Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

Jerry Hausman and Tiemen Woutersen[*,†]
MIT and University of Arizona

February 2012

Abstract. This paper presents a new estimator for counterfactuals in duration models. The counterfactual in a duration model is the length of the spell in case the regressor would have been different. We introduce the structural duration function, which gives these counterfactuals. The advantage of focusing on counterfactuals is that one does not need to identify the mixed proportional hazard model. In particular, we present examples in which the mixed proportional hazard model is unidentified or has a singular information matrix but our estimator for counterfactuals still converges at rate $N^{1/2}$ where N is the number of observations. We apply the structural duration function to simulate important policy effects, including a change in welfare benefits.

Keywords: Transformation Model, Mixed Proportional Hazard Model, Heterogeneity.
JEL Classification: C41, C24, J64

1. Introduction

The estimation of duration models has been the subject of considerable attention in econometrics since the late seventies. Lancaster (1979) introduced the mixed proportional hazard model in which the hazard rate is a function of a regressor X, unobserved heterogeneity v, and a function of time $\lambda(t)$,

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

[*]Correspondence addresses: Massachusetts Institute of Technology, Department of Economics, Building E52-271A, Cambridge, MA 02139, Department of Economics, Eller College of Management, University of Arizona, P.O. Box 210108, Tucson, AZ 85721; Email: jhausman@mit.edu, woutersen@email.arizona.edu.

[†]We thank Matthew Harding and Ketan Patel for excellent research assistance. We thank Andrew Chesher, Nicole Lott, Pierangelo de Pace, and seminar participants at UCL for helpful comments.

where v and X are independent. The function $\lambda(t)$ is often referred to as the baseline hazard. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard rate, $\theta(t|X)$, to decrease with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, v, and genuine duration dependence, $\lambda(t)$.[1] Lancaster (1979) uses functional form assumptions on $\lambda(t)$ and distributional assumptions on v to identify the model. Examples by Lancaster and Nickell (1980) and Heckman and Singer (1984a), however, show the sensitivity to these functional form and distributional assumptions. Elbers and Ridder (1982) and Heckman and Singer (1984b) show that the mixed proportional hazard model is semi-parametrically identified. It took a long time, however, before an estimator was developed that avoided functional form assumptions on $\lambda(t)$ or distributional assumptions on v. Horowitz (1996) shows how to estimate[2] the closely related transformation model,

$$\begin{aligned} H(t) &= X\kappa + \varepsilon \\ \text{where } |\kappa_1| &= 1. \end{aligned} \tag{2}$$

where X is a vector of regressors. This model does not impose the identifying assumption of Elbers and Ridder (1982) that $Ev < \infty$ or the identifying assumption of Heckman and Singer (1984b) that the right tail of the distribution of v decreases at a known rate. For this reason, normalizing κ is necessary for identification, for example, $\kappa = \beta/|\beta_1|$ where β_1 is the first element of the parameter vector β. This implies that one can only estimate the integrated baseline hazard up to an unknown power transformation, $\Lambda(t)^{1/|\beta_1|} = e^{H(t)}$. In particular, one cannot estimate β, the elasticity of the hazard with respect to the regressors, and one usually cannot establish whether $\lambda(t)$ is increasing or decreasing.

Horowitz (1999) assumes that $Ev^3 < \infty$ so that the mixed proportional hazard model is identified and derives a nonparametric estimator for $\lambda(t)$. In particular, Horowitz (1999) uses the fact that $e^{H(T)}$ is distributed as Weibull and uses durations that are close

[1] See Lancaster (1990) and Van den Berg (2001) for overviews and Han and Hausman (1990) and Meyer (1996) for applications.

[2] Ridder (1990) gives identification proofs for the closely linked generalized accelerated failure time model and Hausman and Woutersen (2005) estimate a duration model with time-varying regressors; Honoré and Hu (2010) give a recent review of the transformation model.

to zero[3] to estimate β_1. Under Horowitz' (1999) assumptions, the fastest possible rate of convergence for the integrated baseline hazard and the regressor parameter is $N^{-2/5}$, see Ishwaran (1996), and the rate of convergence of Horowitz' (1999) estimator is arbitrarily close to this rate.

This paper introduces the *structural duration function,* a function that gives the duration of an individual conditional on possible values of the regressor, and conditional on the unobservables. We use the structural duration function to answer counterfactual questions such as "how long would the duration of individual i have lasted if his regressor was X_i^* rather than X_i" and "what is the derivative of the duration with respect to a particular regressor". Thus, the structural duration function is a conditional expectation function for each individual in which we condition on the unobservables[4]. In the linear model $Y = X\beta + \varepsilon$, such conditioning is done implicitly when the marginal effect is calculated as $\triangle Y = (X^* - X)\beta$. In a nonlinear model, one needs to make the conditioning on the stochastic terms explicit.[5] The advantage of this approach is that the structural duration function is identified under milder restrictions than the mixed proportional hazard model. Moreover, singularity of the information matrix of the mixed proportional hazard model does not prevent $N^{-1/2}$ estimation of the structural duration function. The structural duration function is an extension of the transformation model of Horowitz (1996) and we present our estimator in his framework.

We show that, under Lancaster's (1979) assumptions, the counterfactual has the form

$$T^* = T \exp\{(X^* - X)\beta/\alpha\},$$

where α is the Weibull parameter, X is the observed regressor, T is the observed outcome and T^* is the duration that would have happened if X^* were the regressor. The parameters $\{\alpha, \beta\}$ are estimated by Lancaster using maximum likelihood. Lancaster notes that the parameter α is imprecisely estimated for his dataset of more than 500 individuals. Hahn (1994) shows that, without parametric assumptions on the heterogeneity, α cannot be

[3] Van den Berg (2001) argues against relying on very short spells for estimating a duration model.

[4] In the mixed proportional hazard model, one can derive the following equality, $\ln\{\int_0^T \lambda(t)dt\} = -X\beta - \ln(v) + \ln\{-\ln(U)\}$, where U is uniformly distributed on [0,1]; we condition on v and U in this case and we condition on ε when we consider the transformation model of equation (2).

[5] See also Vytlacil (2002).

estimated at rate $N^{-1/2}$. We note that the estimated ratio β/α in Lancaster's study remains constant for several values of α. Moreover, β/α can be estimated at rate $N^{-1/2}$, even without a distributional assumption on the heterogeneity. Thus, applying the idea of the structural duration function to Lancaster's model removes the applied problem of unstable estimates and the theoretical problem of not being able to estimate the object of interest at rate $N^{-1/2}$.

This paper is organized as follows. Section 2 discusses the mixed proportional hazard model and the transformation model. It also shows that our estimator converges at the regular rate and is asymptotically normally distributed. Section 3 suggests several parametric estimators for the structural duration function. Section 4 suggests to estimate the median if spells are censored. Section 5 applies our method to the National Evaluation of Welfare-to-Work Strategies studies. Section 6 concludes and the proofs are in the appendices.

2. The Mixed Proportional Hazard Model and Counterfactuals

Lancaster (1979) introduced the mixed proportional hazard model in which the hazard is a function of a regressor X which does not change over time, unobserved heterogeneity v, and a function of time $\lambda(t)$,

$$\theta(t \mid X, v) = ve^{X\beta}\lambda(t). \tag{3}$$

The function $\lambda(t)$ is often referred to as the baseline hazard. Integrating $\lambda(t)$ with respect to time and taking logarithms gives

$$\ln\{\Lambda(T)\} = -X\beta - \ln(v) + \ln\{-\ln(U)\}, \tag{4}$$

where $\Lambda(t) = \int_0^t \lambda(s)ds$ and U is uniformly distributed on [0,1]. This representation shows the double stochastic nature of the mixed proportional hazard model in the sense that the duration depends on unobserved heterogeneity as well as the uniformly distributed U. In this section, we construct the structural duration function. In particular, we

(i) estimate how long a spell of an individual would have lasted if the regressors were different; and

(ii) estimate the derivative of the duration of an individual with respect to a regressor.

It is important to note that the mixed proportional hazard model is *not* required to be identified in order to calculate counterfactuals. It seems that this observation has not been made before. Horowitz (1996) shows how to estimate a closely related transformation model,

$$\begin{aligned} H(t) &= X\kappa + \varepsilon \\ \text{where } |\kappa_1| &= 1 \end{aligned} \tag{5}$$

where X is a vector. This model does not impose the identifying assumption of Elbers and Ridder (1982) that $Ev < \infty$ or the identifying assumption of Heckman and Singer (1984) that the right tail of the distribution of v decreases at a known rate. For this reason, normalizing κ_1 is necessary for identification.

