Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells*

Jason Abrevaya

University of Chicago

Graduate School of Business

Chicago, IL 60637, U.S.A.

Jerry A. Hausman

Massachusetts Institute of Technology

Department of Economics

Cambridge, MA 02139, U.S.A.

This version: April 1999

Abstract

This paper considers mismeasurement of the dependent variable in a general linear index model, which includes qualitative choice models, proportional and additive hazard models, and censored models as special cases. The monotone rank estimator of Cavanagh and Sherman (1998) is shown to be consistent in the presence of any mismeasurement process that obeys a simple stochastic-dominance condition. The emphasis is on measurement error which is independent of the covariates, but extensions to covariate-dependent measurement error are also discussed. We consider the proportional hazard duration model in detail and apply the estimator to mismeasured unemployment duration data from the Survey of Income and Program Participation (SIPP).

*The authors thank Bruce Meyer, Whitney Newey, Patrick Sevestre, two anonymous referees, and seminar participants at Harvard, Michigan, and MIT for their helpful comments. A preliminary version of this paper was presented at the 1995 NSF Symposium on Duration Models at UC-Berkeley. Jon Gruber provided the SIPP extract and simulated unemployment benefit eligibilities. David Myatt provided excellent research assistance. The first author gratefully acknowledges financial support from the National Science Foundation and the University of Chicago GSB faculty research fund.

1 Introduction

The issue of measurement error has been studied extensively in econometrics but almost exclusively with respect to the independent variables. The studies that have considered mismeasurement of the dependent variable have focused on misclassification of responses in qualitative choice models (e.g., Hausman et. al. (1998) and Poterba and Summers (1995)). Strangely enough, mismeasurement of continuous dependent variables has received almost no attention (aside from the passing remark in econometrics textbooks that additive errors in the classical linear model do not affect the consistency of least squares regression).

This paper provides a treatment of mismeasured dependent variables in a general linear index model, which includes qualitative choice models, proportional and additive hazard models, and censored models as special cases. The emphasis is on measurement error which is independent of the covariates, although extensions to covariate-dependent measurement error are also considered.

Parametric techniques are discussed first. The general conclusion is that parametric estimation results in inconsistent estimates of the parameters of interest if the mismeasurement is incorrectly modeled (or ignored altogether). Once one moves away from a simple model of misclassification (as in the binary choice model), parametric estimation becomes quite cumbersome; moreover, the likelihood of correctly modeling the mismeasurement is greatly reduced.

Semiparametric estimation, using the *monotone rank estimator* of Cavanagh and Sherman (1998), is proposed as an alternative to parametric estimation.[1] The advantage of the semiparametric approach is that the mismeasurement need not be modeled at all. The basic insight is that the monotone rank estimates of the coefficient parameters remain consistent as long as an intuitive sufficient condition is satisfied.

One way of thinking about the measurement error is that the observed dependent variable

[1] Other semiparametric estimators of linear index models, like those of Ichimura (1993) and Powell et. al. (1989), could be used in the presence of mismeasured dependent variables. The rank estimator has been chosen since estimation of the parameters of interest does not require the use of kernels or the choice of bandwidth parameters. The maximum rank correlation estimator of Han (1987) could also be used, under the same sufficient condition presented in this paper. (See Sherman (1993) for the asymptotic distribution of the maximum rank correlation estimator.) The monotone rank estimator is used due to its quicker computation speed.

is a realization from a random variable that depends on the true underlying dependent variable (the "latent dependent variable"). A sufficient condition for consistency of monotone rank estimation is that the random variable associated with a larger latent dependent variable *first-order stochastically dominates* the random variable associated with a smaller latent dependent variable.

The paper is organized as follows. Section 2 describes the general model of interest, formalizes the mismeasurement process, and introduces three illustrative examples. Section 3 describes maximum likelihood estimation methods in the presence of mismeasurement, focusing on the examples from Section 2. Section 4 introduces the semiparametric approach, formalizes the sufficient condition for consistency, and interprets the condition in the context of the examples. Section 5 extends the semiparametric approach to situations in which the measurement error in the dependent variable is dependent upon covariates. Section 6 considers the proportional hazard model in detail. Existing parametric and semiparametric estimation techniques are inconsistent when durations are mismeasured, as illustrated by Monte Carlo simulations. The monotone rank estimator is used to estimate an unemployment duration model using data from the Survey of Income and Program Participation (SIPP) and the results are compared to those obtained using traditional techniques. Finally, Section 7 concludes.

2 The Model

Consider the following model, which is an extension of the *generalized regression model* of Han (1987). The latent dependent variable is described by

$$y^* = g(x\beta_o, \epsilon), \epsilon \text{ i.i.d.}, \tag{1}$$

where g is an unknown function with strictly positive partial derivatives everywhere.[2] The model given by (1) is quite general. For instance, it includes models with nonlinearity on the left-hand side,

$$f(y^*) = x\beta_o + \epsilon, \tag{2}$$

[2] Additional disturbances may be included in (1). As an example, such a specification would allow for unobserved heterogeneity or random right-censoring, as discussed in Section 6.

and models with nonlinearity on the right-hand side,

$$y^* = f(x\beta_o) + \epsilon, \tag{3}$$

where f is strictly increasing. Both of these models have ϵ entering additively, though that is not a restriction made in (1).

