A semi-parametric duration model with heterogeneity and time-varying regressors

Jerry Hausman and Tiemen Woutersen[*]
MIT and Johns Hopkins University

Draft, November 2004

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.

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

Preliminary and Incomplete

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 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

[*]Comments are welcome, jhausman@mit.edu and twouters@mit.edu.

[†]We thank Su-Hsin Chang, Matthew Harding, and Marcel Voia for research assistance. We have received helpful comments from the Harvard-MIT and UCLA econometrics seminars and Moshe Buchinsky.

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 distribution was suggested by Honoré (1990). This estimator assumes a Weibull baseline hazard and only uses very short durations to estimate the Weibull parameter. In the limit, this estimator only uses arbitrarily short durations. Since only a small fraction of

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

the durations is arbitrarily short, this estimator converges at a slow rate. Van den Berg (2001) discusses other disadvantages of focussing only on arbitrarily durations.

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 non-parametrically. 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, 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 difficulties with the 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 very slowly. Thus, the regression coefficient estimates, which are often of primary interest, are often not estimated very precisely.

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 non-parametric 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 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 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 (2000) find that the 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. 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 extremely large.

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 shows that misspecifying the heterogeneity yields inconsistent estimates, even if the baseline hazard is nonparametric. Section 5 presents and empirical example and section 6 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 $\lambda(t)$,

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

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 $\theta(t \mid X, v)$ 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 (2002) 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 (1981) and Heckman and Singer (1984a), 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,

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

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 somewhat abuse notation and also use $x(t)$ to denote the sequence of the regressor from $s = 0$ to $s = t$. The mixed proportional hazard model of equation (2) implies the following survival probabilities,

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

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 for now 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 (2) as follows,

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

where t is a natural number, $\delta_s = \ln\{\int_{s-1}^{s}\lambda(s)ds\}$ and we normalize $\delta_1 = 0$. This specification is similar to Han-Hausman (1990) who specify δ_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 expectation of T depends only 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.[2].

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

$$Q(\beta,\delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)\}. \tag{3}$$

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 the Han objective function $Q(\beta)$. 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 $Z_i(l;\beta,\delta) < Z_j(k;\beta,\delta)$. Vice versa if $Z_i(l;\beta,\delta) > Z_j(k;\beta,\delta)$. 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 and δ, without the restriction of only up to scale as in the Han approach. In particular,

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

We now use a two period outcome to illustrate the necessary conditions for identification.

Note that our approach focuses on the probability than an individual i survives period

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

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 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). 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,

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

The computational burden to calculate[3] $Q(\beta, \delta)$ is proportional to $N \ln(N)$.

Identification, two period model. In order to develop intuition, we first consider a two period model in which we estimate the scale parameter. Suppose we observe a group of N unemployed individuals at the beginning of their unemployment, at the end of the first period and at the end of the second period. In particular, let the data generating process be given by equation (2), let x be exogenous and let v be independent of x. Let $d_{it} = 1\{T_i \geq t\}$. Let x be a scalar and suppose we observe for each i, $\{d_{i1}, d_{i2}, x_{i1}, x_{i2}\}$. Thus, x_{i1}, x_{i2} are the regressors in the first and second period. Note that

$$Ed_{it} = E_v e^{-v \sum_{s=1}^{t} e^{x_{is}\beta + \delta_s}} \text{ where } \delta_1 \text{ is normalized to be zero.}$$

Also note that

$$Ed_{i2} \geq Ed_{j1}$$

[3] 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 $N \ln(N)$.

is equivalent to

$$e^{x_{i1}\beta} + e^{x_{i2}\beta+\delta_2} \leq e^{x_{j1}\beta}$$

where β and δ_2 denote the true parameter values. Let A denote the set of all pairs i, j for which the last equation holds. That is,

$$A = \{i, j : e^{x_{i1}\beta} + e^{x_{i2}\beta+\delta_2} \leq e^{x_{j1}\beta}\}.$$

Let A' denote the following set defined for the parameters (β', δ_2')

$$A' = \{i, j : e^{x_{i1}\beta'} + e^{x_{i2}\beta'+\delta_2'} \leq e^{x_{j1}\beta'}\}.$$

We now give conditions under which $A = A'$ implies $\{\beta' = \beta, \delta_2' = \delta_2\}$.

$$\begin{aligned} A_1 &= \{i, j : x_{i1} = x_{i2},\ e^{x_{i1}\beta}(1 + e^{\delta_2}) \leq e^{x_{j1}\beta}\} \\ &= \{i, j : x_{i1} = x_{i2},\ e^{x_{i1}\beta+c\beta} \leq e^{x_{j1}\beta}\} \end{aligned}$$

where $e^{c\beta} = (1 + e^{\delta_2})$. Thus,

$$A_1 = \{i, j : x_{i1} = x_{i2},\ c \leq x_{j1} - x_{i1}\}.$$

Similarly,

$$\begin{aligned} A_1' &= \{i, j : x_{i1} = x_{i2},\ e^{x_{i1}\beta}(1 + e^{\delta_2}) \leq e^{x_{j1}\beta}\} \\ &= \{i, j : x_{i1} = x_{i2},\ c' \leq x_{j1} - x_{i1}\}. \end{aligned}$$

For these two sets to coincide, $A_1 = A_1'$, for any distribution of the regressors, we need $c' = c$. Thus, c is identified if the density of the regressor is positive in an arbitrarily small neighborhood around x_{i1} and x_{i2} because otherwise if we do not have $c' = c$, the inequality can be reversed for a small change in say x_{j1}. Moreover, for $\tilde{x}_{i1} = \tilde{x}_{i2}$ we have $Ed_{i2} = Ed_{j1}$ if and only if $\tilde{x}_{j1} - \tilde{x}_{i1} = c$. While this explanation is only local, it leads to a proof of global identification because of the global convexity property of $\exp(x_{ik}\beta + \delta_k')$ as we now demonstrate.

