A Simple GMM Estimator for the Semiparametric Mixed Proportional Hazard Model

Govert E. Bijwaard, Geert Ridder and Tiemen Woutersen*
NIDI, University of Southern California, and University of Arizona

February 2013

Abstract. Ridder and Woutersen (2003) have shown that under a weak condition on the baseline hazard, there exist root-N consistent estimators of the parameters in a semiparametric Mixed Proportional Hazard model with a parametric baseline hazard and unspecified distribution of the unobserved heterogeneity. We extend the Linear Rank Estimator (LRE) of Tsiatis (1990) and Robins and Tsiatis (1991) to this class of models. The optimal LRE is a two-step estimator. We propose a simple one-step estimator that is close to optimal if there is no unobserved heterogeneity. The efficiency gain associated with the optimal LRE increases with the degree of unobserved heterogeneity.

JEL Classification: C41, C14

Keywords: mixed proportional hazard, linear rank estimation, counting process.

Corresponding author:
Govert E. Bijwaard
Netherlands Interdisciplinary Demographic Institute (NIDI)
PO Box 11650
NL-2502 AR The Hague
The Netherlands
E-mail: bijwaard@nidi.nl

*We thank Kei Hirano and Nicole Lott for very helpful comments. We also thank seminar participants at the University of Western Ontario, and the Netherlands Interdisciplinary Demographic Institute. This paper replaces the paper Method of Moments Estimation of Duration Models with Exogenous Regressors (2003). Financial support from NORFACE research programme on Migration in Europe - Social, Economic, Cultural and Policy Dynamics is gratefully acknowledged.

1. Introduction

The Mixed Proportional Hazard (MPH) model for duration data was independently introduced by Lancaster (1979) and Manton, Stallard and Vaypel (1981). It has been used quite frequently in empirical work but the standing of this model among econometricians has changed over time. Lancaster noted that the MPH model provided a simple framework for the distinction between unobserved heterogeneity and duration dependence. The question whether these two components of the MPH model are separately identified and estimable with samples of reasonable size has been answered differently. Lancaster's original answer was negative. He gave a simple example in which an observed duration distribution was consistent with an MPH model with duration dependence, but no heterogeneity, and an MPH model with no duration dependence, but with unobserved heterogeneity. Elbers and Ridder (1982) and Heckman and Singer (1984a) showed that to identify unobserved heterogeneity and duration dependence separately, some exogenous variation is needed. Besides exogenous variation, they made an at first sight innocuous assumption on the distribution of the unobserved heterogeneity, namely that this distribution had a finite mean. Heckman and Singer replaced this assumption by a restriction on the tail behavior of the unobserved heterogeneity distribution, in particular that the exponential rate at which this tail went to zero was known.

These results on nonparametric identification led to the development of estimation methods that required fewer parametric assumptions. Heckman and Singer (1984b) used the NPMLE for mixture models that was first characterized by Lindsay (1983) to estimate regression parameters and the parameters of the baseline hazard in an MPH model. A problem with Heckman and Singer's NPMLE is that the speed of convergence and the asymptotic distribution of the estimators are not known. This is not just a theoretical concern. Simulation studies, e.g. the recent study by Baker and Melino (2000), have shown that the NPMLE gives biased estimates of all the parameters in the MPH model if the baseline hazard is left fairly free.

Horowitz (1999) proposed a semiparametric estimator for the MPH model that does not require parametric assumptions either on the unobserved heterogeneity or on the

duration dependence. This estimator is based on Horowitz's (1996) estimator for a semiparametric transformation model. The main problem in the estimation of the parameters of the MPH model is the estimation of a scale parameter. This scale parameter enters the (integrated) baseline hazard as a power and the regression parameter as a multiplicative constant. The scale parameter is identified by the assumption that the mean of the distribution of the unobserved heterogeneity is finite. Because the estimator of the scale parameter only uses information on durations close to zero, the rate of convergence is arbitrarily close to $N^{2/5}$, the fastest possible rate given the model. Honoré (1990) proposed an estimator for the Weibull MPH based on the same idea, and his estimator has the same rate of convergence. The slow rate of convergence of these estimators is an impediment to their use in applied work. It is, however, consistent with the Monte Carlo evidence on the NPMLE and also with a result in Hahn (1994). Hahn shows that in the MPH model with Weibull baseline hazard (but unspecified distribution for the unobserved heterogeneity), the information matrix is singular. This precludes the existence of regular $\sqrt{N}$ consistent estimators of the parameters of this model.

These results suggest that the original idea of using the MPH model to distinguish between unobserved heterogeneity and duration dependence is sound in theory, but that in practice this can be done only in very large samples. However, the situation may not be as bleak. For instance, Ridder and Woutersen (2003) reconsider Hahn's (1994) result. They show that the Weibull example is a worst case, although it is not the only parametric model that gives a singular efficiency bound. They characterize the class of parametric models for the baseline hazard that gives a singular bound, and they show that the defining feature of this class is that the baseline hazard at 0 is either 0 or ∞. Note that this is the case for the Weibull baseline hazard. Although MPH models with Weibull like baseline hazards are identified, their estimation is problematic. Ridder and Woutersen argue that Weibull type behavior near zero is a consequence of a convenient functional form and not of interest in its own right. The distinction between unobserved heterogeneity and duration dependence is more relevant for strictly positive durations. They show that bounding the baseline hazard away from 0 and ∞ at 0 resolves the

problem. Incidentally, this assumption is also sufficient for nonparametric identification of the MPH model and with it the finite mean assumption can be discarded.

Until now we have taken for granted that it is important to make a distinction between unobserved heterogeneity and duration dependence. It has been argued (see e.g. Wooldridge (2005)) that the distinction is irrelevant if one wants to estimate the impact of covariates on the average duration. There are, however, instances where the distinction is important in its own right. Examples are the distinction between heterogeneity and duration dependence as an explanation of the decreasing probability of re-employment for the unemployed (Lancaster (1976), Heckman (1991)). Recently, Chiaporri and Salanie (2000) have argued that the distinction is also important to understand insurance contracts. The distinction is also important if one is interested in the effect of covariates on the quantiles of the duration distribution, which may often be the more interesting effect. In particular, let the waiting time to some event T have a conditional distribution given observed and unobserved covariates with hazard rate

$$\kappa(t|X_h(t), V, \theta) = \lambda(t, \alpha)e^{\beta' X(t)}V,$$

where $\lambda(t, \alpha)$ is the baseline hazard, β is the regressor parameter, and $X_h(t) = \{X(s)|0 \leq s \leq t\}$ is the sample path of the observed covariates. The covariate at time t is denoted by $X(t)$ and V is the multiplicative unobserved heterogeneity. We assume that $X_h(t)$ is independent of V. Since the distribution of V is not specified, the only parameter of the model is $\theta = \{\alpha, \beta\}$. Integrating the baseline hazard with respect to time gives the integrated baseline hazard, $\Lambda(t)$. For an MPH with such time constant covariates, the derivative of the qth quantile $t_q(X)$ with respect to the covariate X is

$$\frac{\partial t_q(X)}{\partial X} = -\beta \frac{\Lambda\big(t_q(X); \alpha\big)}{\lambda\big(t_q(X); \alpha\big)} \tag{1}$$

which is independent of the distribution of the unobserved heterogeneity but depends on the baseline hazard.

In this paper we consider a simple $\sqrt{N}$ consistent estimator for the parameters of a semiparametric MPH model with unspecified distribution of the unobserved heterogeneity. This estimator is a GMM estimator that uses moment conditions to derive estimating

equations. It is based on the linear rank statistic of Prentice (1978). That statistic has been used by Tsiatis (1990) to estimate the parameters of a censored regression model and by Robins and Tsiatis (1992) in the Accelerated Failure Time model. In its simplest form, the estimator does not require nonparametric estimation of unknown densities. Hence, it is simpler than the semiparametric maximum likelihood estimator of Bearse et al. (2007). Moreover, we provide primitive conditions under which our estimator converges while Bearse et al. (2007) assume $\sqrt{N}$ consistency.

Our simple GMM estimator is not efficient. In the case of constant covariates and no censoring it does not reach the Hahn (1994) efficiency bound. Fully efficient estimation requires a second step, in which the hazard of the distribution of the integrated hazard is estimated. This hazard is then used to construct the likelihood function for arbitrarily (noninformatively) censored integrated hazards, and this likelihood is maximized over the parameters of the MPH model. As is evident from the discussion results in Bearse et al. (2007), the second step requires much care, even in the simpler case of no censoring, and achieving the efficiency gain associated with it may be problematic. Therefore, we recommend the simple GMM estimator, which performed well in simulations.

A paper that is related to our work is Hausman and Woutersen (2005), who estimate a related duration model. That paper does not use the identification strategy of Ridder and Woutersen (2003) and requires some regressors to vary over time. This paper allows regressors to vary over time but does not require it. By redefining the regressors to be zero in all but one period, this paper allows the effect of the regressors to have a different coefficient for each period, while Hausman and Woutersen (2005) do not allow for that.

The outline of the article is as follows. In Section 2 discusses the MPH model and a counting process interpretation of the MPH model is given in the appendix. The counting process approach simplifies the definition of predictable time–varying explanatory variables and noninformative censoring. Section 3 presents our version of the linear rank estimator. In Section 4 we derive the asymptotic properties of the (two stage) optimal linear rank estimator. Section 5 discusses the implementation of the estimator. The Monte Carlo experiments of Section 6 give some insight into the small sample behavior of

the estimator. Finally, in Section 7 we apply our estimator to a real data set of cyclical migration. Section 8 summarizes the results and states our conclusion.

2. The Mixed Proportional Hazard model

The waiting time to some event T has a conditional distribution given observed and unobserved covariates with hazard rate

$$\kappa(t|X_h(t), V, \theta) = \lambda(t, \alpha)e^{\beta' X(t)}V. \tag{2}$$

where $X_h(t) = \{X(s)|0 \leq s \leq t\}$ is the sample path of the observed covariates, X, up to and including time t, which without loss of generality is assumed to be left continuous, and V is the multiplicative unobserved heterogeneity. Because V is time constant we assume that its value is determined at time zero. We assume that $X_h(t)$ is independent of V. Note that although we express the hazard at t as a function of $X(t)$, we can allow for lagged covariates by redefining $X(t)$. The positive function $\lambda(t; \alpha)$ is the baseline hazard that is specified up to a vector of parameters α. It reflects the duration dependence of the hazard rate. The other parameter of the model, β, is β is the regressor parameter. The distribution of V is not specified so that the parameter vector is $\theta = \{\alpha, \beta\}$. In the appendix, we give a counting process interpretation of this model.

2.1. Durations and Transformed Durations.

The MPH model in (2) specifies the conditional hazard of the distribution of T given $X_h(t), V$. Because V is not observed, we need to integrate with respect to the conditional distribution of V given $T > t$, $X_h(t)$ to obtain the hazard conditional on $X_h(t)$. An alternative approach is to consider the transformed duration

$$h(t, X_h(t), \theta) = \int_0^t \lambda(s; \alpha)e^{\beta' X(s)}\,\mathrm{d}s. \tag{3}$$

This transformation is the observed integrated hazard, i.e. the integrated baseline hazard except for the unobservable V. A key feature of the MPH model is that in the population

$$h(T, X_h(T), \theta_0) = \frac{A}{V} \stackrel{d}{=} U_0 \tag{4}$$

with A a standard exponential random variable.

Equations (3) and (4) show that the MPH model is essentially a transformation model that transforms the conditional distribution of T given the observable covariates $X_h(.)$ to a positive random variable that is independent of $X_h(.)$ and of the baseline hazard $\lambda(.;\alpha_0)$. This independence is the key to understanding the intuition behind the proposed Linear Rank Estimator (LRE). The fact that the right hand side random variable is the ratio of a standard exponential and a positive random variable only plays a role in the interpretation of the components of the transformation as a baseline hazard and a regression function that multiplies the baseline hazard. For parameter values $\theta \neq \theta_0$, i.e. not equal to the true values, we have

$$h(T, X_h(t), \theta) = U \tag{5}$$

with U a nonnegative random variable. We denote the inverse of $h(T, X_h(t), \theta)$ with respect to its first argument by $h^{-1}(U, X_h(t), \theta)$ and we sometime suppress the last to arguments and use $h(T)$ and $h^{-1}(U)$ for $h(T, X_h(t), \theta)$ and $h^{-1}(U, X_h(t), \theta)$. The hazard rate of $U = h(T)$ is

$$\begin{aligned} \kappa_U(u|V) &= \kappa_T\big(h^{-1}(u)\big)\frac{1}{h'\big(h^{-1}(u)\big)} \\ &= \frac{\lambda\big(h^{-1}(u, X_h(u), \theta), \alpha_0\big)}{\lambda\big(h^{-1}(u, X_h(u), \theta), \alpha\big)} e^{(\beta_0-\beta)' X^U(u,\theta)} V, \end{aligned} \tag{6}$$

where $X^U(u,\theta) = X\big(h^{-1}(u, X_h(u), \theta)\big)$ denotes the process of the time-varying covariate on the transformed duration time.

Like the distribution of T, that of the transformed duration U can be expressed by a (transformed) counting process $\{N^U(u,\theta)|u \geq 0\}$. The relation between the original and transformed counting processes and the observation indicator is

$$N^U(u,\theta) = N\big(h^{-1}(u, X_h(u), \theta)\big) \tag{7}$$

$$Y^U(u,\theta) = Y\big(h^{-1}(u, X_h(u), \theta)\big). \tag{8}$$

The intensity of the transformed counting process (with respect to history $X_h^U(u,\theta), Y_h^U(u,\theta)$

is (see Andersen et al. (1993), page 87, and using (6))

$$\begin{aligned}&\Pr\big(\mathrm{d}N^U(u,\theta)=1\big|X_h^U(u,\theta),Y_h^U(u,\theta)\big)=\\&\quad=Y^U(u,\theta)\frac{\lambda\big(h^{-1}(u,X_h(u),\theta),\alpha_0\big)}{\lambda\big(h^{-1}(u,X_h(u),\theta),\alpha\big)}e^{(\beta_0-\beta)'X^U(u,\theta)}\mathrm{E}\big[V|X_h^U(u,\theta),Y_h^U(u,\theta)\big]\mathrm{d}u,\end{aligned}\tag{9}$$

and we denote this hazard by $\kappa_U\big(u|X_h^U(u,\theta),Y_h^U(u,\theta)\big)$. For the population parameter value θ_0, this becomes

$$\Pr\Big(\mathrm{d}N^U(u,\theta_0)=1\Big|X_h^U(u,\theta_0),Y_h^U(u,\theta_0)\Big)=Y^U(u,\theta_0)\mathrm{E}\Big[V|X_h^U(u,\theta_0),Y_h^U(u,\theta_0)\Big]\mathrm{d}u.\tag{10}$$

If censoring is noninformative, i.e. $Y(t)=I(t\le T)I_C(t)$ with C independent of T (but possibly dependent on X), then U_0 contains all the information concerning V, as implied by equation (4). Therefore

$$\Pr\Big(\mathrm{d}N^U(u,\theta_0)=1\Big|X_h^U(u,\theta_0),Y_h^U(u,\theta_0)\Big)=Y^U(u,\theta_0)\mathrm{E}\Big[V\Big|U_0\ge u\Big]\mathrm{d}u,\tag{11}$$

and the intensity is independent of $X_h^U(u,\theta_0)$. This independence is the basis for the estimation of the parameters of the MPH model. We denote the hazard in (11) by $\kappa_0(u)$.

