Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

JERRY HAUSMAN AND TIEMEN M. WOUTERSEN†
MIT AND JHU
February 2008

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 the counterfactuals. The advantage to focus 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.

KEYWORDS: Transformation Model, Mixed Proportional Hazard Model, Heterogeneity.
JEL CLASSIFICATION: C41, C24, J64

PRELIMINARY, DO NOT DISTRIBUTE

1. Introduction

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

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

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

$$\ln\{\Lambda(t)\} = -X\beta - \ln(v) + \ln(-\ln(u)). \tag{2}$$

*Correspondence addresses: Massachusetts Institute of Technology, Department of Economics, Building E52-271A, Cambridge 02139, Johns Hopkins University, Department of Economics, 3400 N. Charles Street, Baltimore, MD 21218; Email: jhausman@mit.edu, woutersen@jhu.edu.
†We thank Matthew Harding, Ketan Patel for excellent research assistance. We thank Andrew Chesher, Pierangelo de Pace, Su-Hsin Chang and seminar participants at UCL for helpful comments.
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

The last 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 $u$. The popularity of the mixed proportional hazard model is partly due to the fact that it nests two alternative explanations for the hazard $\theta(t \mid X, v)$ to be decreasing with time. In particular, estimating the mixed proportional hazard model gives the relative importance of the heterogeneity, $v$, and genuine duration dependence, $\lambda(t)$, see Lancaster (1990) and Van den Berg (2002) for overviews. Lancaster (1979) uses functional form assumptions on $\Lambda(t)$ and distributional assumptions on $v$ to identify the model. Examples by Lancaster and Nickell (1981) and Heckman and Singer (1984a), however, show the sensitivity to these functional form and distributional assumptions. Elbers and Ridder (1982), and Heckman and Singer (1984b), however, 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 a closely related transformation model,

$$H(t) = X\alpha + u$$

where $|\alpha_1| = 1$.

This model does not impose the identifying assumption of Elbers and Ridder (1982) that $E v < \infty$ 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 $\alpha$ is necessary for identification, for example, $\alpha = \beta/|\beta_1|$ where $\beta_1$ is the first element of the parameter vector $\beta$. 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 $\beta$, 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 $E v^3 < \infty$ so that the mixed proportional hazard model is identified and derives a nonparametric estimator for $\lambda(t)$ and $\Lambda(t)$. In particular, Horowitz (1999) uses the fact that $e^{H(T)}$ is distributed as a Weibull and uses durations that are close to zero to estimate $\beta_1$. Van den Berg (2000, handbook) argues against
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

relying on very short spells for estimating a model. Hahn (1994) that the information matrix of the Weibull model with heterogeneity is singular so that no estimator can exist that converges at rate $N^{-1/2}$.

This paper introduces the structural duration function, a function that gives the duration of an individual conditional on possible values of the regressor, conditional on $\{v, u\}$. We use the structural duration function to answer counterfactual question such as ‘how long would the duration of individual $i$ have lasted if his regressor would have been $x^*$’ 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 $\{v, u\}$. In the linear model $Y_i = X_i\beta + \varepsilon_i$, such conditioning is done implicitly when the marginal effect is calculated as $\Delta Y_i = \Delta X_i \beta$. In a nonlinear model, one needs to make the conditioning on the stochastic terms explicit, see also Vytlacil (2000). 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 following 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 $X^*$ would have been the regressor happened. The parameters $\{\alpha, \beta\}$ are estimated by Lancaster using maximum likelihood. Lancaster notes that the parameter $\alpha$ is imprecisely estimated for his dataset of more than 500 individuals. Hahn (1994) shows that, without parametric assumptions on the heterogeneity, $\alpha$ cannot be estimated at rate $N^{-1/2}$. We note that the ratio estimated ratio $\beta/\alpha$ in Lancaster’s study remains constant for several values of $\alpha$. Moreover, $\beta/\alpha$ can be estimated at rate $N^{-1/2}$, even without 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. Section 3 shows that our estimator converges at
the regular rate and is asymptotically normally distributed. Section 4 shows that mis-
specifying the heterogeneity yields inconsistent estimates, even if the baseline hazard is
nonparametric. 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 $X$, unobserved heterogeneity $v$, and a function of time $\lambda(t)$,

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

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

$$\ln\{\Lambda(t)\} = -X_{\beta} - \ln(v) + \ln\{-\ln(u)\},$$  \hfill (4)

The last 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 his regressors would
have been different.