Without loss of generality, let $\beta > 0$ (if $\beta < 0$, multiply x by -1). Define

$$H(\beta) = e^{x_{i1}\beta} + e^{x_{i2}\beta+\delta_2} - e^{x_{j1}\beta}. \tag{4}$$

For the individuals on the boundary of the set A, we have $H(\beta) = 0$. To make this boundary relevant, we assume that the density of the regressor is positive in a neighborhood of some $\{x_{i1}, x_{i2}, x_{j1}\}$ with $H(\beta) = 0$. We now show that if $H(\beta_0) = 0$ then $H(\beta) = 0$ is uniquely solved for $\beta = \beta_0$. Using $e^{\delta_2} = e^{c\beta} - 1 = e^{(\tilde{x}_{j1} - \tilde{x}_{i1})\beta} - 1$ yields

$$\begin{aligned} H(\beta) &= e^{x_{i1}\beta} + e^{x_{i2}\beta + (\tilde{x}_{j1} - \tilde{x}_{i1})\beta} - e^{x_{i2}\beta} - e^{x_{j1}\beta} \\ &= e^{x_{i1}\beta}\{1 + e^{(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})\beta} - e^{(x_{i2} - x_{i1})\beta} - e^{(x_{j1} - x_{i1})\beta}\}. \end{aligned}$$

Define

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

Consider the derivative of $H^*(\beta)$ with respect to β,

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

Suppose that, for some individual on the boundary of set A, we have $x_{i2} > x_{i1}$. Then $(x_{j1} - x_{i1}) > (\tilde{x}_{j1} - \tilde{x}_{i1})$. Thus,

$$\begin{aligned} \lim_{\beta \downarrow 0} \frac{\partial H^*(\beta)}{\partial \beta} &= (x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1}) - (x_{i2} - x_{i1}) - (x_{j1} - x_{i1}) \\ &= (\tilde{x}_{j1} - \tilde{x}_{i1}) - (x_{j1} - x_{i1}) < 0. \end{aligned}$$

Also note that $(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1}) > (x_{i2} - x_{i1})$ since $\tilde{x}_{j1} > \tilde{x}_{i1}$. Moreover, we have $\tilde{x}_{j1} - \tilde{x}_{i1} = c$ and $x_{j1} - x_{i2} < \tilde{x}_{j1} - \tilde{x}_{i1}$ since $x_{i2} > x_{i1}$. Therefore, $(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1}) > (x_{j1} - x_{i1})$, so that $\frac{\partial H^*(\beta)}{\partial \beta} > 0$ for large β. Consider the second derivative,

$$\frac{\partial^2 H^*(\beta)}{(\partial \beta)^2} = (x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})^2 e^{(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})\beta} - (x_{i2} - x_{i1})^2 e^{(x_{i2} - x_{i1})\beta} - (x_{j1} - x_{i1})^2 e^{(x_{j1} - x_{i1})\beta}$$

We continue to assume that $x_{i2} > x_{i1}$ for some individual. Note that this implies that

(1). $e^{(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})\beta} > e^{(x_{i2} - x_{i1})\beta}$ and $e^{(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})\beta} > e^{(x_{j1} - x_{i1})\beta}$.

(2). $(x_{i2} - x_{i1} + \tilde{x}_{j1} - \tilde{x}_{i1})^2 - (x_{i2} - x_{i1})^2 - (\tilde{x}_{j1} - \tilde{x}_{i1})^2 = 2(x_{i2} - x_{i1})(\tilde{x}_{j1} - \tilde{x}_{i1}) > 0$.

(1)-(2) imply that $\frac{\partial^2 H^*(\beta)}{(\partial \beta)^2} > 0$.

Thus, $H^*(\beta)$ first decreases and then increases in β so that $H^*(\beta) = 0$ is uniquely solved for $\beta = \beta_0$. Similar reasoning applies if $x_{i2} < x_{i1}$ so that the substantive condition is that $P(x_{i1} \neq x_{i2}) > 0$ plus the continuity assumption of the regressor around some points. Identification of $\{\beta, \delta\}$ is equivalent to identification of $\{\beta, c\}$. 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 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[4].

3. Large Sample Properties

In this section, we formalize the example of the last section and derive large sample properties of our estimator. For the two period model, we assume the following.

Assumption 1 (MPH): *Let (i)* $\{T, v, x\}$ *be a random sample where* $x = \{x_1, x_2\}$ *and* x_1, x_2 *are scalars, (ii)* v *and* x *are independent,* x *is exogenous (iii)* $Pr(T \geq 1|x_1, x_2) = E_v e^{-ve^{x_1\beta}}$ *and* $Pr(T \geq 2|x_1, x_2) = E_v e^{-v\{e^{x_1\beta}+e^{x_2\beta+\delta_2}\}}$*, (iv) let* $\{\beta, \delta_2\} \in \Theta$*, which is compact, (v)* $\exists$ *a pair* $\{x_1, x_1', x_2'\}$*,* $x_1' \neq x_2'$*, such that* $Pr(T \geq 1|x_1, x_2) = Pr(T \geq 2|x_1', x_2')$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x_1 *or* $\{x_1', x_2'\}$*, (vi)* $\exists$ *a pair* $\{x_1, x_1', x_2'\}$*,* $x_1' = x_2'$*, such that* $Pr(T \geq 1|x_1, x_2) = Pr(T \geq 2|x_1', x_2')$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x_1 *or* x_1'*.*

Assumption 1 (iii) holds if the data generating process is given by equation (2).

Theorem 1:

Let assumption 1 hold. Let $\{\hat{\beta}, \hat{\delta}_2\} = \underset{\beta,\delta_2}{\operatorname{argmin}} Q(\beta, \delta_2)$ where

$$Q(\beta, \delta) = \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{2} \sum_{l=1}^{2} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l) < Z_j(k)\}.$$

[4] 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.

Then

$$\{\hat{\beta}, \hat{\delta}_2\} \underset{p}{\rightarrow} \{\beta, \delta_2\}.$$

Proof: 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 maximized at the true value of the parameters. The function $Q(\beta, \delta_2)$ is stochastically equicontinuous and the conditions of Newey and McFadden (1994, lemma 2.8) are satisfied so that $Q(\beta, \delta_2)$ converges uniformly to $EQ(\beta, \delta_2)$. Moreover, Θ is assumed to be compact and the data are i.i.d., so that consistency follow from Newey and McFadden (1994, theorem 1). Note that these arguments do not require that there is unoberved heterogeneity; they still hold if all individuals have the same value of v.

Suppose 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. In that case we need to strengthen assumption 1 in order to estimate $\{\delta_3, ..., \delta_K\}$. Let x^l denote $\{x_1, x_2, ..., x_l\}$, which are all scalars.