Example 1 [Piecewise constant hazard and time-varying covariate]**.** *Consider an MPH model with a single time-varying covariate $X(t)$. The baseline hazard is piecewise constant, so*

$$\lambda(t,\alpha)=e^{\alpha}I(0\le t\le t_1)+I(t>t_1).$$

The covariate $X(t)$ is changing, for all individuals, at time $t_2>t_1$ from random variable X_1 to X_2. Thus, the hazard rate of U is

$$\kappa_U(u)=\begin{cases}e^{(\alpha_0-\alpha)+(\beta_0-\beta)X_1}\mathrm{E}[V|U\ge u] & 0\le U\le e^{\alpha+\beta X_1}t_1\\ e^{(\beta_0-\beta)X_1}\mathrm{E}[V|U\ge u] & e^{\alpha+\beta X_1}t_1<U\le e^{\alpha+\beta X_1}t_1+e^{\beta X_2}(t_2-t_1)\\ e^{(\beta_0-\beta)X_2}\mathrm{E}[V|U\ge u] & U>e^{\alpha+\beta X_1}t_1+e^{\beta X_2}(t_2-t_1).\end{cases}\tag{12}$$

For the population parameter value $\theta_0=(\alpha_0,\beta_0)$, this becomes

$$\kappa_0(u)=\mathrm{E}[V|U\ge u].$$

If V has a Gamma distribution with mean 1 and variance σ^2, then

$$\kappa_0(u)=\frac{1}{1+\sigma^2u}.$$

The basis of the LRE is that for the true transformation, and only for the true parameter vector, the hazard rate of the transformed variable is constant if we condition on V. This implies that the unconditional hazard rate (i.e. without conditioning on V) only depends on the distribution of V and not on the regressors. A typical way to test the significance of a covariate on the hazard is the rank test (see Prentice (1978)). This test is based on (possibly weighted) comparisons of the estimated nonparametric hazard rates. It is also equivalent to the score test for significance of a (vector of) coefficient(s) that arises from the Cox partial likelihood. The test rejects the influence of the covariate(s) on the hazard when it is 'close' to zero. Tsiatis (1990) shows that the inverse of the rank test, the value of the (vector of) coefficient(s) that sets the rank test equal to zero, can be used as an estimating equation for Accelerated Failure Time (AFT) models. Here we extend the inverse rank estimation to include the parameters of the duration dependence.

Before we elaborate on the LRE in detail, we first discuss nonparametric identification of the MPH model.

2.2. Identification. Using the counting process framework, we can express an important assumption on the covariate process. We assume that with $\mathrm{d}X(t) = X(t+) - X(t)$

$$\mathrm{d}X(t) \perp N(s), s \geq t | Y_h(t), X_h(t). \tag{13}$$

For the observation process we make a similar assumption. As noted, in all cases of interest we have $Y(t) = I(t \leq T) I_C(t)$ with some random set, e.g. the set $t \leq C$ for right censoring. We assume

$$\mathrm{d}I_C(t) \perp N(s), s \geq t | Y_h(t), X_h(t). \tag{14}$$

In other words, we assume that changes in X and I_C at t are conditionally independent of the occurrence of the event after t. This means that $X(t)$ and $I_C(t)$ are predetermined at t. Note that if $X(t)$ or $I_C(t)$ depends on V, then these assumptions cannot hold.

In (B.1) and the following equations, we condition on the unobserved V. The corresponding unconditional results are obtained by taking the expectation of V given $Y_h(t), X_h(t)$. If $Y(t) = I(t \leq T) I_C(t)$ with $I_C(t)$ independent of V, then we need not condition on $I_C(t)$,

and the conditional expectation is

$$\mathrm{E}\big(V|T \geq t, Y_h(t), X_h(t)\big). \tag{15}$$

The hazard that is not conditional on V is

$$\kappa(t|X_h(t), \theta) = \lambda(t, \alpha)e^{\beta' X(t)}\mathrm{E}\big[V|T \geq t, Y_h(t), X_h(t)\big]. \tag{16}$$

Nonparametric identification of the MPH model has been studied by Elbers and Ridder (1982) and Heckman and Singer (1984a). These results refer to the model in which both the baseline hazard and the distribution of the unobserved heterogeneity are left unspecified. In their proofs, Elbers and Ridder (1982) need the assumption the mean of the distribution of V is finite, and Heckman and Singer (1984a) need the assumption that the tail of that distribution decreases at a fast enough and known rate. Ridder and Woutersen (2003) show that it is possible to replace assumptions on the distribution of V by an assumption on the behavior of the baseline hazard near 0. They show that with time constant covariates the semiparametric MPH model with parametric baseline hazard is identified if the following assumptions hold.

(I1) $0 < \lim_{t\downarrow 0} \lambda(t, \alpha_0) < \infty$. Further $\Lambda(t_0, \alpha_0) = 1$ for some t_0 and $\Lambda(\infty, \alpha_0) = \infty$ with $\Lambda(t, \alpha_0) = \int_0^t \lambda(s, \alpha_0)\,\mathrm{d}s$.

(I2) V and X are stochastically independent.

(I3) There are x_1, x_2 in the support of X with $\beta_0' x_1 \neq \beta_0' x_2$.

(I4) If $\lambda(t, \alpha_0) = \lambda(t, \tilde{\alpha}_0)$ for all $t > 0$, then $\alpha_0 = \tilde{\alpha}_0$, and if $\beta_0' x = \tilde{\beta}_0' x$ for all x in the support of X, then $\beta_0 = \tilde{\beta}_0$.

The key assumptions are the bound on the baseline hazard at 0 in assumption (I1) and assumptions (I2) and (I3). The other assumptions are normalizations (the second part of assumption (I1)) or assumptions that ensure the identification of the parametric functions (assumption (I4)). The main difference with the identification results in Elbers and Ridder and Heckman and Singer is that assumptions on the distribution of V are replaced by an assumption on the baseline hazard at 0. The duality of these two types of

assumptions is a consequence of the Tauber theorem (see Feller (1971), Chapter 13). The assumptions for identification can be weakened if some of the covariates are time-varying, but assumptions (I1)-(I4) are also sufficient in that case.

3. The Linear Rank Estimator

There are a number of estimators for transformation models that transform to an unspecified distribution. Amemiya (1985) has shown that the Nonlinear 2SLS estimator introduced in Amemiya (1974) can be used to estimate both the regression parameters and the parameters in the transformation. Han (1987) proposed an estimator that maximizes the rank correlation between the transformed dependent variable and a linear combination of the covariates (see also Sherman (1993)). These estimators, as well as Khan (2001), Han's (2002), and Khan and Tamer (2007) require that the regressors are time-constant. Amemiya's N2SLS estimator can be used even with time-varying covariates, but not with censored data. The Linear Rank Estimator (LRE) for this transformation model can deal with both time-varying regressors and general noninformative censoring. Moreover, it can be used to estimate the baseline hazard rate, rather than just the transformation.

Before we turn to the general model, we discuss a simple example to provide more insight into the inverse rank estimation approach. Suppose we would like to test whether a covariate X influences the hazard. If the covariate does not influence the hazard, the mean of the covariate among the survivors does not change with the survival time, i.e. $\mathrm{E}[X|T \geq t] = \mathrm{E}[X]$. Then the rank test statistic is (assuming no censoring)

$$\sum_{i}^{n} \left[X_i - \frac{\sum_j Y_j(t_i) X_j}{\sum_j Y_j(t_i)} \right],$$

where the second term is the average of the covariate among those units still alive at t_i. Thus, for each observation of the covariate we compare the observed value with its expected value among those still alive (under the hypothesis of no effect of the covariate) and sum over all observations. If this sum is significantly different from zero, then we reject the null of no influence.

Now assume that the true model is an MPH model without duration dependence with transformed duration $U = e^{\beta X} T$. Then, for the true parameter $\beta = \beta_0$ the hazard of

U does not depend on the covariate X. This implies that the rank statistic for the true parameter on the transformed U–time is zero. However, the β_0 is unknown and an inverse rank estimate $\hat{\beta}$ of β_0 is the value of β for which

$$\sum_i^n \left[X_i - \frac{\sum_j Y_j^U(U_i) X_j}{\sum_j Y_j^U(U_i)} \right] = 0$$

with $U_i = e^{\hat{\beta} X_i} t_i$ and $Y_j^U(u) = I(U_j \geq u)$, the observation indicator on the (transformed) U–time. Tsiatis (1990) used this statistic as an estimating equation for the parameters in a censored linear regression model, and Robins and Tsiatis (1992) employed the same statistic to estimate the parameters in the Accelerated Failure Time (AFT) model with time-varying covariates introduced by Cox and Oakes (1984).

In the general MPH model, we consider a random sample $\tilde{T}_i, \Delta_i, X_{h,i}(T_i), i = 1, \ldots, N$. The indicator Δ_i is 1 if the duration is observed and 0 if it is censored. For some θ this random sample can be transformed to $\tilde{U}_i(\theta), \Delta_i, X_{h,i}^U\big(\tilde{U}_i(\theta)\big), i = 1, \ldots, N$. The rank statistic for these data is

$$S_N(\theta, W) = \sum_{i=1}^N \Delta_i \Big\{ W\Big(\tilde{U}_i(\theta), X_{h,i}^U\big(\tilde{U}_i(\theta)\big)\Big) - W_h\Big(\tilde{U}_i(\theta)\Big) \Big\} \tag{17}$$

with

$$W_h\Big(\tilde{U}_i(\theta)\Big) = \frac{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big) W\Big(\tilde{U}_i(\theta), X_{h,j}^U\big(\tilde{U}_i(\theta)\big)\Big)}{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big)}.$$

In (17) W is a known function of $\tilde{U}_i(\theta)$ and $X_{h,i}^U\big(\tilde{U}_i(\theta)\big)$ with a dimension not smaller than that of θ. The interpretation of S_N is that it compares the weight function for a transformed duration that ends at $\tilde{U}_i(\theta)$ to the average of the weight functions at that time for the units that are under observation. The suggestion is that the difference between the weight function for unit i and the average weight function for the units under observation is 0 at the population parameter value θ_0. In large samples this is correct if we choose, for instance, $W\Big(\tilde{U}_i(\theta), X_{h,i}^U\big(\tilde{U}_i(\theta)\big)\Big) = X_{h,i}^U\big(\tilde{U}_i(\theta)\big)$ because for $\theta = \theta_0$ the transformed duration U_0 is independent of $X_{h,i}^U$. Another choice of W is the indicator function, $W\Big(\tilde{U}_i(\theta), X_{h,i}^U\big(\tilde{U}_i(\theta)\big)\Big) = I(u_k < \tilde{U}_i(\theta) \leq u_{k+1})$ where u_k and u_{k+1} are just two scalars. For $\theta = \theta_0$ the transformed durations U_{0i} are identically distributed, and this implies that the rank statistic is 0 in large samples for this choice of W.

Because $S_N(\theta, W)$ is not continuous in θ (if W is continuous in $\tilde{U}(\theta)$ it need not be a step function either), we may not be able to find a solution to $S_N(\theta, W) = 0$. For that reason, we define the Linear Rank Estimator (LRE) of the parameters of the MPH model by

$$\hat{\theta}_N(W) = \arg\min_{\theta\in\Theta} S_N(\theta, W)' S_N(\theta, W). \tag{18}$$

Lemma 1 below shows that S_N is asymptotically equivalent to a linear (and hence continuous) function in θ.

Example 2 [Continuation of Example 1]**.** *Simple weight functions for this example are*

$$\begin{aligned} W_\beta(u, X) &= X(u) \\ W_\alpha(u, X) &= I\big(0 \le u \le e^\alpha t_1 e^{\beta X(u)}\big) \end{aligned}$$

with $X(u) = X_1$ *when* $h^{-1}(U, X_h(t), \theta) \le t_2$ *and* $X(u) = X_2$ *otherwise. Denote the interval indicator by* $I_1\big(u, X_i(u)\big)$ *The estimation equations become*

$$\begin{aligned} S_{N,\beta}(\theta, W) &= \sum_{i=1}^N \Delta_i \left\{ X_i(\tilde{U}_i) - \frac{\sum_{j=1}^N I(\tilde{U}_j \ge U_i) X_j(\tilde{U}_i)}{\sum_{j=1}^N I(\tilde{U}_j \ge \tilde{U}_i)} \right\} \\ S_{N,\alpha}(\theta, W) &= \sum_{i=1}^N \Delta_i \left\{ I_1\big(\tilde{U}_i, X_i(\tilde{U}_i)\big) - \frac{\sum_{j=1}^N I(\tilde{U}_j \ge \tilde{U}_i) I_1\big(\tilde{U}_i, X_j(\tilde{U}_i)\big)}{\sum_{j=1}^N I(\tilde{U}_j \ge \tilde{U}_i)} \right\}. \end{aligned}$$

The expression for the rank statistic simplifies if we order the observations by increasing transformed duration

$$\tilde{U}_{(1)}(\theta) \le \tilde{U}_{(2)}(\theta) \le \ldots \le \tilde{U}_{(N)}(\theta).$$

In the ordered transformed durations, we obtain

$$\begin{aligned} S_{N,\beta}(\theta, W) &= \sum_{i=1}^N \Delta_{(i)} \left\{ X_{(i)}(\tilde{U}_{(i)}) - \frac{\sum_{j=i}^N X_{(j)}\big(\tilde{U}_{(i)}\big)}{N - i + 1} \right\} \\ S_{N,\alpha}(\theta, W) &= \sum_{i=1}^N \Delta_{(i)} \left\{ I_1\big(\tilde{U}_{(i)}, X_{(i)}(\tilde{U}_{(i)})\big) - \frac{\sum_{j=i}^N I_1\big(\tilde{U}_{(i)}, X_{(j)}(\tilde{U}_{(i)})\big)}{N - i + 1} \right\}. \end{aligned}$$

Thus, $S_{N,\beta}$ *compares the value of* $X_{(i)}$ *at transformed duration* $\tilde{U}_{(i)}$ *(which is either drawn from* X_1 *or from* X_2*) to the average value of* $X_{(j)}$ *of all* $j > i$ *at* $\tilde{U}_{(i)}$ *and takes the sum over all (uncensored) units.* $S_{N,\alpha}$ *compares the value of the indicator function,* $I\big(\tilde{U}_{(i)}, X_{(i)}(\tilde{U}_{(i)})\big)$*, at transformed duration* $\tilde{U}_{(i)}$ *(which is either 1 or 0) to the average value of the indicator functions,* $I\big(\tilde{U}_{(i)}, X_{(j)}(\tilde{U}_{(i)})\big)$ *of all* $j > i$ *at* $\tilde{U}_{(i)}$*.*

The functions $S_{N,\beta}$ and $S_{N,\alpha}$ are not continuous in $\theta = (\alpha, \beta)$. The points of discontinuity are values of θ that make e.g. $\tilde{U}_{(k)}(\theta) = \tilde{U}_{(k+1)}(\theta)$. If $\Delta_{(k)} = \Delta_{(k+1)} = 1$, the discontinuity is

$$\frac{X_{(k+1)}\big(\tilde{U}_{(k)}(\theta)\big) - X_{(k)}\big(\tilde{U}_{(k)}(\theta)\big)}{N-k} \tag{19}$$

$$\frac{I\Big(\tilde{U}_{(k)} \leq e^{\alpha} t_1 \exp\Big[\beta X_{(k+1)}\big(\tilde{U}_{(k)}(\theta)\big)\Big]\Big) - I\Big(\tilde{U}_{(k)} \leq e^{\alpha} t_1 \exp\Big[\beta X_{(k)}\big(\tilde{U}_{(k)}(\theta)\big)\Big]\Big)}{N-k}, \tag{20}$$

and this difference goes to 0 if N increases for both $W_\beta(u, X)$ and $W_\alpha(u, X)$.

