonably long to do so.[23]

Pattern search methods have been available for many decades and rigorous convergence results have become available in recent years (Lewis and Torczon, 1999; Audet and Dennis, 2003). Although anecdotal evidence on the performance of these algorithms often suggests slow convergence we find that the convergence of the objective function at 4 decimal places for the specification with 13 periods takes about 20 minutes while the specification with 24 periods takes approximately 50 minutes to convergence.

We shall now provide a brief introduction to the mechanism of pattern search.[24] For some given real valued objective function $Q(\gamma)$ defined on the n-dimensional Euclidean space, let γ_0 be the initial guess. In our case we use $\gamma_0 = [\ \widehat{\beta}, \widehat{\delta}]_{Gamma}$, the parameter estimates from the HHM Gamma Heterogeneity model estimated using a quasi-Newton derivative based method. Additionally define a *forcing function* $\rho(t)$ to be a continuous function such that $\rho(t)/t \to 0$ as t$\to$ 0. Let Δ_k control the step length at each iteration.

Search patterns for some initial starting value γ_0 are drawn from a given *generating set.* A minimal generating set corresponds to some positive spanning set for the n-dimensional space, where the number of dimensions corresponds to the number of parameters to be

[22] There is still an issue regarding the treatment of ties in the Rank function but it seems to matter little in practice.

[23] Convergence of the objective function to 4 decimal places may take as long as 9 hours to compute.

[24] For a more detailed review and convergence proofs see Kolda, Lewis and Torczon (2003).

estimated. The defining requirement for a generating set is that any vector in $\mathbb{R}^n$ may be written as a linear combination of elements in the generating set using positive coefficients only. A generating set will thus contain at least $n+1$ elements. To illustrate the generating set for $n = 2$ is

$$G = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}. \tag{14}$$

Alternatively we could use the set of $2n$ coordinate directions as the elements of our generating set. In our application however we have found computational performance to be superior under the setup with $n+1$ directions. Additionally, heuristic additions to the generating set may be implemented in order to improve speed and performance. These heuristic additions allow the algorithm to evaluate other points in the same direction as the last successful search, but further away from the starting point than the standard elements of the generating set would allow for, thus allowing for the possibility that if the correct direction of improvement was found, several computation steps will be skipped and the search converges more rapidly. Random polling vectors also provide heuristic evaluations of the objective function without compromising the convergence properties of the algorithm which only depend on the minimal generating set.

We use the standard errors of the HHM estimation to construct a "bounding box" that is then used to bound the parameter space for the optimization under the semi-parametric setup. We start with a bounding[25] box of $\pm$ 3 standard errors.

At each iteration the algorithm evaluates the objective function for all vectors $g_k \in G$ and compares $Q(\gamma_k + \Delta_k g_k)$ with $Q(\gamma_k) - \rho(\Delta_k)$. If an improvement is found $\gamma_{k+1} = \gamma_k + \Delta_k g_k$ and Δ_k is increased to Δ_{k+1}. If no improvement is found then $\gamma_{k+1} = \gamma_k$ and Δ_k is decreased to Δ_{k+1}. This process is iterated to convergence.

Since the true parameter values are not guaranteed to lie within this bounding box it may be that the algorithm constrained by the location and size of the bounding box only reaches a local optima. In order to correct for this possibility we gradually expand the bounding box as long as the estimated parameters change with a larger bounding box. A large bounding box however may imply that the estimates have only low precision, since

[25] We would increase the number of standard errors if the sample size would be larger.

the algorithm visits every point in the domain with a probability decreasing in the size of the bounding box. In order to improve accuracy, once the desired size of the bounding box has been reached, the bounding box is re-centered on the new parameter estimates from the semi-parametric setup. The size of the bounding box is then sequentially decreased in order to verify the accuracy of the obtained estimates. Refinements are made if an improvement is found.

We use the estimated values $\widehat{\delta}_{Pattern}$ to compute an estimate of the survival probability at each time period. Using the delta method we compute the associated estimates of the standard error of the survival probability in each period. Interpretation is made easier by smoothing the pair $(P(T \geq t_i), t_i)$ for all time periods t_i using a local polynomial method. The neighborhood of t_i is defined as a percentage of the total number of periods under consideration and may be chosen using cross-validation techniques. Each point in the neighborhood $N(t_i)$ is assigned two sets of weights. One set of weights is inversely proportional to the standard error of the survivor estimate as given by the pattern search optimization. The other set of weights is provided by the *tri-cubic weight function* and weighs the impact of distant data points on the smoothing estimate of one particular observation. The tri-cubic weight function involved in the smoothing of point t_i places the following weight on observation t_j:

$$W(t_i, t_j) = \left(1 - \left(\frac{|t_i - t_j|}{\max_{t_j \in N(t_i)} |t_i - t_j|}\right)^3\right)^3 \mathbf{1}\left\{0 \leq \frac{|t_i - t_j|}{\max_{t_j \in N(t_i)} |t_i - t_j|} < 1\right\}. \quad (15)$$