ASSUMPTION 1' (MPH): *Let (i)* $\{T, v, x\}$ *be a random sample* $x = \{x_1, ..., x^K\}$, $x_1, ..., x_K$ *are scalars, (ii)* v *and* x *are independent,* x *is exogenous (iii)* $Pr(T \geq l|x) = E_v e^{-v \sum_{s=1}^{l} e^{x_s \beta + \delta_s}}$ for $l = 1, ..., K$, *(iv)* δ_1 *is normalized to be zero, let* $\{\beta, \delta_2\} \in \Theta$, *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* $\{x^l, x^k\}$ *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) $\exists$ *a pair of regressors,* $\{x^r, x^s, \}$, $r \neq s$, *such that* $Pr(T \geq r|x^r) = Pr(T \geq s|x^s)$ *where the density of the regressor is positive in an arbitrarily small neighborhood around* x^r *or* x^s [moreover, let $x_1 = x_2 = ...x_r$ or $x_1 = x_2 = ... = x_s$].

Condition (i)-(v) ensure identification up to scale. Condition (i)-(vi) ensures complete identification[5]. We hope to remove the condition between square brackets, [..], in a future version. Similar to theorem 1, $\{\hat{\beta}, \hat{\delta}_2, ..., \hat{\delta}_K\}$ converges to $\{\beta, \delta_2, ..., \delta_K\}$ under assumption 1'.

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 estimate 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. In order to estimate β up to scale, we assume the following.

ASSUMPTION 2 (MRC): Let (i) β *be contained in a compact subset* $\widetilde{B}$ *of* $\mathcal{R}^q$, (ii) $Pr(T \geq l|x) = E_v e^{-v \sum_{s=1}^{l} e^{x_s \beta + \delta_s}}$ (iii) $\{T, v, x\}$ be a random sample (iv) *the support of the distribution of* x, S_x, *is not contained in any proper linear subspace of* $\Re^q$, (v) $\beta_1 \neq 0$, *and for almost every* $\tilde{x}_{i1} \equiv (x_{i1,2}, ..., x_{i1,q})'$, *the distribution of* $x_{i1,1}$ *conditional on* $\tilde{x}_{i1}$ *has everywhere positive density with respect to the Lesbesgue measure (vi)* v and x are independent.

Assumption 2 is sufficient to estimate β up to scale. In particular, under assumption 2, β can be estimated up to scale using

$$Q'(\beta, \delta) = \frac{1}{N(N-1)} \sum_i \sum_j [1\{T_i \geq 1\} - 1\{T_j \geq 1\}] 1\{Z_i(1) < Z_j(1)\}. \tag{5}$$

Assumption 2 can be replaced by any assumption that ensures that β is consistently estimated using the following objective function,

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

In particular, $Q''(\beta, \delta)$ uses K restrictions and the resulting estimator for β is no longer an MRC estimator and uses more data then just applying the MRC estimator to the first

[5] 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 density on the whole real line.

period. Combining the last two assumption ensures consistency of $\{\hat{\beta}, \hat{\delta}_2, ..., \hat{\delta}_K\}$.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. 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 $\theta = \{\beta, \delta\}$,

$$\begin{aligned}
Q_N(\theta) &= \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} \sum_{k=1}^{K} [1\{T_i \geq l\} - 1\{T_j \geq k\}] 1\{Z_i(l) < Z_j(k)\} \\
&= \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{L} [1\{T_i \geq l\} \sum_{l=1}^{K} 1\{Z_i(l) < Z_j(k)\} \\
&- \frac{1}{N(N-1)} \sum_i \sum_j \sum_{l=1}^{K} 1\{T_j \geq k\} \sum_{l=1}^{L} 1\{Z_i(l) < Z_j(k)\} \\
&= \frac{\sum_i}{N} \sum_{l=1}^{L} 1\{T_i \geq l\} \frac{\sum_j}{N-1} \sum_{k=1}^{K} [1\{Z_i(l) < Z_j(k)\} - 1\{Z_i(l) > Z_j(k)\}] \\
&= \frac{\sum_i}{N} \sum_{l=1}^{L} 1\{T_i \geq l\} \frac{\sum_j}{N-1} \sum_{k=1}^{K} [2 * 1\{Z_i(l) < Z_j(k)\} - 1] \\
&= \frac{\sum_i}{N} \sum_{l=1}^{L} 1\{T_i \geq l\} K[1 - 2\hat{F}_Z\{Z_i(l)\}]. \qquad (7)
\end{aligned}$$

where $\hat{F}_Z\{Z_i(l)\} = \frac{\sum_j}{N-1} \frac{\sum_{k=1}^{K}}{K} 1\{Z_i(l) < Z_j(k)\}$. Note that $E[\hat{F}_Z\{Z_i(l)\}|Z_i(l)] = F_Z\{Z_i(l)\}$ where F_Z is the cdf of $Z_i(l)$ for $l = 1, ..., K$ and $i = 1, ..., N$. Define H to be the second derivative of $Q_0(\beta, \delta) = E\{Q_N(\beta, \delta)\}$, evaluated at θ_0, i.e.

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

We assume the following.

ASSUMPTION 3 (INTERIOR): *Let $\theta_0 = (\beta_0, \delta_0) \in$ Interior(Θ), where Θ is compact.*

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

ASSUMPTION 4: *Let (i) $Q_0(\theta)$ be twice continuously differentiable at θ_0 with nonsingular derivative H; (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 3 is a standard regularity condition and supports an argument based on a Taylor expansion[6].

Theorem 3 (Asymptotic Normality)

Let assumption 1', 2-4 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

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

Proof:

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

The assumption $|f_Z(z)\frac{\partial Z}{\partial \theta}| < M$ and the random sample assumption of assumption 1 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

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

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

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 $u \neq 0$.

[6] 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.