For consistency and asymptotic normality of the MPH LRE estimator, we make the several assumptions that we present in the appendix. Under these assumptions, the linear rank statistic is asymptotically equivalent to a statistic that is linear in the parameters. This linearity causes the estimators to be asymptotically normally distributed. Here is our linearity result where $\bar{K}$ denotes a positive constant.

Lemma 1

Under assumptions (A1)–(A4) for all $\bar{K} > 0$

$$\sup_{|\theta - \theta_0| \leq C N^{-\frac{1}{2}}} N^{-\frac{1}{2}} \Big| S_N(\theta, W) - \tilde{S}_N(\theta, W) \Big| \xrightarrow{p} 0 \tag{21}$$

with

$$\tilde{S}_N(\theta, W) = \sum_{i=1}^{N} \int_0^{\tau} \Big(W\big(u, X^U_{h,i}(u, \theta_0)\big) - W_h(u, \theta_0) \Big) \, dM_i^0(u) + B(\theta_0) N(\beta - \beta_0) + A(\theta_0) N(\alpha - \alpha_0). \tag{22}$$

Proof: See Appendix.

From Lemma 1, we obtain the asymptotic distribution of the LRE.

Theorem 1

Under assumptions (A1)–(A4) we have with $D(\theta_0) = \big[A(\theta_0) B(\theta_0)\big]$

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, D(\theta_0)^{-1} V(\theta_0) D'(\theta_0)^{-1}\big) \tag{23}$$

with

$$\frac{1}{N}\sum_{i=1}^{N}\int_0^\tau \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big)\Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big)' \cdot \\ \cdot\, Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u \xrightarrow{p} V(\theta_0). \quad (24)$$

Proof: See Appendix.

The function $S_N(\theta, W)$ in lemma 1 is an 'approximate derivative' and an 'influence function' in the terminology of Newey and McFadden (1994). It allows us to view the asymptotic behavior of an estimator as an average, multiplied by $\sqrt{N}$. Moreover, as Horowitz (2001, theorem 2.2) shows, bootstrapping an asymptotically normally distributed estimator that can be represented by an influence function yields a consistent variance-covariance matrix and consistent confidence intervals.[1]

The variance matrix of the LRE is the limit of (we suppress the dependence on $X_{h,i}^U(u,\theta_0)$ and $Y_{h,i}^U(u,\theta_0)$ and use a subscript i instead)

$$\left[\frac{1}{N}\sum_{i=1}^{N}\int_0^\tau \big(W_i(u) - W_h(u,\theta_0)\big)\frac{\partial \ln \kappa_{U_i}}{\partial \theta'} Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right]^{-1} \cdot \\ \left[\frac{1}{N}\sum_{i=1}^{N}\int_0^\tau \big(W_i(u) - W_h(u,\theta_0)\big)\big(W_i(u) - W_h(u,\theta_0)\big)' Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right] \cdot \\ \left[\frac{1}{N}\sum_{i=1}^{N}\int_0^\tau \big(W_i(u) - W_h(u,\theta_0)\big)\frac{\partial \ln \kappa_{U_i}}{\partial \theta'} Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right]'^{-1} \quad (25)$$

By the Cauchy-Schwartz inequality, this matrix is minimal if

$$W_{0i}\big(u, X_{h,i}^U(u,\theta_0)\big) = \frac{\partial \ln \kappa_U\big(u|X_{h,i}^U(u,\theta_0)\big)}{\partial \theta}. \quad (26)$$

With this weighting matrix, $V(\theta_0) = D(\theta_0)$ and the variance matrix of the LRE with the optimal weighting matrix is $V(\theta_0)$. A consistent estimator of this matrix is

$$\frac{1}{N}\sum_{i=1}^{N}\int_0^\tau \big(W_{0i}(u) - W_{h,0}(u,\theta_0)\big)\big(W_{0i}(u) - W_{h,0}(u,\theta_0)\big)'\,\mathrm{d}N(u), \quad (27)$$

which is just the average over the uncensored population transformed durations U_0.

[1] Horowitz (2001, Theorem 2.2) averages $g_n(X_i)$; the STATA program on our website is sufficiently fast to apply the bootstrap to most survey datasets.

The optimal weighting function depends on the distribution of U_0 through its hazard and the derivative of that hazard. In the Appendix, we find from (B.29) and (B.30) that

$$\begin{aligned}\frac{\partial \ln \kappa_U(u,\theta)}{\partial \alpha} &= -\frac{\kappa_0'(u)}{\kappa_0(u)}\int_0^u \frac{\partial \ln \lambda}{\partial \alpha}\big(h_0^{-1}(s),\alpha_0\big)\mathrm{d}s - \\ &\qquad -\frac{\partial \ln \lambda}{\partial \alpha}\big(h_0^{-1}(u),\alpha_0\big) \qquad (28)\\ \frac{\partial \ln \kappa_U(u,\theta)}{\partial \beta} &= -\frac{\kappa_0'(u)}{\kappa_0(u)}\int_0^u X\big(h_0^{-1}(s)\big)\mathrm{d}s - X\big(h_0^{-1}(u)\big). \qquad (29)\end{aligned}$$

Note that the inverse of the transformed duration is also needed, so that a closed form of this inverse is desirable.

Example 3 [Continuation of Example 1]**.** *By* (B.29) *and* (B.30) *the optimal weighting functions are*

$$\begin{aligned}W_{0\beta}(u,X) &= -\Big(1+u\frac{\kappa_0'(u)}{\kappa_0(u)}\Big)X(u)\\ W_{0\alpha}(u,X) &= -\Big(1+u\frac{\kappa_0'(u)}{\kappa_0(u)}\Big)I(0\le u\le e^{\alpha}t_1e^{\beta X(u)}).\end{aligned}$$

If U_0 is unit–exponentially distributed, i.e. if there is no unobserved heterogeneity, then we obtain the weighting functions in Example 2. In general, this weighting function is a feasible but suboptimal choice. Note that factor in W_0 depends on the distribution of V. If V has a Gamma distribution with mean 1 and variance σ^2, then

$$1+u\frac{\kappa_0'(u)}{\kappa_0(u)} = \frac{1}{1+\sigma^2 u}.$$

Hence the weight decreases with the transformed duration.

4. The Linear Rank Estimator with an Estimated weight function

First, we simplify the notation by suppressing the dependence of the weight function on the covariate history. Instead we make the dependence of this function on the parameters θ_0 and the hazard of U_0, κ_0, explicit. With this change, the LRE estimating equation is

$$S_N(\theta,W) = \sum_{i=1}^N \Delta_i\Big\{W_i\big(\tilde{U}_i(\theta),\theta_0,\kappa_0\big) - W_h\big(\tilde{U}_i(\theta),\theta_0,\kappa_0\big)\Big\} \qquad (30)$$

with

$$W_h\big(\tilde{U}_i(\theta),\theta_0,\kappa_0\big) = \frac{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big)W_j\big(\tilde{U}_i(\theta),\theta_0,\kappa_0\big)}{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big)}.$$

The optimal weight functions are given in (28) and (29). We obtain an estimated weight function by substituting the consistent first-stage estimates $\hat{\beta}_N, \hat{\alpha}_N$ for the parameters and by using a nonparametric estimator for the hazard κ_0 of U_0 and its derivative. This complicates the asymptotic analysis of the estimator because the estimated weight function is not predictable, i.e. at (transformed duration) time u it depends on values of the transformed durations beyond u.

To deal with this problem, we use a method that was first used by Lin and Ying (1991). They suggested to split the sample $i = 1, \ldots, N$ randomly into two subsamples of size N_1 and N_2 with $N_1 + N_2 = N$ and $N_1 = O(N), N_2 = O(N)$. Sample 1 is used to obtain consistent, but not necessarily efficient, estimators of α, β which we denote by $\hat{\beta}_{N_1}, \hat{\alpha}_{N_1}$ and the corresponding transformed durations $U_{1i}(\hat{\theta}_{N_1}), i = 1, \ldots, N_1$. The residuals are used in a nonparametric estimator of the hazard of $U(\theta_0), \hat{\kappa}_{0N_1}$, and this nonparametric estimator and the estimated parameters are substituted in (28) and (29) to obtain the estimated weight function $W_i(u, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})$. The same steps for subsample 2 gives the estimated weight function $W_i(u, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2})$. The estimated weight function $W_i(u, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})$ is used in the estimating equation for subsample 2

$$S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) = \sum_{i=1}^{N_2} \Delta_i \Big\{ W_i\big(\tilde{U}_{2i}(\theta), \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1}\big) - W_h\big(\tilde{U}_{2i}(\theta), \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1}\big) \Big\}. \tag{31}$$

In the same way, the estimated weight function derived from subsample 2 is used in the estimating equation for subsample 1, $S_{1N_1}\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2})\big)$. The efficient LRE estimator makes the combined estimating equation

$$S_N\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2}), W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) = S_{1N_1}\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2})\big) + S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) \tag{32}$$

equal to zero, or because the S_N is a step function, the efficient LRE is defined by

$$\hat{\theta}_N(W) = \arg\min_{\theta \in \Theta} \Big| S_N\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2}), W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) \Big|^2. \tag{33}$$

The advantage of the sample splitting is that the estimated weight function $W_i(u, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})$ does not depend on the transformed durations $U_{2i}(\theta), i = 1, \ldots, N_2$ that enter in $S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big)$. We can think of the parameters $\hat{\theta}_{N_1}$ and the estimated transformed durations $U_{1i}(\hat{\theta}_{N_1}), i = 1, \ldots, N_1$ as determined at time 0 in the analysis of

$S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big)$, and the usual operations can be performed to derive e.g. its variance (conditional on $\hat{\theta}_{N_1}$) and the estimated transformed durations $U_{1i}(\hat{\theta}_{N_1}), i = 1, \ldots, N_1$. The linearization lemma applies to random, but predictable weight functions that converge uniformly to a nonstochastic function. To prove uniform convergence of the weight function, we must establish the uniform convergence of the nonparametric estimator of κ_0 based on the estimated transformed durations (see Lemmas 2 and 3). We need to know the uniform rate of convergence because we need to modify the nonparametric hazard estimator to avoid a zero denominator in the weight function.

The nonparametric hazard estimator is the kernel estimator of Ramlau-Hansen (1983). If we were to observe the possibly censored transformed durations $\tilde{U}_i(\theta_0), i = 1, \ldots, N$ the kernel estimator is

$$\hat{\kappa}_N(u, \theta_0) = \frac{1}{b_N} \sum_{i=1}^{N} \Delta_i \frac{I\big(Y_N^U(\tilde{U}_i(\theta_0), \theta_0) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\theta_0), \theta_0)} K\left(\frac{u - \tilde{U}_i(\theta_0)}{b_N}\right) \tag{34}$$

with $Y_N^U(u, \theta_0) = \sum_{i=1}^{N} Y_i^U(u, \theta_0)$ and $Y_{h,N}^U(u, \theta_0) = Y_N^U(u, \theta_0)/N$.

The properties of the kernel hazard estimator have been studied by Ramlau-Hansen (1983) and Andersen et al. (1993). In particular, Theorem IV.2.2. of Andersen et al. (1993) gives a sufficient condition for uniform convergence. Inspection of their proof shows that the same method gives Lemma 2 that we present in the appendix.

Also, note that Lin and Ying (1995) propose an alternative inference procedure that avoids numerical derivatives. However, Lin and Ying (1995) do not allow for a weighting matrix. For that case, we recommend the bootstrap. However, if the bootstrap is computationally infeasible than the procedure by Lin and Ying (1995) may be worthwhile. Finally, our GMM estimator can be extended to the case in which some of the covariates are endogenous using techniques similar to what Bijwaard (2009) uses for the AFT model.

5. Practical implementation

To obtain the LRE we use an iterative procedure of two methods for finding the roots of a non–differentiable (multidimensional) function. We iterate between the Powell method and the Nelder-Mead method. The Powell method (see Press et al.(1986, §10.5) and Powell (1964)) minimizes the function by a bi-directional search along each search vector, in turn. The new position can then be expressed as a linear combination of the search

vectors. The new displacement vector becomes a new search vector, and is added to the end of the search vector list. Meanwhile the search vector which contributed most to the new direction, i.e. the one which was most successful, is deleted from the search vector list. The algorithm iterates an arbitrary number of times until no significant improvement is made. The basic algorithm is simple, the complexity is in the linear searches along the search vectors, which can be achieved via Brent's method.[2]

A nice feature of our estimation procedure is that it provides a convergence test because the solution of the estimation equations implies that a small change of the value of any element of the parameter leads to a sign change in the S-statistic. One iteration of the Powell-method often does not converge to the solution. Then, starting from the last obtained parameter estimate, a search using the Nelder-Mead method is used. We iterate between these two methods till convergence. Still, a solution is not always found. Then, 10 iterations of a pseudo Newton Raphson method, based on the pseudo derivative, provides a new starting point for the iterative procedure. A STATA (STATA 11.0) procedure, LRE, is available on our website.[3]

6. Monte Carlo experiments

In this section we show that estimating a hazard regression with NPMLE can lead to biased inference if we allow for duration dependence and unobserved heterogeneity when they are not present in the DGP. The LRE does not suffer from this misspecification.

