DISCUSSION PAPER SERIES

IZA DP No. 4543

# A Simple GMM Estimator for the Semi-Parametric Mixed Proportional Hazard Model

Govert E. Bijwaard
Geert Ridder

November 2009

Forschungsinstitut
zur Zukunft der Arbeit
Institute for the Study
of Labor

# A Simple GMM Estimator for the Semi-Parametric Mixed Proportional Hazard Model

**Govert E. Bijwaard**
*Netherlands Interdisciplinary Demographic Institute (NIDI)*
*and IZA*

**Geert Ridder**
*University of Southern California*

Discussion Paper No. 4543
November 2009

IZA

P.O. Box 7240
53072 Bonn
Germany

Phone: +49-228-3894-0
Fax: +49-228-3894-180
E-mail: iza@iza.org

IZA Discussion Paper No. 4543
November 2009

# ABSTRACT

## A Simple GMM Estimator for the Semi-Parametric Mixed Proportional Hazard Model

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 first-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

# 1 Introduction

Although the Mixed Proportional Hazard (MPH) model for duration data that was independently introduced by Lancaster (1979) and Manton, Stallard, and Vaupel (1981) has been used quite frequently in empirical work, 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 identied 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) (see also Heckman and Singer (1984b) ) 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 0 was known.

These results on nonparametric identification led to the development of estimation methods that required fewer parametric assumptions. Heckman and Singer (1984a) used the NPMLE for mixture models that was first characterized by Linsay (1983), to estimate regression parameters and the parameters of the baseline hazard in an MPH model. Biostatisticians who are reluctant to make parametric assumptions on the baseline hazard introduced a method the assumes a parametric distribution for the unobserved heterogeneity, but is nonparametric with respect to the duration dependence , Nielsen et al. (1992). A problem with Heckman and Singer's NPMLE is that the speed of convergence and the asymptotic distribution of the estimators is 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 semi-parametric estimator for the MPH model that does not require parametric assumptions neither on the unobserved heterogeneity nor on the duration dependence. This estimator is based on Horowitz (1996) estimator for a semi-parametric 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 0, the rate of convergence is $N^{1/3}$. Honoré (1990) proposed an estimator for the Weibull MPH based an 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) who shows that in the MPH model with Weibull baseline hazard (but unspecified distribution for the unobserved heterogeneity) the efficiency bound is singular. This precludes the existence of regular $\sqrt{N}$ consistent estimators of the parameters of this model. He also shows that $\sqrt{N}$ consistent estimators may exist if there are repeated spells on the same individual, and there seems to be an emerging consensus that unobserved heterogeneity and duration dependence can only be distinguished if one has access to multiple spells for each individual.

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 a common feature of this class is that the baseline hazard in 0 is either 0 or $\infty$. 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 0 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 from 0 and $\infty$ in 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 instances that 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. For an MPH with time constant covariates the derivative of the $q$th 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 semi-parametric 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 non-parametric estimation of unknown densities. Hence, it is simpler than the semi-parametric maximum likelihood estimator of Bearse et al. (2007). Both the simple estimator in this paper and the Bearse et al. estimator are based on the idea that population distribution of the integrated baseline hazard is independent of the covariates. Woutersen (2000) and Ridder and Woutersen (2003) use the same idea to obtain an estimator that does not require parametric assumptions on the baseline hazard. The GMM estimator can be extended to the case that some of the covariates are endogenous (Bijwaard (2009) uses the estimator in such a case). The 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 (non-informatively) censored integrated hazards, and this likelihood is maximized over the parameters of the MPH model. As is evident from the simulation results in Bearse et al., 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.

The outline of the article is as follows. In Section 2 a counting process interpretation of the MPH model is given. The counting process approach simplifies the definition of predictable time–varying explanatory variables and noninformative censoring. Within the framework of counting processes, the asymptotic properties of our estimator, which is introduced in section 3, can be elegantly justified by martingale theory. In Section 4 we derive the asymptotic properties of the, two stage, optimal LRE. The weight functions of this estimator are obtained by substituting consistent first stage estimators for the parameters and by using a nonparametric estimator for the hazard and its derivative of the transformed durations. The Monte Carlo experiments of Section 5 give some insight in the (small) sample behaviour of the estimator. Finally, in section 6 we apply our estimator on a real data set of unemployment durations. Section 7 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

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

In (2) $\overline{X}(t) = \{X(s)|0 \leq s \leq t\}$ is the sample path of $X$ up to 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 0. We assume that $\overline{X}(t)$ is independent of $V$ . Note that, although we express the hazard at $t$ is 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 $\alpha$.It reflects the duration dependence of the hazard rate.

### 2.1 A Counting process approach

The cdf and pdf of the distribution of the duration $T$ can be expressed as functions of the hazard rate. These expressions can be used to obtain the MLEs of the parameters of the model. To understand all the features of the Linear Rank Estimator the counting process approach provides a better framework. The counting process approach has increasingly become the standard framework for analyzing duration data. Andersen et al. (1993) have provided an excellent survey of counting processes. Less technical surveys have been given by Klein and Moeschberger (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 equal to 1 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 $Y(t)\kappa(t)$, with $\kappa(t)$ 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 with $g(T, C) = \min(T, C)$. 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 $t$, $\overline{Y}(t)$ is assumed to be a left continuous function of $t$. The history of the whole process also includes the history of the covariate process, $\overline{X}(t)$, and $V$. Thus, we have

$$\Pr\big(\mathrm{d}N(t) = 1|\overline{Y}(t), \overline{X}(t), V\big) = Y(t)\kappa(t|\overline{X}(t), V, \theta) \tag{3}$$

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-Meier decomposition, allows us to write

$$\mathrm{d}N(t) = Y(t)\kappa(t|\overline{X}(t), V, \theta)\mathrm{d}t + \mathrm{d}M(t) \tag{4}$$

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

$$\begin{aligned} \mathrm{E}\big(\mathrm{d}M(t)|V, \overline{Y}(t), \overline{X}(t)\big) &= 0 & (5) \\ \mathrm{Var}\big(\mathrm{d}M(t)|V, \overline{Y}(t), \overline{X}(t)\big) &= Y(t)\kappa(t|\overline{X}(t), V, \theta)\mathrm{d}t & (6) \end{aligned}$$

The (conditional) mean and variance of the counting process are equal, so that the disturbances in (4) are heteroscedastic. The probability in (3) is 0, if the unit is no longer under observation. A counting process can be considered as a sequence of Bernoulli experiments, because if d$t$ is small, (5) and (6) give the mean and variance of a Bernoulli random variable. The relation between the counting process and the sequence of Bernoulli experiments is given in (4), that 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-Meier decomposition is the key to the derivation of the distribution of the estimators, because the asymptotic behavior of partial sums of martingales is well-known.

## 2.2 Durations and Transformed Durations

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

$$h(T, \overline{X}(t), \theta) = \int_0^T \lambda(t, \alpha)e^{\beta' X(t)}\,\mathrm{d}t \tag{7}$$

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

$$h(T,\overline{X}(t),\theta_0)=\frac{A}{V}\overset{d}{=}U_0 \tag{8}$$

with $A$ a standard exponential random variable.