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

$$\hat{F}(Z_i(l)) = \sum_i \sum_{k=1}^{K} G\{\frac{Z_i(l) - Z_j(k)}{h}\}. \tag{8}$$

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

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

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

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

$$q_N(\theta) - q_N(\theta_0) = 2\frac{\sum_i}{N} \sum_{l=1}^{L} 1\{T_i \geq l\} \sum_{k=1}^{K} \{\tilde{F}(Z_{0,i}(l)) - \tilde{F}(Z_i(l))\}. \tag{9}$$

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,

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

Let $q_0(\theta) = E\{q_N(\theta)\}$, 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 $h \to 0$,

$$\begin{aligned}
r_N(\theta) &= \{\frac{\partial q_N(\theta)}{\partial \theta}|_{\theta=\bar{\theta}} - \frac{\partial q_0(\theta)}{\partial \theta}|_{\theta=\bar{\theta}}\}(\theta-\theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L} 1\{T_i \geq l\}\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta-\theta_0) \\
&- 2E[\frac{\sum_i}{N}\sum_{l=1}^{L} 1\{T_i \geq l\}\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta-\theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L} [1\{T_i \geq l\} - E(1\{T_i \geq l\}|X)]\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta-\theta_0) \\
&- 2\frac{\sum_i}{N}\sum_{l=1}^{L} [E(1\{T_i \geq l\}|X)]\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\} \\
&- E[E(1\{T_i \geq l\}|X)\{\frac{1}{h}\frac{\exp(Z_i(l)/h)}{\{1+\exp(Z_i(l)/h)\}^2}\frac{\partial Z_i(\theta)}{\partial \theta}\}]]|_{\theta=\bar{\theta}}(\theta-\theta_0) \\
&= 2\frac{\sum_i}{N}\sum_{l=1}^{L} [1\{T_i \geq l\} - E(1\{T_i \geq l\}|X)]\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\}|_{\theta=\bar{\theta}}(\theta-\theta_0) \\
&- 2\frac{\sum_i}{N}\sum_{l=1}^{L} [E(1\{T_i \geq l\}|X)\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\} \\
&- E[E(1\{T_i \geq l\}|X)\{f_Z(Z_i(l))\frac{\partial Z_i(\theta)}{\partial \theta}\}]]|_{\theta=\bar{\theta}}(\theta-\theta_0) + o_p(\frac{||\theta-\theta_0||}{\sqrt{N}}) \\
&= D_N + o_p(\frac{||\theta-\theta_0||}{\sqrt{N}}). \qquad (10)
\end{aligned}$$

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 $D_N(\theta - \theta_0)$. Define

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

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.

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, j th 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 (Estimating H)

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

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

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 given with probability p_h to individuals who survived h periods seems 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 exogenous. 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 very 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.

The last theorem 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 given with probability p_h to individuals who survived h periods seems 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 exogenous. 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

$$\widehat{\Lambda(t)} = \sum_{s=1}^{s=t} \exp(\hat{\delta}_s) \text{ where } t \in \{0, 1, ..., K\}. \tag{11}$$

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,

$$\widehat{\lambda(s)} = \exp(\hat{\delta}_t) \text{ for } t - 1 < s < t. \tag{12}$$

This approach is similar to Han-Hausman (1990). If the durations are measured on a very fine grid, then one could also approximate the hazard by numerically differentiating the integrated hazard $\widehat{\Lambda(t)}$.

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[7] 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). 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

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

time[8] $\bar{t}$. 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, the conditional expectation of $1\{T_i \geq l\}$, conditional on $\{v, Z_i(l), X\}$, does not depend on X.

$$E[1\{T_i \geq l\}|v, Z_i(l), X = 0] = E_v\{e^{-vZ_i(l)}|v, Z_i(l), X = 0\} = E_v\{e^{-vZ_i(l)}|v, Z_i(l)\},$$

so that

$$E[1\{T_i \geq l\}|v, Z_i(l), X = 0] = E[1\{T_j \geq l\}|v, Z_j(l) = Z_i(l), X = 1].$$

However, 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,

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

In particular, one can view the indicators $1\{T_i \geq l\}$ and $1\{T_j \geq k\}$ as estimators of survival functions. In order to ensure that the unobserved heterogeneity distribution is the same for i and j so that we do not have to explicitly model the distribution of the heterogeneity, we need to choose a set to condition on.

Suppose that individuals are treated in period $\bar{t}$. In order to avoid survival bias, we condition on survival up to $\bar{t}$ and also on the index at $\bar{t} - 1$, $Z(\bar{t} - 1)$. Thus, a duration model is a natural framework to handle survival selection. Let R denote the treatment intention, X the actual treatment and $R, X \in \{0, 1\}$. For now, we assume that $R = 0$ implies $X = 0$. Below, we present a simulations in which treatments happens right after 4 weeks and in which unemployment is another time-varying regressor. Without loss of generality, we can assume that R is given at the beginning of the spell. Denote this by

[8] One could also assume that R is assigned at time the beginning of the duration spell; for given $P(R = 1)$, this is equivalent to assuming that R is assigned at time $\bar{t}$.

R^*. Let $P(R^*=0)=P(R^*=1)=\frac{1}{2}$. Let G_i denote the group of individual i and assume that there are M groups. If $G_i=G_j$ then individual i and individual j belong to the same group. For each group, we define

$$\hat{F}_{g,11|C}=\frac{\sum_{l=5}^{26}1\{T_i\geq l\}1\{X_i=1\}}{\sum_{l=5}^{26}1\{T_i\geq 4\}1\{X_i=1\}}$$

$$\begin{aligned}\hat{F}_{g,00|C} &= \frac{\sum_{l=5}^{26}[1\{T_i\geq l\}1\{R_i=0\}1\{X_i=0\}-1\{T_i\geq l\}1\{R_i=1\}1\{X_i=0\}]}{\sum_{l=5}^{26}[1\{T_i\geq 4\}1\{R_i=0\}1\{X_i=0\}-1\{T_i\geq 4\}1\{R_i=1\}1\{X_i=0\}]}\\ &= \frac{\sum_{l=5}^{26}1\{T_i\geq l\}1\{X_i=0\}[1\{R_i=0\}-1\{R_i=1\}]}{\sum_{l=5}^{26}1\{T_i\geq 4\}1\{X_i=0\}[1\{R_i=0\}-1\{R_i=1\}]}.\end{aligned}$$

We will use this estimator of the survival function instead of $1\{T_i\geq l\}$ and $1\{T_j\geq k\}$. Define $Z_{g,11}(l)$ as the index of the treated individuals of group g (i.e. those with $X=1$). Define $Z_{g,00}(l)$ as the index of the untreated individuals of group g (i.e. those with $X=0$). We then use

$$Q^*(\beta,\delta)=Q_1^*(\beta,\delta)+Q_2^*(\beta,\delta) \tag{13}$$

where

$$\begin{aligned}Q_1^*(\beta,\delta) &= \frac{\sum_i}{(N-1)}\sum_{l=1}^{l=26-22R_i^*}1\{T_i\geq l\} \qquad (14)\\ *[\frac{\sum_j}{N}1\{R_i^* &= 0\}\sum_{k=1}^{k=26}[1\{Z_i(l)<Z_j(k)\}-1\{Z_i(l)>Z_j(k)\}\\ +\frac{\sum_j}{N}1\{R_i^* &= 1\}\sum_{k=1}^{k=4}[1\{Z_i(l)<Z_j(k)\}-1\{Z_i(l)>Z_j(k)\}].\end{aligned}$$

and

$$Q_2^*(\beta,\delta)=\frac{\sum_g}{M}\sum_{l=5}^{26}\sum_{k=5}^{26}[\hat{F}_{g,11|C}-\hat{F}_{g,00|C}]1\{Z_{g,11}(l)<Z_{g,00}(k)\}.$$