6.1. Sample design. We try to resemble the simulation experiments by Baker and Melino (2000) who choose true hazards that match those typically observed in unemployment duration data. They assume a discrete time duration model, while we consider a continuous time model. First we consider the very simple exponential model without unobserved heterogeneity (and no duration dependence) and one explanatory variable, that

[2] The Brent's method combines the bisection method, the secant method and inverse quadratic interpolation. The idea is to use the secant method or inverse quadratic interpolation if possible, because they converge faster, but to fall back to the more robust bisection method if necessary. The secant method can be thought of as a finite difference approximation of the Newton-Raphson method. The Powell method extends the Brent method by searching in a specific direction, rather than changing one parameter at the time.

[3] See http://publ.nidi.nl/output/other/LRE.zip for the program and http://publ.nidi.nl/output/other/LRE_help.pdf for the help file.

is

$$\lambda(t|X_i) = \exp(X_i\beta + \beta_0), \tag{35}$$

where X is normally distributed with mean zero and variance 0.5. The true value of the regression parameter, β, is 1. The true value of the intercept, β_0, is $\ln(0.05)$. The variance of X and the regression parameter determine the relative importance of the unobserved heterogeneity; they determine how accurate we can estimate β and whether we can distinguish duration dependence from unobserved heterogeneity. We choose the variance of X such that the R^2 from a regression of the log duration on X is 0.13, close to values typically observed in practice. This implies that the average duration is 22.5, say weeks. In practice the durations are often censored, that is only observed up to a certain time. We choose a moderate censoring scheme that censors all durations lasting more than 40 (weeks). This implies a censoring rate of 16%. We generated 100 random samples of size 5000 for this DGP and stored it.

We are interested in the effect of wrongly assuming duration dependence and/or unobserved heterogeneity. We therefore consider estimating a flexible duration dependence despite the fact that the DGP has no duration dependence. In the estimation we assume three alternative specifications for the duration dependence: none, a piecewise constant duration dependence on four intervals and a piecewise constant duration dependence on 10 intervals. This implies the following baseline hazard

$$\lambda_0(t) = \sum_{k=1}^{K} e^{\alpha_k} I_k(t) \tag{36}$$

with $K = 4$ or 10 and $I_k(t) = I(t_{k-1} \leq t < t_k)$, which is one if the duration falls between t_{k-1} and t_k. For the 4 interval piecewise constant duration dependence, we choose $t_0 = 0$, $t_1 = 5$, $t_2 = 10$, $t_3 = 20$ and $t_4 = \infty$, such that each interval contains about a quarter of the durations. For the 10 interval piecewise constant duration dependence, we have $t_0 = 0$, $t_1 = 2$, $t_2 = 4$, $t_3 = 6$, $t_4 = 10$, $t_5 = 13$, $t_6 = 16$, $t_7 = 20$, $t_8 = 25$, $t_9 = 30$ and $t_{10} = \infty$, such that each interval contains about 10% of the durations. The parameter of the first interval, α_1, is fixed to zero. The remaining α's now reflect the proportional shift in the baseline hazard in each interval compared to the first, base, interval. This facilitates the comparison between the MLE results and the LRE results.

The effect of wrongly assuming unobserved heterogeneity is investigated by estimating an MPH model with discrete unobserved heterogeneity using a maximum likelihood procedure. In one approach, we assume a fixed number of two support points for the distribution of the unobserved heterogeneity, (MLE two points)[4]. The other approach estimates the NPMLE of Heckman and Singer (1984b) where the number of support points is determined by the Gateaux derivative[5]. Note that multiplicative unobserved heterogeneity does not influence the LRE procedure.

For the LRE, we use the most simple weight functions, X_i for β and the interval indicator on the transformed time scale, $I_k(u) = I\big(m_{k-1}(X,t) \leq u < m_{k-1}(X,t)\big)$ with $m_k(X,t) = e^{\beta X}\int_0^{t_k}\lambda(s)\,ds$, for α_k. These weight functions might be inefficient but it simplifies the estimation. In Section 6.3 we elaborate on estimating efficient LRE in just one additional step. To obtain the LRE, we need to solve the minimizer of the quadratic form of the estimation equations in (18). However the statistic $S_n(\theta;W)$ is a multi–dimensional step–function and the standard Newton–Raphson algorithm cannot be used to solve this.

We also investigate the effect of sample size on our estimations. We consider three values for the number of observations in the sample: 500, 1000 and 5000. The experiments involving a sample size of 500 are constructed using the first 500 observations of the 5000 observations generated by the true DGP. For the experiments involving a sample size of 1000, we add to the observations in the experiments the next 500 generated observations.

For each of the alternative duration dependences and each sample size, we apply four different estimation procedures: MLE of MPH without unobserved heterogeneity (PH-model), MLE two points, NPMLE and LRE. Thus in total we have 36 experiments in our sample design constructed from 1 DGP, 3 specifications for the duration dependence, 3 sample sizes and 4 different estimation techniques.

[4] In the MLE for models with duration dependence, we do not need the standard identification restriction that the unobserved heterogeneity term has mean one because the baseline hazard is normalized to be equal to one in the first interval.

[5] The Gâteaux derivative is a directional derivative; let $x \in \mathbb{R}^K$ and $f(x) \in \mathbb{R}$, $\eta \in \mathbb{R}$, and $\eta > 0$ then $df(x,a) = \lim_{\eta \downarrow 0}[\{f(x+a\eta) - f(x)\}/\eta]$.

6.2. Monte Carlo Results. In Table 1 we report the average bias and standard deviation of the average for the estimates of β in the 36 experimental settings.[6] For each of the 3 sample sizes, we took the 100 simulated samples and estimated β using each of the three alternative duration dependence specifications and the four different estimation procedures.[7]

The results indicate that assuming a discrete unobserved heterogeneity distribution when it is absent leads to well behaved estimates when it is known that there is no duration dependence. The LRE is also unbiased and the efficiency of the LRE is close to the MLE.

Assuming duration dependence when it is absent also leads to well behaved estimators of β when it is known that there is no unobserved heterogeneity. However, the combination of a flexible duration dependence and the distribution of the unobserved heterogeneity leads to a systematic positive bias for the maximum likelihood estimates of β that declines very slowly with sample size. This is in line with the results from Baker and Melino (2000). The LRE continues to provide unbiased estimates of β despite assuming duration dependence that is not present.

If β is not estimated well, this is reflected in the estimates of the parameters of the duration dependence (see Table A.1 and Table A.2 in Appendix A). Assuming unobserved heterogeneity when it is absent leads to a positive duration dependence that declines very slowly with the sample size. Baker and Melino (2000) also find that an overestimation of β is accompanied by a positive bias in the estimated duration dependence. Note that the MLE of the model without unobserved heterogeneity also leads to a bias in the estimated duration dependence in small samples. The LRE estimates the nonexistent duration dependence well, although at the expense of efficiency loss.

6.3. Duration dependence and efficiency. Two remaining interesting issues are estimating duration dependence that is truly present and the efficiency of the (optimal) LRE. If unobserved heterogeneity is present, the optimal LRE should be more efficient than the first stage LRE (see example 3). To this end we simulate four different random samples from a gamma-mixture with different types of duration dependence. We assume

[6] Our calculations were done in Gauss 6.0 on 3 parallel computers: a Pentium 2.1 PC, a Pentium 2.8 PC and a Pentium 2.0 laptop. The calculations took about 9 weeks of CPU time.

[7] The LRE with a duration dependence on 10 intervals for a sample size of 500 did not converge in 7 of the experiments. The average is therefore base on 93 experiments instead of 100.

Table 1: Average bias of estimates of β across the experiments

Duration dependence	estimation method	Sample size		
		500	1000	5000
No duration dependence	MLE no hetero	0.0017	0.0051	-0.0010
		(0.0115)	(0.0080)	(0.0035)
	MLE 2 points	0.0198	0.0247*	0.0038
		(0.0122)	(0.0086)	(0.0040)
	NPMLE	0.0191	0.0165*	0.0046
		(0.0118)	(0.0082)	(0.0037)
	LRE	0.0028	0.0045	-0.0008
		(0.0122)	(0.0084)	(0.0038)
4 piecewise constant	MLE no hetero	0.0022	0.0048	-0.0022
		(0.0115)	(0.0082)	(0.0036)
	MLE 2 points	0.0599*	0.0531*	0.0144*
		(0.0153)	(0.0120)	(0.0044)
	NPMLE	0.1142*	0.0765*	0.0241*
		(0.0160)	(0.0116)	(0.0045)
	LRE	0.0286	0.0179	-0.0041
		(0.0172)	(0.0128)	(0.0057)
10 piecewise constant	MLE no hetero	0.0005	0.0038	-0.0022
		(0.0116)	(0.0082)	(0.0036)
	MLE 2 points	0.0734*	0.0571*	0.0273*
		(0.0162)	(0.0127)	(0.0052)
	NPMLE	0.2376*	0.1519*	0.0592*
		(0.0247)	(0.0162)	(0.0067)
	LRE	-0.0161	-0.0124	-0.0040
		(0.0247)	(0.0192)	(0.0092)

*$p < 0.05$

Based on 93 experiments, because in 7 experiments the estimation procedure did not convergence

a piecewise constant baseline hazard on 3 intervals, 0–5, 5–20 and 20 and over, with $\lambda_0(t) = \sum_{k=1}^{3} e^{\alpha_k} I_k(t)$ and $\alpha_1 = 0$ with the following four types of duration dependence:

1 Positive duration dependence: $\alpha_2 = 0.2$ and $\alpha_3 = 0.5$;

2 Negative duration dependence: $\alpha_2 = -0.2$ and $\alpha_3 = -0.4$;

3 U-shaped duration dependence: $\alpha_2 = -0.2$ and $\alpha_3 = 0.2$;

4 Inverse U-shaped duration dependence: $\alpha_2 = 0.2$ and $\alpha_3 = -0.2$.

Again we assume that we have only one explanatory variable X that is normally distributed with mean zero and variance 0.5. The true value of the regression parameter, β, is 1. The variance of the gamma mixture is 0.75. For each DGP, we create 100 samples of 1000 observations and store them. We estimate the regression parameter and the parameters of the duration dependence by the following six alternative methods (*i*) MLE for a gamma-mixture (the true model); (*ii*) MLE no unobserved heterogeneity; (*iii*) MLE with discrete

unobserved heterogeneity and two points of support; (*iv*) NPMLE where the number of support points is determined by the Gateaux derivative; (*v*) LRE and (*vi*) Optimal LRE. We estimate the parameters using both the uncensored sample and a sample in which the durations are artificially censored at 30. This implies a censoring rate of around 15%.

For the first stage LRE we use, again, the weight functions, X_i for β and the interval indicator on the transformed time scale, $I_k(u)$ for α_k. For calculating the optimal LRE, we need to know the distribution of U_0 because the optimal weighting function depends on the distribution of U_0 through its hazard and the derivative of that hazard (see (28) and (29)). We use the method with an estimated weight function described in Section 4 to obtain the efficient optimal LRE. First we randomly split each sample into two subsamples. Then, for each subsample, we estimate the parameters and the corresponding transformed durations using LRE. Based on the transformed durations of the first subsample, we estimate the weights in the second subsample and vice versa. We use the kernel estimator of Ramlau-Hansen (1983) to obtain these functionals. The efficient LRE is now obtained from the combined estimating equation (32) and equal is given in (33), see Section 4.

Table 2: Average bias, standard error and RMSE of estimates of β across the experiments

Duration dependence	estimation method	bias	std error	RMSE
positive duration dependence	MLE gamma	-0.0074	0.0222	0.0234
	MLE no hetero	-0.3884^*	0.0232	0.3889
	MLE 2 points	-0.2656^*	0.0202	0.2664
	NPMLE	-0.0036	0.0216	0.0219
	LRE	-0.0264	0.0245	0.0360
	LRE-opt	-0.0205	0.0238	0.0314
negative duration dependence	MLE gamma	0.0331	0.0206	0.0390
	MLE no hetero	-0.3963^*	0.0270	0.3970
	MLE 2 points	-0.2797^*	0.0242	0.2808
	NPMLE	0.0382	0.0230	0.0446
	LRE	0.0341	0.0238	0.0416
	LRE-opt	0.0296	0.0231	0.0375
U-shaped duration dependence	MLE gamma	-0.0208	0.0192	0.0283
	MLE no hetero	-0.3707^*	0.0299	0.3711
	MLE 2 points	-0.2895^*	0.0170	0.2900
	NPMLE	-0.0088	0.0203	0.0221
	LRE	-0.0138	0.0231	0.0269
	LRE-opt	-0.0124	0.0206	0.0240
inverse U duration dependence	MLE gamma	0.0248	0.0184	0.0309
	MLE no hetero	-0.3798^*	0.0165	0.3806
	MLE 2 points	-0.2743^*	0.0174	0.2748
	NPMLE	0.0341	0.0191	0.0391
	LRE	0.0190	0.0205	0.0280
	LRE-opt	0.0195	0.0202	0.0281

$^*p < 0.05$. For each DGP (gamma mixture) 100 simulations with 1000 observations each.

In Table 2, we report the average bias, the standard deviation of the average bias and the RMSE for the estimates of β in the four experimental settings. Table 3 gives the results for the censored sample.[8] The results indicate that ignoring the unobserved heterogeneity leads to a severe bias. Using a two point discrete unobserved heterogeneity distribution to approximate the true gamma heterogeneity distribution still leads to biased estimation results. The MLE based on the true gamma mixture DGP is, not surprisingly, the most efficient estimation procedure.

For two of the four DGP's the RMSE of the NPMLE is higher than the RMSE of the LRE. In particular, for both the negative and the inverse U-shaped duration dependence, the NPMLE is biased if the sample is censored. The optimal LRE is 5% to 25% (uncensored U-shaped duration dependence) more efficient than the LRE.

Table 3: Average bias, standard error and RMSE of estimates of β across the experiments, **censored sample**

Duration dependence	estimation method	bias	std error	RMSE
positive duration dependence	MLE gamma	-0.0098	0.0228	0.0248
	MLE no hetero	-0.3420^*	0.0158	0.3424
	MLE 2 points	-0.1204^*	0.0236	0.1227
	NPMLE	0.0048	0.0238	0.0243
	LRE	-0.0277	0.0249	0.0372
	LRE-opt	-0.0253	0.0247	0.0353
negative duration dependence	MLE gamma	0.0398	0.0213	0.0451
	MLE no hetero	-0.3164^*	0.0151	0.3668
	MLE 2 points	-0.0527^*	0.0241	0.0579
	NPMLE	0.0550^*	0.0228	0.0595
	LRE	0.0419	0.0231	0.0478
	LRE-opt	0.0406	0.0229	0.0466
U-shaped duration dependence	MLE gamma	-0.0171	0.0194	0.0259
	MLE no hetero	-0.3289^*	0.0144	0.3292
	MLE 2 points	-0.1346^*	0.0226	0.1365
	NPMLE	-0.0094	0.0203	0.0224
	LRE	-0.0330	0.0198	0.0385
	LRE-opt	-0.0298	0.0196	0.0356
inverse U duration dependence	MLE gamma	0.0265	0.0185	0.0323
	MLE no hetero	-0.3311^*	0.0126	0.3321
	MLE 2 points	-0.0632^*	0.0203	0.0664
	NPMLE	0.0395^*	0.0193	0.0440
	LRE	0.0297	0.0194	0.0355
	LRE-opt	0.0263	0.0191	0.0325

For each DGP 100 (gamma mixture) simulations with 1000 observations each. 10-18% censored.
$^*p < 0.05$

[8] The results for the parameters of the piecewise constant duration dependence, α_2 and α_3, are given in Table A.3 and Table A.4 in Appendix A.