The smoothed estimates of the survivor function are then computed as the predicted values of the weighted linear regression of second degree for each point in the corresponding neighborhood using the two sets of weights. The choice of the span of the neighborhood at each point using cross-validation tends to matter little in this case.

The pattern search method we employed to derive estimates of the model parameters seems to perform well, both in terms of accuracy and computational time. Nevertheless, the nature of the objective function and the dependency of our use of the pattern search method on a good estimate of the relevant bounding box, raises the question to what extent a global optimum has been reached for our objective function. Since it is possible to conceive of our optimization problem as a stochastic optimization problem we consider

the implementation of a *genetic optimization* procedure as a global optimizer capable of overcoming the nondifferentiability of the objective function, as discussed by Spall (2003). Few applications of this procedure to econometrics exist in spite of numerous reported successful implementations in other areas of science (Haupt and Haupt, 1998; Reeves and Rowe, 2003).

Genetic optimization methods describe a number of processes based on principles from biological sciences aimed at generating a population of parameter values which optimizes its *fitness* defined as the corresponding value of the objective function. The core idea involves the use of stochastic perturbations in the population of potential optimizing parameters so as to improve the optimality of the solution. This approach mirrors the biological concept of evolution. The use of a population of parameters as the primary building block of the algorithm aims at avoiding convergence towards one local optimum.

Since the outcome of a genetic optimization procedure is not dependent on the initial population we use as starting values for the population unit-uniform random numbers. The objective function is evaluated for each member of the population. Members of the population with the best values are selected as candidates for the generation of individuals of the subsequent population through the processes of elitism, crossover or mutation. A (small) number of the successful members of a population are simply copied over in the next generation of the population, a process termed elitism. The crossover process randomly combines values of the parameter vector of two evolutionary successful individuals to obtain a new individual for the next population. The process of mutation adds random noise from a Normal distribution to the parameter values of one successful individual to create a new individual in the next generation. Since with each additional generation we are more likely to close-in on the optimum, we shrink the variance of the mutation process at each generation.

Convergence for the genetic optimization process tends to be much slower than that of the pattern search procedure. Nevertheless, the algorithm can be used to confirm the global optimality of the point estimates obtained by pattern search. Our results using genetic optimization are the same as the pattern search algorithm to 4 significant digits

for the objective function.

References

[1] Abrevaya, J. (1999): Computation of the maximum rank correlation estimator, *Economics Letters* 62, 279–285

[2] Audet, C. and J. E. Dennis, Jr. (2003): "Pattern Search Algorithms for Mixed Variable Programming," *SIAM Journal on Optimization* 11: 573-594.

[3] Baker, M. and A. Melino (2000): "Duration Dependence and Nonparametric Heterogeneity: A Monte Carlo Study," *Journal of Econometrics*, 96, 357-93.

[4] Bijwaard, G. and G. Ridder (2002): "Efficient Estimation of the Semi-parametric Mixed Proportional Hazard Model", in preparation.

[5] Cavanagh, C., R. P. Sherman (1998): "Rank Estimators for monotonic index models", *Journal of Econometrics*, 84, 351-381

[6] Cox, D. R. (1972): "Regression models and life tables (with discussion)", *Journal of the Royal Statistical Society* B, 34: 187-220.

[7] Chen, S. (2002): "Rank Estimation of Transformation Models", *Econometrica*, 70, 1683-96.

[8] Elbers, C. and G. Ridder (1982): "True and Spurious Duration Dependence: The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, 49, 402-409.

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

[10] 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, 175-205.

[11] Han, A. K. (1987): "Non-parametric Analysis of a Generalized Regression Model, the Maximum Rank Correlation Estimator", *Journal of Econometrics, 35, 303-316.*

[12] Han, A. K. and J. A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics.*

[13] Haupt, R.L. and S.E. Haupt (1998) *Practical Genetic Algorithms,* Wiley-Interscience.

[14] Hausman, J. A. (1978): "Specification Tests in Econometrics", *Econometrica*, 46, 1251-72.

[15] Hausman, J. A., and D. A. Wise (1979): "Attrition Bias in Experimental and Panel Data: The Gary Income Maintenance Experiment", *Econometrica,* 47, 455-474.