The objective function $Q_1^*(\beta,\delta)$ can be interpreted as the outcomes for the two groups, in terms of treatment assignment considered at the beginning of the unemployment spell. The first group is considered up through the end of the period since is not assigned a treatment while the second group is consider up to the fourth week when it would be offered treatment. The objective function $Q_2^*(\beta,\delta)$ can be interpreted as conditioning on both survival up to the end of the fourth period as well as $z(4)$ which removes possible

dependence between treatment and the unobserved heterogeneity term. This DGP 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 th situation of non-compliance 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 heterogeneity. We illustrate this using two examples.

Example 1:

Suppose we observe $\bar{F}(t \mid x)$ for $x = 0, 1$. Moreover suppose we estimate the following model,

$$\bar{F}(t \mid x) = e^{-\phi^x \Lambda(t)}.$$

Then,

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

For a given $\Lambda(t) = -\ln \bar{F}(t \mid x = 0)$, the MLE of ϕ can be derived,

$$\begin{aligned} f(t \quad &| \quad x = 1) = \phi\lambda(t)e^{-\phi\Lambda(t)} \\ L(\phi) \quad &= \quad \ln\phi + \ln\lambda(t) - \phi\Lambda(t) \\ \frac{\partial L(\phi)}{\partial\phi} \quad &= \quad \frac{1}{\phi} - \Lambda(t) \Rightarrow \\ \hat{\phi}^{-1} \quad &= \quad E\{\Lambda(t)|x = 0\} = -E\ln\bar{F}_0(t_1), \end{aligned}$$

where $\bar{F}_0$ is the survival function for $x = 0$. If $v \sim$ Gamma(α, α), then $\bar{F}_0 = \frac{1}{\left(1+\frac{\Lambda(t)}{\alpha}\right)^\alpha}$ so that $-\ln F_0(t) = \alpha \ln\left(1 + \frac{\Lambda(t)}{\alpha}\right)$ $\phi\Lambda(t_1) = \frac{z}{v}$ where $z \sim \exp(1)$ and $v \sim$Gamma(α, α).

Thus, $\Lambda(t) = \frac{z}{v\phi}$. This yields

$$\hat{\phi}^{-1} = -E \ln \bar{F}_0(t_1) = E\alpha \ln \left(1 + \frac{z}{v\phi\alpha}\right).$$

As a result, $\hat{\phi}^{-1}$ is not an consistent estimator for ϕ^{-1}. In particular, for $\phi = 2$ and $\phi = 10$ we find the following,

True ϕ	True α	plim $\hat{\phi}$
$\phi = 2$	$\alpha = 1$	$\hat{\phi} = 1.46$
$\phi = 2$	$\alpha = 2$	$\hat{\phi} = 1.089$
$\phi = 10$	$\alpha = 1$	$\hat{\phi} = 4.04$
$\phi = 10$	$\alpha = 2$	$\hat{\phi} = 3.197$

Note that, without loss of generality, we can write the integrated baseline hazard as follows,

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

where $H(t)$ is unrestricted and $\alpha > 0$. Horowitz (1996) and Chen (2002) show how to estimate $H(t)$ at rate $\sqrt{N}$. Suppose one first estimates $H(t)$ using one of these methods. Estimating α is then like estimating a Weibull model. We therefore consider the following.

Example 2: Consider the Weibull model with a Gamma mixing distribution,

$$\begin{aligned}
\theta\left(t \mid v, x\right) &= v e^{x\beta} \alpha t^{\alpha-1} \\
v &\sim \text{Gamma}\left(\gamma, \delta\right) \\
\bar{F}\left(t_i \mid v\right) &= e^{-v e^{x\beta} \epsilon^{\alpha}} \\
\bar{F}(t_i) &= E v e^{-v e^{x\beta} t^{\alpha}} = \frac{1}{\left(1 + \frac{e^{x\beta} t^{\alpha}}{\delta}\right)^{\gamma}} \\
f(t_i) &= \frac{\alpha\gamma e^{x\beta} t_i^{\alpha}}{\delta} \frac{1}{\left(1 + \frac{e^{x_i\beta} t_i^{\alpha}}{\delta}\right)^{\gamma+1}}
\end{aligned}$$

$$
\begin{aligned}
L &= \sum_i \ln\alpha + \ln\gamma + x_i\beta + \alpha\ln t_i - \ln\delta - (\gamma+1)\ln\left(1+\frac{e^{x_i\beta}t_i^{\alpha}}{\delta}\right)\\
L_\alpha &= \sum_i \frac{1}{\alpha} + \ln t_i - \frac{(\gamma+1)}{\delta}\frac{e^{x_i\beta}t_i^{\alpha}\ln t_i}{!+\frac{e^{x_i\beta}t^{\alpha}}{\delta}}\\
L_\beta &= \sum_i x_i - \frac{(\gamma+1)}{\delta}\frac{x_i e^{x_i\beta}t^{\alpha}}{1+\frac{e^{x\beta}t^{\alpha}}{\delta}}\\
L_\gamma &= \sum_i \frac{1}{\gamma} - \ln\left(1+\frac{e^{x\beta}t_i^{\alpha}}{\delta}\right) \Rightarrow \gamma = \frac{\sum}{N}\ln\left(1+\frac{e^{x\beta}t_i^{\alpha}}{\delta}\right) = \frac{\sum}{N}\ln\left(1+\frac{(e^{x\eta t})^{\alpha}}{\delta}\right)\\
L_\delta &= \sum -\frac{1}{\delta} + (\gamma+1)\frac{e^{x\beta}t^{\alpha}\cdot\frac{1}{\delta^2}}{1+\frac{e^{x\beta}t^{\alpha}}{\delta}}
\end{aligned}
$$

$$
\begin{aligned}
\ln t^{\alpha} &= -x\beta - \ln v + \ln z\\
\ln t &= -x\eta - \frac{\ln v}{\alpha} + \frac{\ln z}{\alpha} Ew = \psi(\alpha) - \ln\beta uaw = \psi'(\alpha)\\
E\ln t &= -x\eta - \frac{1}{\alpha}\psi(\gamma) + \frac{1}{\alpha}\ln\delta + \frac{\psi(1)}{\alpha}\\
v_\alpha \ln t &= \frac{1}{\alpha^2}\left(\psi'(\gamma) + \psi'(1)\right)\\
Ee^{x\beta}t^{\alpha} &= E\frac{1}{v} = \frac{\beta}{\alpha-1}\\
Ee^{\eta}t &= Es^{1/\alpha} = \frac{1}{v^{1/\alpha}}Ez^{1/\alpha}\\
\int z^{1/\alpha}e^{-z}\alpha z \underbrace{=}_{a=\frac{1}{\alpha}+1} \Gamma\left(\frac{1}{\alpha}+1\right) &= \frac{1}{\alpha}!
\end{aligned}
$$