6.4. Time-varying covariates. One advantage of the LRE is that it can handle time-varying covariates. In the next Monte Carlo study, we show that LRE performs rather well when regressors vary with time. Note that a hazard model is a very natural way to model time-varying regressors. To this end we simulate random samples from a gamma-mixture with positive duration dependence that includes a time-varying covariate. We assume a piecewise constant baseline hazard on three intervals: 0–5, 5–20 and 20 and over, with $\lambda_0(t) = \sum_{k=1}^{3} e^{\alpha_k} I_k(t)$, $\alpha_1 = 0$, $\alpha_2 = 0.2$ and $\alpha_3 = 0.5$.
Now we assume that we have two explanatory variables, a time-constant variable X_0 that is normally distributed with mean zero and variance 0.5 with a true regression parameter β_0 of 0.6 and a time-varying variable $X_1(t)$ that is also normally distributed with mean zero and variance 0.5, but it changes value after $t = 5$ and $t = 20$. These changes are exogenous to the process. The true regression parameter of the time-varying covariate is 0.4. The variance of the gamma mixture is 0.75. We create 500 samples of 1000 observations and store them. We estimate the regression parameter and the parameters of the duration dependence using the LRE and the Optimal LRE, both on the uncensored sample and a sample in which the durations are artificially censored at 30. This implies a censoring rate of around 42%.

For the first stage LRE, we use the weight functions, X_{0i} and $X_{1i}(u)$ for β_0 and β_1 and the interval indicator on the transformed time scale, $I_k(u)$ for α_k. For calculating the optimal LRE, we need to know the distribution of U_0, because the optimal weighting function depends on the distribution of U_0 through its hazard and the derivative of that hazard. We use a Ramlau-Hansen kernel method to obtain these functionals necessary to estimate the weight function of the efficient optimal LRE.

In Table 4 we report the average bias, the standard deviation of the average bias and the RMSE for the estimates of the two regression parameters β_0 (time-constant covariate) and β_1 (time-varying covariate) and the parameters of the baseline hazard. The results show that the LRE and optimal LRE give consistent estimates of all parameters. The optimal LRE is slightly more efficient, although the efficiency gain is rather small.

Table 4: Average bias, standard error and RMSE of estimates of the parameters for a model with time-varying covariates

parameter	estimation method	bias	std error	RMSE
	Uncensored			
β_0	LRE	0.0014	0.0049	0.0051
	LRE-opt	0.0011	0.0048	0.0050
β_1	LRE	-0.0011	0.0035	0.0037
	LRE-opt	-0.0005	0.0034	0.0035
α_2	LRE	0.0044	0.0090	0.0100
	LRE-opt	0.0016	0.0088	0.0090
α_3	LRE	0.0038	0.0128	0.0133
	LRE-opt	-0.0022	0.0127	0.0130
	Censored			
β_0	LRE	-0.0034	0.0049	0.0060
	LRE-opt	-0.0040	0.0048	0.0059
β_1	LRE	-0.0017	0.0033	0.0037
	LRE-opt	-0.0012	0.0032	0.0035
α_2	LRE	-0.0049	0.0091	0.0104
	LRE-opt	-0.0055	0.0089	0.0104
α_3	LRE	-0.0151	0.0131	0.0199
	LRE-opt	-0.0142	0.0130	0.0193

500 simulations with 1000 observations each.

7. Empirical Application

Much of the theoretical and empirical literature on the economics of migration views migrations as permanent. This is a convenient assumption and often facilitates the analysis of immigrant behavior and the impact of migration on the host country. The life–cycle theories imply that assimilation in the host country and migration decisions are correlated over time. It is therefore more appropriate to base the analysis of migration on a dynamic model that takes the timing of migration moves into account. The literature on the timing of out-migration is rather scarce. Bijwaard (2010) shows that recent migrants to the Netherlands leave rather fast. After 5 years about 40% of the labour migrants have left the country. We use a subset of this data by choosing a particular group of migrants. However, the data now includes information on labour market status and income. The observation window is also extended with two more years.

We have data on recent immigrants to the Netherlands (1999-2007). All immigration by non-Dutch citizens, immigrants who do not hold the Dutch nationality, who legally entered The Netherlands is registered in the Central Register Foreigners (Centraal Register Vreemdelingen, CRV), using information from the Immigration Police (Vreemdelingen Politie) and the Immigration and Naturalization Service (Immigratie- en Naturalisatie Dienst, IND). For all these immigrants without the Dutch nationality we know when their migration move(s) took place and what their migration motive was to enter the Netherlands. For people with a nationality that implies a visa to enter The Netherlands, their migration motive can be directly derived from their legal entry status. People with other, Western nationalities, fill in their migration motive at their mandatory registration at their municipality of residence. The data further contain information on the timing of migration moves, both on the timing of immigration and on the timing of (return) emigration. This enables us to construct the duration till out-migration (or the end of the observation window).

The CBS, Statistics Netherlands, has linked these data to the Municipal Register of Population (Gemeentelijke Basisadministratie, GBA) and to their Social Statistical database (SSB). The GBA data contain basic demographic characteristics of the migrants, such as age, gender, marital status and country of origin. From the SSB we have information (on a monthly basis) on the labour market position and income. The most important income source determines the labour market position.

For this article we use a subsample of the data used in Bijwaard et al. (2012). Bijwaard et al. (2012) only used data labor migrants aged 18 to 64 at entry. We further restrict the data to migrants born in EU-countries who entered the Netherlands in 1999 and who

have a monthly income above 1000 per month at entry. We end up with 4418 individual migrants, 3109 males and 1309 females, with 62,699 records (m 42,548; f 20,151), an average of 14.2 observations per migrant (m 13.7; f 15.4), due to changes in time-varying information. From these migrants 3080 (m 2256; f 824), around 70% (m 73%; f 63%), have left the Netherlands before the end of the observation period (December 31, 2007).

Table 5 and 6 provide the estimates of the out-migration intensity of these labor migrants. Self-employed migrants have higher investments and are therefore less prone to leave the Netherlands. Marriage has only an impact on the behavior of female migrants with married migrants remaining longer in the country. The income of the migrant plays an important role in explaining the out-migration. It has a U-shaped effect, as both low income and high income lead to faster out-migration.

As usual the inclusion of unobserved heterogeneity in the intensity (compare the PH to NPMLE) leads to more pronounced positive duration dependence and regression parameters further away from zero. The main difference between the NPMLE (and PH) estimates and the LRE is a change in the duration dependence. The LRE and optimal-LRE show a much less pronounced duration dependence and this duration dependence is insignificant for females. This change in duration dependence is also the main difference with the results in Bijwaard et al. (2012) using a more extensive sample of all recent labor migrants to the Netherlands. The regression parameters also change when using the LRE procedures, but only a little. The optimal LRE is only slightly more efficient than the LRE.

Table 5: Estimates of out-migration intensity, males

	PH	NPMLE	LRE	optimal LRE
Self-employed	-1.438^{**}	-1.595^{**}	-1.574^{**}	-1.582^{**}
	(0.273)	(0.286)	(0.264)	(0.260)
Inactive	0.120	0.117	0.174	0.197^{+}
	(0.109)	(0.101)	(0.097)	(0.091)
Married	-0.050	-0.070	-0.059	-0.061
	(0.053)	(0.060)	(0.051)	(0.050)
Divorced	-0.378^{**}	-0.424^{+}	-0.429^{**}	-0.426^{**}
	(0.135)	(0.165)	(0.143)	(0.142)
Age	-0.010	-0.009	-0.012	0.009
	(0.045)	(0.051)	(0.046)	(0.045)
Age-squared	0.004	0.001	0.005	-0.002
	(0.024)	(0.029)	(0.017)	(0.016)
Income <1000	1.598^{**}	1.811^{**}	1.701^{**}	1.625^{**}
	(0.122)	(0.124)	(0.115)	(0.112)
Income 1000-2000	-0.021	0.049	-0.007	-0.032
	(0.101)	(0.106)	(0.098)	(0.093)
Income 3000-4000	0.333^{**}	0.341^{**}	0.350^{**}	0.331^{**}
	(0.099)	(0.104)	(0.098)	(0.095)
Income 4000-5000	0.229	0.264^{+}	0.254^{+}	0.230^{+}
	(0.124)	(0.129)	(0.120)	(0.118)
Income 5000-6000	0.504^{**}	0.554^{**}	0.542^{**}	0.511^{**}
	(0.135)	(0.142)	(0.131)	(0.128)
Income >6000	0.549^{**}	0.586^{**}	0.602^{**}	0.541^{**}
	(0.094)	(0.106)	(0.095)	(0.092)
duration dependence				
α_2	1.299^{**}	1.314^{**}	0.872^{**}	0.858^{**}
	(0.202)	(0.203)	(0.262)	(0.257)
α_3	1.917^{**}	2.098^{**}	1.511^{**}	1.477^{**}
	(0.189)	(0.196)	(0.299)	(0.289)
α_4	1.843^{**}	2.285^{**}	1.661^{**}	1.605^{**}
	(0.187)	(0.216)	(0.314)	(0.303)
α_5	1.492^{**}	2.074^{**}	1.423^{**}	1.336^{**}
	(0.189)	(0.228)	(0.327)	(0.311)

Notes: SE in parentheses. $^{+}: p < 0.05$ and $^{**}: p < 0.01$.

Table 6: Estimates of out-migration intensity, females

	PH	NPMLE	LRE	optimal LRE
Self-employed	-1.572^{**}	-1.929^{**}	-1.684^{**}	-1.647^{**}
	(0.558)	(0.633)	(0.653)	(0.652)
Inactive	0.667^{**}	0.706^{**}	0.776^{**}	0.772^{**}
	(0.162)	(0.167)	(0.173)	(0.039)
Married	-0.562^{**}	-0.650^{**}	-0.645^{**}	-0.644^{**}
	(0.100)	(0.119)	(0.121)	(0.120)
Divorced	-0.536^{+}	-0.644^{+}	-0.639^{+}	-0.642^{+}
	(0.223)	(0.287)	(0.282)	(0.281)
Age	0.050	0.048	0.043	0.013
	(0.075)	(0.082)	(0.086)	(0.083)
Age-squared	0.039	0.054	0.055	0.058
	(0.044)	(0.060)	(0.060)	(0.058)
Income <1000	1.051^{**}	1.275^{**}	1.173^{**}	1.149^{**}
	(0.179)	(0.193)	(0.189)	(0.188)
Income 1000-2000	-0.154	-0.133	-0.138	-0.158
	(0.139)	(0.146)	(0.144)	(0.140)
Income 3000-4000	0.119	0.172	0.151	0.157
	(0.157)	(0.162)	(0.159)	(0.158)
Income 4000-5000	0.443^{+}	0.534^{+}	0.511^{+}	0.500^{+}
	(0.201)	(0.209)	(0.211)	(0.209)
Income 5000-6000	0.751^{**}	0.851^{**}	0.854^{**}	0.897^{**}
	(0.252)	(0.247)	(0.243)	(0.242)
Income >6000	1.027^{**}	1.151^{**}	1.176^{**}	1.150^{**}
	(0.184)	(0.198)	(0.208)	(0.200)
duration dependence				
α_2	1.113^{**}	1.134^{**}	0.456	0.437
	(0.302)	(0.304)	(0.481)	(0.471)
α_3	1.517^{**}	1.661^{**}	0.986	0.972
	(0.280)	(0.287)	(0.520)	(0.503)
α_4	1.463^{**}	1.838^{**}	1.192^{+}	1.163^{+}
	(0.275)	(0.305)	(0.552)	(0.512)
α_5	1.144^{**}	1.686^{**}	1.048	1.010
	(0.279)	(0.329)	(0.571)	(0.553)

Notes: SE in parentheses. $^{+}: p < 0.05$ and $^{**}: p < 0.01$.

8. Conclusion

In this paper, we have discussed and implemented a simple $\sqrt{N}$ consistent estimator for the parameters of a semiparametric MPH model with an unspecified distribution of the unobserved heterogeneity. This Linear Rank Estimator (LRE) is a GMM estimator that uses moment conditions to derive estimating equations. It is based on the linear rank statistic. We have derived the asymptotic properties of the LRE and of the two-stage optimal LRE.

We presented Monte Carlo evidence that the LRE performs well in samples of moderate size. In contrast to the commonly applied Nonparametric MLE of Heckman and Singer (1984b), the LRE provides asymptotically unbiased estimates of the regression coefficients despite allowing for nonexistent duration dependence. Moreover, we derive the asymptotic distribution of the LRE estimators (so that we can derive confidence intervals) while the rate of convergence and the asymptotic distribution of the Nonparametric MLE are unknown.