Equations (7) and (8) show that the MPH model is essentially a transformation model that transforms the conditional distribution of $T$ given the observable covariates $X(.)$ to a positive random variable that is independent of $X(.)$ and of the baseline hazard $\lambda(.,\alpha_0)$. This independence is the key to understand 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

$$h(T,\overline{X}(t),\theta)=U \tag{9}$$

with $U$ a nonnegative random variable. The hazard rate of $U=h(T)$ is

$$\begin{aligned}\kappa_U(u|V) &= \kappa_T\big(h^{-1}(u)\big)\frac{1}{h'\big(h^{-1}(u)\big)}\\ &= \frac{\lambda\big(h^{-1}(u,\overline{X}(u),\theta),\alpha_0\big)}{\lambda\big(h^{-1}(u,\overline{X}(u),\theta),\alpha\big)}e^{(\beta_0-\beta)'X^U(u,\theta)}V \end{aligned}\tag{10}$$

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

Just as 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 process and observation indicator is

$$N^U(u,\theta) = N\big(h^{-1}(u,\overline{X}(u),\theta)\big) \tag{11}$$

$$Y^U(u,\theta) = Y\big(h^{-1}(u,\overline{X}(u),\theta)\big) \tag{12}$$

The intensity of the transformed counting process (with respect to history $\overline{X}^U(u,\theta),\overline{Y}^U(u,\theta)$ is (see Andersen et al. (1993), p. 87, and using (10))

$$\Pr\bigl(\mathrm{d}N^U(u,\theta)=1\big|\overline{X}^U(u,\theta),\overline{Y}^U(u,\theta)\bigr)=$$
$$=Y^U(u,\theta)\frac{\lambda\bigl(h^{-1}(u,\overline{X}(u),\theta),\alpha_0\bigr)}{\lambda\bigl(h^{-1}(u,\overline{X}(u),\theta),\alpha\bigr)}e^{(\beta_0-\beta)'X^U(u,\theta)}\mathrm{E}\bigl[V|\overline{X}^U(u,\theta),\overline{Y}^U(u,\theta)\bigr]\mathrm{d}u \quad (13)$$

and we denote this hazard by $\kappa_U\bigl(u|\overline{X}^U(u,\theta),\overline{Y}^U(u,\theta)\bigr)$. For the population parameter value $\theta_0$ this becomes

$$\Pr\Bigl(\mathrm{d}N^U(u,\theta_0)=1\Big|\overline{X}^U(u,\theta_0),\overline{Y}^U(u,\theta_0)\Bigr)=Y^U(u,\theta)\mathrm{E}\Bigl[V|\overline{X}^U(u,\theta_0),\overline{Y}^U(u,\theta_0)\Bigr]\mathrm{d}u \quad (14)$$

If censoring is non-informative, i.e. $Y(t)=I(t\leq T)I_C(t)$ with $C$ independent of $T$ (but possibly dependent on $X$), then

$$\Pr\Bigl(\mathrm{d}N^U(u,\theta_0)=1\Big|\overline{X}^U(u,\theta_0),\overline{Y}^U(u,\theta_0)\Bigr)=Y^U(u,\theta)\mathrm{E}\Bigl[V\Big|U_0\geq u\Bigr]\mathrm{d}u \quad (15)$$

and the intensity is independent of $\overline{X}^U(u,\theta_0)$. This independence is the basis for the estimation of the parameters of the MPH model. We denote the hazard in (15) 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

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

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

$$\kappa_U(u)=\begin{cases}e^{(\alpha_0-\alpha)+(\beta_0-\beta)X_1}\mathrm{E}[V|U\geq u] & 0\leq U\leq e^{\alpha+\beta X_1}t_1\\ e^{(\beta_0-\beta)X_1}\mathrm{E}[V|U\geq u] & e^{\alpha+\beta X_1}t_1<U\leq e^{\alpha+\beta X_1}t_1+e^{\beta X_2}(t_2-t_1)\\ e^{(\beta_0-\beta)X_2}\mathrm{E}[V|U\geq u] & U>e^{\alpha+\beta X_1}t_1+e^{\beta X_2}(t_2-t_1)\end{cases} \quad (16)$$

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

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

If $V$ has a Gamma distribution with mean 1 and variance $\sigma^2$, 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 only dependents on the distribution of $V$. 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 non-parametric 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 estimation equation for 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 non-parametric identification of the MPH model.

## 2.3 Identification

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

$$\mathrm{d}X(t) \perp N(s), s \geq t | \overline{Y}(t), \overline{X}(t) \tag{17}$$

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 a some random set, e.g. the set $t \leq C$ for right censoring. We assume

$$\mathrm{d}I_C(t) \perp N(s), s \geq t | \overline{Y}(t), \overline{X}(t) \tag{18}$$

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 (3) and the following equations we condition on the unobserved $V$. The corresponding unconditional results are obtained by taking the expectation of $V$ given $\overline{Y}(t), \overline{X}(t)$. 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

$$\mathrm{E}\big(V | T \geq t, \overline{Y}(t), \overline{X}(t)\big) \tag{19}$$

The hazard that is not conditional on V is

$$\kappa(t|\overline{X}(t),\theta) = \lambda(t,\alpha)e^{\beta' X(t)}\mathrm{E}\big[V|T \geq t, \overline{Y}(t), \overline{X}(t)\big] \tag{20}$$

Non-parametric identification of the MPH model has been studied by Elbers and Ridder (1982) and Heckman and Singer (1984b). Their results refer to the model in which both the baseline hazard and the distribution of the unobserved heterogeneity are left unspecified. In their proofs they need the assumption the mean of the ditribution of V is finite, Elbers and Ridder (1982), or the assumption that the tail of that distribution decreases at a fast enough rate, Heckman and Singer (1984b). 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 semi-parametric 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 $\Lambda(t,\alpha_0) = 1$ 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 in 0 in assumption (I1) and assumptions (I2) and (I3). The other assumptions are normalizations (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 in 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 the 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) ). Han's estimator can be used, if the regressors are time constant and if the durations are not censored and the same is true for more recent estimators that are based on rank correlation, e.g. Khan (2001), Chen (2002) and Khan and Tamer (2007). 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 with time varying regressors and general non-informative censoring.

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, 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 $\beta_0$ is unknown and an inverse rank estimate $\hat{\beta}$ of $\beta_0$ is the value of $\beta$ 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).

### 3.1 The Linear Rank Estimator

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

$$S_N(\theta, W) = \sum_{i=1}^{N} \Delta_i \Big\{ W\Big(\tilde{U}_i(\theta), \overline{X}_i^U\left(\tilde{U}_i(\theta)\right)\Big) - \overline{W}\Big(\tilde{U}_i(\theta)\Big) \Big\} \tag{21}$$

with

$$\overline{W}\Big(\tilde{U}_i(\theta)\Big) = \frac{\sum_{j=1}^{N} Y_j^U\left(\tilde{U}_i(\theta)\right) W\Big(\tilde{U}_i(\theta), \overline{X}_j^U\left(\tilde{U}_i(\theta)\right)\Big)}{\sum_{j=1}^{N} Y_j^U\left(\tilde{U}_i(\theta)\right)}$$

In (21) $W$ is a known function of $\tilde{U}_i(\theta)$ and $\overline{X}_i^U\left(\tilde{U}_i(\theta)\right)$ of a dimension that not smaller than that of $\theta$. 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 $\theta_0$.

Because $S_N(\theta, W)$ is not continuous in $\theta$ (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

$$\hat{\theta}_N(W) = \arg\min_{\theta \in \Theta} S_N(\theta, W)' S_N(\theta, W) \tag{22}$$

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

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 $\theta_0$. In large samples this is correct if we choose for instance $W\Big(\tilde{U}_i(\theta), \overline{X}_i^U\left(\tilde{U}_i(\theta)\right)\Big) = \overline{X}_i^U\left(\tilde{U}_i(\theta)\right)\Big)$, because for $\theta = \theta_0$ the transformed duration $U_0$ is independent of $\overline{X}_i^U$. Another choice of $W$ is $W\Big(\tilde{U}_i(\theta), \overline{X}_i^U\left(\tilde{U}_i(\theta)\right)\Big) = I(u_k < \tilde{U}_i(\theta) \leq u_{k+1})$. 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 $W$.

*Example* 2 (Continuation of Example 1). Simple weight functions for this example are

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

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

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

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

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

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 $\theta$ that make e.g. $\tilde{U}_{(k)}(\theta)$ and $\tilde{U}_{(k+1)}(\theta)$ equal. If $\Delta_{(k)} = \Delta_{(k+1)} = 1$, the discontinuity is

$$\frac{X_{(k+1)}\big(\tilde{U}_{(k)}(\theta)\big) - X_{(k)}\big(\tilde{U}_{(k)}(\theta)\big)}{N-k} \tag{23}$$

$$\frac{I\bigg(\tilde{U}_{(k)} \leq e^{\alpha} t_1 \exp\Big[\beta X_{(k+1)}\big(\tilde{U}_{(k)}(\theta)\big)\Big]\bigg) - I\bigg(\tilde{U}_{(k)} \leq e^{\alpha} t_1 \exp\Big[\beta X_{(k)}\big(\tilde{U}_{(k)}(\theta)\big)\Big]\bigg)}{N-k} \tag{24}$$

and this 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 following assumptions. To simplify the expressions we use the notation $h_i(t, \theta) = h(t, \overline{X}_i(t), \theta)$.