Let $T^*(X, X^*, T)$ denote[6] the duration if the regressors were X^* instead of the observed X. We condition on the unobserved ε. Then

$$\begin{aligned} H(T^*) &= X^*\kappa + \varepsilon \\ &= X^*\kappa + H(T) - X\kappa. \end{aligned}$$

Consider the structural duration function $D(X^*, \varepsilon)$ to calculate T^*. In particular,

$$T^* = D(X^*, \varepsilon) = H^{-1}\{X^*\kappa - X\kappa + H(T)\}. \tag{6}$$

Thus, for known $H(.)$, $H^{-1}(.)$ and κ, one can construct the counterfactual T^*. We estimate $D(X^*, \varepsilon)$ by replacing $H(.)$, $H^{-1}(.)$ and κ by estimators. There are several estimators for κ; see Horowitz (1996) and Chen (2002). In the application, we use the estimators of Han (1987) and Cavanagh and Sherman (1998). Horowitz (1996) and Chen (2002) assume that $H(.)$ is strictly increasing. We also assume this to ensure that $H(.)$ is invertible. We use Chen's (2002) estimator for $H(.)$. This estimator, denoted by $\hat{H}(.)$, is piece-wise constant and we therefore define an estimator of the inverse of $\hat{H}(.)$ as follows. Let $t = \hat{G}(y)$ where t is the smallest s for which $\hat{H}(s) = y$. We use $\hat{D}(X^*, \hat{\varepsilon})$ to denote our estimator for $D(X^*, \varepsilon)$ and define it as follows,

$$\begin{aligned} \hat{D}(X^*, \hat{\varepsilon}) &= \hat{D}\{X^*, \hat{H}(T) - X\hat{\kappa}\} \\ &= \hat{G}\{(X - X^*)\hat{\kappa} + \hat{H}(T)\}, \end{aligned}$$

[6] We often surpress the arguments and write T^* rather then $T^*(X, X^*, T)$.

where $\hat{\kappa}$ is an estimator for κ.

Chen (2002) shows that the following assumptions ensure $\sqrt{N}$-convergence of $\widehat{H}(t)$.

A1. Let $H(T) = X\kappa + \varepsilon$. *Let* $\{T_i, X_i, w_i, i = 1, ..., n\}$ *be a random sample of* $\{T, X, \varepsilon\}$ *and let* ε *be independent of* X.

A2. (a) $|\kappa_1| = 1$, *(b) the distribution of the first component of* X *conditional on* $\widetilde{X} = \widetilde{x}$ *is absolutely continuous with respect to the Lebesque measure, (c) the support of* X *is not contained in any proper linear subspace of* $\mathbb{R}^p$, *where* p *is the number of exogenous regressors.*

A3. $H(.)$ *is strictly increasing,* $H(t_0) = 0$, $[H(t_1 - \epsilon), H(t_2 + \epsilon)] \subset M_H$, *for a small* $\epsilon > 0$, *for some* t_0, t_1 *and* t_2 *in the support of* T, *with* M_H *a compact interval.*

A4. The conditional density of $X\kappa$ *given* $\widetilde{X} = \widetilde{x}$ *and the density of* ε *at* s, $f_{X\kappa}(s|\widetilde{x})$ *and* $p(s)$, *are twice differentiable in* s, *the derivatives are uniformly bounded, and* $\widetilde{X}$ *has finite third-order moments.*

Define $r_1\{t_1, x_1, t_2, x_2, H(t)\} = [1\{t_1 \geq t\} - 1\{t_2 \geq t_0\}] \cdot 1\{x_1\kappa - x_2\kappa \geq H(t)\}$ and $r_2\{t_1, x_1, t_2, x_2, H(t)\} = [1\{t_2 \geq t\} - 1\{t_1 \geq t_0\}] \cdot 1\{x_2\kappa - x_1\kappa \geq H(t)\}$.

A5. $V(t) = \frac{1}{2}E\left[\frac{\partial^2 r_1\{t_1,x_1,t_2,x_2,H(t)\}}{\partial\{H(t)\}^2} + \frac{\partial^2 r_2\{t_1,x_1,t_2,x_2,H(t)\}}{\partial\{H(t)\}^2}\right]$ *is negative for each* $t \in [u, \bar{t}]$ *for any* $u > 0$, *and uniformly bounded away from zero.*

A6. The first step estimator $\hat{\kappa}$ *is* $\sqrt{N}$ *− consistent, i.e.* $\sqrt{N}(\hat{\kappa} - \kappa) = O_p(1)$.

We also assume that the following assumption holds.

A7. $H(t)$ *is twice continuously differentiable on* $t \in (0, \bar{t}]$ *and the counterfactual regressor* X^* *is independent of* ε.

Define $h(t) = \frac{\partial H(t)}{\partial t}$ for all $t \in (0, \bar{t}]$ and define $q = H(T^*) = H(T) - (X^* - X)\kappa$ and $H^{-1}(q)$ as the inverse of $H(T^*)$. We now state the theorems concerning the counterfactual duration.

Theorem 1

Under assumptions A1-A7, the estimated counterfactual duration is defined as

$$\widehat{T}^* = \widehat{H}^{-1}(q), \tag{7}$$

with the following properties

(i) for any $u > 0$,

$\sup_{q \in [H^{-1}(u), H^{-1}(\bar{t})]} N^{1/2} \left| \widehat{H}^{-1}(q) - H^{-1}(q) + \frac{1}{h(H^{-1}(q))} \left(\widehat{H}\left(H^{-1}(q)\right) - q \right) \right| \xrightarrow{p} 0.$

(ii) $\widehat{H}^{-1}(q)$ is a uniformly consistent estimator of $H^{-1}(q)$ i.e. for any $u > 0$,

$\sup_{q \in [H^{-1}(u), H^{-1}(\bar{t})]} \left| \widehat{H}^{-1}(q) - H^{-1}(q) \right| = o_p(1).$

Proof: See the appendix.

Theorem 2 (Structural Duration Function)

Let assumptions A1-A7 hold. Let

$$\widehat{T^*}(X, X^*, T) = \hat{D}(X^*, \hat{\varepsilon}) = \widehat{H}^{-1}\{\widehat{H}(T) + (X^* - X)\hat{\kappa}\}$$

and

$$\begin{aligned} T^*(X, X^*, T) &= D(X^*, \varepsilon) = H^{-1}\{X^*\kappa + \varepsilon\} \\ &= H^{-1}\{H(T) + (X^* - X)\kappa\}. \end{aligned}$$

Then

$$\hat{D}(X^*, \hat{\varepsilon}) \underset{p}{\rightarrow} D(X^*, \varepsilon).$$

Let assumptions A1-A7 hold and let $X\kappa \neq X^*\kappa$. Then

$$\sqrt{N}\{\hat{D}(X^*, \hat{\varepsilon}) - D(X^*, \varepsilon)\} \underset{d}{\rightarrow} N(0, \Omega),$$

where $\Omega = E[\{\hat{D}(X^*, \hat{\varepsilon}) - D(X^*, \varepsilon)\}^2]$.

Proof: See the appendix.

Another function that is of interest in applied work is the *derivative function*, the partial derivative of a potential outcome T_i^* with respect to a regressor X_k, $\frac{\partial T_i^*}{\partial X_k}$. Thus, $\frac{\partial T_i^*}{\partial X_k}$ is the partial derivative of $D_{X_k}(X^*, \varepsilon_i)$ with respect to X_k. Let $D_{X_k}(X^*, \varepsilon_i)$ denote the partial derivative of $D(X, \varepsilon_i)$ with respect to X_k, evaluated at X^*,

$$\frac{\partial T_i^*}{\partial X_k} = D_{X_k}(X^*, \varepsilon_i) = \frac{\partial D_{X_k}(X, \varepsilon_i)}{\partial X_k}|_{X=X^*}.$$

Consider writing the error term ε as a function of the duration T and the regressor X,

$$\varepsilon(T, X) = H(T) - X\kappa.$$

Total differentiating with respect to T and X_k gives

$$\frac{\partial T}{\partial X_k} = \frac{\partial\varepsilon(T, X)/\partial X}{\partial\varepsilon(T, X)/\partial T} = -\frac{\kappa_k}{H'(T)}.$$

Consider the estimator

$$\widehat{(\frac{\partial T}{\partial X_k})} = D_{X_k,N}(X^*, \varepsilon) = \frac{\hat{\kappa}_k}{\widehat{H'}(T^*)},$$

where $\widehat{H'}(.)$ is an estimator for the derivative of $H(.)$ with respect to its argument. We base an estimator for $H'(T)$ on Chen's (2002) estimator for $H(T)$. Consider

$$\widehat{H'(T^*)} = \frac{1}{\sigma_N}\int_{-\infty}^{\infty} w(\frac{s}{\sigma_N})\hat{H}(s + T^*)ds,$$

where $w(s)$ is a derivative kernel as in Härdle (1990). In particular, we assume the following.