(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,
\[ H(t) = X\alpha + u \]  \hspace{1cm} (5)
where \(|\alpha_1| = 1\).

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 \(\alpha_1\) is necessary for identification.

Let \(T^*\) denote that duration if the regressors would have been \(X^*\) instead of the observed \(X\). We condition on \(U\). Then
\[ H(T^*) = X^*\alpha + U \]
\[ = X^*\alpha + H(T) - X\alpha. \]

Consider the structural duration function \(D(X^*, U)\) to calculate \(T^*\). In particular,
\[ T^* = D(X^*, U) = H^{-1}(X^*\alpha - X\alpha + H(T)). \] \hspace{1cm} (6)

Thus, for known \(H(\cdot), H^{-1}(\cdot)\) and \(\alpha\), one can construct the counterfactual \(T^*\). We approximate \(D(X^*, U)\) by replacing \(H(\cdot), H^{-1}(\cdot)\) and \(\alpha\) by estimators. Let \(D_N(X^*, U_N)\) denote and estimator for \(D(X^*, U)\). There are several estimator for \(\alpha\), 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) derive estimators for \(H(T)\).

Horowitz (1996) and Chen (2002) assume that \(H(\cdot)\) is strictly increasing. We assume it to ensure that \(H(\cdot)\) is invertible. Chen (2002) estimator is piece-wise constant and we therefore define an estimator of the inverse of \(H(\cdot)\) as follows. Let \(t = G_N(y)\) where \(t\) is the smallest \(s\) for which \(H_N(s) = y\). We now define \(D_N(X^*, U)\) as follows,
\[ D_N(X^*, U_N) = D_N\{X^*, H_N(T) - X\alpha_N\} \]
\[ = G_N\{(X - X^*)\alpha_N + H(T)\} \]

where \(\alpha_N\) is an estimator for \(\alpha\).

Chen (2002) shows that the following assumptions ensure \(\sqrt{N}\)-convergence of \(\hat{H}(t)\).
A1. \{T_i, X_i, w_i, i = 1, ..., n\} is a random sample of \{T, X, w\} in (4) and \varepsilon is independent of \(X\).

A2. (a) \(|\beta_1| = 1\), (b) the distribution of the first component of \(X\) conditional on \(\bar{X} = \bar{x}\) is absolutely continuous with respect to the Lebesgue measure, (c) the support of \(X\) is not contained in any proper linear subspace of \(R^p\), where \(p\) is the number of exogenous regressors.

A3. \(H(.)\) is strictly increasing, \(H(t_0) = 0, [H(t_1 - \varepsilon), H(t_2 + \varepsilon)] \subset M_H\), for a small \(\varepsilon > 0\), 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 \(X\beta\) given \(\bar{X} = \bar{x}\) and the density of \(\varepsilon\) at \(s\), \(f_{X|\beta}(s|\bar{x})\) and \(p(s)\), are twice differentiable in \(s\), the derivatives are uniformly bounded, and \(\bar{X}\) has finite third-order moments. The same assumption holds for \(w\). 

A5. \(V(t) = \frac{1}{2}E \left[ \frac{\partial^2 E_0[(1{\{t_1 \geq t\} - 1{\{t_2 \geq t_0\})}1{\{x_1 \leq t_0 \}}+((1{\{t_2 \geq t\} - 1{\{t_1 \geq t_0\})}1{\{x_2 \leq t_0 \}})\}}}{\partial H(t)^2} \right]\)

is negative for each \(t \in [u, \overline{t}]\) for any \(u > 0\), and uniformly bounded away from zero.