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

$$\kappa(t|\overline{X}(t), V, \theta) = \lambda(t, \alpha) e^{\beta' X(t)} V \tag{25}$$

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

(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. We also assume

$$\mathrm{E}\big(V|T \geq t, \overline{Y}(t), \overline{X}(t)\big) \tag{26}$$

The hazard that is not conditional on V is

$$\kappa(t|\overline{X}(t), \theta) = \lambda(t, \alpha) e^{\beta' X(t)} \mathrm{E}\big[V|T \geq t, \overline{Y}(t), \overline{X}(t)\big] \tag{27}$$

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

$$\mathrm{d}I(t \leq C) \perp N(s), s \geq t | \overline{Y}(t), \overline{X}(t) \tag{28}$$

The density of $C$ is bounded.

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

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

$$\overline{W}\Big(\tilde{U}_i(\theta)\Big) = \frac{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big) W\Big(\tilde{U}_i(\theta), \overline{X}_j^U\big(\tilde{U}_i(\theta)\big)\Big)}{\sum_{j=1}^N Y_j^U\big(\tilde{U}_i(\theta)\big)}$$

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

$$\sup_{\theta \in \Theta, u \leq \tau + \psi} \big|\overline{W}(u, \theta) - \mu(u, \theta)\big| \xrightarrow{p} 0 \tag{29}$$

and

$$\sup_{\theta \in \Theta, u \leq \tau + \psi} \left| \frac{1}{N} \sum_{i=1}^N \Big( W\big(u, \overline{X}_i^U(u, \theta)\big) - \overline{W}(u, \theta) \Big) Y_i^U(u, \theta) X_i^U(s, \theta)' - V_\beta(u, s, \theta) \right| \xrightarrow{p} 0 \tag{30}$$

and

$$\sup_{\theta \in \Theta, u \leq \tau + \psi} \left| \frac{1}{N} \sum_{i=1}^N \Big( W\big(u, \overline{X}_i^U(u, \theta)\big) - \overline{W}(u, \theta) \Big) Y_i^U(u, \theta) \frac{\partial \ln \lambda}{\partial \alpha'} \big( h_i^{-1}(s, \theta) - V_\alpha(u, s, \theta) \right| \xrightarrow{p} 0 \tag{31}$$

Define

$$\begin{aligned} B(\theta_0) &= -\int_0^\tau \int_0^u V_\beta(u, s, \theta) \kappa_0'(u)\, ds\, \mathrm{d}u - \int_0^\tau V_\beta(u, u, \theta) \kappa_0(u)\, \mathrm{d}u && (32) \\ A(\theta_0) &= -\int_0^\tau \int_0^u V_\alpha(u, s, \theta) \kappa_0'(u)\, ds\, \mathrm{d}u - \int_0^\tau V_\alpha(u, u, \theta) \kappa_0(u)\, \mathrm{d}u && (33) \end{aligned}$$

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 semi-parametric information bound is nonsingular (see below). Assumption A2 states that the covariates and the observation indicator are predetermined. 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

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

If $\mathrm{E}(V) = \infty$, this density is not bounded in 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 $\tau$ 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 $\tau < \infty$.

The next lemma shows that the linear rank statistic is asymptotically equivalent to a statistic that is linear in the parameters.

**Lemma 1**

Under assumptions (A1)–(A4) for all $C > 0$

$$\sup_{|\theta-\theta_0|\leq CN^{-\frac{1}{2}}} N^{-\frac{1}{2}}\Big|S_N(\theta, W) - \tilde{S}_N(\theta, W)\Big| \xrightarrow{p} 0 \tag{34}$$

with

$$\tilde{S}_N(\theta, W) = \sum_{i=1}^{N} \int_0^{\tau} \Big(W\big(u, \overline{X}_i^U(u, \theta_0)\big) - \overline{W}(u, \theta_0)\Big)\, dM_i^0(u) \\ + B(\theta_0)N(\beta - \beta_0) + A(\theta_0)N(\alpha - \alpha_0) \tag{35}$$

**Proof**: See Appendix.

From Lemma 1 we obtain the asymptotic distribution of the LRE

**Theorem 1**

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

$$\sqrt{N}(\hat{\theta}_N - \theta_0) \xrightarrow{d} \mathcal{N}\big(0, D(\theta_0)^{-1}V(\theta_0)D'(\theta_0)^{-1}\big) \tag{36}$$

with

$$\frac{1}{N}\sum_{i=1}^{N} \int_0^{\tau} \Big(W\big(u, \overline{X}_i^U(u, \theta_0)\big) - \overline{W}(u, \theta_0)\Big)\Big(W\big(u, \overline{X}_i^U(u, \theta_0)\big) - \overline{W}(u, \theta_0)\Big)' \cdot \\ \cdot\, Y_i^U(u, \theta_0)\kappa_0(u)\, \mathrm{d}u \xrightarrow{p} V(\theta_0) \tag{37}$$

**Proof**: By van der Vaart (1998), Theorem 5.45 we have from Lemma 1

$$\sqrt{N}(\hat{\theta}_N - \theta_0) = D(\theta_0)^{-1}\frac{1}{\sqrt{N}}\int_0^\tau\Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big)\mathrm{d}M_{i0}$$

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. Andersen et al. (1993)), the sum on the right-hand side converges to a normal distribution with the variance matrix in (37).

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

$$\left[\frac{1}{N}\sum_{i=1}^N\int_0^\tau\big(W_i(u) - \overline{W}(u,\theta_0)\big)\frac{\partial\ln\kappa_{U_i}}{\partial\theta'}Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right]^{-1}\cdot$$
$$\left[\frac{1}{N}\sum_{i=1}^N\int_0^\tau\big(W_i(u) - \overline{W}(u,\theta_0)\big)\big(W_i(u) - \overline{W}(u,\theta_0)\big)'Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right]\cdot$$
$$\left[\frac{1}{N}\sum_{i=1}^N\int_0^\tau\big(W_i(u) - \overline{W}(u,\theta_0)\big)\frac{\partial\ln\kappa_{U_i}}{\partial\theta'}Y_i^U(u,\theta_0)\kappa_0(u)\,\mathrm{d}u\right]^{\prime -1} \quad (38)$$

By the Cauchy-Schwartz inequality this matrix in minimal if

$$W_{0i}\big(u,\overline{X}_i^U(u,\theta_0)\big) = \frac{\partial\ln\kappa_U\big(u|\overline{X}_i^U(u,\theta_0)\big)}{\partial\theta} \quad (39)$$

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

$$\frac{1}{N}\sum_{i=1}^N\int_0^\tau\big(W_{0i}(u) - \overline{W}_0(u,\theta_0)\big)\big(W_{0i}(u) - \overline{W}_0(u,\theta_0)\big)'\,\mathrm{d}N(u) \quad (40)$$

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

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.1) and (B.2)

$$\begin{aligned}\frac{\partial\ln\kappa_U(u,\theta)}{\partial\alpha} &= -\frac{\kappa_0'(u)}{\kappa_0(u)}\int_0^u\frac{\partial\ln\lambda}{\partial\alpha}\big(h_0^{-1}(s),\alpha_0\big)\mathrm{d}s - \quad (41)\\ &\quad - \frac{\partial\ln\lambda}{\partial\alpha}\big(h_0^{-1}(u),\alpha_0\big)\\ \frac{\partial\ln\kappa_U(u,\theta)}{\partial\beta} &= -\frac{\kappa_0'(u)}{\kappa_0(u)}\int_0^u X\big(h_0^{-1}(s)\big)\mathrm{d}s - X\big(h_0^{-1}(u)\big) \quad (42)\end{aligned}$$

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.1) and (B.2) 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 and this is a feasible, but in general suboptimal choice. Because the optimal weighting function factorizes the optimal linear rank statistic is a weighted version of the linear rank statistic based on $W_1$.