A8 Let $w(s)$ be a twice continuously differentiable function with bounded derivatives and $\int_{-\infty}^{\infty} w(s)ds = 0$, $\int_{-\infty}^{\infty} sw(s)ds = 1$, and $\int_{-\infty}^{\infty} s^2 w(s)ds = 0$.

Theorem 3 (Derivative of structural duration function)

Let assumptions A1-A8 hold. Let $\sigma_N = c \cdot N^{-1/5}$ for $c > 0$. Then

$$N^{2/5}\{D_{X_k,N}(X^*, \varepsilon) - \frac{\partial T^*}{\partial X_k}\} = O_p(1).$$

Undersmoothing (i.e. choosing a σ_N that goes to zero at a rate slower than $N^{-1/5}$) yields a normally distributed estimator for the derivative. As we now demonstrate by example,the mixed proportional hazard model is *not* required to be identified in order to calculate counterfactuals or derivatives. It seems that this observation has not been made before. In particular, consider the following two data generating processes.

DGP I: The baseline hazard is constant and there is no heterogeneity, so

$$\theta(t|x, v) = e^{X\beta},$$

where X is a scalar exogenous regressor that is continuously distributed and $\beta \neq 0$. This hazard yields the following survival function:

$$\bar{F}(t|X) = e^{-e^{X\beta}t}.$$

DGP II: Consider the same regressor as in DGP I and let

$$\begin{aligned} \theta(t|X, v) &= 2ve^{2X\beta}t, \\ \text{where } p(v) &\propto v^{-3/2}e^{-1/(4v)}. \end{aligned}$$

This Lèvy mixing distribution is discussed by Woutersen (2003) and this model yields the same survival function as DGP I.

Note that the survival function $\bar{F}(t|x) = e^{-e^{X\beta}t}$ does not identify the mixed proportional hazard model. For simplicity, we assume that $\beta > 0$. In large samples, the empirical survival function converges to $\bar{F}(t|x) = e^{-e^{X\beta}t}$. Without imposing parametric assumptions on v, our estimate for $H(t)$ converges to

$$\widehat{H(t)} = \frac{1}{\beta}\ln(t).$$

We now use $\widehat{T^*}(X, X^*, T) = \widehat{H}^{-1}\{\widehat{H}\,(T) + (X^* - X)\hat{\kappa}\}$ to find counterfactuals. Note that X is a scalar and that $\hat{\kappa}_1$ is normalized and assumed to be positive, so that $\kappa_1 = 1$. Also note that $\widehat{H}^{-1}(q) = \exp(\beta \cdot q)$. Thus,

$$\begin{aligned} \widehat{T^*}(X, X^*, T) &= \exp[\beta\{\frac{1}{\beta}\ln(T) + (X^* - X)\}] \\ &= T\exp\{\beta\{(X^* - X)\}. \end{aligned}$$

In particular, consider individual i and consider what the duration of this individual would have been if the regressor was X_i^* instead of X_i,

$$\widehat{T^*}(X_i, X_i^*, T_i) = T_i \exp\{\beta\{(X_i^* - X_i)\}.$$

Moreover,

$$\frac{\partial T_i}{\partial X_i} = -\beta e^{X_i\beta}T_i.$$

We thus have estimated counterfactuals without calculating the baseline hazard or the mixing distribution, both of which are unidentified in this example.

So far, we assumed that the researcher has one particular value of the counterfactual in mind for individual i, X_i^*. Off course, the researcher may be interested in a distribution of the counterfactual regressor, denoted by $p_{X_i^*(x_i^*)}$. In that case, the researcher can sample from $p_{X_i^*(x_i^*)}$ and calculate $\widehat{T^*}(X_i, X_i^*, T_i)$ for each realization of X_i^*. Given that $H(t)$ and its inverse are differentiable, the distribution of $\widehat{T^*}(X_i, X_i^*, T_i)$ converges to the distribution of $T^*(X_i, X_i^*, T_i)$. This may be useful in some applications where the treatment can only be extended to a subset of the population.[7]

3. Parametric estimators of the Mixed Proportional Hazard Model and the Structural Duration Function

The structural duration function can also be estimated when $H(.)$ and/or F is parametric. The advantage of the structural duration function is that (i) it can remove an empirical identification problem when both $H(.)$ and F are parametric, and (ii) it can increase the rate of convergence for a moment estimator if F is nonparametric. Assuming a parametric $H(.)$ can be a reasonable choice if the sample size is small or the transformation model fails to be nonparametrically identified, e.g. in case no regressor is continuously distributed.

3.1. Parametric H, parametric F. Lancaster (1979) introduced the mixed proportional hazard model; he used a Weibull function for the baseline hazard and modelled the heterogeneity as a gamma distribution. Writing his model as a transformation model, we have

$$\ln(T^\alpha) = X\beta - \ln(v) + \ln(z)$$

where v has a gamma distribution and z has a unit exponential distribution. Thus,

$$\begin{aligned} \ln(T) &= \frac{X\beta}{\alpha} - \ln(v') + \ln(z') \\ &= X\gamma - \ln(v') + \ln(z') \end{aligned}$$

where v' and z' have a generalized gamma distribution. It is noteworthy that in Lancaster's (1979) application, the Weibull parameter α is weakly identified in the sense that the standard error of the parameter estimate for α is "quite large" (Lancaster, page 954). Lancaster (1979) reports his results in Tables IV and V. In Table IV, Lancaster restricts

[7] We thank an anonymous referee for stressing the importance of a stochastic X_i^*.

α to be one. In Table V, α is estimated and the estimated values of β decrease by about 10%. For the transformation model, the restriction that $\alpha = 1$ implies that v is a generalized gamma and z is a unit exponential random variable. Apparently, allowing for generalized gamma distributions and imposing the restriction $\alpha = 1$ hardly changes the estimate of $\gamma = \frac{\beta}{\alpha}$. The structural duration function relies only on γ. Therefore, the counterfactuals based on Lancaster's (1979) Table IV nearly coincide with the counterfactuals based on Lancaster's (1979) Table V. Thus, the approach that uses the counterfactual duration function reconciles and somewhat explains the different estimates Lancaster (1979) obtains in Tables IV and V.

3.2. Parametric H, nonparametric F. Honoré (1990) introduces an estimator for the Weibull model that only requires the mixing distribution to have finite mean. This estimator only uses durations that are very short and this estimator converges at a rate slower than $N^{-1/3}$. Honoré (1990) suggests that his estimator could be used as a first step and that, in a second step, the coefficients of the regressors would be estimated. Thus, in the second step, $\ln(T^{\alpha})$ would be regressed on X. This second step is equivalent to estimating the model

$$\ln(T) = \frac{X\beta}{\hat{\alpha}} + \varepsilon,$$

where X and ε are independent and $\hat{\alpha}$ denotes Honoré's (1990) estimator for the Weibull coefficient. Applying the approach of this paper would imply estimating γ using the following model:

$$\ln(T) = X\gamma + \varepsilon.$$

The parameter γ can be estimated at rate $N^{-1/2}$ and counterfactuals would be constructed using estimates for γ. Note that the rate of convergence of $\frac{\hat{\beta}}{\hat{\alpha}}$ is slower than $N^{-1/3}$, which is slower than the rate of convergence of $\hat{\gamma}$. Thus, as far as counterfactuals are concerned, one may want to avoid estimating α using arbitrarily short durations.

Kiefer and Wolfowitz (1956) consider likelihood models and account for heterogeneity by extending their likelihood model to include a discrete mixture. Heckman and Singer (1984a) apply Kiefer and Wolfowitz (1956) to the mixed proportional hazard model and derive an estimator that uses a parametric hazard rate. However, the rate of convergence of the estimator of Kiefer and Wolfowitz (1956) is unknown.

For any integrated baseline hazard rate, we have the equality

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

Choosing a parametric form for $\Lambda(t)$ implies that $H(t)$ is a parametric function. Let $H(t;\theta)$ denote this parametric function so that

$$H(t;\theta) = X\gamma + \varepsilon.$$

The following moments can be used to estimate θ :

$$g(\theta) = X\{H(T;\theta) - X\gamma\}.$$

Identification of the mixed proportional hazard model implies that $e^{H(t;\theta)}$ is *not* closed under the power transformation.[8] Therefore, under regularity conditions of Ridder and Woutersen (2003, proposition 2), the information matrix is regular. As a consequence, the parameter θ, $H(t;\theta)$ and $H(t;\theta)^{-1}$ can all be estimated at rate $N^{-1/2}$.