A. Appendix: Additional tables

Table A.1: Average bias of estimates of the log α's across the experiments with a piecewise constant duration dependence on 4 intervals

Estimation method		Sample size		
		500	1000	5000
MLE no hetero	α_2	-0.0480^*	-0.0319^*	-0.0095^*
		(0.0150)	(0.0103)	(0.0042)
	α_3	-0.0082	-0.0127	-0.0094^*
		(0.0132)	(0.0088)	(0.0041)
	α_4	-0.0149	-0.0102	-0.0079
		(0.0127)	(0.0089)	(0.0046)
MLE 2 points	α_2	0.0282	0.0257	0.0140^*
		(0.0194)	(0.0158)	(0.0053)
	α_3	0.1131^*	0.0713^*	0.0257^*
		(0.0237)	(0.0175)	(0.0064)
	α_4	0.1480^*	0.1013^*	0.0438^*
		(0.0273)	(0.0213)	(0.0076)
NPMLE	α_2	0.0785^*	0.0495^*	0.0211^*
		(0.0210)	(0.0152)	(0.0050)
	α_3	0.2011^*	0.1207^*	0.0389^*
		(0.0275)	(0.0183)	(0.0059)
	α_4	0.2835^*	0.1782^*	0.0612^*
		(0.0339)	(0.0228)	(0.0079)
LRE	α_2	-0.0333	-0.0234	-0.0074
		(0.0230)	(0.0184)	(0.0066)
	α_3	0.0391	0.0158	-0.0087
		(0.0306)	(0.0224)	(0.0093)
	α_4	0.0536	0.0264	-0.0109
		(0.0383)	(0.0287)	(0.0128)

$^*p < 0.05$

Table A.2: Average bias of estimates of the log α's across the experiments with a piecewise constant duration dependence on 10 intervals

	Sample size			Sample size		
	500	1000	5000	500	1000	5000
	MLE no hetero			MLE 2 points		
α_2	−0.0240	−0.0098	0.0068	0.0704*	0.0498*	0.0464*
	(0.0216)	(0.0153)	(0.0063)	(0.0230)	(0.0176)	(0.0080)
α_3	−0.0162	−0.0089	−0.0090	0.1096*	0.0740*	0.0420*
	(0.0241)	(0.0157)	(0.0061)	(0.0283)	(0.0195)	(0.0086)
α_4	−0.0609*	−0.0378*	−0.0069	0.0958*	0.0627*	0.0590*
	(0.0207)	(0.0135)	(0.0054)	(0.0273)	(0.0204)	(0.0098)
α_5	0.0073	−0.0035	−0.0115	0.1991*	0.1229*	0.0690*
	(0.0206)	(0.0144)	(0.0069)	(0.0305)	(0.0231)	(0.0117)
α_6	−0.0097	−0.0024	−0.0059	0.1986*	0.1348*	0.0766*
	(0.0207)	(0.0127)	(0.0067)	(0.0340)	(0.0226)	(0.0123)
α_7	−0.0593*	−0.0464*	−0.0074	0.1617*	0.0971*	0.0823*
	(0.0226)	(0.0154)	(0.0072)	(0.0364)	(0.0269)	(0.0135)
α_8	−0.0144	−0.0130	−0.0023	0.2161*	0.1491*	0.0963*
	(0.0204)	(0.0151)	(0.0070)	(0.0360)	(0.0277)	(0.0141)
α_9	−0.0209	−0.0076	−0.0120	0.2309*	0.1616*	0.0964*
	(0.0243)	(0.0149)	(0.0075)	(0.0388)	(0.0284)	(0.0137)
α_{10}	−0.0383	−0.0217	−0.0078	0.2324*	0.1658*	0.1068*
	(0.0206)	(0.0153)	(0.0071)	(0.0379)	(0.0287)	(0.0154)
	NPMLE			LRE		
α_2	0.1790*	0.1157*	0.0703*	−0.0648*	−0.0460*	0.0088
	(0.0267)	(0.0184)	(0.0088)	(0.0298)	(0.0221)	(0.0106)
α_3	0.3039*	0.1880*	0.0871*	−0.0784	−0.0664*	−0.0070
	(0.0397)	(0.0239)	(0.0099)	(0.0446)	(0.0315)	(0.0136)
α_4	0.3730*	0.2298*	0.1181*	−0.1236*	−0.0942*	−0.0041
	(0.0466)	(0.0298)	(0.0120)	(0.0514)	(0.0387)	(0.0166)
α_5	0.5390*	0.3248*	0.1372*	−0.0554	−0.0605	−0.0093
	(0.0554)	(0.0343)	(0.0146)	(0.0599)	(0.0443)	(0.0203)
α_6	0.5848*	0.3649*	0.1573*	−0.0716	−0.0617	−0.0050
	(0.0583)	(0.0383)	(0.0151)	(0.0646)	(0.0496)	(0.0220)
α_7	0.5910*	0.3554*	0.1692*	−0.1230	−0.1079*	−0.0078
	(0.0646)	(0.0413)	(0.0170)	(0.0698)	(0.0530)	(0.0245)
α_8	0.6916*	0.4232*	0.1884*	−0.0844	−0.0792	−0.0042
	(0.0678)	(0.0429)	(0.0179)	(0.0782)	(0.0570)	(0.0258)
α_9	0.7346*	0.4594*	0.1918*	−0.0921	−0.0819	−0.0157
	(0.0734)	(0.0441)	(0.0191)	(0.0782)	(0.0578)	(0.0278)
α_{10}	0.7758*	0.4816*	0.2123*	−0.1230	−0.1038	−0.0117
	(0.0736)	(0.0486)	(0.0209)	(0.0803)	(0.0637)	(0.0309)

For sample size of 500 based on 93 experiments, because in 7 experiments the estimation procedure did not convergence . $^*p < 0.05$

Table A.3: Average bias, standard error and RMSE of estimates of parameters of piecewise constant baseline hazard across the experiments, **Second set of Monte Carlo experiments**

Duration dependence	estimation method		bias	std error	RMSE
positive duration dependence	MLE gamma	α_2	0.0069	0.0096	0.0118
		α_3	−0.0149	0.0206	0.0255
	NPMLE	α_2	0.0205	0.0157	0.0258
		α_3	0.0091	0.0283	0.0298
	LRE	α_2	−0.0130	0.0200	0.0238
		α_3	−0.0645	0.0329	0.0724
	LRE-opt	α_2	−0.0134	0.0195	0.0236
		α_3	−0.0533	0.0327	0.0625
negative duration dependence	MLE gamma	α_2	0.0211	0.0111	0.0239
		α_3	0.0553*	0.0229	0.0598
	NPMLE	α_2	0.0345*	0.0174	0.0386
		α_3	0.1079*	0.0310	0.1123
	LRE	α_2	0.0369*	0.0179	0.0410
		α_3	0.0643*	0.0315	0.0716
	LRE-opt	α_2	0.0358*	0.0178	0.0400
		α_3	0.0627*	0.0314	0.0701
U-shaped duration dependence	MLE gamma	α_2	−0.0009	0.0097	0.0097
		α_3	−0.0338*	0.0173	0.0379
	NPMLE	α_2	0.0385*	0.0155	0.0416
		α_3	0.0149	0.0251	0.0292
	LRE	α_2	0.0334	0.0186	0.0383
		α_3	−0.0215	0.0271	0.0346
	LRE-opt	α_2	0.0261	0.0183	0.0319
		α_3	−0.0247	0.0263	0.0361
inverse U duration dependence	MLE gamma	α_2	0.0102	0.0104	0.0146
		α_3	−0.0047	0.0232	0.0237
	NPMLE	α_2	0.0232	0.0140	0.0271
		α_3	0.0327	0.0295	0.0440
	LRE	α_2	0.0335	0.0183	0.0381
		α_3	0.0400	0.0336	0.0522
	LRE-opt	α_2	0.0321	0.0182	0.0369
		α_3	0.0344	0.0336	0.0481

For each DGP (gamma mixture) 100 simulations with 1000 observations each. $^*p < 0.05$

Table A.4: Average bias, standard error and RMSE of estimates of parameters of piecewise constant baseline hazard across the experiments, **Second set of Monte Carlo experiments, censored sample**

Duration dependence	estimation method		bias	std error	RMSE
positive duration dependence	MLE gamma	α_2	0.0010	0.0135	0.0135
		α_3	−0.0267	0.0269	0.0379
	NPMLE	α_2	0.0120	0.0177	0.0213
		α_3	−0.0204	0.0310	0.0371
	LRE	α_2	−0.0148	0.0199	0.0248
		α_3	−0.0656*	0.0329	0.0734
	LRE-opt	α_2	−0.0138	0.0199	0.0242
		α_3	−0.0599	0.0328	0.0683
negative duration dependence	MLE gamma	α_2	0.0347*	0.0131	0.0371
		α_3	0.0633*	0.0277	0.0691
	NPMLE	α_2	0.0417*	0.0184	0.0456
		α_3	0.0898*	0.0325	0.0956
	LRE	α_2	0.0378*	0.0182	0.0420
		α_3	0.0539	0.0329	0.0631
	LRE-opt	α_2	0.0375*	0.0181	0.0416
		α_3	0.0501	0.0327	0.0598
U-shaped duration dependence	MLE gamma	α_2	0.0052	0.0133	0.0143
		α_3	−0.0269	0.0225	0.0350
	NPMLE	α_2	0.0308	0.0173	0.0353
		α_3	−0.0159	0.0292	0.0333
	LRE	α_2	0.0266	0.0184	0.0323
		α_3	−0.0321	0.0254	0.0410
	LRE-opt	α_2	0.0263	0.0182	0.0320
		α_3	−0.0315	0.0253	0.0404
inverse U duration dependence	MLE gamma	α_2	0.0137	0.0123	0.0184
		α_3	−0.0030	0.0263	0.0264
	NPMLE	α_2	0.0183	0.0149	0.0236
		α_3	0.0283	0.0305	0.0416
	LRE	α_2	0.0340	0.0185	0.0387
		α_3	0.0360	0.0335	0.0491
	LRE-opt	α_2	0.0313	0.0183	0.0363
		α_3	0.0290	0.0333	0.0441

For each DGP (gamma mixture) 100 simulations with 1000 observations each. $^*p < 0.05$

B. Appendix: Proofs and technical details

B.1. Technical details section 2: A Counting process approach. The counting process approach is a very useful framework for analyzing duration data since an indicator can be used to denote whether a transition happened or not. Andersen et al. (1993) have provided an excellent survey of counting processes. Less technical surveys have been given by Moeschberger and Klein (1997), Therneau and Grambsch (2000), and Aalen et al. (2009). The main advantage of this framework is that it allows us to express the duration distribution as a regression model with an error term that is a martingale difference. Regression models with martingale difference errors are the basis for inference in time series models with dependent observations. Hence, it is not surprising that inference is much simplified by using a similar representation in duration models.

To start the discussion, we first introduce some notation. A counting process $\{N(t)|t \geq 0\}$ is a stochastic process describing the number of events in the interval $[0, t]$ as time proceeds. The process contains only jumps of size $+1$. For single duration data, the event can only occur once because the units are observed until the event occurs. Therefore we introduce the observation indicator $Y(t) = I(T \geq t)$ that is equal to one if the unit is under observation at time t and zero after the event has occurred. The counting process is governed by its random intensity process, $Y(t)\kappa(t)$, where $\kappa(t)$ is the hazard in (2). If we consider a small interval $(t - dt, t]$ of length dt, then $Y(t)\kappa(t)$ is the conditional probability that the increment $\mathrm{d}N(t) = N(t) - N(t-)$ jumps in that interval given all that has happened until just before t. By specifying the intensity as the product of this observation indicator and the hazard rate, we effectively limit the number of occurrences of the event to one. It is essential that the observation indicator only depends on events up to time t.

Usually we do not observe T directly. Instead we observe $\tilde{T} = g(T, C)$ with g a known function and C a random vector. The most common example is right censoring, where $g(T, C) = \min(T, C)$. By defining the observation indicator as the product of the indicator $I(t \leq T)$ and, if necessary, an indicator of the observation plan, we capture when a unit is at risk for the event. In the case of right censoring $Y(t) = I(t \leq T)I(t \leq C)$, and in all cases of interest we have $Y(t) = I(t \leq T)I_A(t)$ with A a random set that may depend on random variables. We assume that C and T are conditionally independent given X. The history up to and including t, $Y_h(t)$ is assumed to be a left continuous function of t. The history of the whole process also includes the history of the covariate process, $X_h(t)$, and

V. Thus, we have

$$\Pr\big(\mathrm{d}N(t) = 1|Y_h(t), X_h(t), V\big) = Y(t)\kappa(t|X_h(t), V, \theta). \tag{B.1}$$

The sample paths of the conditioning variables should be up to $t-$, but because these paths are left continuous we can take them up to t. A fundamental result in the theory of counting processes, the Doob-Meyer decomposition[9], allows us to write

$$\mathrm{d}N(t) = Y(t)\kappa(t|X_h(t), V, \theta)\mathrm{d}t + \mathrm{d}M(t), \tag{B.2}$$

where $M(t), t \geq 0$ is a martingale with conditional mean and variance given by

$$\begin{aligned} \mathrm{E}\big(\mathrm{d}M(t)|V, Y_h(t), X_h(t)\big) &= 0 & \text{(B.3)} \\ \mathrm{Var}\big(\mathrm{d}M(t)|V, Y_h(t), X_h(t)\big) &= Y(t)\kappa(t|X_h(t), V, \theta)\mathrm{d}t. & \text{(B.4)} \end{aligned}$$

The (conditional) mean and variance of the counting process are equal, so the disturbances in (B.2) are heteroscedastic. The probability in (B.1) is zero, if the unit is no longer under observation. A counting process can be considered as a sequence of Bernoulli experiments because if $\mathrm{d}t$ is small, (B.3) and (6) give the mean and variance of a Bernoulli random variable. The relation between the counting process and the sequence of Bernoulli experiments given in (B.2) can be considered as a regression model with an additive error that is a martingale difference. This equation resembles a time-series regression model. The Doob-Meyer decomposition is very helpful to the derivation of the distribution of the estimators because the asymptotic behavior of partial sums of martingales is well-known.

B.2. Technical details section 3: Assumptions 1-4. To simplify the expressions, we use the notation $h_i(t, \theta) = h(t, X_{h,i}(t), \theta)$.

(A1) The conditional distribution of T given $X(\cdot)$ and V has hazard rate

$$\kappa(t|X_h(t), V, \theta) = \lambda(t, \alpha)e^{\beta' X(t)}V \tag{B.5}$$

with $X(\cdot)$ a K covariate bounded stochastic process that is independent of V and such that if the probability of the event $\{c_1' X(t) + c_2 \ln \lambda(t, \alpha_0) = 0, t \in S\}$ some set S with positive measure and for some constants c_1, c_2, then $c_1 = c_2 = 0$. For the baseline hazard, $0 < \lim_{t \downarrow 0} \lambda(t, \alpha_0) < \infty$.

[9] The Doob–Meyer decomposition theorem is a theorem in stochastic calculus stating the conditions under which a submartingale may be decomposed in a unique way as the sum of a martingale and a continuous increasing process, see Meyer (1963) and Protter (2005).

(A2) For the covariate process $X(t), t \geq 0$, we assume that the sample paths are piecewise constant, i.e. its derivative with respect to t is 0 almost everywhere, and left continuous. The hazard that is not conditional on V is

$$\kappa(t|X_h(t),\theta) = \lambda(t,\alpha)e^{\beta' X(t)}\mathrm{E}\big[V|T \geq t, Y_h(t), X_h(t)\big]. \tag{B.6}$$

The observation process is $Y(t), t \geq 0$ with $Y(t) = I(t \leq T)I(t \leq C)$ and we assume

$$\mathrm{d}I(t \leq C) \perp N(s), s \geq t | Y_h(t), X_h(t). \tag{B.7}$$

The support of C is bounded.

(A3) The parameter vector $\theta = (\beta', \alpha')'$ is an M vector with β a K vector and α an L vector. The parameter space Θ is convex. The baseline hazard $\lambda(t,\alpha) > 0$ and is twice differentiable and the second derivative is bounded in α (in the parameter space) and t.