[16] Hausman, J. A., and W. E. Taylor (1981): "Panel Data and Unobservable Individual Effects", *Econometrica,* 49, 1377-1398.

[17] Heckman, J. J. (1978): "Simple statistical models for discrete panel data developed and applied to test the hypothesis of true state dependence against the hypothesis of spurious state dependence", annales de l'insee, 30-31, page 227-269.

[18] Heckman, J. J. (1991): "Identifying the Hand of the Past: Distinguishing State Dependence from Heterogeneity," *American Economic Review*, 81, 75-79.

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

[20] Heckman, J. J., R. J. LaLonde, and J. Smith (1999): "The Economics and Econometrics of Active Labor Market Programmes" in the *Handbook of Labor Economics, Volume 3A.*

[21] Honoré, B. E. (1990): "Simple Estimation of a Duration Model with Unobserved Heterogeneity," *Econometrica*, 58, 453-473.

[22] Honoré, B. E. (1993a): "Identification Results for Duration Models with Multiple Spells or Time-Varying Regressors," Northwestern working paper.

[23] Honoré, B. E. (1993b): "Identification Results for Duration Models with Multiple Spells," *Review of Economic Studies*, 60, 241-246.

[24] Horowitz, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica*, 64, 103-107.

[25] Horowitz, J. L. (1999): "Semiparametric Estimation of a Proportional Hazard Model with Unobserved Heterogeneity" *Econometrica*, 67, 1001-1028.

[26] Horowitz, J. L. (2001): "The Bootstrap" in *Handbook of Econometrics,* Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.

[27] Ishwaran, H. (1996a): "Identifiability and Rates of Estimation for Scale Parameters in Location Mixture Models," *The Annals of Statistics*, 24, 1560-1571.

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

[29] Kendall, M. G. (1938): "A new measure for rank correlation", *Biometrica*, 30, 81-93.

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

[31] Kolda, T. G., Lewis, R. M., and Torczon, V. (2003): "Optimization by direct search: New perspectives on some classical and modern methods", *SIAM Review*, 45: 383–482.

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

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

[34] Lancaster, T. and S. J. Nickell, (1980): "The Analysis of Re-employment Probabilities for the Unemployed", *Journal of the Royal Statistical Society*, A, 143, 141-165.

[35] Lewis, R. M. and V. Torczon (1999): "Pattern search algorithms for bound constrained minimization", *SIAM Journal on Optimization*, 9: 1082–1099.

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

[37] Mundlak, Y. (1961): "Empirical Production Function Free of Management Bias," *Journal of Farm Economics,* 43, 44-56.

[38] Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics,* Vol. 4, ed. by R. F. Engle and D. MacFadden. Amsterdam: North-Holland.

[39] Newey, W. K. (1991): "Uniform Convergence in Probability and Stochastic Equicontinuity", *Econometrica*, 59, 1161-1167.

[40] Nielsen, G.G., Gill, R.D., Andersen, P.K. & Sørensen, T.I.A. (1992): "A counting-process approach to maximum likelihood estimation in frailty models", *Scandanavian Journal of Statistics*. 19, 25-43

[41] Reeves, C.R. and J.E. Rowe (2003) *Genetic Algorithms - Principles and Perspectives: A Guide to GA Theory*, Kluwer Academic.

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

[43] Ridder, G. and T. M. Woutersen (2003): "The Singularity of the Information Matrix of the Mixed Proportional Hazard Model" *Econometrica*, 71, 1579-1589.

[44] Sherman, R. P. (1993): "The Limiting Distribution of the Maximum Rank Correlation Estimator", *Econometrica*, 61, 123-137.

[45] Spall, J. (2003) *Introduction to Stochastic Search and Optimization: Estimation, Simulation and Control*, Wiley-Interscience.

[46] Van den Berg, G. J. (2001): "Duration Models: Specification, Identification, and Multiple Duration," in *Handbook of Econometrics,* Vol. 5, ed. by J. J. Heckman and E. Leamer. Amsterdam: North-Holland.