3.3. Nonparametric H, parametric F. Meyer (1990), Meyer (1996), and Han and Hausman (1990) approximate the baseline hazard using a piece-wise constant. This gives a flexible parametrization of $\Lambda(t) = e^{\alpha H(t)}$. The mixing distribution is then approximated using the gamma distribution. The idea behind these estimators is that the flexibility in $\Lambda(t) = e^{\alpha H(t)}$ makes up for the restricted distributional form of the heterogeneity. As discussed above, α may not be well identified in the Weibull model with gamma heterogeneity and changes if one switches from a gamma heterogeneity to a generalized gamma heterogeneity.

3.4. Nonparametric H, nonparametric F. Horowitz (1999) introduces an estimator for the mixed proportional hazard model that allows for a nonparametric hazard and nonparametric heterogeneity. In particular, Horowitz (1999) considers $\Lambda(t) = e^{\alpha H(t)}$. He estimates $H(t)$ using Horowitz (1996) and estimates α using an estimator that is similar

[8] A set of functions $\mathcal{H}$ is closed under the power transformation if $f(t) \in \mathcal{H}$ implies $\{f(t)\}^{\alpha} \in \mathcal{H}$ for every $\alpha > 0$. Here, however, we have $e^{H(t)} \in \mathcal{H}$ implies $\{e^{H(t)}\}^{\alpha} \notin \mathcal{H}$ for every $\alpha > 0$, $\alpha \neq 1$.

to Honoré (1990, 1998). The rate of convergence of Horowitz' (1999) estimator is determined by the rate of convergence of the estimator for α. The advantage of the structural duration function is that it avoids estimating α and, therefore, converges at rate $N^{-1/2}$.

3.5. Weibull hazard, Gamma Mixing Distribution. We showed that, under Lancaster's (1979) assumptions, the counterfactual has the form

$$T^* = T \exp\{(X^* - X)\beta/\alpha\},$$

where X is the observed regressor, T is the observed outcome and T^* is the duration that would have happened if X^* was the regressor. The parameters $\{\alpha, \beta\}$ are estimated by Lancaster using maximum likelihood. Lancaster notes that the estimate of α depends a lot on which regressors he includes and this "worries" him. Hahn (1994) shows that, without parametric assumptions on the heterogeneity, α cannot be estimated at rate $N^{-1/2}$. We note that the estimated ratio β/α in Lancaster's study depends less on which regressors are included than when β or α are compared separately. Moreover, β/α can be estimated at rate $N^{-1/2}$.

4. Structural Duration Function for unfinished spells

For completed spells, we can condition on the unobserved error term ε and calculate the structural duration function. The intuition for this result is that, for known κ and $H(t)$, we can calculate $\varepsilon = H(T) - X\kappa$ and $T^* = H^{-1}(X^*\kappa + \varepsilon)$. However, for censored or unfinished spells, we do not observe the duration T and therefore cannot calculate ε. Suppose that duration spells are censored at C and that we observe the minimum of C and T. Note that,for known κ and known $H(t)$, $T \geq C$ implies that $\varepsilon \geq H(C) - X\kappa$ so that we can sample from $p(\varepsilon \geq H(C) - X\kappa)$. Thus, we need to ensure that the estimators for κ and $H(.)$ are still consistent and asymptotically normally distributed in the presence of censoring. Fortunately, the estimator for κ is still consistent and asymptotically normally distributed as long as assumption A10 below holds. Also, Chen (2002) also considers the model estimation for censored data and adds three assumptions to cover this case[9] . In

[9] Gørgens and Horowitz (1996) also derive an estimator for $H(.)$ for this case of censored data.

particular, the model we use for the censored observations is

$$H(T^c) = X\kappa + \varepsilon, \tag{8}$$

where T^c is unobserved. We observe instead $T = \min\{T^c, C\}$, and X, with C a random censoring variable. As before, we normalize normalize $H(t_0) = 0$ for a finite t_0 and use a scale normalization for the coefficient of the first component of X, $|\kappa_1| = 1$. Assume $d_{it} = 1\{T_i \geq t\}$, $d_{jt_0} = 1\{T_j \geq t_0\}$, and $G(t) = Pr(C > t)$ for any t, $\widetilde{X}$ is the vector containing all but the first component of X, and $\delta_i = 1\{T^c \leq C\}$. Following Chen (2002), the estimator of the transformation model $H(T)$ for censored data is

$$\hat{H}_{cens}(t) = \arg\max_{H(t)} Q\left(H(t)\right) = \arg\max_{H(t)} \frac{1}{n(n-1)} \sum_{i \neq j} \left(\frac{d_{it}}{\hat{G}(t)} - \frac{d_{jt_0}}{\hat{G}(t_0)} \right) 1\{X_i\hat{\kappa} - X_j\hat{\kappa} \geq H(t)\}, \tag{9}$$

for a given $t \in [t_1, t_2]$, $\hat{\kappa}$ a consistent estimator for κ and $\hat{G}$, the Kaplan-Meier estimator or the product limit estimator for the survival function G. The large sample properties of the $\hat{H}_{cens}(t)$ estimator, when the data is censored, are presented by Chen (2002) in Theorem 2. For the censored case, Chen (2002) uses three additional assumptions.

A9. $\{X_i, T_i, i = 1, 2, ..., n\}$ *is a random sample of* $\{X, T\}$ *in (8) and* ε *is independent of* X.

A10. The censoring variable C *is independent of* (X, T^c) *and continuously distributed with positive density on interval containing* t_0, t_1, t_2.

Define $r_{1,cens}\{t_1, x_1, t_2, x_2, H(t)\} = \left(\left(\frac{1\{t_1 \geq t\}}{G(t)} - \frac{1\{t_2 \geq t_0\}}{G(t_0)}\right)\right) 1\{x_1\hat{\kappa} - x_2\hat{\kappa} \geq H(t)\}$ and $r_{2,cens}\{t_1, x_1, t_2, x_2, H(t)\} = \left(\left(\frac{1\{t_2 \geq t\}}{G(t)} - \frac{1\{t_1 \geq t_0\}}{G(t_0)}\right)\right) 1\{x_2\hat{\kappa} - x_1\hat{\kappa} \geq H(t)\}$.

A11. $V_{cens}(t) = \frac{1}{2}E\left[\frac{\partial^2 r_{1,cens}\{t_1,x_1,t_2,x_2,H(t)\}}{\partial\{H(t)\}^2} + \frac{\partial^2 r_{2,cens}\{t_1,x_1,t_2,x_2,H(t)\}}{\partial\{H(t)\}^2}\right]$ *is negative for each* $t \in [t_1, t_2]$ *and uniformly bounded away from zero.*

Chen's (2002) Theorem 2 states that under assumptions A2-A4, A6 and A9-A11,

(i) $\sup_{t_1 \leq t \leq t_2} \left|\hat{H}_{cens}(t) - H_0(t)\right| = o_p(1)$,

(ii) uniformly over $t \in [t_1, t_2]$,

$$\sqrt{N}\left(\hat{H}_{cens}(t) - H_0(t)\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J^c_{t_0,t}(X_i, T_i, \delta_i) + o_p(1), \tag{10}$$

and $N^{\frac{1}{2}}\{\hat{H}_{cens}(t) - H_0(t)\} \Longrightarrow G^c_H(t_0, t)$ where $G^c_H(t_0, t)$ is a Gaussian process with mean 0 and covariance function $E\{G^c_H(t_0, t)\, G^c_H(t_0, t')\} = E\{J^c_{t_0,t}(X, T, \delta)\, J^c_{t_0,t'}(X, T, \delta)\}$ with $J^c_{t_0,t}(X, T)$ as in Chen (2002).

Thus, after making stronger assumptions that ensure that the estimated transformation function, $\hat{H}_{cens}(t)$, is well behaved, we can calculate ε for the uncensored observations and establish that $\varepsilon \geq \hat{H}_{cens}(C) - X\hat{\kappa}$ for the censored observations. Suppose that we can estimate the distribution of ε for some value of the regressors, then we can sample from $p(\varepsilon|\varepsilon \geq \hat{H}_{cens}(C) - X\hat{\kappa})$ for the censored observations. If none of the observations were censored, then we have the same estimator as before. If some observations are censored, then we draw from $p(\varepsilon|\varepsilon \geq \hat{H}_{cens}(C) - X\hat{\kappa})$ many times for each censored duration, calculate T^* each time, and then we average over these draws. This method allows us to calculate the counterfactual policy estimates and we present an example in the next section.