(A4) The weight function $W\big(u, X_h^U(u)\big)$ is an M vector of bounded and left continuous functions. If

$$W_h\Big(\tilde{U}_i(\theta)\Big) = \frac{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big) W\Big(\tilde{U}_i(\theta), X_{h,j}^U\big(\tilde{U}_i(\theta)\big)\Big)}{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big)}, \tag{B.8}$$

then there are functions $\mu(u,\theta)$ (an M vector), $V_\beta(u,s,\theta)$ (an $M \times K$ matrix), and $V_\alpha(u,s,\theta)$ (an $M \times L$ matrix) such that

$$\sup_{\theta\in\Theta, u\leq\tau+\psi} \big|W_h(u,\theta) - \mu(u,\theta)\big| \xrightarrow{p} 0 \tag{B.9}$$

and

$$\sup_{\substack{\theta\in\Theta, u\leq\tau+\psi \\ s\leq\tau+\psi}} \left| \frac{1}{N}\sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta)\big) - W_h(u,\theta)\Big) Y_i^U(u,\theta) X_i^U(s,\theta)' - V_\beta(u,s,\theta)\right| \xrightarrow{p} 0 \tag{B.10}$$

and

$$\sup_{\substack{\theta\in\Theta, u\leq\tau+\psi \\ s\leq\tau+\psi}} \left| \frac{1}{N}\sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta)\big) - W_h(u,\theta)\Big) Y_i^U(u,\theta) \frac{\partial \ln\lambda}{\partial\alpha'}\big(h_i^{-1}(s,\theta) - V_\alpha(u,s,\theta)\right| \xrightarrow{p} 0. \tag{B.11}$$

Define

$$B(\theta_0) = -\int_0^\tau\int_0^u V_\beta(u,s,\theta)\kappa_0'(u)\,ds\,\mathrm{d}u - \int_0^\tau V_\beta(u,u,\theta)\kappa_0(u)\,\mathrm{d}u \tag{B.12}$$

$$A(\theta_0) = -\int_0^\tau\int_0^u V_\alpha(u,s,\theta)\kappa_0'(u)\,ds\,\mathrm{d}u - \int_0^\tau V_\alpha(u,u,\theta)\kappa_0(u)\,\mathrm{d}u. \tag{B.13}$$

We assume that the $M \times M$ matrix $\big[B(\theta_0)A(\theta_0)\big]$ is nonsingular.

The restriction on the baseline hazard in Assumption A1 ensures identification (see Section 3) and guarantees that the semiparametric information bound is nonsingular (see below). Assumption A2 states that the covariates and the observation indicator are predetermined. Assumption A4 is about smoothness: Suppose that one censors all the data at $u = \tau + \psi$, then the expressions in equation (30) and (31) do not change if the value of ψ varies. The derivation of the asymptotic distribution of the LR estimator follows the proof in Tsiatis (1990). Tsiatis requires that the density of U_0 is bounded. For the MPH model, this density is

$$f(u_0) = \mathrm{E}\big[Ve^{-u_0 V}\big].$$

If $\mathrm{E}(V) = \infty$, this density is not bounded at $u_0 = 0$. Inspection of Tsiatis' proof shows that this does not change the result, and we do not need to impose the restriction that $\mathrm{E}(V)$ is finite. The transformed durations are observed up to τ with $\tau < \infty$ such that for some $\psi, \eta > 0$

$$\Pr\big[\min(U_0, C) > \tau + \psi\big] \geq \eta.$$

In the MPH model, this is just an assumption on the distribution of C because for U_0 it is satisfied for all $\tau < \infty$.

B.3. Technical details section 4: Lemma 2-3. Lemma 2:

If the derivative κ' is bounded on $[0, \tau]$, then for $\epsilon > 0$ with

$$\inf_{0 \leq u \leq \tau} b_N^2 N^{1-\epsilon} Y_{h,N}^U(u, \theta_0) \xrightarrow{p} \infty \tag{B.14}$$

and

$$b_N N^{1-\epsilon} \to \infty, \tag{B.15}$$

we have

$$\sup_{u_1 \leq u \leq u_2} N^{\epsilon} \big|\hat{\kappa}_N(u, \theta_0) - \kappa_0(u)\big| \xrightarrow{p} 0 \tag{B.16}$$

for u_1, u_2 with $0 < u_1 < u_2 < \tau$.

If $Y_{h,N}(t)$ is bounded away from zero on $[0, \tau]$ for large N, then (B.14) and (B.15) imply that if $b_N = N^{-c}$ for $\epsilon < c < \frac{1}{2} - \epsilon$, then $\epsilon < \frac{1}{4}$. Note that the uniform convergence holds on a compact subset of $[0, \tau]$. Although this can be generalized to uniform convergence on $[0, \tau]$, the variable kernels that are needed for this generalization complicate the asymptotic analysis. In practice, estimation of the hazard is inaccurate near the endpoints, and it may be preferable to exclude observations that are close to the endpoints. Note that the

observations near the endpoints are used in the estimation of the hazard. Also, using a bandwidth proportional to $N^{-1/5}$ and $\varepsilon = \frac{1}{11}$ satisfies all the assumptions of this paper.

We do not observe the transformed duration $\tilde{U}_0(\theta_0)$ but rather an estimate $\tilde{U}_0(\hat{\theta}_N)$ of this transformed duration, and hence we consider the kernel estimator

$$\hat{\kappa}_N(u, \hat{\theta}_N) = \frac{1}{b_N} \sum_{i=1}^{N} \Delta_i \frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} K\left(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N}\right). \tag{B.17}$$

Lemma 3

The kernel K is positive and bounded on $[-1, 1]$ (and zero elsewhere) and satisfies a Lipschitz condition on this interval. The covariate process $X(t)$ is bounded on $[0, \tau]$ and so is $\left|\frac{\partial \lambda(t,\alpha)}{\partial \alpha}\right|$ for all α in an open neighborhood of α_0. Moreover

$$\frac{I\big(Y_N^U(u, \theta) > 0\big)}{Y_{h,N}^U(u, \theta)} \xrightarrow{p} H(u, \theta) \tag{B.18}$$

uniformly for $0 \leq u \leq \tau, \theta \in N(\theta_0)$ and H has derivatives that are bounded for $0 \leq u \leq \tau, \theta \in N(\theta_0)$. Then for $\epsilon > 0$ such that

$$b_N^2 N^{\frac{1}{2}-\epsilon} \to \infty, \tag{B.19}$$

we have

$$\sup_{0 \leq u \leq \tau} N^{\epsilon} \left|\hat{\kappa}_N(u, \hat{\theta}_N) - \hat{\kappa}_N(u, \theta_0)\right| \xrightarrow{p} 0. \tag{B.20}$$

Proof: See below.

Note that the conditions on b_N are determined in Lemma 2 and that a bandwidth proportional to $N^{-1/5}$ and $\varepsilon = \frac{1}{11}$ satisfies all the assumptions of this paper. The fact that we use estimated transformed durations does not change the restrictions on the bandwidth choice.

At this point we consider the condition in (B.18) more closely. With $h(T, \theta) = \int_0^T \lambda(t, \alpha) e^{\beta' X(t)} \mathrm{d}t$, if the duration T is (right) censored at C, $Y(t) = I(T \geq t) I(C \geq t)$, so

$$Y^U(u, \theta) = I\big(h(T, \theta) \geq u\big) \cdot I\big(h(C, \theta) \geq u\big).$$

If the censoring time and the duration are conditionally independent given the history up to t, i.e.

$$I(T \geq t) \perp I(C \geq t) \big| Y(s), X(t), 0 \leq s \leq t, \tag{B.21}$$

then

$$I\big(h(T, \theta) \geq u\big) \perp I\big(h(C, \theta) \geq u\big) \big| Y^U(s), X^U(t), 0 \leq s \leq u. \tag{B.22}$$

If $N(\theta_0)$ is an open neighborhood of θ_0, X_i and C_i are i.i.d., and

$$\sup_{\theta\in N(\theta_0)0\leq u\leq\tau} \Pr\big(h(T,\theta)<u\big)<1 \tag{B.23}$$

$$\sup_{\theta\in N(\theta_0)0\leq u\leq\tau} \Pr\big(h(C,\theta)<u\big)<1, \tag{B.24}$$

then

$$\inf_{\theta\in N(\theta_0)0\leq u\leq\tau} I\big(Y_N^U(u,\theta)>0\big)\xrightarrow{p}0 \tag{B.25}$$

and by the uniform law of large numbers

$$Y_{h,N}^U(u,\theta)\xrightarrow{p}\Pr\big(h(T,\theta)\geq u\big)\cdot\Pr\big(h(C,\theta)\geq u\big) \tag{B.26}$$

uniformly for $\theta\in N(\theta_0)$ and $0\leq u\leq\tau$. Because by (B.23) the limit is bounded away from zero, we have

$$\frac{I\big(Y_N^U(u,\theta)>0\big)}{Y_{h,N}^U(u,\theta)}\xrightarrow{p}H(u,\theta) \tag{B.27}$$

uniformly for $\theta\in N(\theta_0)$ and $0\leq u\leq\tau$ with

$$H(u,\theta)=\frac{1}{\Pr\big(h(T,\theta)\geq u\big)\cdot\Pr\big(h(C,\theta)\geq u\big)}. \tag{B.28}$$

Because $h(T,\theta_0)=U_0$, (B.19) holds for $\theta=\theta_0$ if $\kappa_0(u)$ is bounded for $0\leq u\leq\tau$. From the expression for $\kappa_U(u,\theta)$ in (9), a sufficient condition for $\kappa_U(u,\theta)$ to be bounded for all θ in a neighborhood of θ_0 and $0\leq t\leq\tau$ is that $\lambda(t,\alpha)>0$ for all t and on a neighborhood of α_0. In the same way, (B.20) holds if the hazard of C is bounded and $\lambda(t,\alpha)$ is bounded away from zero in a neighborhood around α_0.

B.4. Proof of Lemma 1.

$\tilde{S}_N(\theta)$ is a linearization of $\tilde{S}_N(\theta)$. Because $S_N(\theta)$ is not continuous in θ, it is not possible to linearize this function by a first order Taylor series expansion. Instead we linearize the hazard rate of the transformed durations $U(\theta)$. From (4) and (5) we obtain

$$U=h\big(h_0^{-1}(U_0),\theta\big).$$

This relates the hazard of the distribution of $U(\theta)$ to that of U_0

$$\kappa_U(u,\theta)=\kappa_0\Big(h_0\big(h^{-1}(u,\theta)\big)\Big)\frac{\lambda\big(h^{-1}(u,\theta),\alpha_0\big)}{\lambda\big(h^{-1}(u,\theta),\alpha\big)}e^{(\beta_0-\beta)'X\left(h^{-1}(u,\theta)\right)}.$$

Because $h\big(h^{-1}(u,\theta),\theta\big)=u$, we have

$$\frac{\partial h^{-1}}{\partial\theta}(u,\theta)=-\frac{\frac{\partial h}{\partial\theta}\big(h^{-1}(u,\theta),\theta\big)}{\frac{\partial h}{\partial t}\big(h^{-1}(u,\theta),\theta\big)}.$$

The derivatives of $\kappa_U(u,\theta)$ with respect to θ are

$$\begin{aligned}\frac{\partial\kappa_U(u,\theta)}{\partial\alpha}\Big|_{\theta=\theta_0} &= -\kappa_0'(u)\int_0^{h_0^{-1}(u)}\frac{\partial\lambda}{\partial\alpha}(t,\alpha_0)e^{\beta_0'X(t)}\mathrm{d}t-\kappa_0(u)\frac{\partial\ln\lambda}{\partial\alpha}\big(h_0^{-1}(u),\alpha_0\big)\\ &= \kappa_0'(u)\int_0^u\frac{\partial\ln\lambda}{\partial\alpha}\big(h_0^{-1}(s),\alpha_0\big)\mathrm{d}s-\kappa_0(u)\frac{\partial\ln\lambda}{\partial\alpha}\big(h_0^{-1}(u),\alpha_0\big),\end{aligned}\tag{B.29}$$

where the last equality follows from a change of variables in the integral. In the same way, we obtain with a change of variable in the integral

$$\begin{aligned}\frac{\partial\kappa_U(u,\theta)}{\partial\beta}\Big|_{\theta=\theta_0} &= -\kappa_0'(u)\int_0^{h_0^{-1}(u)}\lambda(t,\alpha_0)e^{\beta_0'X(t)}\mathrm{d}t-\kappa_0(u)X\big(h_0^{-1}(u)\big)\\ &= \kappa_0'(u)\int_0^u X\big(h_0^{-1}(s),\alpha_0\big)\mathrm{d}s-\kappa_0(u)X\big(h_0^{-1}(u)\big).\end{aligned}\tag{B.30}$$

The proof consists of checking the conditions for asymptotic linearity of $S_N(\theta)$ in Tsiatis (1990) and a computation of the coefficients in the linear approximation. In Tsiatis' proof the covariate in the estimating equation is X_i. We have $W\big(u,X_{h,i}^U(u,\theta)\big)$ and hence the requirement that this is a vector of bounded functions. The equations (B.9), (B.10) and (B.11) are stability conditions (see also Andersen et al. (1993)). Instead of a mean and variance condition as in Tsiatis (1990), we have a mean and two covariance conditions. Note that by setting $s=u$, we obtain conditions for uniform convergence to $V_\alpha(u,u)$ and $V_\beta(u,u)$. The final condition for linearization is that for $u\le\tau$

$$\Big|\kappa_U(u,\theta)-\kappa_0(u)-\frac{\partial\kappa_U}{\partial\theta'}(u,\theta_0)(\theta-\theta_0)\Big|\le|\theta-\theta_0|^2h(u).\tag{B.31}$$

The assumptions that $\lambda(t,\alpha)$ is bounded away from zero for all $t\ge0$ and α in the parameter space, that $\big|\frac{\partial^2\lambda}{\partial\alpha\partial\alpha'}(t,\alpha)\big|<\infty$ for all $t\ge0$ and α in the parameter space, and that $X(t)$ is bounded, imply that the second derivative of $\kappa_U(u,\theta)$ with respect to θ is bounded for all $u\le\tau$ and $\theta\in\Theta$. This is sufficient for (B.31) if the parameter space is convex.