A6. The first step estimator \(b\) is \(\sqrt{N} - \) consistent, i.e. \(\sqrt{N}(b - \beta) = O_p(1)\).

We also assume that the following assumption holds.

A7. \(H(t)\) is twice continuously differentiable on \(t \in (0, \overline{t})\). \(X^*\) is independent of \(\varepsilon\).

Define \(q = H(T^*) = H(T) - Xb + wb\) and \(H^{-1}(q)\) as the inverse of \(H(T^*)\). Adjust the notation as: \(H_T^{-1}(q) = H^{-1}(q), H_T(t^*) = H(t^*), \hat{H}_T^{-1}(q) = \hat{H}^{-1}(q), \hat{H}_T(t^*) = \hat{H}(t^*)\).

**Theorem 1**

Under assumptions 1 to 7, the estimated counterfactual duration is defined as

\[
\hat{T}^* = \hat{H}^{-1}_T(q),
\]

with the following properties

(i) for any \(u > 0\), \(\sup_{q \in [H^{-1}(u), H^{-1}(\overline{q})]} \sup_q N^{1/2} \left| \hat{H}_T^{-1}(q) - H_T^{-1}(q) \right| + \frac{1}{h_T(H_T^{-1}(q))} \left( \hat{H}_T(H_T^{-1}(q)) - q \right) \overset{P}{\rightarrow} 0.

(ii) \(\hat{H}_T^{-1}(q)\) is a uniformly consistent estimator of \(H_T^{-1}(q)\) i.e. for any \(u > 0\), \(\sup_{q \in [H^{-1}(u), H^{-1}(\overline{q})]} \left| \hat{H}_T^{-1}(q) - H_T^{-1}(q) \right| \overset{P}{\rightarrow} 0\).
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

\( (iii) \sqrt{N} \left( \hat{H}^{-1}(q) - H^{-1}(q) \right) \) is Gaussian with mean zero and asymptotic variance-covariance matrix \( \left\{ N \left( \frac{\partial h_T(t^*)}{\partial t} \right)^{-1} h_T(t^*) h_T(t^*)' \left( \frac{\partial h_T(t^*)}{\partial t} \right)^{-1} \right\} \) where \( h(t) = \frac{\partial H(t)}{\partial t} \) and \( t^* = H^{-1}(q) \).

Proof: See appendix.

**Theorem 2 (Structural Duration Function)**

Let assumptions A1-A7 hold. Let

\[ \hat{T}^*(X, X^*, T) = D_N(X^*, U_N) = \hat{H}_T^{-1} \{ \hat{H}(T) + (X^* - X)\beta \} \]

and

\[ T^*(X, X^*, T) = D(X^*, U) = H_T^{-1} \{ X^*\beta + U \} \]

\[ = H_T^{-1} \{ H(T) + (X^* - X)\beta \} \]

Then

\[ D_N(X^*, U_N) \rightarrow D(X^*, U). \]

Let assumptions A1-A7 hold and let \( X\beta \neq X^*\beta \). Then

\[ \sqrt{N} \{ D_N(X^*, U_N) - D(X^*, U) \} \rightarrow N(0, \Omega). \]

Proof: See 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 \( D_{X_k}(X^*, U_i) \) with respect to \( X_k \). Let \( D_{X_k}(X^*, U_i) \) denote the partial derivative of \( D(X, U_i) \) with respect to \( X_k \), evaluated at \( X^* \),

\[ \frac{\partial T^*_i}{\partial X_k} = D_{X_k}(X^*, U_i) = \frac{\partial D_{X_k}(X, U_i)}{\partial X_k} |_{X=X^*}. \]

Consider writing the error term \( U \) as a function of the duration \( T \) and the regressor \( X \),

\[ U(T, X) = H(T) - X\beta. \]