5. Empirical Results

The data used in the empirical analysis are from the National Evaluation of Welfare-to-Work Strategies (NEWWS) study, a study undertaken by the US Department of Health and Human Services to determine the efficacy of various welfare-to-work programs.[10] The NEWWS study was quite broad, testing eleven welfare-to-work strategies across seven cities over the course of several years. Not all the strategies, however, were tested in all cities, and we consider only data from Riverside, CA in this analysis.

The study in Riverside accepted participants from June 1991 to June 1993. Once accepted into the study, participants' welfare grants, employment status, and earnings were tracked for 5 years. In addition, data on a variety of demographic characteristics were collected when participants entered the study.

Participants in Riverside were randomly assigned into either one of two treatment groups or a control group. The two treatment groups were the Labor Force Attachment (LFA) group and the Human Capital Development (HCD) group.[11] Members of the LFA

[10] See U.S. Department of Health and Human Services (HHS), "National Evaluation of Welfare-to-Work Strategies: How Effective are Different Welfare-to-Work Approaches?" (2000) for more information about the study's design. See this document for a bibliography of previous research using these data.

[11] These strategies were also tested in Atlanta, GA and Grand Rapids, MI.

group were urged to obtain a job as quickly as possible with the rationale that, once the participant found a job, the subject would remain in the labor force. Members of the HCD group were offered free access to education. It was hoped that the skills these participants gained through educational programs would make them attractive to employers, allowing them improved access into the labor force.

In order to be a participant in this study, subjects had to satisfy an array of requirements.[12] Chief among these requirements was that subjects had to be eligible for and apply for Aid to Families with Dependant Children (AFDC) grants or had to be already receiving AFDC payments. If they satisfied this criterion, among others, the subject was directed to attend an orientation with the threat of decreased welfare benefits if they did not comply.[13] At these orientations, the subject was randomly sorted into one of the three groups (LFA, HCD, and control).

California had existing welfare regulations governing educational assistance. Because of these existing regulations, only those subjects in need of basic education could be sorted into the HCD group. This group consisted of subjects who did not have a high school degree or GED, were not fluent in English, or scored below 215 on any part of the CASAS. Because this restricted the sample of people to be sorted into the control and LFA groups, the study designers oversampled the LFA and control groups in Riverside. Sample sizes for the three groups are provided in Table 1.[14]

Table 1: Sample sizes

Group	Subjects
Labor Force Attachment	3384
Human Capital Development	1596
Control	3342
Total:	8322

After sorting participants into one of the three groups, the study tracked them for five years, mainly through the reports of state-level government agencies. Among the data

[12] US HHS op. cit.

[13] Sanctions were implemented by individual welfare offices, and not all subjects who failed to attend an orientation were sanctioned. The NEWWS study found no clear relationship between sanctioning rates within a city and participation rates in welfare-to-work programs.

[14] Brock and Harknett (1998) found that about 66% of the people directed to attend an orientation in Riverside actually attended.

collected are quarterly information about AFDC grant levels, food stamp grants, employment status, and earnings. Demographic characteristics as of the random assignment are also provided; however, many of these characteristics were grouped in order to preserve anonymity (i.e. age is reported only as an age group). Table 2 presents summary statistics of relevant variables.

Table 2: Summary statistics

Variable	Description	Mean	Std. Dev.
Non-White	= 1 if Black or Hispanic	0.446	0.497
HS Education	= 1 if HS degree or GED	0.580	0.494
Young Child	= 1 if child $<$ 5	0.572	0.495
1 Child	= 1 if 1 child	0.390	0.488
2 Children	= 1 if 2 children	0.324	0.468
3+ Children	= 1 if 3+ children	0.285	0.452
Age $<$ 30	= 1 if age $<$ 30	0.402	0.490
Age 30-39	= 1 if age 30-39	0.465	0.499
Age 40+	= 1 if age 40+	0.133	0.340
log(Avg. welfare)	log of avg. FS + AFDC	8.238	1.165
Employed prev. year	= 1 if employed in year before entering study	0.396	0.489
N = 2881			

As participation in the AFDC program was required to be in the study, the overwhelming majority of study participants were females. Females account for nearly 90% of the subjects. Because the number of male subjects is too small to yield useful information, we estimate parameters only for female subjects.

For the purposes of this analysis, we define non-White to be Blacks and Hispanics and welfare benefits to be the sum of food stamp and AFDC grants.[15] Welfare levels generally do not vary sufficiently across time to allow for identification of a duration model with time varying regressors. Therefore, we use average welfare receipt from the time that a subject enters the study until the start of their first employment spell. The demographic characteristics provided in the data set are as of the date of random assignment and do not vary across time. As AFDC participation is a requirement for inclusion in the study, all participants receive some amount of welfare between their random assignment date and the start of their first spell of employment.

[15] We group Asians with Whites because their employment and earnings characteristics are more similar to that of Whites than Blacks or Hispanics.

We now discuss estimation results. First, we estimate the Han-Hausman (1990) and Meyer (1990) (HHM) model that allows for a non-parametric baseline hazard and gamma heterogeneity. The other two estimators are based on the approach of this paper which uses semi-parametric estimation of the unknown coefficients in the regression hazard model and non-parametric estimation of the integrated baseline hazard. The estimators used for the regression hazard model are the Han (1987) maximum rank correlation (MRC) estimator and the Cavanagh-Sherman (1998) (CS) rank estimator.

The estimated coefficients in the regression hazard model are given in Table 3.

Table 3: Estimated Coefficients

	HHM	Scaled HHM	MRC	Cavanagh-Sherman log(t)
Non-White	0.185	0.254	0.329	0.326
	(0.061)	(0.088)	(0.097)	(0.097)
High School Education	0.188	0.260	0.422	0.424
	(0.063)	(0.092)	(0.072)	(0.084)
Young Child	-0.104	-0.143	-0.029	-0.027
	(0.069)	(0.094)	(0.108)	(0.103)
Second Child	0.189	0.260	0.182	0.174
	(0.073)	(0.107)	(0.100)	(0.095)
Third Child	0.435	0.599	0.482	0.439
	(0.081)	(0.136)	(0.133)	(0.122)
Age 35	-0.129	-0.178	-0.023	-0.023
	(0.071)	(0.099)	(0.119)	(0.110)
Age 45	-0.453	-0.624	-0.453	-0.373
	(0.106)	(0.159)	(0.146)	(0.126)
Log Average Welfare	-0.692	-0.954	-0.542	-0.364
	(0.031)	(0.121)	(0.042)	(0.043)
Previously Employed	0.726	1.000	1.000	1.000
	(0.073)	-	-	-
ln(Variance)	-0.567	-0.781		
	(0.151)	(0.178)		
Observations	2881	2881	2881	2881

The HHM model finds a significant role for heterogeneity with the variance estimate found to be large and highly significant. We find log of average welfare to be a significant disincentive to exiting into the labor force. In the second column, we rescale the HHM estimates using the coefficient on previously employed so we can compare these estimates to the MRC and CS results. The MRC and CS estimates should be close to each other as they have the same probability limit. We find all of the estimates are quite close to each other except for the coefficient of log of average welfare, which is the variable of most interest! The MRC estimate is considerably higher than the CS estimate, although the MRC estimate is still lower than the HHM estimate. Thus, considerable uncertainty exists about the magnitude of the effect of welfare, although all these models find large and significant results.

In Figure 1, we display the results of the Chen (2002) estimator of the transformation model along with two standard error bounds derived from bootstrap estimation. The estimates are quite precise. In Figure 2, we use the Chen (2002) estimates to estimate a local third degree polynomial along with two standard error bounds using the same approach we previously used in Hausman-Woutersen (2005). Note that the amount of uncertainty becomes considerably greater as the durations become longer. In Figure 3, we calculate the derivative of the transformation function. Note the non-monotonic features of the estimates. Ridder and Woutersen (2003) discuss that the slope of the baseline hazard function cannot be derived from the derivative of the transformation function. Lastly, in Figure 4 we calculate the inverse of the Chen (2002) estimates which we now use to do policy simulations to estimate the counterfactual durations using the structural model approach of equation (11).

We now consider a policy simulation using the estimates in Table 3 along with the estimates of the inverse of the transformation function for use in equation (11). We present the changes in the average welfare amount in Table 4.