Table 1: Data Description and Summary Statistics

Variable	*Description*	*Mean*	*Standard Deviation*
Race	= 1 if UI recipient is Black or African-American	0.1172	0.3217
Age	= 1 if UI recipient is over 50 years old at the start of the benefit year	0.1776	0.3822
Replacement Rate	= Weakly Benefit Amount divided by UI recipient's base period earnings	0.0129	0.0076
State Unemployment Rate	= Unemployment rate of the state from which the individual received UI benefits during the period in which the individual filed for benefits		
	Week 1	4.6863	1.0875
	Week 2	4.6726	1.0834
	Week 3	4.6603	1.0794
	Week 4	4.6453	1.0747
	Week 5	4.6301	1.0698
	Week 6	4.6211	1.0649
	Week 7	4.6164	1.0665
	Week 8	4.5981	1.0641
	Week 9	4.5710	1.0616
	Week 10	4.5382	1.0615
	Week 11	4.5318	1.0630
	Week 12	4.5091	1.0678
	Week 13	4.4832	1.0751
	Week 14	4.4620	1.0802
	Week 15	4.4604	1.0756
	Week 16	4.4490	1.0735
	Week 17	4.4400	1.0675
	Week 18	4.4407	1.0557
	Week 19	4.4316	1.0546
	Week 20	4.4207	1.0452
	Week 21	4.4240	1.0337
	Week 22	4.4315	1.0298
	Week 23	4.4364	1.0240
	Week 24	4.4414	1.0156
	Week 25	4.4424	1.0121

Table 2: HHM Gamma Heterogeneity Model, Period 1 normalized to zero.

		6 periods		13 periods		24 periods	
		Parameters	**s.e.**	**Parameters**	**s.e.**	**Parameters**	**s.e.**
alpha		0.9307	2.1675	0.1089	0.0120	0.0993	0.0182
gamma		7.9607	0.2383	0.3164	0.0773	0.1655	0.6082
State Unemployment Rate		-0.1019	0.0246	-0.2762	0.0341	-0.3875	0.0393
Race		-0.0350	0.0653	-0.2167	0.1155	-0.2061	0.1370
Age>50		-0.2047	0.0623	-0.4290	0.0932	-0.4317	0.1557
Replacement Rate		-0.5393	0.0497	-0.5498	0.0562	-0.5059	0.1493
Period	2	-0.3259	0.0747	-0.0494	0.0787	0.0010	0.1576
	3	0.0198	0.0814	0.5517	0.0905	0.6479	0.1342
	4	-0.3032	0.0939	0.4661	0.1157	0.6053	0.1222
	5	0.1430	0.1026	1.1678	0.1275	1.3511	0.1532
	6	-0.3780	0.1256	0.8858	0.1553	1.1134	0.1979
	7			1.4905	0.1811	1.7608	0.1879
	8			1.3001	0.2086	1.6111	0.2144
	9			1.7490	0.2228	2.0944	0.2359
	10			1.7326	0.2486	2.1103	0.2753
	11			2.2152	0.2661	2.6362	0.3007
	12			2.3336	0.2870	2.7970	0.3510
	13			2.6485	0.3108	3.1545	0.3966
	14					3.4413	0.3856
	15					3.8034	0.4204
	16					3.7589	0.5024
	17					4.3672	0.5399
	18					4.4417	0.5073
	19					4.9485	0.5167
	20					4.9909	0.5785
	21					5.3740	0.5845
	22					5.4392	0.6022
	23					5.9363	0.6546
Period	24					6.0436	0.6891
Number of observations		**15,491**		**15,491**		**15,491**	
Likelihood		**0.6664**		**1.2242**		**1.0131**	

Table 3: New Duration Model, Period 1 normalized to zero.

		6 periods		13 periods		24 periods	
		Parameters	**s.e.**	**Parameters**	**s.e.**	**Parameters**	**s.e.**
State Unemployment Rate		-1.4672	0.0965	-1.4643	0.0832	-1.3953	0.0483
Race		-0.5663	0.2728	-0.5928	0.2444	-0.5656	0.2105
Age>50		-1.0701	0.2146	-1.0712	0.1974	-0.8067	0.1770
Replacement Rate		-2.2347	0.1778	-2.2693	0.1588	-0.5372	0.1097
Period	2	2.7287	0.1295	2.6191	0.1604	2.0707	0.2422
	3	3.8869	0.1298	4.1002	0.1812	3.2261	0.2451
	4	5.0912	0.1276	5.4381	0.1657	4.2821	0.2116
	5	5.6051	0.1440	5.9834	0.1737	4.7376	0.2132
	6	6.5985	0.1380	7.1400	0.1704	5.7784	0.2028
	7			7.1200	0.2092	5.6905	0.2444
	8			7.9306	0.1860	6.5007	0.1955
	9			8.2543	0.2017	6.7297	0.2212
	10			8.3960	0.2382	6.5937	0.3050
	11			8.7536	0.2265	7.1753	0.2422
	12			9.4656	0.2218	7.8302	0.2218
	13			9.7804	0.2361	8.3342	0.2227
	14					8.1757	0.3352
	15					8.4889	0.3058
	16					9.1671	0.2548
	17					9.5479	0.2597
	18					9.8108	0.2818
	19					10.0790	0.2968
	20					10.6790	0.3018
	21					10.7060	0.3229
	22					10.9360	0.3409
	23					10.9230	0.3419
Period	24					11.3860	0.3437
Number of observations		**15,491**		**15,491**		**15,491**	
Objective Function		**30.221**		**122.050**		**332.890**	