Total differentiating with respect to \( T \) and \( X_k \) gives

\[ \frac{\partial T}{\partial X_k} = \frac{\partial U(T, X)}{\partial X} \frac{\partial X}{\partial T} = -\frac{\beta_k}{H'(T)}. \]
Consider the following estimator,

\[
\frac{\partial T}{\partial X_k} = D_{X_k,N}(X^*, U) = \frac{\hat{\beta}_k}{H'_N(T^*)}
\]

where \(H'_N(\cdot)\) is an estimator for the derivative of \(H(\cdot)\) with respect to its argument. We base an estimator for \(H'(T)\) on Chen’s (2002) estimator for \(H(T)\). Consider,

\[
H'(T^*) = \frac{1}{\sigma_N} \int_{-\infty}^{\infty} w(\frac{s}{\sigma_N})H_N(s + T^*)ds
\]

where \(w(s)\) is twice continuously differentiable with bounded derivatives and \(w(s) = -w(-s)\). In particular, \(w(s)\) is a derivative kernel as in Härdle (1990).

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

**Theorem 3 (derivative of structural duration function)**

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

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

Proof: Use properties of Gaussian process and properties of derivative kernel.

Undersmoothing (i.e. choosing a \(\sigma_N\) that goes to zero at a rate slower than \(N^{-1/5}\) yields a normally distributed estimator for the derivative.

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. In particular, consider the following two data generating processes.

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

\[
\theta(t|x, v) = e^{x\beta},
\]

where \(x\) is an exogenous regressor that is continuously distributed and \(\beta \neq 0\). This hazard yields the following survival function,

\[
\tilde{F}(t|x) = e^{-e^{x\beta}t}.
\]
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

DGP II:

\[ \theta(t|x, v) = 2ve^{2x \beta t}, \]

where \( p(v) \propto v^{-3/2}e^{-1/(4v)} \).

This mixing distribution was introduced by Woutersen (2002) and this model yields the same survival function as DGP I.

Note that the survival function \( \bar{F}(w|x) = h(w) \) does not identify the mixed proportional hazard model. For simplicity, we assume that \( x \) is a scalar and that \( \beta > 0 \). In large samples, the empirical survival function converges to \( \bar{F}(t|x) = e^{-e^{x \beta t}} \). Without imposing smoothness or parametric assumptions on \( \delta(t) \), our estimate for \( \delta(t) \) converges to

\[ \hat{\delta}(t) = t^{1/\beta}. \]

We now use \( \hat{z}_i(t) = e^{x_i t_i^{1/\beta}} \) to find counterfactuals. In particular, consider the duration if the regressor would have been \( x_i^* \) instead of \( x_i \). Equating \( e^{x_i^* (t_i^*)^{1/\beta}} \) to \( \hat{z}_i(t) \) yields

\[ e^{x_i (t_i^*)^{1/\beta}} = \hat{z}_i(t) = e^{x_i t_i^{1/\beta}}. \]

This gives

\[ (t_i^*)^{1/\beta} = e^{x_i - x_i^*} t_i^{1/\beta} \quad \text{or} \quad t_i^* = e^{(x_i - x_i^*) \beta} t_i. \]

Moreover,

\[ \frac{\partial t_i}{\partial x_i} = -\beta e^{x_i \beta} t_i. \]

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


The structural duration function can also be estimated for \( H(.) \) and/or \( F \) being parametric. The advantage of the structural duration function is that is (i) can remove en empirical identification problem in case both \( H(.) \) and \( F \) are parametric, (ii) 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 in case the transformation model fails to be nonparametrically identified, e.g. in case no regressor is continuously distributed.

2.2. Parametric $H$, parametric $F$. Lancaster (1979) introduced the mixed proportional hazard model and used a Weibull functional for the baseline hazard and modelled the heterogeneity as a gamma distribution. Writing the 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,

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

where $v'$ and $z'$ have a generalized gamma distribution. It is noteworthy, that in Lancaster’s (1979) application, the Weibull parameter $\alpha$ is weakly identified in the sense that standard error of the parameter estimate for $\alpha$ is “quite large” (Lancaster, page 954).