Table 4: Simulation of Policy Change					
HHM					
	Change in Welfare Benefits				
	-20.00%	-10.00%	0.00%	10.00%	20.00%
Unemployment Duration					
Mean	3.85	5.14	8.41	9.35	8.92
Median	2.90	2.90	4.00	3.90	3.90
Standard deviation	2.88	4.54	8.33	8.59	8.89
MRC					
	Change in Welfare Benefits				
	-20.00%	-10.00%	0.00%	10.00%	20.00%
Unemployment Duration					
Mean	6.64	7.09	8.41	8.78	9.92
Median	2.90	2.90	4.00	3.90	3.90
Standard deviation	6.61	7.12	8.33	8.81	8.21
Cavanagh-Sherman (CS)					
	Change in Welfare Benefits				
	-20.00%	-10.00%	0.00%	10.00%	20.00%
Unemployment Duration					
Mean	7.05	7.69	8.41	8.70	8.80
Median	2.90	2.90	4.00	3.90	3.90
Standard deviation	7.12	7.93	8.33	8.07	8.12

In Table 4, the HHM estimator predicts a larger change in duration for the same change in welfare benefits. For example, for a 10% increase in welfare benefits the HHM model predicts a 6.5% larger change in the mean duration than does the model based on the MRC estimates and a 7.5% larger change than the estimates based on CS. To the extent that the HHM estimates are inconsistent because of model misspecification arising from the assumption of gamma heterogeneity, our new approach that does not require estimation of heterogeneity may provide more reliable results. Also note that the estimates of policy changes based on the MRC and CS estimates are considerably closer to each other than the estimates based on HHM.

6. Conclusion

In this paper, we propose a new method to estimate the effect of a counterfactual policy in the transformation model. That is, a method that predicts the effect of new policies, and, also, the effect of an existing policy on a new population. We also propose an estimator for the counterfactual policy and show that this estimator converges at the $\sqrt{N}-$ rate, even in cases where the underlying duration model is unidentified or has a singular information matrix. We do not need to specify the heterogeneity distribution parametrically and we do not need stringent identification assumptions. We use this structural duration function to simulate important policy effects, including a change in welfare benefits.

Appendix: The Transformation Model

Consider the following transformation model

$$H(T) = X\kappa + \varepsilon, \tag{11}$$

where T is the dependent variable (e.g. duration), $H(t)$ is a strictly increasing function, X is the set of explanatory variables (strictly exogenous), κ is the vector of associated coefficients and ε is the unobserved error term.

The transformation model (1) generates a significant number of econometric models. The rank estimation of equation (11) was presented in Chen (2002). He considers the model estimation for uncensored and censored data.

He normalizes $H(t_0) = 0$ for a finite t_0 and uses a scale normalization for the coefficient of the first component of X, $|\kappa_1| = 1$.

Assume $d_{it} = 1\{T_i \geq t\}$, $d_{jt_0} = 1\{T_j \geq t_0\}$, M_H is a compact set, $\widetilde{X}$ is the vector containing all but the first component of X, and $\delta_i = 1\{T^c \leq C\}$.

The estimator of the transformation model $H(t)$ for uncensored data is

$$\hat{H}(t) = \arg\max_{H(t)} [Q(H(t))] = \arg\max_{H(t)} \frac{1}{N(N-1)} \sum_{i \neq j} (d_{it} - d_{jt_0}) 1\{X_i\hat{\kappa} - X_j\hat{\kappa} \geq H(t)\}, \tag{12}$$

for a given $t \in [t_1, t_2]$ and where $\hat{\kappa}$ is a consistent estimator for κ.

The large sample properties of the $\hat{H}(t)$ estimator, when the data is uncensored, are presented by Chen (2002) in Theorem 1.

Appendix: Proof of Theorem 1 and 2

We use Chen's (2002) estimator for $H(.)$ and as an estimator for $H^{-1}(q)$ we use

$$t^* = \widehat{H}^{-1}(q) = \min\left\{t \in \mathbb{R}^+ : \widehat{H}(t) \geq q\right\}. \tag{13}$$

The following Lemmas are useful to prove Theorem 1:

Lemma A1

Under assumptions A1-A7, $h(t) = \frac{\partial H(t)}{\partial t} > 0$ for all $t \in (0, \bar{t}]$ and $\frac{\partial H(t)}{\partial t} H^{-1}(q)$ is a strictly increasing function for $q \in (H(0), H(\bar{t})]$.

Proof:

Note that $H(.)$ is strictly increasing by assumption A3 and twice continuously differentiable by assumption A7 so that the first derivative $h(t) = \frac{\partial H(t)}{\partial t} > 0$ for all $t \in (0, \bar{t}]$. Moreover, under assumptions A3 and A7,

$$\frac{\partial H^{-1}(q)}{\partial q} = \frac{1}{h\left(H^{-1}(q)\right)} > 0$$

so that $H^{-1}(q)$ is a strictly increasing function.

Lemma A2

Let assumptions A1-A6 hold.[16] Then, for any $u > 0$, $\eta \geq 0$,

$$\sup_{t \in [u, \bar{t}]} \left|\widehat{H}(t) - \widehat{H}(t-\eta) - \{H(t) - H(t-\eta)\}\right| = O_p\left(\sqrt{\frac{\eta}{N}}\right).$$

Proof:

First, let only those individuals with $||x|| < M$ for some $M < \infty$ be used for the estimation of $H(t)$ where $||.||$ is the Euclidian norm. Define $\Delta = \widehat{H}(t) - \widehat{H}(t-\eta) - \{H(t) - H(t-\eta)\}$. The assumption that only individuals with $||x|| < M$ for some $M < \infty$ ensures that the variation of $\widehat{H}(t)$ exists. Note that

$$Var(\Delta) = Var\left(\widehat{H}(t)\right) + Var\left(\widehat{H}(t-\eta)\right) - 2Cov\left(\widehat{H}(t), \widehat{H}(t-\eta)\right).$$

Chen (2002) gives expressions for the variance and covariance function and it follows that $Var(\Delta) = O(\frac{\eta}{N})$. The expectation of Δ is $o(\sqrt{\frac{\eta}{N}})$ so that the variance determines the

[16] For the censored case, we need assumptions A1-A7 plus Chen's (2002) A8-A9.

order of Δ and $\Delta = O_p(\sqrt{\frac{\eta}{N}})$ for every t. The functions $\widehat{H}(t)$ and Δ are stochastically equicontinuous so that uniform convergence follows from Newey (1991).

Lemma A3

Define $t^* = \widehat{H}^{-1}\left(\widehat{H}(t)\right)$ and let the assumptions of Lemma A1 hold. Let $\eta = \frac{1}{\sqrt{N}\ln N}$. Then $\sup_t \sqrt{N}\,|t^* - t| \xrightarrow{p} 0$.

Proof:

Note that $t^* \leq t$. By definition of $\widehat{H}^{-1}(q)$ we have

$$\begin{aligned} t - t^* &< \eta \Longleftrightarrow \widehat{H}(t) > \widehat{H}(t-\eta) \\ &\Longleftrightarrow \widehat{H}(t) - \widehat{H}(t-\eta) = H(t) - H(t-\eta) + W > 0 \end{aligned}$$

where W is a random variable and $W = O_p\left(\sqrt{\frac{\eta}{N}}\right)$. Using a Taylor expansion around t yields

$$\begin{aligned} t - t^* &< \eta \\ &\Longleftarrow \underline{h}\eta + W + o(\eta) > 0 \\ &\Longleftrightarrow \underline{h} + \frac{W}{\eta} + o(1) > 0 \end{aligned}$$

which holds with probability 1 since $\frac{W}{\eta} = O_p\left(\sqrt{\frac{1}{\eta N}}\right) = O_p\left(\sqrt{\frac{\ln N}{\sqrt{N}}}\right) = o_p(1)$. Thus $\sup_t \sqrt{N}\,|t^* - t| \xrightarrow{p} 0$.

Lemma A4

Let the assumptions of Lemma A2 hold. Let $\eta = N^{-1/2}$. Then

$$\sup_t |\widehat{H}(t) - \widehat{H}(t-\eta) - \eta h(t)| = o_p(1).$$

Proof: Note that $H(t) - H(t-\eta) = h(t)\eta + o(\eta)$ and the result follows.

Lemma A5

For any $\delta < \frac{1}{2}$,

$$\sup_q N^{\delta} \left| \widehat{H}^{-1}(q) - H^{-1}(q) \right| \xrightarrow{p} 0. \qquad (14)$$

Proof:

Define $\frac{\partial H^{-1}(q)}{\partial q} = \frac{1}{h(H^{-1}(q))} = \frac{1}{h}$.

By the triangle inequality:

$\sup_q N^{\delta} \left| \widehat{H}^{-1}(q) - H^{-1}(q) \right| = \sup_{t^*} N^{\delta} |t^* - t|$

$\leq \frac{1}{\underline{h}} \sup_{t^*} N^{\delta} |H(t^*) - H(t)|$, with $\frac{1}{\underline{h}}$the upper bound of $\frac{1}{h}$.