MLE:

$$
\begin{aligned}
p(u) &= e^{c-v} \qquad v \geq c\\
\int_c^{\infty} e^{-v}dv = -e^{-v}\big|_c^{\infty} &= e^{-c}
\end{aligned}
$$

DGP:

$$\theta\left(t \mid x, v\right) = e^{x}$$

Thus, the true value of β is one.

c	β	γ	δ	$\beta;\gamma=2,\delta=1$
0	1	1	1	1
0.1	1.11	1.12	0.96	1.06
0.2	1.154	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

$N = 10,000$

The simulation results do not depend on the distribution of x.

Lemma C1: Let $\theta(t \mid v, x) = ve^{x\beta}\lambda(t)$ where $v \perp x$. Let $v - c \mid T \geq 0 \sim \text{Gamma}(\gamma, \delta)$. 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.

5.1. Role of the Mixing Distribution in Applied Research. We have developed an econometric approach that permits estimation of both the regression parameters and the integrated hazard parameters at a rate of $\text{N}^{-1/2}$ and determined the asymptotic distribution of the estimators. Thus, while pervious approach to estimating the mixed proportion hazard model either had to specify a parametric form of the unknown (heterogeneity) mixing function to achieve $\text{N}^{1/2}$ consistency or relied on a non-parametric estimator for the unknown mixing distribution to permit estimation of the regression parameters which ruled out $\text{N}^{1/2}$ consistency, our approach seemingly eliminates any role for the mixing distribution. Thus, if we return to the basic model specification:

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

and the associated survival probability:

$$P(T \geq t|x(t)) = E_v\{\bar{F}(t|x(t), v)\} = E_v\{\exp(-v \int_0^t e^{x(s)\beta}\lambda(s)ds)\}. \tag{15}$$

we see that we can calculate elasticities of the underlying model and survival probability without any requirement to estimate the unknown mixing distribution. In this sense, the mixing distribution has been returned to the role of "nuisance" parameters, similar to the role of individual effects parameters in panel data models where they are often "conditioned out" as in e.g. Cox (1982) and Hausman, Hall and Griliches (1984). We are able to estimate the regression and integrated hazard parameters and the associated elasticities because the mixing distribution is assumed to be independent of the regression variables. Many previous estimation approaches induced a dependence between the unknown mixing distribution and the regression variables, which created the requirement

to specify a parametric form of the mixing distribution or use deconvolution estimation approaches with their associated slow convergence properties.

While most questions that arise in applied research are answered by using the above approach, a set of questions do arise that create a dependence between the unknown mixing distribution and the regression variables. Suppose that conditional on continued unemployment for τ weeks, we are interested in the elasticity with respect to a component of $x(\tau)$ of the conditional survival probability:

$$P(T \geq t|x(t), \nu, T > \tau) = E_v\{\exp(-v \int_\tau^t e^{x(s)\beta}\lambda(s)ds)\}.$$

The expectation of ν now will depend on the history of the $x(s)$ for a particular individual j. For example, with a scalar $x(s)$ and a positive β high $x(s)'s$ will imply a larger expectation of ν_j than low $x(s)'s$. However, we can continue to evaluate the elasticities without the need to estimate the distribution of ν. Return to the notation $Z_i(l;\beta,\delta) = \sum_{s=1}^{l} e^{X_{is}\beta+\delta_s}$ for individual i. We set $l = \tau$ so that for individual i we have $Z_i(\tau;\beta,\delta)$ and consider a set of individuals indexed by k who have a similar value of $Z_k(\omega;\beta,\delta)$ to $Z_i(\tau;\beta,\delta)$ for some ω. (to be completed)

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.[9]

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

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

> 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 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:

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.

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

Figure 1:

Figure 2:

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.

In Figures 3 and 4 we plot the survival curves for the 13 week and 24 week gamma heterogeneity estimates. 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.[10] The estimated survival curves fit the data quite well with only the first period not being fit well by the local polynomial estimation.