Lancaster (1979) reports his results in table IV and V. In the table IV, Lancaster restricts $\alpha$ to be one. In table V, $\alpha$ is estimated and the estimated values of $\beta$ decrease by about 10%. For the transformation model, the restriction that $\alpha = 1$ implies that $v$ is a gamma and $z$ is a unit exponential random variable. Allowing for generalized gamma distributions hardly changes estimates of $\gamma = \frac{2}{\alpha}$, see the appendix for details. The structural duration function only relies on $\gamma$. 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 explains and reconciles the different estimates of Lancaster (1979) estimates in table IV and V.

2.3. 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,
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

in the second step, \( \ln(T^\alpha) \) would be regressed on \( X \). This second step is equivalent to estimating the following model,

\[
\ln(T) = \frac{X\beta}{\hat{\alpha}} + \varepsilon
\]

where \( X \) and \( \varepsilon \) are independent and \( \hat{\alpha} \) denotes Honoré (1990) estimator for the Weibull coefficient. Applying the approach of this paper would imply to estimate \( \gamma \) using the following model,

\[
\ln(T) = X\gamma + \varepsilon.
\]

The parameter \( \gamma \) can be estimated at rate \( N^{-1/2} \) and counterfactuals would be constructed using estimates for \( \gamma \). 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 \( \alpha \) 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 (1984) 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 (1984) is unknown.

For any integrated baseline hazard rate, we have the equality

\[
\Lambda(t) = e^{H(t)}.
\]

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

\[
H(t; \theta) = X\gamma + U.
\]

The following moments can be used to estimate \( \theta \),

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

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

\(^1\) A set of functions \( \mathcal{H} \) is closed under the power transformation if \( f(t) \in \mathcal{H} \) implies \( \{f(t)^\alpha \}^\alpha \in \mathcal{H} \) for every \( \alpha > 0 \). Here, however, we have \( e^{H(t)} \in \mathcal{H} \) implies \( \{e^{H(t)^\alpha} \}^\alpha \notin \mathcal{H} \) for every \( \alpha > 0 \), \( \alpha \neq 1 \).
2.4. 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 gamma distribution. The idea behind of 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, $\alpha$ may not be well identified in the Weibull model with gamma heterogeneity and changed if one switched from a gamma heterogeneity to a generalized gamma heterogeneity. Below we show that misspecifying the heterogeneity yields inconsistent estimates for all parameters and the having a flexible integrated baseline hazard does not have much effect. We therefore advise against using a flexible $\Lambda(t)$ as a tool to get robustness against misspecification of the heterogeneity.

2.5. 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) consider $\Lambda(t) = e^{\alpha H(t)}$ and estimates $H(t)$ using Horowitz (1996) and estimates $\alpha$ using an estimator that is very similar to Honoré (1990). The rate of convergence of Horowitz (1999) estimator is determined by the rate of convergence of the estimator for $\alpha$. The advantage of the structural duration function is that it avoids estimating and, therefore, converges at rate $N^{-1/2}$.

3. Gamma Mixing Distribution

We show that, under Lancaster’s (1979) assumptions, the counterfactual has the following 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 $X^*$ would have been the regressor. The parameters $\{\alpha, \beta\}$ are estimated by Lancaster using maximum likelihood. Lancaster notes that the estimate of $\alpha$ depends a lot on which regressors he includes and this “worries” him. Hahn (1994) shows that, without parametric assumptions on the heterogeneity, $\alpha$ cannot be estimated at rate $N^{-1/2}$. We note that the ratio estimated ratio $\beta/\alpha$ in Lancaster’s study depends less on which regressors are included than when $\beta$ or $\alpha$ are compared separately. Moreover, $\beta/\alpha$
can be estimated at rate $N^{-1/2}$.