This is because, by making use of a Taylor expansion and of the intermediate value theorem,

$$H(t^*) = H(t) + h\left(\widetilde{t}\right)(t^* - t), \text{ with } \widetilde{t} \in (t, t^*)$$
$$\Longrightarrow (t^* - t) = \frac{H(t^*) - H(t)}{h\left(\widetilde{t}\right)} \leq \frac{H(t^*) - H(t)}{\underline{h}}.$$

By Lemma A3, $\sup_t \sqrt{N} |t^* - t| \xrightarrow{p} 0$.

This implies that $\frac{1}{\underline{h}} \sup_{t^*} N^{\delta} |H(t^*) - H(t)| = \frac{h(\widetilde{t})}{\underline{h}} \sup_{t^*} N^{\delta} |t^* - t| \xrightarrow{p} 0, \forall 0 < \delta < \frac{1}{2}$.

Thus $\sup_q N^{\delta} \left| \widehat{H}^{-1}(q) - H^{-1}(q) \right| \xrightarrow{p} 0$.

The remainder of the proof of theorem 1 follows the proof of Athey and Imbens's (2006), Lemma 8.5 and Theorem 5.4 (using our lemma A1-A5).

Proof of **Theorem 2**

Note that

$$\begin{aligned} \widehat{T^*}(X, X^*, T) &= \widehat{H}^{-1}\{\widehat{H}(T) + (X^* - X)\hat{\kappa}\} \\ &= H^{-1}(\hat{q}) + \frac{1}{h(H^{-1}(\hat{q}))}\left(\widehat{H}\left(H^{-1}(\hat{q})\right) - \hat{q}\right) + o_p(1) \end{aligned}$$

by Theorem 1 where $\hat{q} = \widehat{H}(T) + (X^* - X)\hat{\kappa}$. Also note that

$$\widehat{T^*}(X, X^*, T) = H^{-1}(q) + \frac{1}{h(H^{-1}(q))}(\hat{q} - q) + \frac{1}{h(H^{-1}(q))}\left(\widehat{H}\left(H^{-1}(q)\right) - q\right) + o_p(1),$$

where $q = H(T) + (X^* - X)\kappa$. By Chen (2002), $\sqrt{N}\left(\widehat{H}(t) - H(t)\right) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_{t_0,t}(X_i, T_i) + o_p(1)$.

Thus,

$$\widehat{H}(t+\eta) - \widehat{H}(t) - (H(t+\eta) - H(t)) = \frac{\sum_{i=1}^{N} J_{t_0,t+\eta}(X_i, T_i) - \sum_{i=1}^{N} J_{t_0,t}(X_i, T_i)}{N} + o_p\left(\frac{1}{\sqrt{N}}\right).$$

The remainder of the proof of theorem 2 follows the proof of Athey and Imbens's (2006) Theorem 5.4 (where the normality result follows from the normality of $\frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_{t_0,t}(X_i, T_i)$).

References

[1] Athey, S. and G. Imbens, (2006): "Identification and Inference in Nonlinear Difference-In-Differences Models", *Econometrica,* 74, 431 - 497.

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

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

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

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

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

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

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

[9] Härdle, W. (1990): *Applied Nonparametric Regression.* Cambridge: Cambridge University Press.

[10] Hausman, J., and T. Woutersen (2005): Estimating a Semi-Parametric Duration Model with Heterogeneity and Time-Varying Regressors, MIT mimeo.

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

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

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

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

[15] Honoré, B. E., and L. Hu (2010): "Estimation of a transformation model with truncation, interval observation and time-varying covariates", Econometric Journal, Reviews, 13, nr 1, 127–144.

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

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

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

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

[20] Kiefer and Wolfowitz (1956): Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters, *Annals of Mathematical Statistics,* volume 27, 4, 887-906.

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

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

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

[24] Kaplan, E. L. and P. Meier (1958): "Nonparametric Estimation from Incomplete Observations," Journal of the American Statistical Association, 53, 457-481.

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

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

[27] Newey, W. K. (1991): “Uniform Convergence in Probability and Stochastic Equicontinuity”, *Econometrica,* Vol. 59, 4, 1161-1167

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

[29] Ridder, G., and T. Woutersen (2002): “Semiparametric and Method of Moments Estimation of Duration Models with Exogenous and Endogenous Regressors”, UWO working paper.

[30] Ridder, G., and T. Woutersen (2003): “The Singularity of the information matrix of the mixed proportional hazard model”, *Econometrica.*

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

[32] Van den Berg, G. J. (2001): “Duration Models: Specification, Identification, and Multiple Duration,” in *Handbook of Econometrics,* Vol. 5. Amsterdam: North-Holland.

[33] Vytlacil, E (2002): "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," Econometrica, 70, 331-341.

[34] Woutersen, T. (2003): “A New Probability Distribution with some Stability Properties”, MIT mimeo.

Figure 1: Chen Estimator for the Riverside Control Group

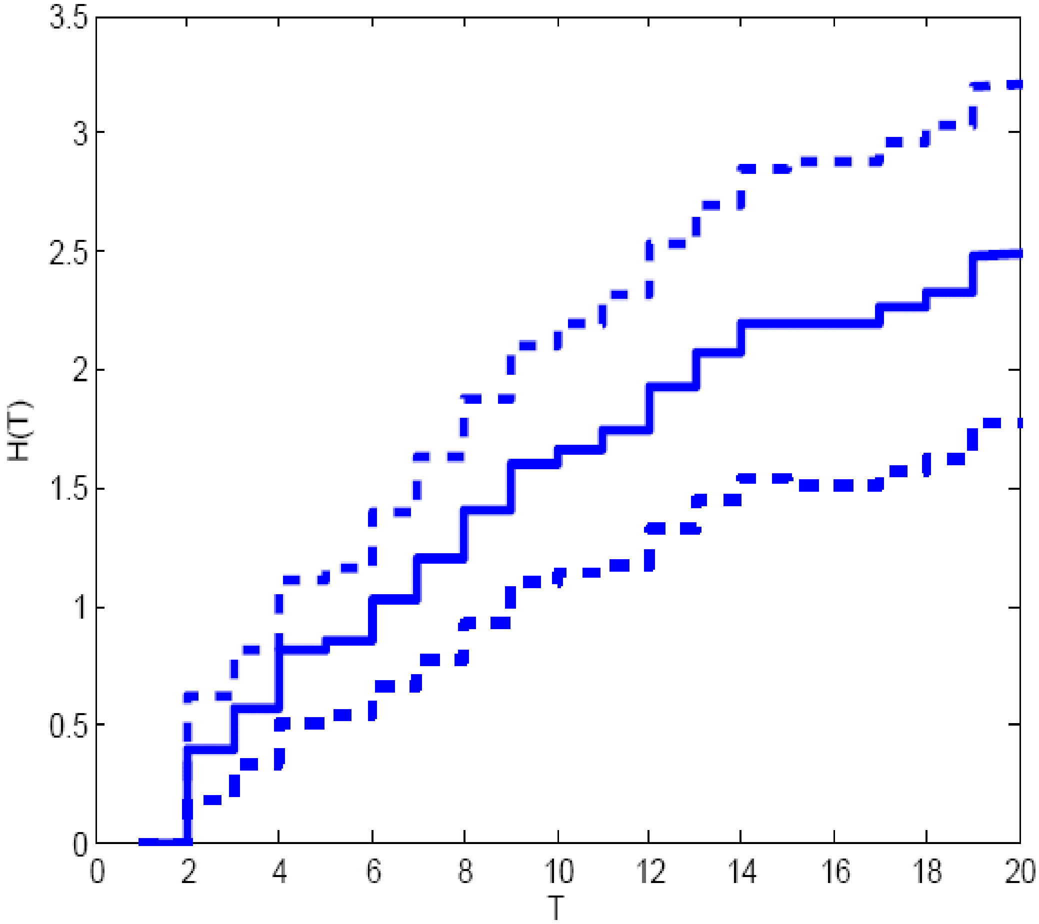


Figure 2: Smoothed Integrated Hazard.
Error bounds based on 250 bootstrap repetitions.