The factor in $W_0$ depends on the distribution of $V$. If $V$ has a Gamma distribution with mean 1 and variance $\sigma^2$, then

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

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 $\theta_0$ and the hazard of $U_0$, $\kappa_0$ explicit. With this change, the LRE estimating equation is

$$S_N(\theta,W)=\sum_{i=1}^{N}\Delta_i\Big\{W_i\big(\tilde U_i(\theta),\theta_0,\kappa_0\big)-\overline{W}\big(\tilde U_i(\theta),\theta_0,\kappa_0\big)\Big\} \tag{43}$$

with

$$\overline{W}\big(\tilde U_i(\theta),\theta_0,\kappa_0\big)=\frac{\sum_{j=1}^{N}Y_j^U\big(\tilde U_i(\theta)\big)W_j\big(\tilde U_i(\theta),\theta_0,\kappa_0\big)}{\sum_{j=1}^{N}Y_j^U\big(\tilde U_i(\theta)\big)}$$

The optimal weight functions are given in (41) and (42). 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 $\kappa_0$ 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 Lai 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 $\alpha, \beta$ 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 (41) and (42) 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

$$S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) = \sum_{i=1}^{N_2} \Delta_i \Big\{ W_i\big(\tilde{U}_{2i}(\theta), \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1}\big) - \overline{W}\big(\tilde{U}_{2i}(\theta), \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1}\big) \Big\} \tag{44}$$

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

$$S_N\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2}), W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) = S_{1N_1}\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2})\big) + S_{2N_2}\big(\theta, W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) \tag{45}$$

equal to 0 or because the $S_N$ is a step function the efficient LRE is defined by

$$\hat{\theta}_N(W) = \arg\min_{\theta \in \Theta} \Big| S_N\big(\theta, W(\cdot, \hat{\theta}_{N_2}, \hat{\kappa}_{0N_2}), W(\cdot, \hat{\theta}_{N_1}, \hat{\kappa}_{0N_1})\big) \Big|^2 \tag{46}$$

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 $\kappa_0$ based on the estimated transformed durations (see Lemma 2 and 3). We need to know the uniform rate of convergence because we need to modify the nonparametric hazard estimator to avoid a 0 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

$$\hat{\kappa}_N(u, \theta_0) = \frac{1}{b_N} \sum_{i=1}^{N} \Delta_i \frac{I\big(Y_N^U(\tilde{U}_i(\theta_0), \theta_0) > 0\big)}{\overline{Y}_N^U(\tilde{U}_i(\theta_0), \theta_0)} K\bigg(\frac{u - \tilde{U}_i(\theta_0)}{b_N}\bigg) \tag{47}$$

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

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**

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

$$\inf_{0 \le u \le \tau} b_N^2 N^{1-\epsilon} \overline{Y}_N^U(u, \theta_0) \xrightarrow{p} \infty \tag{48}$$

and

$$b_N N^{1-\epsilon} \to \infty \tag{49}$$

we have

$$\sup_{u_1 \le u \le u_2} N^{\epsilon} \big| \hat{\kappa}_N(u, \theta_0) - \kappa_0(u) \big| \xrightarrow{p} 0 \tag{50}$$

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

If $\overline{Y}_N(t)$ bounded from 0 on $[0, \tau]$ for large $N$, then (48) and (49) imply that if $b_N = N^{-c}, \epsilon < c < \frac{1}{2} - \epsilon$, and hence $\epsilon < \frac{1}{4}$. 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.

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

$$\hat{\kappa}_N(u, \hat{\theta}_N) = \frac{1}{b_N} \sum_{i=1}^{N} \Delta_i \frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N) > 0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N), \hat{\theta}_N)} K\bigg(\frac{u - \tilde{U}_i(\hat{\theta}_N)}{b_N}\bigg) \tag{51}$$

We have

**Lemma 3**

The kernel $K$ is positive and bounded on $[-1,1]$ (0 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 $\alpha$ in an open neighborhood of $\alpha_0$. Moreover

$$\frac{I\big(Y_N^U(u,\theta)>0\big)}{\overline{Y}_N^U(u,\theta)} \xrightarrow{p} H(u,\theta) \tag{52}$$

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

$$b_N^2 N^{\frac{1}{2}-\epsilon}\to\infty \tag{53}$$

we have

$$\sup_{0\le u\le\tau} N^{\epsilon}\big|\hat{\kappa}_N(u,\hat{\theta}_N)-\hat{\kappa}_N(u,\theta_0)\big| \xrightarrow{p} 0 \tag{54}$$

**Proof**: See Appendix.

Note that the conditions on $b_N$ are determined in Lemma 2. 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 (52) more closely. With $h(T,\theta)=\int_0^T\lambda(t,\alpha)e^{\beta' X(t)}\mathrm{d}t$ we have if the duration $T$ is (right) censored at $C$ that $Y(t)=I(T\ge t)I(C\ge t)$ so that

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

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

$$I(T\ge t)\perp I(C\ge t)\big|Y(s),X(t),0\le s\le t \tag{55}$$

then also

$$I\big(h(T,\theta)\ge u\big)\perp I\big(h(C,\theta)\ge u\big)\big|Y^U(s),X^U(t),0\le s\le u \tag{56}$$

We have with $N(\theta_0)$ an open neighborhood of $\theta_0$, and if $X_i$ and $C_i$ i.i.d. and

$$\sup_{\theta\in N(\theta_0) 0\le u\le\tau}\Pr\big(h(T,\theta)<u\big)<1 \tag{57}$$

$$\sup_{\theta\in N(\theta_0) 0\le u\le\tau}\Pr\big(h(C,\theta)<u\big)<1 \tag{58}$$

that

$$\inf_{\theta\in N(\theta_0)0\leq u\leq\tau} I\big(Y_N^U(u,\theta)>0\big)\xrightarrow{p}0 \tag{59}$$

and by the uniform law of large numbers

$$\overline{Y}_N^U(u,\theta)\xrightarrow{p}\Pr\big(h(T,\theta)\geq u\big)\cdot\Pr\big(h(C,\theta)\geq u\big) \tag{60}$$

uniformly for $\theta\in N(\theta_0)0\leq u\leq\tau$. Because by (57) the limit is bounded from 0, we have

$$\frac{I\big(Y_N^U(u,\theta)>0\big)}{\overline{Y}_N^U(u,\theta)}\xrightarrow{p}H(u,\theta) \tag{61}$$

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

$$H(u,\theta)=\frac{1}{\Pr\big(h(T,\theta)\geq u\big)\cdot\Pr\big(h(C,\theta)\geq u\big)} \tag{62}$$

Because $h(T,\theta_0)=U_0$ (53) 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 (13) a sufficient condition for $\kappa_U(u,\theta)$ being bounded for all $\theta$ in a neighborhood of $\theta_0$ and $0\leq t\leq\tau$ is that $\lambda(t,\alpha)>0$ for all $t$ and on a neighborhood of $\alpha_0$. In the same way (54) holds if the hazard of $C$ is bounded and $\lambda(t,\alpha)$ is bounded from 0 on a neighborhood of $\alpha_0$.

## 5 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 it is not present in de DGP. The LRE does not suffer from this misspecification.

### 5.1 Sample design

We try to resemble the simulation experiments by Baker and Melino (2000), who choose a 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 is

$$\lambda(t|X_i)=\exp(X_i\beta+\beta_0) \tag{63}$$

where $X$ is normally distributed with mean zero and variance 0.5. The true value of the regression parameter, $\beta$, is 1. The true value of the intercept, $\beta_0$, is ln(0.05). The variance of $X$ and the regression parameter determine the relative importance of the observed heterogeneity and they determine how accurate we can estimate $\beta$ 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 GDP 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 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

$$\lambda_0(t) = \sum_{k=1}^{K} e^{\alpha_k} I_k(t) \tag{64}$$

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, $\alpha_1$, is fixed to zero. The remaining $\alpha$'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 a Mixed proportional hazards models with a discrete unobserved heterogeneity with 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).[1] The other approach estimates

[1] In the MLE for models with duration dependence we do not need the standard identification restriction that

the NPMLE of Heckman and Singer (1984b) where the number of support points is determined by the Gateaux derivative. Note that multiplicative unobserved heterogeneity does not influence the LRE procedure.