Next we linearize $S_N(\theta)$. Because

$$\mathrm{d}N_i^U(u,\theta)=\mathrm{d}M_i^U(u,\theta)+Y_i^U(u,\theta)\kappa_{U_i}(u,\theta)\mathrm{d}u,$$

we have if $|\theta-\theta_0|$ is small

$$\begin{aligned}S_N(\theta)&=\sum_{i=1}^N\int_0^\tau\Big(W\big(u,X_{h,i}^U(u,\theta_0)\big)-W_h(u,\theta_0)\Big)\,dM_i^0(u)+\\ &+\Big[\sum_{i=1}^N\int_0^\tau\Big(W\big(u,X_{h,i}^U(u,\theta_0)\big)-W_h(u,\theta_0)\Big)Y_i^0(u)\frac{\partial\kappa_{U_i}}{\partial\theta'}(u,\theta_0)\mathrm{d}u\Big](\theta-\theta_0)+o(|\theta-\theta_0|).\end{aligned}\tag{B.32}$$

The second term is after substitution of (B.29), and (B.30)

$$
\begin{aligned}
&- \Big[\int_0^\tau \int_0^u \sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big) Y_i^0(u) \frac{\partial \ln \lambda}{\partial \alpha'}\big(h_{0i}^{-1}(s),\alpha_0\big)\kappa_0'(u)\mathrm{d}s\mathrm{d}u + \\
&+ \int_0^\tau \sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big) Y_i^0(u) \frac{\partial \ln \lambda}{\partial \alpha'}\big(h_{0i}^{-1}(u),\alpha_0\big)\kappa_0(u)\mathrm{d}u\Big](\alpha-\alpha_0) - \\
&\quad - \Big[\int_0^\tau \int_0^u \sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big) Y_i^0(u) X\big(h_{0i}^{-1}(s),\alpha_0\big)\kappa_0'(u)\mathrm{d}s\mathrm{d}u + \\
&\quad + \int_0^\tau \sum_{i=1}^N \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big) Y_i^0(u) X\big(h_{0i}^{-1}(u),\alpha_0\big)\kappa_0(u)\mathrm{d}u\Big](\beta-\beta_0)
\end{aligned}
\tag{B.33}
$$

The normalized vectors of coefficients converge to (B.12) and (B.13) if (B.10) and (B.11) hold. This proves the lemma.

B.5. Proof of Theorem 1. By van der Vaart (1998) Theorem 5.45, we have from Lemma 1

$$
\sqrt{N}(\hat\theta_N - \theta_0) = D(\theta_0)^{-1}\frac{1}{\sqrt{N}}\int_0^\tau \Big(W\big(u, X_{h,i}^U(u,\theta_0)\big) - W_h(u,\theta_0)\Big)\mathrm{d}M_{i0}
$$

with M_0 the martingale associated with the counting process N_0 for U_0. By the central limit theorem for integrals of predetermined functions with respect to a martingale, (see e.g. Anderson et al. (1993)), the sum on the right-hand side converges to a normal distribution with the variance matrix in (24).

B.6. Proof of Lemma 2 and 3. We have

$$
\begin{aligned}
&N^\epsilon\big|\hat\kappa_N(u,\hat\theta_N) - \hat\kappa_N(u,\theta_0)\big| \leq \\
&\Big|\frac{N^\epsilon}{Nb_N}\sum_{i=1}^N \Delta_i\Big(\frac{I\big(Y_N^U(\tilde U_i(\hat\theta_N),\hat\theta_N)>0\big)}{Y_{h,N}^U(\tilde U_i(\hat\theta_N),\hat\theta_N)} - \frac{I\big(Y_N^U(\tilde U_i(\theta_0),\theta_0)>0\big)}{Y_{h,N}^U(\tilde U_i(\theta_0),\theta_0)}K\Big(\frac{u-\tilde U_i(\theta_0)}{b_N}\Big)\Big| + \\
&\quad + \Big|\frac{N^\epsilon}{Nb_N}\sum_{i=1}^N \Delta_i\Big(K\Big(\frac{u-\tilde U_i(\hat\theta_N)}{b_N}\Big) - K\Big(\frac{u-\tilde U_i(\theta_0)}{b_N}\Big)\Big)\frac{I\big(Y_N^U(\tilde U_i(\hat\theta_N),\hat\theta_N)>0\big)}{Y_{h,N}^U(\tilde U_i(\hat\theta_N),\hat\theta_N)}\Big|.
\end{aligned}
\tag{B.34}
$$

We first consider the second term. Because K is Lipschitz this is bounded by

$$
\frac{CN^\epsilon}{Nb_N^2}\sum_{i=1}^N \Delta_i\big|\tilde U_i(\hat\theta_N) - \tilde U_i(\theta_0)\big|\frac{I\big(Y_N^U(\tilde U_i(\hat\theta_N),\hat\theta_N)>0\big)}{Y_{h,N}^U(\tilde U_i(\hat\theta_N),\hat\theta_N)}. \tag{B.35}
$$

Moreover by the mean value theorem, we have that for some intermediate $\overline{\beta}_{iN}, \overline{\alpha}_{iN}$

$$\begin{aligned}\tilde{U}_i(\hat{\theta}_N) - \tilde{U}_i(\theta_0) &= \int_0^{\tilde{T}_i} \lambda(t, \overline{\alpha}_{iN}) e^{\overline{\beta}'_{iN} X_i(s)} X_i(s)' \, \mathrm{d}s(\hat{\beta}_N - \beta_0) + \\ &\quad + \int_0^{\tilde{T}_i} e^{\overline{\beta}'_{iN} X_i(s)} \frac{\partial \lambda(t, \overline{\alpha}_{iN})}{\partial \alpha'} \, \mathrm{d}s(\hat{\alpha}_N - \alpha_0).\end{aligned} \tag{B.36}$$

Because $X_i(t)$ is bounded on $[0, \tau]$ and so is $\left|\frac{\partial \lambda(t,\alpha)}{\partial \alpha}\right|$ for all α in an open neighborhood of α_0, (B.36) is bounded by $|c_1'(\hat{\beta}_N - \beta_0)| + |c_2'(\hat{\alpha}_N - \alpha_0)|$ and substitution in (B.35) gives the upper bound

$$\frac{\bar{K}}{N} \sum_{i=1}^{N} \Delta_i \frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} \left(\frac{N^\epsilon |c_1'(\hat{\beta}_N - \beta_0)|}{b_N^2} + \frac{N^\epsilon |c_2'(\hat{\alpha}_N - \alpha_0)|}{b_N^2} \right). \tag{B.37}$$

Because the estimator $\hat{\theta}_N$ is $\sqrt{N}$ consistent, the upper bound converges to 0 in probability if $b_N^2 N^{\frac{1}{2}-\epsilon} \to \infty$.

Next we consider the first term in (B.34). By subtraction and addition of expected values, this term is bounded by

$$\begin{aligned}&\left| \frac{N^\epsilon}{N b_N} \sum_{i=1}^{N} \Delta_i \left[\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} K\left(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N} \right) - \right.\right. \\ &\qquad \left. - \mathrm{E}\left(\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} K\left(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N} \right) \middle| \Delta_i = 1 \right) \right] \Bigg| + \\ &\qquad + \left| \frac{N^\epsilon}{N b_N} \sum_{i=1}^{N} \Delta_i \left[\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0), \theta_0) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\theta_0), \theta_0)} K\left(\frac{u - \tilde{U}_i(\theta_0)}{b_N} \right) - \right.\right. \\ &\qquad \left. - \mathrm{E}\left(\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0), \theta_0) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\theta_0), \theta_0)} K\left(\frac{u - \tilde{U}_i(\theta_0)}{b_N} \right) \middle| \Delta_i = 1 \right) \right] \Bigg| + \\ &\quad + \frac{N^\epsilon}{N b_N} \sum_{i=1}^{N} \Delta_i \left| \mathrm{E}\left[\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} K\left(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N} \right) \middle| \Delta_i = 1 \right) \right] - \\ &\qquad \left. \mathrm{E}\left[\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0), \theta_0) > 0\big)}{Y_{h,N}^U(\tilde{U}_i(\theta_0), \theta_0)} K\left(\frac{u - \tilde{U}_i(\theta_0)}{b_N} \right) \middle| \Delta_i = 1 \right) \right] \right|.\end{aligned} \tag{B.38}$$

The first and second terms converge to 0 in probability if $b_N N^{\frac{1}{2}-\epsilon} \to \infty$. Because of (B.18) the final term converges in probability to

$$\frac{N^\epsilon}{N b_N} \sum_{i=1}^{N} \Delta_i \left| \mathrm{E}\left[H\big(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N\big) K\left(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N} \right) \right] - \mathrm{E}\left[H\big(\tilde{U}_i(\theta_0), \theta_0\big) K\left(\frac{u - \tilde{U}_i(\theta_0)}{b_N} \right) \right] \right| \tag{B.39}$$

This expression is bounded (both H and K are bounded) by

$$\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i \mathrm{E}\Big[\Big|H\big(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N\big)-H\big(\tilde{U}_i(\theta_0),\theta_0\big)\Big|\Big]+$$
$$+\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i \mathrm{E}\Big[\Big|K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)\Big|\Big]. \quad \text{(B.40)}$$

The first term goes to 0 in probability if $b_N N^{\frac{1}{2}-\epsilon}\to\infty$ and the second if $b_N^2 N^{\frac{1}{2}-\epsilon}\to\infty$. This completes the proof.

References

[1] Aalen, O. O., O. Borgan, and H. K. Gjessing (2009). Survival and Event History Analysis. Springer Verlag, New York.

[2] Amemiya, T. (1974). The nonlinear two–stage least–squares estimator. Journal of Econometrics 2, 105–110.

[3] Amemiya, T. (1985). Instrumental variable estimation for the nonlinear errors–in–variables model. Journal of Econometrics 28, 273–289.

[4] Andersen, P. K., O. Borgan, R. D. Gill, and N. Keiding (1993). Statistical Models Based on Counting Processes. Springer Verlag, New York.

[5] Andersen, P. K. and R. D. Gill (1982). Cox's regression model for counting processes: A large sample study. Annals of Statistics 10, 1100–1120.

[6] Baker, M. and A. Melino (2000). Duration dependence and nonparametric heterogeneity: A Monte Carlo study. Journal of Econometrics 96, 357–393.

[7] Bearse, P., J. Canals-Cerd´a, and P. Rilstone (2007). Efficient semiparametric estimation of duration models with unobserved heterogeneity. Econometric Theory 23, 281–308.

[8] Bijwaard, G. E. (2009). Instrumental variable estimation for duration data. In H. Engelhardt, H.-P. Kohler, and A. Fürnkranz-Prskawetz (Eds.), Causal Analysis in Population Studies: Concepts, Methods, Applications, pp. 111–148. Springer Verlag, New York.

[9] Bijwaard, G. E. (2010). Immigrant migration dynamics model for The Netherlands. Journal of Population Economics, *23*, 1213–1247.

[10] Bijwaard, G. E. and G. Ridder (2005). Correcting for selective compliance in a re–employment bonus experiment. Journal of Econometrics 125, 77–111.

[11] Bijwaard, G. E., C. Schluter, and J. Wahba (2012). The impact of labour market dynamics on the return–migration of immigrants. Discussion Paper No. 27/12, CReAM.

[12] Borjas, G. J. and B. Bratsberg (1996). Who leaves? The outmigration of the foreign-born. The Review of Economics and Statistics, 78, 165–176.

[13] Chen, S. (2002). Rank estimation of transformation models. Econometrica 70, 1683–1697.

[14] Chiaporri, P. A. and B. Salanie (2000). Testing for asymmetric information in insurance markets. Journal of Political Economy 108, 56–78.

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

[16] Elbers, C. and G. Ridder (1982). True and spurious duration dependence: The identifiability of the proportional hazard model. Review of Economic Studies 49, 403–410.

[17] Feller, W. (1971). An introduction to probability theory and its applications (third edition). John Wiley and Sons.

[18] Hahn, J. (1994). The efficiency bound of the mixed proportional hazard model. Review of Economic Studies 61, 607–629.

[19] Han, A. K. (1987). Non–parametric analysis of a generalized regression model: The maximum rank correlation estimator. Journal of Econometrics 35, 303–316.

[20] Hausman, J. A. and T. Woutersen (2005). Estimating a semi–parametric duration model without specifying heterogeneity. CeMMAP, working paper, CWP11/05.

[21] Heckman, J. J. (1991). Identifying the hand of the past: Distinguishing state dependence from heterogeneity. American Economic Review 81, 75–79.

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

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

[24] Honoré, B. E. (1990). Simple estimation of a duration model with unobserved heterogeneity. Econometrica 58, 453–473.

[25] Horowitz, J. L. (1996). Semiparametric estimation of a regression model with an unknown transformation of the dependent variable. Econometrica 64, 103–137.

[26] Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. Econometrica 67, 1001–1018.

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

[28] Khan, S. (2001). Two stage rank estimation of quantile index models. Journal of Econometrics 100, 319–355.

[29] Khan, S. and E. Tamer (2007). Partial rank estimation of duration models with general forms of censoring. Journal of Econometrics 136, 251–280.

[30] Klein, J. P. and M. L. Moeschberger (1997). Survival Analysis: Techniques for Censored and Truncated Data. Springer Verlag, New York.

[31] Lai, T. L. and Z. Ying (1991). Rank regression methods for left–truncated and right-censored data. Annals of Statistics 19, 531–556.

[32] Lancaster, T. (1976). Redundancy, unemployment and manpower policy: A comment. Economic Journal 86, 335–338.

[33] Lancaster, T. (1979). Econometric methods for the duration of unemployment. Econometrica 47, 939–956.

[34] Lin, D. Y. and Ying, Z (1995). Semiparametric inference for the accelerated life model with time-dependent covariates. Journao of Statistical Planning and Inference, 44, 47-63.

[35] Linsay, B. G. (1983). The geometry of mixture likelihoods: A general theory. Annals of Statistics 11, 86–94.

[36] Manton, K. G., E. Stallard, and J. W. Vaupel (1981). Methods for the mortality experience of heterogeneous populations. Demography 18, 389–410.

[37] Meyer, P (1963). Decomposition of supermartingales: the uniqueness theorem. Illinois Journal of Mathematics 7: 1–17.

[38] Moeschberger, M. L., and J. P. Klein (1997): Survival Analysis: Techniques for Censored and Truncated Data, Springer Verlag, New York.

[39] 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.

[40] Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. The Computer Journal 7, 155–162.

[41] Prentice, R. L. (1978). Linear rank tests with right censored data. Biometrika 65, 167–179.

[42] Press,W. H., B. P. Flannert, S. A. Teukolsky, and W. T. Vetterling (1986). Numerical Recipes: The Art of Scientific Computing. Cambridge: Cambridge University Press.

[43] Protter, P. (2005). Stochastic Integration and Differential Equations. Springer Verlag, New York. 107–113.

[44] Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. Annals of Statistics 11, 453–466.

[45] Ridder, G. and T. Woutersen (2003). The singularity of the efficiency bound of the mixed proportional hazard model. Econometrica 71, 1579–1589.

[46] Robins, J. M. and A. A. Tsiatis (1992). Semiparametric estimation of an accelerated failure time model with time–dependent covariates. Biometrika 79, 311–319.

[47] Sherman, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. Econometrica 61, 123–137.

[48] Therneau, T. and P. Grambsch (2000). Modeling Survival Data: Extending the Cox Model. Springer Verlag, New York.

[49] Tsiatis, A. A. (1990). Estimating regression parameters using linear rank tests for censored data. Annals of Statistics 18, 354–372.

[50] van der Vaart, A. W. (1998). Asymptotic Statistics. Cambridge University Press.

[51] Wooldridge, J. M. (2005). Unobserved heterogeneity and estimation of average partial effects. In D. W. K. Andrews and J. H. Stock (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg, pp. 27–55. Cambridge University Press.

[52] Woutersen, T. (2000). Consistent estimators for panel duration data with endogenous censoring and endogenous regressors. Dissertation Brown University.