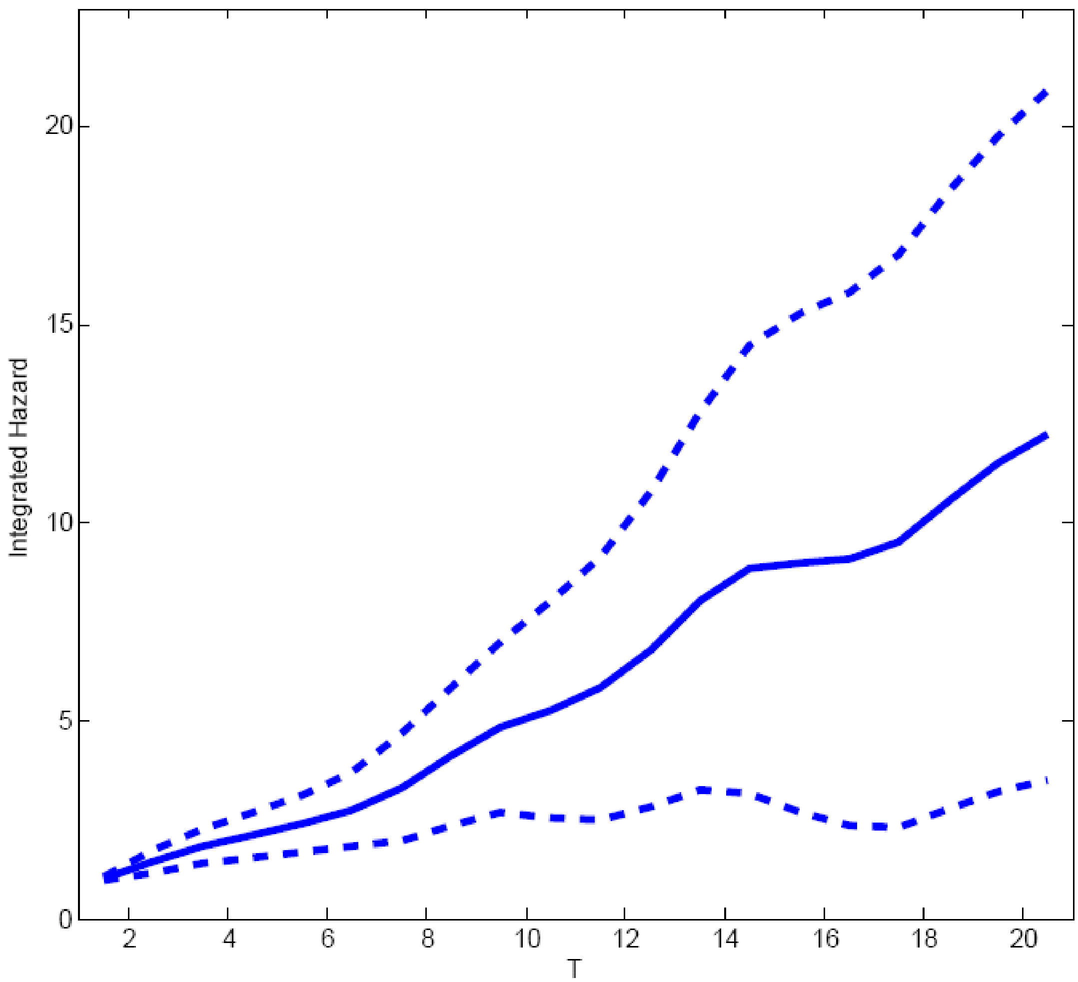
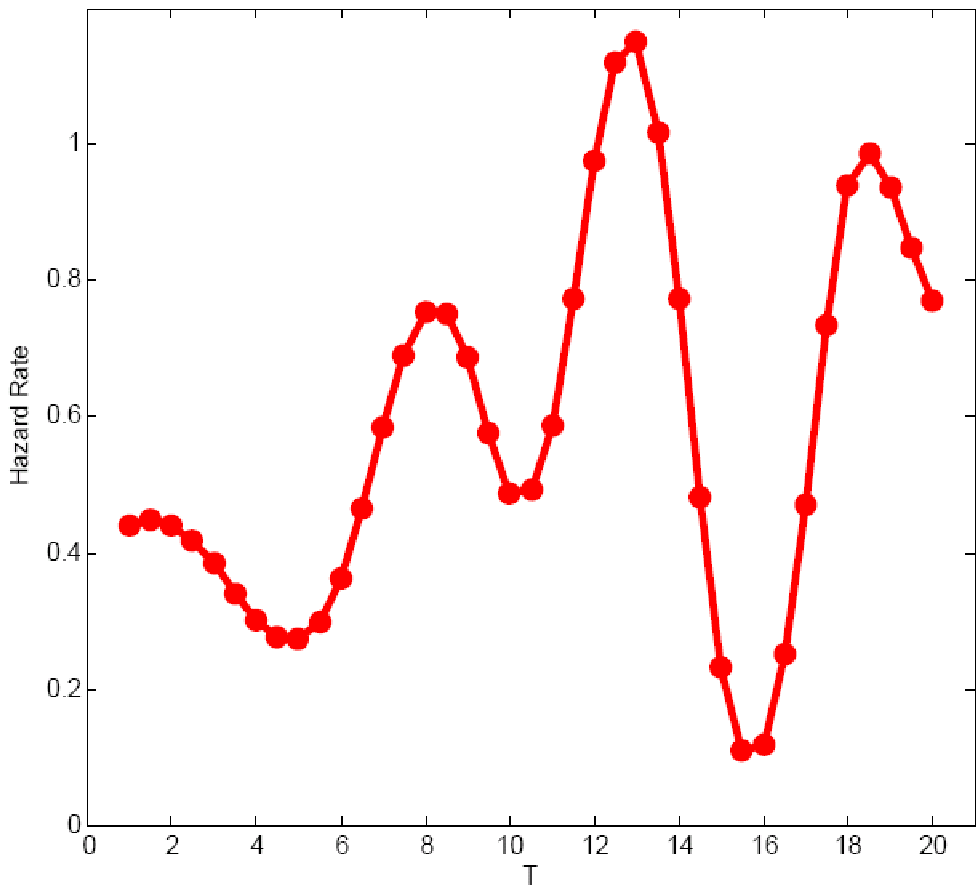


Figure 3: Estimated hazard rate based on the derivative
of the smoothed integrated hazard.



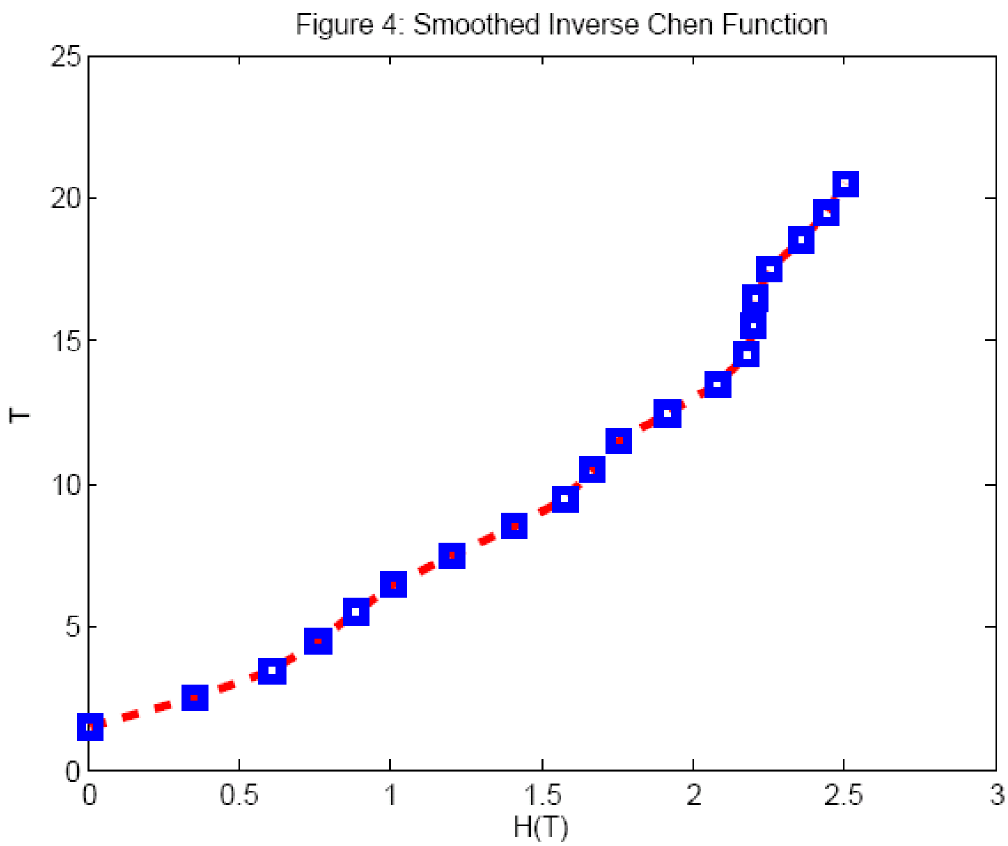


Figure 4: Smoothed Inverse Chen Function
T
H(T)
0
5
10
15
20
25
0
0.5
1
1.5
2
2.5
3