4. **Structural Duration Function for unfinished spells**

For completed spells, we can condition on $u$ and calculate the structural duration function. The intuition for this result is that, for known $\beta$ and $H(t)$, we can calculate $U = H(T) - X\beta$ and $T^* = H^{-1}(X^*\beta + U)$. However, for censored or unfinished spells, we do not observe the duration $T$ and therefore cannot calculate $T^*$. However, for known $\beta$ and $H(t)$, $T \geq \hat{t}$ implies that $U \geq H(\hat{t}) - X\beta$. Suppose the censoring is not so severe that one can still calculate the median of $U$ conditional on $U \geq H(\hat{t}) - X\beta$. We denote this conditional median by $U_{med}$ and calculate $T^*_{med} = H^{-1}(X^*\beta + U_{med})$. We can also calculate other quantiles of $T^*$. Replacing $H^{-1}(\cdot)$, $\beta$, and $U_{med}$ by estimates yields the following theorem.

**Theorem 4 (median of unfinished spells)**

Let $T^*_{med} = H^{-1}(X^*\beta + U_{med})$. Let ...

Then $T^*_{med} - T^*_{med}$ is $O_p(T^{-1/2})$.

Other quantiles can be estimated at the same rate. Theorem 3 is useful to predict that median unemployment duration of an individual that has been unemployed for a while.

In particular, the median duration can be estimated for different values of the regressor. In particular, let individual $i$ have been unemployed for $t$ weeks and have regressor $X_i$. Then $U_i$ let job search training since $X\beta$ can be absorbed.

Chen also considers the model estimation for censored data. The model used for the censored observations is

$$H(T^c) = X\beta + \varepsilon,$$

(8)

where $T^c$ is unobserved. We observe instead $T = \min\{T^c, C\}$, and $X$, with $C$ a random censoring variable. He uses the following normalization $H(t_0) = 0$ for a finite $t_0$ and a scale normalization for the coefficient of the first component of $X$, $|\beta_1| = 1$.

Some notation is useful $d_{it} = 1\{T_i \geq t\}$, $d_{jt0} = 1\{T_j \geq t_0\}$, $M_H$ is a compact set, $G(t) = Pr(C > t)$ for any $t$, $\bar{X}$ is the vector containing all but first component of $X$, $\delta_i = 1 \{T^c \leq C\}$. 

The estimator of the transformation model $H(T)$ for censored data is

$$H_n(t) = \arg \max_{b \in H(t)} Q(H(t)) = \arg \max_{H(t)} \frac{1}{n(n-1)} \sum_{i \neq j} \left( \frac{d_{it}}{G_n(t)} - \frac{d_{j0}}{G_n(t_0)} \right) 1\{X_i b - X_j b \geq H(t)\},$$

(9)

for a given $t \in [t_1, t_2]$, $b$ a consistent estimator for $\beta$ and $G_n$, the Kaplan-Meier estimator or the product limit estimator for the survival function $G$.

The large sample properties of the $H_n(t)$ estimator, when the data is censored, are presented by Chen in Theorem 2. For the censored case Chen defines other three assumptions

A9. $\{X_i, T_i, \delta_i, i = 1, 2, ..., n\}$ is a random sample of $\{X, T, \delta\}$ in (2) and $\varepsilon$ 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 $V^c(t) = \frac{1}{2} E \left[ \frac{\partial^2}{\partial H(t)^2} \left( \left( \frac{1 - 1_{t \geq t_2}}{C(t)} \right) \right) 1\{X b - X b \geq H(t)\} + \left( \frac{1_{t < t_1}}{C(t)} \right) \right] 1\{X b - X b \geq H(t)\} \right]$,  

A11. $V^c(t)$ is negative for each $t \in [t_1, t_2]$ and uniformly bounded away from zero.