For the LRE we use the most simple weight-functions, $X_i$ for $\beta$ and the interval indicator on the transformed time-scale, $I_k(u) = I\big(m_{k-1}(X,t) \leq u < m_{k-1}(X,t)\big)$ for $\alpha_k$, with $m_k(X,t) = e^{\beta X} \int_0^{t_k} \lambda(s)\, ds$. These weight-functions might be inefficient but it simplifies the estimation. In Section 5.3 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 (22). 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. One of the alternative methods for finding the roots of a non–differentiable function is the Powell-method. This method (see Press et al. (1986, §10.5) and Powell (1964)) is a multidimensional version of the Brent algorithm.

The Powell-method does not always stop at a parameter value that makes the S-statistic was close to zero. 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. Thus, when the Powell method stopped before reaching convergence we reiterated the method untill convergence is found.

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 observations of the generated observations.

For each of the alternative duration dependences and each sample size we apply four different estimation procedure: 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.

the unobserved heterogeneity term has mean one, because the restricted the baseline hazard in the first interval.

### 5.2 Monte Carlo Results

In Table 1 we report the average bias and standard deviation of the average for the estimates of $\beta$ in the 36 experimental settings.[2] For each of the 3 sample sizes we took the 100 simulated samples and estimated $\beta$ using each of the three alternative duration dependence specifications and the four different estimation procedures.[3]

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 $\beta$ 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 $\beta$ 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 $\beta$ despite assuming duration dependence that is not present.

If $\beta$ 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 $\beta$ 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, non-existing, duration dependence well, although at the expense of efficiency loss.

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

[3] 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 $\beta$ 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[a] | -0.0161 | -0.0124 | -0.0040 |
| | | (0.0247) | (0.0192) | (0.0092) |

$^*p < 0.05$

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

### 5.3 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 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, $\beta$, is 1. The variance of the gamma mixture is 0.75. For each GDP we create 100 samples of 1000 observations and stored it. 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 $\beta$ and the interval indicator on the transformed time-scale, $I_k(u)$ for $\alpha_k$. 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 (41) and (42). 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 to obtain these functionals. The efficient LRE is now obtained from the combined estimation equation (45) and equal is given in (46), see Section 4.

Table 2: Average bias, standard error and RMSE of estimates of $\beta$ 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 $\beta$ in the 4 experimental settings. Table 3 gives the results for the censored sample.[4] The results indicate that ignoring the unobserved heterogeneity leads to a severe bias. Using a 2 point discrete unobserved heterogeneity distribution to approximate the

[4] The results for the parameters of the piecewise constant duration dependence, $\alpha_2$ and $\alpha_3$, are given in Table A.3 and Table A.4 in Appendix A.

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.

The NPMLE is more efficient than the rank estimators. However, for two of the four DGP's the RMSE of the NPMLE is higher. In particular, for both a 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 $\beta$ 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.
$^*p < 0.05$

# 6 Empirical Application

Between mid–1984 and mid–1985, the Illinois Department of Employment Security conducted two controlled social experiments. These experiments were conducted to evaluate the potential of using cash bonus offers to induce early return to work by unemployment insurance (UI) claimants[5]. These experiments provide the opportunity to explore, within a controlled experimental setting, whether bonuses paid to Unemployment Insurance (UI) beneficiaries or their employers reduce the unemployment of beneficiaries relative to a randomly selected control group. Both treatments consisted of a \$500 bonus payment, which was about four times the average weekly unemployment insurance benefit.

In another article we focus on estimating the effects of these bonus payments on the duration of unemployment in an MPH (Bijwaard and Ridder (2005) and Bijwaard (2009)). The extra complication for the analysis of these treatment effects is that some of the UI claimants did not comply with their assigned treatment. They were free to choose not to become eligible for the bonus. The choice whether or not to comply may depend on unobserved characteristics that also influence the duration. Then the censoring times are not independent of $U_0$ anymore for all observed transformed durations In the articles mentioned above we explain how we can solve this problem.

Here, we only use the data on those people who were assigned to the control group. This group consisted of 3952 individuals, who were excluded from participation in the experiment. In fact, they did not know that the experiment took place. We shall estimate the parameters of an MPH model for these data using Linear Rank Estimators and an NPMLE. The efficient LRE is obtained using the steps described in Section 4. We include the following (all time–invariant) explanatory variables: age and age squared, the logarithm of the pre–unemployment earnings (LNBPE), gender (MALE= 1), ethnicity (BLACK= 1), and the logarithm of the weekly amount of UI benefits plus dependence allowance (LNBEN). Thus, we have six regression parameters to estimate.

We assume that the duration dependence can be approximated by a piecewise constant function. The maximum unemployment duration in our sample is 26 weeks. We assume the

[5] A complete description of the experiment and a summary of its results can be found in Woodbury and Spiegelman (1987).

hazard is constant on each two–week interval. The last interval is the reference interval, i.e. $\alpha_{13} = 0$.

The results are presented in Table 4. The re-employment hazard is the lowest at age 44. Blacks have a lower and males (not significant for the optimal LRE) a higher re-employment hazard. Higher pre-employment earnings increase the hazard and higher dependence allowance decrease the hazard. For the NPMLE we could not find an indication of unobserved heterogeneity. Thus the results from the NPMLE do not differ from the results of a PH-model. This may indicate that unobserved heterogeneity is only a minor issue in these data. However, as Bijwaard and Ridder (2005) point out, even in large samples inference on the unobserved heterogeneity using the NPMLE is inaccurate. We find that the NPMLE differs substantially from the LRE. The NPMLE seems to overestimate (in absolute value) the effect of the covariates. The U-shaped duration dependence is more pronounced for the LRE and the optimal LRE.

Table 4: Linear Rank estimates for the regression coefficients of the control group of the Illinois data

| | NPMLE | LRE | optimal LRE |
|---|---|---|---|
| age | $-0.1598^*$ | $-0.1188^*$ | $-0.0968^*$ |
| | (0.0346) | (0.0302) | (0.0257) |
| age-squared | $0.0720^*$ | $0.0541^*$ | $0.0448^*$ |
| | (0.0280) | (0.0225) | (0.0195) |
| LNBPE | $0.2494^*$ | $0.1830^*$ | $0.1480^*$ |
| | (0.0700) | (0.0558) | (0.0480) |
| Black | $-0.5216^*$ | $-0.3758^*$ | $-0.3188^*$ |
| | (0.0849) | (0.0777) | (0.0676) |
| Male | $0.1026^*$ | $0.0744^*$ | 0.0564 |
| | (0.0454) | (0.0359) | (0.0309) |
| LNBEN | $-0.4886^*$ | $-0.3598^*$ | $-0.2961^*$ |
| | (0.1211) | (0.1024) | (0.0880) |
| *duration dependence* | | | |
| $\alpha_1$ (0–2 weeks) | $-0.4789^*$ | $-0.4783^*$ | $-0.5214^*$ |
| | (0.1153) | (0.1700) | (0.1699) |
| $\alpha_2$ (2–4 weeks) | $-0.6525^*$ | $-0.9161^*$ | $-1.0662^*$ |
| | (0.1560) | (0.2177) | (0.2138) |
| $\alpha_3$ (4–6 weeks) | $-0.7296^*$ | $-1.1278^*$ | $-1.3048^*$ |
| | (0.1890) | (0.2466) | (0.2393) |
| $\alpha_4$ (6–8 weeks) | $-0.8085^*$ | $-1.2742^*$ | $-1.4951^*$ |
| | (0.2186) | (0.2683) | (0.2586) |
| $\alpha_5$ (8–10 weeks) | $-0.9378^*$ | $-1.4367^*$ | $-1.6605^*$ |
| | (0.2435) | (0.2854) | (0.2739) |
| $\alpha_6$ (10–12 weeks) | $-0.8814^*$ | $-1.4115^*$ | $-1.6707^*$ |
| | (0.2639) | (0.3003) | (0.2871) |
| $\alpha_7$ (12–14 weeks) | $-1.0729^*$ | $-1.6317^*$ | $-1.8864^*$ |
| | (0.2806) | (0.3134) | (0.2985) |
| $\alpha_8$ (14–16 weeks) | $-1.0455^*$ | $-1.6380^*$ | $-1.8967^*$ |
| | (0.2940) | (0.3241) | (0.3082) |
| $\alpha_9$ (16–18 weeks) | $-0.9847^*$ | $-1.6084^*$ | $-1.9253^*$ |
| | (0.3090) | (0.3350) | (0.3176) |
| $\alpha_{10}$ (18–20 weeks) | $-0.7121^*$ | $-1.3741^*$ | $-1.7078^*$ |
| | (0.3181) | (0.3447) | (0.3249) |
| $\alpha_{11}$ (20–22 weeks) | $-0.8654^*$ | $-1.5498^*$ | $-1.8852^*$ |
| | (0.3321) | (0.3584) | (0.3351) |
| $\alpha_{12}$ (22–24 weeks) | $-1.4938^*$ | $-2.1748^*$ | $-2.4569^*$ |
| | (0.3312) | (0.3614) | (0.3388) |