Figure 1: Design with 13 Periods

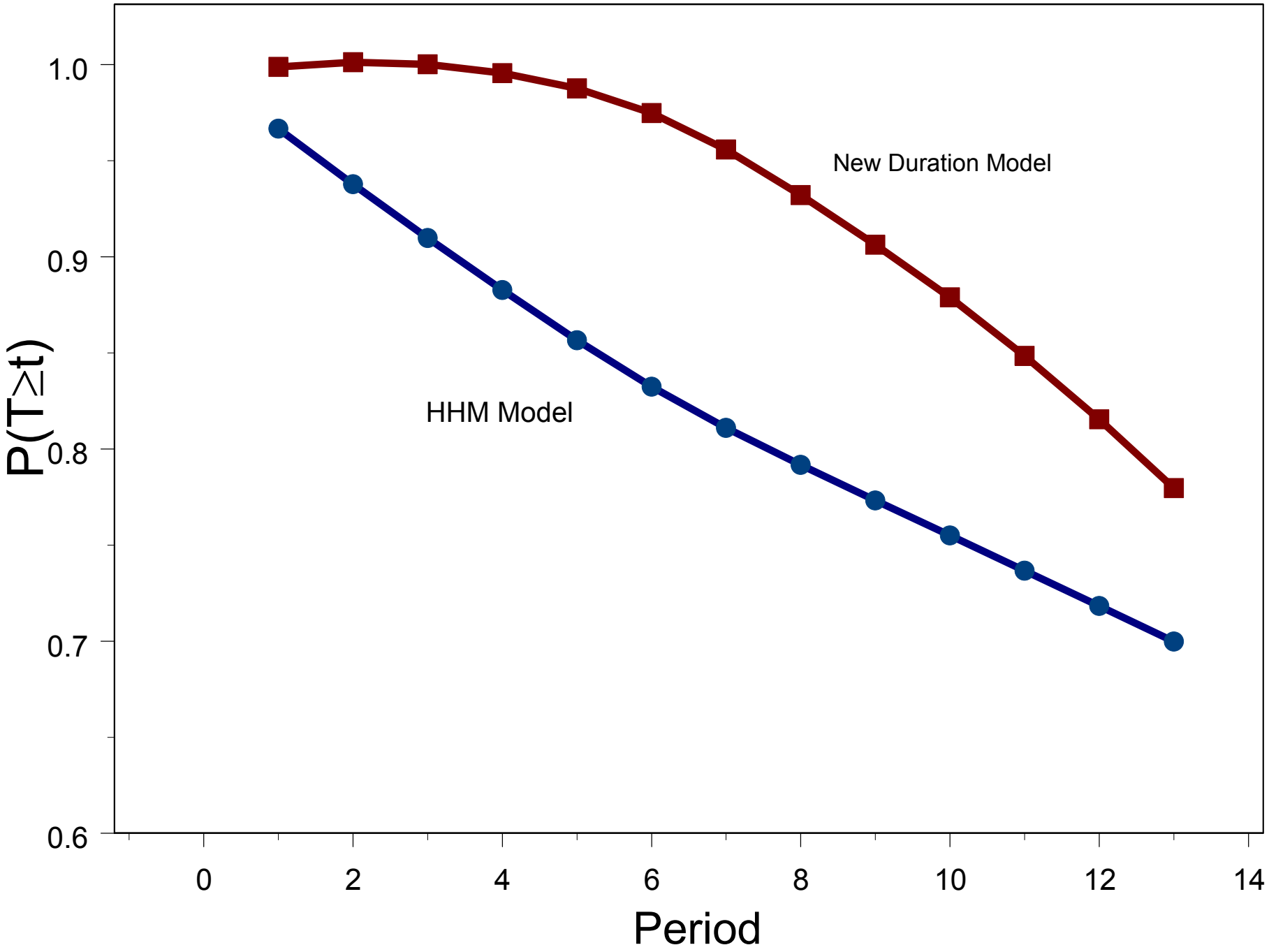

Figure 2: Design with 24 Periods