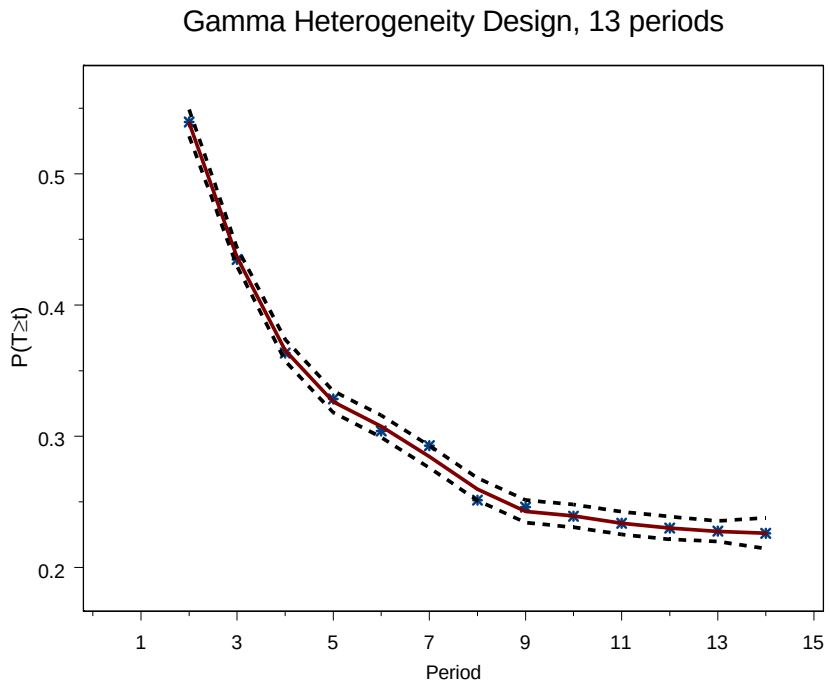

Figure 3: Gamma Heterogeneity 13 Week Survival Curve

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

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

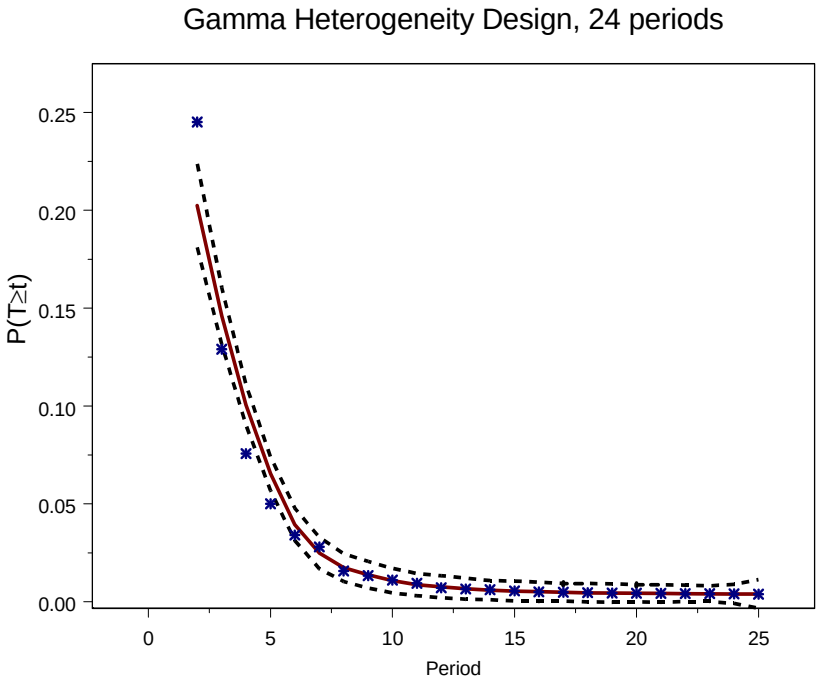

Figure 4: Gamma Heterogeneity 24 Week Survival Curve

of the objective function can now create a problem because of its lack of smoothness. Usual Newton-type gradient methods are not applicable in this situation. To date we have found that generalized pattern search algorithms perform best.[11] We use the patternsearch 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.

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

[11] 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.

Figure 5:

finding may a function 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. In Figures **??** and **??** we plot the survival curves for the 13 week and 24 week duration model estimates. We again 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 3. The estimated survival curves fit the data quite well with only the first period not being fit well by the local polynomial estimation.

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.

These results should be taken as tentative since we need to explore the performance of the semi-parametric estimator more before we can be confident in terms of its accuracy.

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 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 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 baseline hazard only converges at this set of points.

Appendix: Proofs of Theorems 1 and 2

Proof of **Theorem 1**: $EQ(\beta, \delta_2)$ is maximized at the true value of the parameters. The conditions of Newey and McFadden (1994, lemma 2.4) are satisfied so that $Q(\beta, \delta_2)$ converges uniformly to $EQ(\beta, \delta_2)$. Note that Newey and McFadden (1994, lemma 2.4) does not require continuity of $Q(\beta, \delta_2)$. Newey and McFadden (1994, theorem 1) implies consistency.

Proof of **Theorem 2**: Define $\alpha = \beta/|\beta_1|$ and note that Han's (1987) identification result applies to α. Reasoning similar as in the texts yields that $\{\beta, \delta\}$ is identified. Thus, $EQ(\beta, \delta)$ is maximized at the true value of the parameters. The conditions of Newey and McFadden (1994, lemma 2.4) are satisfied so that $Q(\beta, \delta)$ converges uniformly to $EQ(\beta, \delta)$. Note that Newey and McFadden (1994, lemma 2.4) does not require continuity of $Q(\beta, \delta)$. Newey and McFadden (1994, theorem 1) implies consistency.

Appendix: Algebra Identification

Identification

Note that

$$P\{T_i > l|, Z_i(l; \theta_0)\} = P\{Y_i > l|, Z_i(l; \theta_0)\} = E_v[\exp\{-ve^{Z_i(l;\theta_0)}\}]$$

is decreasing in $Z_i(T_j; \theta_0)$. Also note

Lemma 1: Let $\{x, T\}$ be a random sample. Let the data generating process be given by equation (2). Let the first element of x be denoted by x_1 and the second by x_2. For some x^*, let $x_{i1} = x_{i2} = x^* - c$ with positive probability and let the density of x_1, be continuous for $x_1 \in [x^* - \eta, x^* + \eta]$ for some $\eta > 0$. Let $\{\beta, c\} \in \Theta$, which is compact. Then c is identified.

Lemma 2 (Identification scalar case):

Let the assumptions of lemma 1 hold. Let $P(x_1 < x_2) > 0$. Then $\{\beta, c\}$ is identified.

Proof: Let $f(x)$ be continuously differentiable over a bounded interval $I(x)$. Let $f'(x) > 0$ if $f(x) = 0$. Then $f(x) = 0$ has only one solution for $x \in I(x)$.

Lemma 2A (alternative conditions)

Assumption 2: $x_{i2} = x_{j1}$ for some pairs i, j.

Let A_2 denote all elements of A for which $x_{i2} = x_{j1}$,

$$A_2 = \{i, j : x_{i2} = x_{j1},\ e^{(x_{i1} - x_{j1})\beta} + e^{\delta_2} > 1\}.$$

Similarly,

$$A_2' = \{i, j : x_{i2} = x_{j1},\ e^{(x_{i1} - x_{j1})\beta'} + e^{\delta_2'} > 1\}.$$

Lemma 3 (Consistency scalar case):

Let the assumptions of lemma 1 and lemma 2 hold. Let

$$\{\hat{\beta}, \hat{\delta}_2\} = \underset{\beta, \delta_2}{\arg\max} Q(\beta, \delta_2).$$

Then

$$\{\hat{\beta}, \hat{\delta}_2\} \underset{p}{\to} \{\beta, \delta_2\}.$$

Appendix: Section Estimating the Mixing Distribution

{to be completed} (**point out that we can use deconvolution methods similar to Horowitz to estimate the mixing distribution.)