Chen’s Theorem 2 states that under assumptions 2-4 and 9-11:

(i) $\sup_{t_1 \leq t \leq t_2} |H_n^c(t) - H_0(t)| = o_p(1)$;

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

$$\sqrt{N} (H_n^c(t) - H_0(t)) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} J_{t_0, t}^c (X_i, T_i, \delta_i) + o_p(1),$$

(10)

and $N^{1/2} (H_n^c(t) - H_0(t)) \Rightarrow \mathcal{G}_H^c (t_0, t)$ where $\mathcal{G}_H^c (t_0, t)$ is a Gaussian process with mean 0 and covariance function $E \mathcal{G}_H^c (t_0, t) G_H^c (t_0, t') = \mathbb{E}J_{t_0, t'}^c (X, T, \delta) J_{t_0, t'}^c (X, T, \delta)$ with

$$J_{t_0, t}^c (X, T) = -V^c (t_0, t)^{-1} \left[ \frac{\partial E \left( \left( \frac{1_{t_1 \geq t}}{C(t)} \right) \right) 1\{X b - X b \geq H(t)\} + \left( \frac{1_{t_2 \geq t_1}}{C(t)} \right) \right] 1\{X b - X b \geq H(t)\} + \xi_i (t) \right]$$

with $\xi_i (t)$ defined as in Chen.

5. Empirical Results

The data used in the empirical analysis are from the Nation 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\(^2\). The

NEWWWS 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 at the time entry into 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. Members of the LFA 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. 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. It was at these orientations that the subject was randomly sorted into one of the three groups (LFA, HCD, and control).

California had existing welfare regulations governing the offering of educational assistance. Because of these existing regulations, only those subjects in need of basic education could be sorted into the HCD group. This group consistent of subjects who did not have a high school degree or GED, were not fluent in English, or scored below 215 on any part

---

3 These strategies were also tested in Atlanta, GA and Grand Rapids, MI.
4 US HHS op. cit.
5 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.
of the CASAS were eligible to be sorted into the HCD group. 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 1.6

<table>
<thead>
<tr>
<th>Group</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labor Force Attachment</td>
<td>3384</td>
</tr>
<tr>
<td>Human Capital Development</td>
<td>1596</td>
</tr>
<tr>
<td>Control</td>
<td>3342</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td><strong>8322</strong></td>
</tr>
</tbody>
</table>

After sorting participants into one of the three groups, the study tracked them for five
years, mainly through state-level government agency reporting. Among the data collected
are quarterly information about AFDC grant levels, food stamp grants, employment sta-
tus, and earnings. Demographic characteristics as of the random assignment are also pro-
vided; 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>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-White</td>
<td>= 1 if Black or Hispanic</td>
<td>0.446</td>
<td>0.497</td>
</tr>
<tr>
<td>HS Education</td>
<td>= 1 if HS degree or GED</td>
<td>0.580</td>
<td>0.494</td>
</tr>
<tr>
<td>Young Child</td>
<td>= 1 if child &lt; 5</td>
<td>0.572</td>
<td>0.495</td>
</tr>
<tr>
<td>1 Child</td>
<td>= 1 if 1 child</td>
<td>0.390</td>
<td>0.488</td>
</tr>
<tr>
<td>2 Children</td>
<td>= 1 if 2 children</td>
<td>0.324</td>
<td>0.368</td>
</tr>
<tr>
<td>3+ Children</td>
<td>= 1 if 3+ children</td>
<td>0.285</td>
<td>0.452</td>
</tr>
<tr>
<td>Age &lt; 30</td>
<td>= 1 if age &lt; 30</td>
<td>0.402</td>
<td>0.490</td>
</tr>
<tr>
<td>Age 30-39</td>
<td>= 1 if age 30-39</td>
<td>0.465</td>
<td>0.499</td>
</tr>
<tr>
<td>Age 40+</td>
<td>= 1 if age 40+</td>
<td>0.133</td>
<td>0.340</td>
</tr>
<tr>
<td>log(Avg. welfare)</td>
<td>log of avg. FS + AFDC</td>
<td>8.238</td>
<td>1.165</td>
</tr>
<tr>
<td>Employed prev. year</td>
<td>= 1 if employed in year before entering study</td>
<td>0.396</td>
<td>0.489</td>
</tr>
</tbody>
</table>

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

---

6Brock and Harknett (1998) found that about 66% of the people directed to attend an orientation in
Riverside actually attended.
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. 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.