Standard error in brackets. The age is centered by its mean value (33) and divided by ten. Both LNBPE and LNBEN are centered by their mean value. $^*p < 0.05$

# 7 Conclusion

In this paper we have discussed and implemented a simple $\sqrt{N}$ consistent estimator for the parameters of a semi-parametric MPH model with 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 the LRE provides unbiased estimates of the regression coefficients despite assuming nonexistent duration dependence.

# References

Aalen, O. O., O. Borgan, and H. K. Gjessing (2009). *Survival and Event History Analysis.* Springer–Verlag.

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

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

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

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.

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

Bearse, P., J. Canals-Cerdá, and P. Rilstone (2007). Efficient semiparametric estimation of duration models with unobserved heterogeneity. *Econometric Theory 23*, 281–308.

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.

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

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

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

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

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.

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

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

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

Hausman, J. A. and T. M. Woutersen (2005). Estimating a semi–parametric duration model without specifying heterogeneity. C*e*MMAP, working paper, CWP11/05.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Nielsen, G. G., R. D. Gill, P. K. Andersen, and T. A. I. Sørensen (1992). A counting process approach to maximum likelihood estimation in frailty models. *Scandinavian Journal of Statistics 19*, 25–43.

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.

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

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

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

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

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

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

Therneau, T. and P. Grambsch (2000). *Modeling Survival Data: Extending the Cox Model.* Springer–Verlag.

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

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge UP.

Woodbury, S. A. and R. G. Spiegelman (1987). Bonuses to workers and employers to reduce unemployment: Randomized trials in Illinois. *American Economic Review 77*, 513–530.

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 UP.

Woutersen, T. M. (2000). Consistent estimators for panel duration data with endogenous censoring and endogenous regressors. memeo Brown University.

# A Appendix: Additional tables

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

| Estimation method | | Sample size | | |
|---|---|---|---|---|
| | | 500 | 1000 | 5000 |
| MLE no hetero | $\alpha_2$ | $-0.0480^*$ | $-0.0319^*$ | $-0.0095^*$ |
| | | (0.0150) | (0.0103) | (0.0042) |
| | $\alpha_3$ | $-0.0082$ | $-0.0127$ | $-0.0094^*$ |
| | | (0.0132) | (0.0088) | (0.0041) |
| | $\alpha_4$ | $-0.0149$ | $-0.0102$ | $-0.0079$ |
| | | (0.0127) | (0.0089) | (0.0046) |
| MLE 2 points | $\alpha_2$ | 0.0282 | 0.0257 | $0.0140^*$ |
| | | (0.0194) | (0.0158) | (0.0053) |
| | $\alpha_3$ | $0.1131^*$ | $0.0713^*$ | $0.0257^*$ |
| | | (0.0237) | (0.0175) | (0.0064) |
| | $\alpha_4$ | $0.1480^*$ | $0.1013^*$ | $0.0438^*$ |
| | | (0.0273) | (0.0213) | (0.0076) |
| NPMLE | $\alpha_2$ | $0.0785^*$ | $0.0495^*$ | $0.0211^*$ |
| | | (0.0210) | (0.0152) | (0.0050) |
| | $\alpha_3$ | $0.2011^*$ | $0.1207^*$ | $0.0389^*$ |
| | | (0.0275) | (0.0183) | (0.0059) |
| | $\alpha_4$ | $0.2835^*$ | $0.1782^*$ | $0.0612^*$ |
| | | (0.0339) | (0.0228) | (0.0079) |
| LRE | $\alpha_2$ | $-0.0333$ | $-0.0234$ | $-0.0074$ |
| | | (0.0230) | (0.0184) | (0.0066) |
| | $\alpha_3$ | 0.0391 | 0.0158 | $-0.0087$ |
| | | (0.0306) | (0.0224) | (0.0093) |
| | $\alpha_4$ | 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 $\alpha$'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 | | |
| $\alpha_2$ | −0.0240 | −0.0098 | 0.0068 | 0.0704* | 0.0498* | 0.0464* |
| | (0.0216) | (0.0153) | (0.0063) | (0.0230) | (0.0176) | (0.0080) |
| $\alpha_3$ | −0.0162 | −0.0089 | −0.0090 | 0.1096* | 0.0740* | 0.0420* |
| | (0.0241) | (0.0157) | (0.0061) | (0.0283) | (0.0195) | (0.0086) |
| $\alpha_4$ | −0.0609* | −0.0378* | −0.0069 | 0.0958* | 0.0627* | 0.0590* |
| | (0.0207) | (0.0135) | (0.0054) | (0.0273) | (0.0204) | (0.0098) |
| $\alpha_5$ | 0.0073 | −0.0035 | −0.0115 | 0.1991* | 0.1229* | 0.0690* |
| | (0.0206) | (0.0144) | (0.0069) | (0.0305) | (0.0231) | (0.0117) |
| $\alpha_6$ | −0.0097 | −0.0024 | −0.0059 | 0.1986* | 0.1348* | 0.0766* |
| | (0.0207) | (0.0127) | (0.0067) | (0.0340) | (0.0226) | (0.0123) |
| $\alpha_7$ | −0.0593* | −0.0464* | −0.0074 | 0.1617* | 0.0971* | 0.0823* |
| | (0.0226) | (0.0154) | (0.0072) | (0.0364) | (0.0269) | (0.0135) |
| $\alpha_8$ | −0.0144 | −0.0130 | −0.0023 | 0.2161* | 0.1491* | 0.0963* |
| | (0.0204) | (0.0151) | (0.0070) | (0.0360) | (0.0277) | (0.0141) |
| $\alpha_9$ | −0.0209 | −0.0076 | −0.0120 | 0.2309* | 0.1616* | 0.0964* |
| | (0.0243) | (0.0149) | (0.0075) | (0.0388) | (0.0284) | (0.0137) |
| $\alpha_{10}$ | −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[a] | | |
| $\alpha_2$ | 0.1790* | 0.1157* | 0.0703* | −0.0648* | −0.0460* | 0.0088 |
| | (0.0267) | (0.0184) | (0.0088) | (0.0298) | (0.0221) | (0.0106) |
| $\alpha_3$ | 0.3039* | 0.1880* | 0.0871* | −0.0784 | −0.0664* | −0.0070 |
| | (0.0397) | (0.0239) | (0.0099) | (0.0446) | (0.0315) | (0.0136) |
| $\alpha_4$ | 0.3730* | 0.2298* | 0.1181* | −0.1236* | −0.0942* | −0.0041 |
| | (0.0466) | (0.0298) | (0.0120) | (0.0514) | (0.0387) | (0.0166) |
| $\alpha_5$ | 0.5390* | 0.3248* | 0.1372* | −0.0554 | −0.0605 | −0.0093 |
| | (0.0554) | (0.0343) | (0.0146) | (0.0599) | (0.0443) | (0.0203) |
| $\alpha_6$ | 0.5848* | 0.3649* | 0.1573* | −0.0716 | −0.0617 | −0.0050 |
| | (0.0583) | (0.0383) | (0.0151) | (0.0646) | (0.0496) | (0.0220) |
| $\alpha_7$ | 0.5910* | 0.3554* | 0.1692* | −0.1230 | −0.1079* | −0.0078 |
| | (0.0646) | (0.0413) | (0.0170) | (0.0698) | (0.0530) | (0.0245) |
| $\alpha_8$ | 0.6916* | 0.4232* | 0.1884* | −0.0844 | −0.0792 | −0.0042 |
| | (0.0678) | (0.0429) | (0.0179) | (0.0782) | (0.0570) | (0.0258) |
| $\alpha_9$ | 0.7346* | 0.4594* | 0.1918* | −0.0921 | −0.0819 | −0.0157 |
| | (0.0734) | (0.0441) | (0.0191) | (0.0782) | (0.0578) | (0.0278) |
| $\alpha_{10}$ | 0.7758* | 0.4816* | 0.2123* | −0.1230 | −0.1038 | −0.0117 |
| | (0.0736) | (0.0486) | (0.0209) | (0.0803) | (0.0637) | (0.0309) |