Appendix: Computational Issues

by Matthew Harding, Jerry Hausman, and Tiemen Woutersen

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

$$Q(\beta, \delta) = \sum_{i=1}^{n} \sum_{l=1}^{L} 1\{T_i \geq l\} \sum_{j=1}^{n} \sum_{k=1}^{K} [1\{Z_i(l) < Z_j(k)\} - 1\{Z_i(l) > Z_j(k)\}]. \tag{16}$$

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 re-write $Q(\beta, \delta)$ using these vectors and the Rank function:

$$Q(\beta, \delta) = \frac{1}{N(N-1)} \sum_{i=1}^{Nk} d(i) \left[2Rank(Z(i)) - Nk\right]. \tag{17}$$

This simpler yet numerically identical representation[12] will be more efficient to evaluate numerically because 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 spurious local optima makes the Nelder-Mead method computationally unstable. The Nelder-Mead algorithm fails to converge or takes unreasonably long to do so.[13]

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.[14] For some given real valued objective function $Q(\gamma)$ defined on the n-dimensional Euclidean

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

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

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

space, let γ_0 be the initial guess. In our case we use $\gamma_0 = [\ \widehat{\beta}, \widehat{\delta}]_{Gamma}$ be 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 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

$$G = \left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} -1 \\ -1 \end{pmatrix}, \begin{pmatrix} -1 \\ 1 \end{pmatrix} \right\}. \tag{18}$$

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 will hopefully contain the parameter estimates under the semi-parametric setup. A bounding box is required to increase the probability that the true optimum will be visited with increased probability. For most cases a bounding box of $\pm 3 s.e.$ seems to be sufficient for convergence.

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.

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 survivor curve. 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:

$$W(t_i, t_j) = \left(1 - \left(\frac{|t_i - t_j|}{\max_{t_j \in N(t_i)} |t_i - t_j|}\right)^3\right)^3 \mathbf{1}\left\{0 \leq \frac{|t_i - t_j|}{\max_{t_j \in N(t_i)} |t_i - t_j|} < 1\right\}. \quad (19)$$

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.

References

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

Cox, D. R. and D. Oakes (1984): *Analysis of Survival Data.* London:Chapman and Hall.

Chamberlain, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, ed. by J. J. Heckman and B. Singer. Cambridge: University Press.

Chen (2002, July, Econometrica)

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.

Gørgens, T. and J. L. Horowitz (1996): "Semiparametric Estimation of a Censored Regression Model with Unknown Transformation of a dependent variable," *Journal of Econometrics*, 90, 155-191.

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

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.

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

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

Hărdle, W. (1990): *Applied Nonparametric Regression.* Cambridge: Cambridge Univer-

sity Press.

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

Heckman, J. J., and G. J. Borjas (1980): "Does Unemployment Cause Future Unemployment? Definitions, Questions and Answers for a Continuous Time Model of Heterogeneity and State Dependence," *Economica*, 47, 247-283.

Heckman, J. J., and B. Singer (1982): "The Identification Problem in Econometric Models for Duration Data," in *Advances in Econometrics,* ed. by W. Hildenbrand. New York: Cambridge University Press.

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.

Heckman, J. J., and B. Singer (1984a): "The Identifiability of the Proportional Hazard Model," *Review of Economic Studies*, 60, 231-243.

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*.

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

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

Honoré, B. E. (1998): "A Note of the Rate of Convergence of Estimators of Mixtures of Weibulls," Working paper, Princeton University.

Honoré, B. E., and J. L. Powell (1998): "Pairwise Difference Estimators for Non-Linear Models," Working paper, Princeton University.

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

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

Kaplan, E. L. and P. Meier (1958): "Nonparametric estimation from incomplete observations". *Journal of of the American Statistical Association*, 53, 457-481.

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

Kiefer and Wolfowitz (1956): Annals of Mathematical Statistics

Lancaster, T. (1976): "Redundancy, Unemployment and Manpower Policy: a Comment," *Economic Journal*, 86, 335-338.

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

Lancaster, T. (1985): "Generalized Residuals and Heterogeneous Duration Models - With Applications to the Weibull Model," *Journal of Econometrics*, 28, 113-126.

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

Lancaster, T. (1997): "Orthogonal Parameters and Panel Data," Working paper, Brown University.

Lancaster, T. (1999): "Some Econometrics of Scarring," Brown manuscript, Brown University.

Lancaster, T. (2000): "The Incidental Parameters since 1948," *Journal of Econometrics*, 95, 391-413.

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.

Lecoutre, J. P. (1983): "Almost complete convergence of the statistically equivalent blocks estimator of the regression function," in *Probability and statistical theory, Proceedings of the 4th Pannonian Symposium on Mathematical Statistics*, ed. by F. Konecny, J. Mogyorodi and W. Wertz. Dortrecht: Reidel.

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

Meyer, B. D. (1996): "Implications of the Illinois Reemployment Bonus Experiments for Theories of Unemployment and Policy Design," *Journal of Labor Economics*, 14, 26-51.

Mortensen, D. (1986): "Job Search and Labor Market Analysis," in *Handbook of Labor Economics*, vol. 2, ed. by O. Ashenfelter adn R. Layard. Amsterdam: North-Holland.

Mortensen, D. (1987): "The Effects of a UI Bonus on Job Search." Unpublished manuscript. Evanston, IL, Northwestern University.

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

Neyman, J., and E. L. Scott (1948): "Consistent Estimation from Partially Consistent Observations," *Econometrica*, 16, 1-32.

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.

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

Ridder, G., and I. Tunali (2000): "Stratified Partial Likelihood Estimation," *Journal of Econometrics.*

Robins, J.M. (1998): "Structural Nested Failure Time Models", in: *Survival Analysis*, ed. by P.K. Andersen and N. Keiding (section eds.), part of *Encyclopedia of Biostatistics*, ed. by P. Armitage and T. Colton. Wiley: Chichester, UK, 3472-4389.

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

Van den Berg, G. J. (2000): "Duration Models: Specification, Identification, and Multiple Duration," in *Handbook of Econometrics,* Vol. 5, forthcoming. Amsterdam: North-Holland.

Yamaguchi, K. (1986): "Alternative approaches to unobserved heterogeneity in the analysis of repeatable events" in *Sociological Methodology,* Vol. 16, ed. by N. B. Tuma. San Francisco: Jossey-Bass.

Woutersen, T. M. (2000): "Consistent Estimators for Panel Duration Data with Endoge-

nous Censoring and Endogenous Regressors" Brown Dissertation and UWO manuscript, UWO, London, Canada.