We now discuss estimation of three estimators: 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 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:

---

7 We group Asians with Whites because their employment and earnings characteristics are more similar to that of Whites than Blacks or Hispanics.
Table 3: Estimated Coefficients

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

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 of previously employed so we can compare the coefficient estimates to the MRC and CS estimators. The MRC and CS estimates should be close to each other as they have the same limiting distribution. We find all of the estimates to 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 estimator of the integrated baseline hazard along with two standard errors bounds derived from bootstrap estimation. The estimates are quite precise. In Figure 2 we use the Chen 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
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

becomes considerably greater as the durations become longer. In figure 3 we calculate the hazard rate as the derivative of the integrated hazard estimates. Note the non-monotonic features of the estimates. Lastly, in Figure 4 we calculate the inverse Chen 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 suing the estimates in Table 3 along with the estimates of the inverse integrated baseline hazard for use in equation (11). We consider change in the average welfare amount in Table 4:

<table>
<thead>
<tr>
<th></th>
<th>Change in Welfare Benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>HHM</strong></td>
<td>-20.00%  -10.00%  0.00%   10.00%  20.00%</td>
</tr>
<tr>
<td>Unemployment Duration</td>
<td>Mean   3.85   5.14   8.41   9.35   9.92</td>
</tr>
<tr>
<td></td>
<td>Median  2.90   2.90   4.00   3.90   3.90</td>
</tr>
<tr>
<td></td>
<td>Standard deviation  2.88   4.54   8.33   8.59   8.89</td>
</tr>
<tr>
<td><strong>MRC</strong></td>
<td>-20.00%  -10.00%  0.00%   10.00%  20.00%</td>
</tr>
<tr>
<td>Unemployment Duration</td>
<td>Mean  6.64   7.09   8.41   8.78   9.92</td>
</tr>
<tr>
<td></td>
<td>Median  2.90   2.90   4.00   3.90   3.90</td>
</tr>
<tr>
<td></td>
<td>Standard deviation  6.61   7.12   8.33   8.81   8.21</td>
</tr>
<tr>
<td><strong>Cavanagh-Sherman (CS)</strong></td>
<td>Change in Welfare Benefits</td>
</tr>
<tr>
<td>Unemployment Duration</td>
<td>-20.00%  -10.00%  0.00%   10.00%  20.00%</td>
</tr>
<tr>
<td>Mean</td>
<td>7.05   7.69   8.41   8.70   8.80</td>
</tr>
<tr>
<td>Median</td>
<td>2.90   2.90   4.00   3.90   3.90</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>7.12   7.93   8.33   8.07   8.12</td>
</tr>
</tbody>
</table>

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 7.0% large change in duration than does the model based on the MRC estimates and a 7.8% large change the estimates based on the CS estimates. 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 require estimation of
Estimating the Derivative Function and Counterfactuals in Duration Models with Heterogeneity

Heterogeneity may provide more reliable results. Also note that the estimates of policy changes for the estimates based on MRC and CS are considerably closer to each other than the estimates based on HHM.

6. Conclusion

In conclusion we have derived a model and estimator that allows for root-$N$ consistent estimates of the unknown parameters of a duration models that permits counterfactual policy estimates without the need of specifying a parametric heterogeneity distribution or adopting stringent identification assumptions. In a subsequent draft we will demonstrate how to estimate the treatment response effects using this setup even in the presence of non-random takeup or non-random attrition.

Appendix: The Transformation Model

Consider the following transformation model

$$H(T) = X\beta + \varepsilon,$$  \hspace{1cm} (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), $\beta$ is the vector of associated coefficients and $\varepsilon$ 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 uses the following normalization $H(t_0) = 0$ for a finite $t_0$ and a scale normalization for the coefficient of the first component of $X$, $\|\beta_1\| = 1$.

Some notation is useful $d_{it} = 1\{T_i \geq t\}$, $d_{j0} = 1\{T_j \geq t_0\}$, $M_H$ is a compact set, $\tilde{X}$ is the vector containing all but first component of $X$, $\delta_i = 1\{T^c \leq C\}$.

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

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