[a] 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 | $\alpha_2$ | 0.0069 | 0.0096 | 0.0118 |
| | | $\alpha_3$ | −0.0149 | 0.0206 | 0.0255 |
| | NPMLE | $\alpha_2$ | 0.0205 | 0.0157 | 0.0258 |
| | | $\alpha_3$ | 0.0091 | 0.0283 | 0.0298 |
| | LRE | $\alpha_2$ | −0.0130 | 0.0200 | 0.0238 |
| | | $\alpha_3$ | −0.0645 | 0.0329 | 0.0724 |
| | LRE-opt | $\alpha_2$ | −0.0134 | 0.0195 | 0.0236 |
| | | $\alpha_3$ | −0.0533 | 0.0327 | 0.0625 |
| negative duration dependence | MLE gamma | $\alpha_2$ | 0.0211 | 0.0111 | 0.0239 |
| | | $\alpha_3$ | 0.0553* | 0.0229 | 0.0598 |
| | NPMLE | $\alpha_2$ | 0.0345* | 0.0174 | 0.0386 |
| | | $\alpha_3$ | 0.1079* | 0.0310 | 0.1123 |
| | LRE | $\alpha_2$ | 0.0369* | 0.0179 | 0.0410 |
| | | $\alpha_3$ | 0.0643* | 0.0315 | 0.0716 |
| | LRE-opt | $\alpha_2$ | 0.0358* | 0.0178 | 0.0400 |
| | | $\alpha_3$ | 0.0627* | 0.0314 | 0.0701 |
| U-shaped duration dependence | MLE gamma | $\alpha_2$ | −0.0009 | 0.0097 | 0.0097 |
| | | $\alpha_3$ | −0.0338* | 0.0173 | 0.0379 |
| | NPMLE | $\alpha_2$ | 0.0385* | 0.0155 | 0.0416 |
| | | $\alpha_3$ | 0.0149 | 0.0251 | 0.0292 |
| | LRE | $\alpha_2$ | 0.0334 | 0.0186 | 0.0383 |
| | | $\alpha_3$ | −0.0215 | 0.0271 | 0.0346 |
| | LRE-opt | $\alpha_2$ | 0.0261 | 0.0183 | 0.0319 |
| | | $\alpha_3$ | −0.0247 | 0.0263 | 0.0361 |
| inverse U duration dependence | MLE gamma | $\alpha_2$ | 0.0102 | 0.0104 | 0.0146 |
| | | $\alpha_3$ | −0.0047 | 0.0232 | 0.0237 |
| | NPMLE | $\alpha_2$ | 0.0232 | 0.0140 | 0.0271 |
| | | $\alpha_3$ | 0.0327 | 0.0295 | 0.0440 |
| | LRE | $\alpha_2$ | 0.0335 | 0.0183 | 0.0381 |
| | | $\alpha_3$ | 0.0400 | 0.0336 | 0.0522 |
| | LRE-opt | $\alpha_2$ | 0.0321 | 0.0182 | 0.0369 |
| | | $\alpha_3$ | 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 | $\alpha_2$ | 0.0010 | 0.0135 | 0.0135 |
| | | $\alpha_3$ | −0.0267 | 0.0269 | 0.0379 |
| | NPMLE | $\alpha_2$ | 0.0120 | 0.0177 | 0.0213 |
| | | $\alpha_3$ | −0.0204 | 0.0310 | 0.0371 |
| | LRE | $\alpha_2$ | −0.0148 | 0.0199 | 0.0248 |
| | | $\alpha_3$ | −0.0656* | 0.0329 | 0.0734 |
| | LRE-opt | $\alpha_2$ | −0.0138 | 0.0199 | 0.0242 |
| | | $\alpha_3$ | −0.0599 | 0.0328 | 0.0683 |
| negative duration dependence | MLE gamma | $\alpha_2$ | 0.0347* | 0.0131 | 0.0371 |
| | | $\alpha_3$ | 0.0633* | 0.0277 | 0.0691 |
| | NPMLE | $\alpha_2$ | 0.0417* | 0.0184 | 0.0456 |
| | | $\alpha_3$ | 0.0898* | 0.0325 | 0.0956 |
| | LRE | $\alpha_2$ | 0.0378* | 0.0182 | 0.0420 |
| | | $\alpha_3$ | 0.0539 | 0.0329 | 0.0631 |
| | LRE-opt | $\alpha_2$ | 0.0375* | 0.0181 | 0.0416 |
| | | $\alpha_3$ | 0.0501 | 0.0327 | 0.0598 |
| U-shaped duration dependence | MLE gamma | $\alpha_2$ | 0.0052 | 0.0133 | 0.0143 |
| | | $\alpha_3$ | −0.0269 | 0.0225 | 0.0350 |
| | NPMLE | $\alpha_2$ | 0.0308 | 0.0173 | 0.0353 |
| | | $\alpha_3$ | −0.0159 | 0.0292 | 0.0333 |
| | LRE | $\alpha_2$ | 0.0266 | 0.0184 | 0.0323 |
| | | $\alpha_3$ | −0.0321 | 0.0254 | 0.0410 |
| | LRE-opt | $\alpha_2$ | 0.0263 | 0.0182 | 0.0320 |
| | | $\alpha_3$ | −0.0315 | 0.0253 | 0.0404 |
| inverse U duration dependence | MLE gamma | $\alpha_2$ | 0.0137 | 0.0123 | 0.0184 |
| | | $\alpha_3$ | −0.0030 | 0.0263 | 0.0264 |
| | NPMLE | $\alpha_2$ | 0.0183 | 0.0149 | 0.0236 |
| | | $\alpha_3$ | 0.0283 | 0.0305 | 0.0416 |
| | LRE | $\alpha_2$ | 0.0340 | 0.0185 | 0.0387 |
| | | $\alpha_3$ | 0.0360 | 0.0335 | 0.0491 |
| | LRE-opt | $\alpha_2$ | 0.0313 | 0.0183 | 0.0363 |
| | | $\alpha_3$ | 0.0290 | 0.0333 | 0.0441 |

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

# B Appendix: Proofs

## B.1 Proof of Lemma 1

$\tilde{S}_N(\theta)$ is a linearization of $\tilde{S}_N(\theta)$. Because $S_N(\theta)$ is not continuous in $\theta$ 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 (8) and (9) 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 $U_0$

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

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 $\theta$ are

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

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

$$\begin{aligned}\frac{\partial \kappa_U(u, \theta)}{\partial \beta}\Big|_{\theta=\theta_0} &= -\kappa_0'(u)\int_0^{h_0^{-1}(u)}\lambda(t, \alpha_0)e^{\beta_0'X(t)}\mathrm{d}t - \kappa_0(u)X\big(h_0^{-1}(u)\big) \\ &= \kappa_0'(u)\int_0^{u}X\big(h_0^{-1}(s), \alpha_0\big)\mathrm{d}s - \kappa_0(u)X\big(h_0^{-1}(u)\big) \qquad \text{(B.2)}\end{aligned}$$

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 $W\big(u, \overline{X}_i^U(u, \theta)\big)$ and hence the requirement that this is a vector of bounded functions. The equations (29), (30) and (31) are stability conditions (see also Andersen and Gill (1982)). 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 $u \leq \tau$

$$\left|\kappa_U(u,\theta) - \kappa_0(u) - \frac{\partial \kappa_U}{\partial \theta'}(u,\theta_0)(\theta-\theta_0)\right| \leq |\theta-\theta_0|^2 h(u) \tag{B.3}$$

The assumptions that $\lambda(t,\alpha)$ is bounded from 0 for all $t \geq 0$ and $\alpha$ in the parameter space, that $\left|\frac{\partial^2 \lambda}{\partial\alpha\partial\alpha'}(t,\alpha)\right| < \infty$ for all $t \geq 0$ and $\alpha$ in the parameter space, and that $X(t)$ is bounded, imply that the second derivative of $\kappa_U(u,\theta)$ with respect to $\theta$ is bounded for all $u \leq \tau$ and $\theta \in \Theta$. This is sufficient for (B.3) 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

$$\begin{aligned} S_N(\theta) \approx \sum_{i=1}^N \int_0^\tau \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big)\, dM_i^0(u) + \\ + \Big[\sum_{i=1}^N \int_0^\tau \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big) Y_i^0(u) \frac{\partial \kappa_{U_i}}{\partial \theta'}(u,\theta_0)\mathrm{d}u\Big](\theta-\theta_0) \end{aligned} \tag{B.4}$$

The second term is after substitution of (B.1), and (B.2)

$$\begin{aligned} -\Big[\int_0^\tau\int_0^u \sum_{i=1}^N \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big) Y_i^0(u) \frac{\partial \ln\lambda}{\partial\alpha'}\big(h_{0i}^{-1}(s),\alpha_0\big)\kappa_0'(u)\mathrm{d}s\mathrm{d}u + \\ + \int_0^\tau \sum_{i=1}^N \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big) Y_i^0(u) \frac{\partial \ln\lambda}{\partial\alpha'}\big(h_{0i}^{-1}(u),\alpha_0\big)\kappa_0(u)\mathrm{d}u\Big](\alpha-\alpha_0) - \\ -\Big[\int_0^\tau\int_0^u \sum_{i=1}^N \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big) Y_i^0(u) X\big(h_{0i}^{-1}(s),\alpha_0\big)\kappa_0'(u)\mathrm{d}s\mathrm{d}u + \\ + \int_0^\tau \sum_{i=1}^N \Big(W\big(u,\overline{X}_i^U(u,\theta_0)\big) - \overline{W}(u,\theta_0)\Big) Y_i^0(u) X\big(h_{0i}^{-1}(u),\alpha_0\big)\kappa_0(u)\mathrm{d}u\Big](\beta-\beta_0) \end{aligned} \tag{B.5}$$

The normalized vectors of coefficients converge to (32) and (33) if (30) and (31) hold. This proves the lemma.

## B.2 Proof of Lemma 2 and 3

We have

$$
\begin{aligned}
&N^\epsilon\big|\hat{\kappa}_N(u,\hat{\theta}_N)-\hat{\kappa}_N(u,\theta_0)\big| \leq \\
&\quad \left|\frac{N^\epsilon}{Nb_N}\sum_{i=1}^N \Delta_i\left(\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}-\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0),\theta_0)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\theta_0),\theta_0)}K\left(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\right)\right|+ \\
&\qquad +\left|\frac{N^\epsilon}{Nb_N}\sum_{i=1}^N \Delta_i\left(K\left(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\right)-K\left(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\right)\right)\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}\right| \qquad \text{(B.6)}
\end{aligned}
$$

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

$$
\frac{CN^\epsilon}{Nb_N^2}\sum_{i=1}^N \Delta_i\big|\tilde{U}_i(\hat{\theta}_N)-\tilde{U}_i(\theta_0)\big|\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)} \qquad \text{(B.7)}
$$

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

$$
\begin{aligned}
\tilde{U}_i(\hat{\theta}_N)-\tilde{U}_i(\theta_0) &= \int_0^{\tilde{T}_i}\lambda(t,\overline{\alpha}_{iN})e^{\overline{\beta}'_{iN}X_i(s)}X_i(s)'\,\mathrm{d}s(\hat{\beta}_N-\beta_0)+ \qquad \text{(B.8)}\\
&\quad +\int_0^{\tilde{T}_i}e^{\overline{\beta}'_{iN}X_i(s)}\frac{\partial\lambda(t,\overline{\alpha}_{iN})}{\partial\alpha'}\,\mathrm{d}s(\hat{\alpha}_N-\alpha_0)
\end{aligned}
$$

Because $X_i(t)$ is bounded on $[0,\tau]$ and so is $\big|\frac{\partial\lambda(t,\alpha)}{\partial\alpha}\big|$ for all $\alpha$ in an open neighborhood of $\alpha_0$, (B.8) is bounded by $|c_1'(\hat{\beta}_N-\beta_0)|+|c_2'(\hat{\alpha}_N-\alpha_0)\big|$ and substitution in (B.7) gives the upperbound

$$
\frac{C}{N}\sum_{i=1}^N \Delta_i\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}\left(\frac{N^\epsilon|c_1'(\hat{\beta}_N-\beta_0)|}{b_N^2}+\frac{N^\epsilon|c_2'(\hat{\alpha}_N-\alpha_0)|}{b_N^2}\right) \qquad \text{(B.9)}
$$

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

Next we consider the first term in (B.6). By subtraction and addition of expected values this term is bounded by

$$
\begin{aligned}
&\Bigg|\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\Bigg[\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)- \\
&\qquad\qquad -\mathrm{E}\Big(\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)\Big|\Delta_i=1\Big)\Bigg]\Bigg|+ \\
&\qquad\qquad +\Bigg|\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\Bigg[\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0),\theta_0)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\theta_0),\theta_0)}K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)- \\
&\qquad\qquad -\mathrm{E}\Big(\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0),\theta_0)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\theta_0),\theta_0)}K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)\Big|\Delta_i=1\Big)\Bigg]\Bigg|+ \\
&\qquad +\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\Bigg|\mathrm{E}\Bigg[\frac{I\big(Y_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N)}K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)\Big|\Delta_i=1\Big)\Bigg]- \\
&\qquad\qquad\qquad \mathrm{E}\Bigg[\frac{I\big(Y_N^U(\tilde{U}_i(\theta_0),\theta_0)>0\big)}{\overline{Y}_N^U(\tilde{U}_i(\theta_0),\theta_0)}K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)\Big|\Delta_i=1\Big)\Bigg]\Bigg|
\end{aligned}
\tag{B.10}
$$

The first and second term converge to 0 in probability if $b_N N^{\frac{1}{2}-\epsilon}\to\infty$. Because of (52) the final term converges in probability to

$$
\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\Bigg|\mathrm{E}\Bigg[H\big(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N\big)K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)\Bigg]-\mathrm{E}\Bigg[H\big(\tilde{U}_i(\theta_0),\theta_0\big)K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)\Bigg]\Bigg| \tag{B.11}
$$

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

$$
\begin{aligned}
&\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\mathrm{E}\Big[\Big|H\big(\tilde{U}_i(\hat{\theta}_N),\hat{\theta}_N\big)-H\big(\tilde{U}_i(\theta_0),\theta_0\big)\Big|\Big]+ \\
&\qquad\qquad +\frac{N^{\epsilon}}{Nb_N}\sum_{i=1}^{N}\Delta_i\mathrm{E}\Big[\Big|K\Big(\frac{u-\tilde{U}_i(\hat{\theta}_N)}{b_N}\Big)K\Big(\frac{u-\tilde{U}_i(\theta_0)}{b_N}\Big)\Big|\Big]
\end{aligned}
\tag{B.12}
$$

The first time 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.