If there is no mismeasurement of the dependent variable, the observed dependent variable y would be a deterministic function of y^*. Let $d: \mathcal{R} \to \mathcal{R}$ be the (weakly) increasing function defining y in terms of y^*. For instance, the binary choice model has $d(z) = 1(z > 0)$, the traditional censored model has $d(z) = z \cdot 1(z > 0)$, and a model with no censoring has $d(z) = z$. This deterministic specification is the one considered by Han (1987).

To introduce the possibility of mismeasurement, one can instead model y as a stochastic function of the underlying y^*, where the distribution of y has the conditional c.d.f. given by

$$F_{y|y^*}(v|t) = \Pr(y \leq v | y^* = t). \tag{4}$$

For most of this paper, it is assumed that the mismeasurement is independent of x:

$$\Pr(y \leq v | y^* = t, x) = \Pr(y \leq v | y^* = t). \tag{5}$$

Extension to covariate-dependent measurement error is considered in a later section.

The case of perfectly measured dependent variables corresponds to a c.d.f. with a single jump from zero to one. If d denotes the deterministic function described above, then

$$F_{y|y^*}(v|t) = 1(v > d(t)). \tag{6}$$

We consider three simple models of mismeasurement below. In each example, the observed dependent variable takes on a different form. The first and second examples consider the case of discrete-valued dependent variables, with the first example focusing on the binary-choice case. The third example considers the case of a continuous dependent variable. The duration model application considered in Section 6 can be thought of as a hybrid of the second and third examples. The unemployment durations take on integer values (corresponding to the number of weeks of unemployment), but the range of possible durations is large enough that viewing the dependent variable as continuous is a good approximation.

Example 1: Binary Choice with Misclassification

Following Hausman et. al. (1998), assume that there is some probability (independent of x) that the binary response will be misclassified. The latent dependent variable is $y^* = x\beta_o + \epsilon$, and the misclassification errors are

$$\alpha_0 \equiv \Pr(y = 1 | y^* < 0) \tag{7}$$

$$\alpha_1 \equiv \Pr(y = 0 | y^* > 0). \tag{8}$$

The conditional c.d.f. $F_{y|y^*}$ is

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 0 \\ 1 - \alpha_0 & \text{if } v \in [0, 1) \quad \text{if } t < 0 \\ 1 & \text{if } v \geq 1 \end{cases} \tag{9}$$

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 0 \\ \alpha_1 & \text{if } v \in [0, 1) \quad \text{if } t \geq 0. \\ 1 & \text{if } v \geq 1 \end{cases} \tag{10}$$

No matter how negative y^* is, there is a positive probability (equal to α_0) that the response will be misclassified as a one. Thus, for negative y^*, $F_{y|y^*}$ jumps from 0 to $1 - \alpha_0$ at 0 and from $1 - \alpha_0$ to 1 at 1. The conditional c.d.f. for positive y^* also has two jumps.

This model is a bit simplistic since one might want to allow the probability of misclassification to depend on the level of y^*. In addition, the model has a discontinuity at $y^* = 0$ for $\alpha_0 \neq \alpha_1$. The misclassification could instead be modeled with the function $\alpha : \mathcal{R} \to [0, 1]$, defined by

$$\alpha(t) \equiv \Pr(y = 1 | y^* = t) \text{ for } t \in \mathcal{R}. \tag{11}$$

Example 2: Mismeasured Discrete Dependent Variable

The binary choice framework can be extended to handle (ordered) discrete dependent variables with more than two possible values. Without loss of generality, assume that the dependent variable can take on any integer value between 1 and K. The (continuous) latent variable $y^* = x\beta_o + \epsilon$ will belong to one of K subsets $S_1, S_2, \ldots, S_K$ of the real line. In the absence of mismeasurement, the value of y corresponds to the subscript of the subset containing y^*; i.e., $y = t$ if and only if $y^* \in S_t$.

To introduce mismeasurement, parametrize the misclassification probabilities by $\alpha_{s,t}$ for each s and t in $\{1,\ldots,K\}$, where

$$\alpha_{s,t} = \Pr(y = t | y^* \in S_s). \tag{12}$$

Then, $\alpha_{s,s}$ is the probability that a response is correctly classified, and $\sum_t \alpha_{s,t} = 1$ by definition. We can represent this misclassification with a (transition) matrix

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \cdots & \alpha_{1,K} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \cdots & \alpha_{2,K} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \alpha_{K,1} & \cdots & \cdots & \alpha_{K,K-1} & \alpha_{K,K} \end{bmatrix} \tag{13}$$

with the elements of each row adding up to one.

The conditional c.d.f. $F_{y|y^*}$ is

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 1 \\ \sum_{i=1}^{\lfloor v \rfloor} \alpha_{s,i} & \text{if } v \in [1, K] \quad \text{for } t \in S_s \\ 1 & \text{if } v > K \end{cases} \tag{14}$$

where $\lfloor v \rfloor$ denotes the integer part of v.

As in Example 1, a more complicated model could allow the misclassification probabilities to depend on the actual level of y^* rather than just the subset S_s to which it belongs.

Example 3: Mismeasured Continuous Dependent Variable

Consider a setting in which the dependent variable is continuous and can take on any real value. Assume that the observed dependent variable is a function of the latent dependent variable and a random disturbance,

$$y = h(y^*, \eta), \tag{15}$$

where η is i.i.d. (independent of x and ϵ) and h is an unknown function satisfying $h_{y^*} > 0$ and $h_\eta > 0$. This model is quite general, including additive and multiplicative i.i.d. measurement errors as special cases.

The conditional c.d.f. $F_{y|y^*}$ is

$$F_{y|y^*} = \Pr(h(t, \eta) \le v). \tag{16}$$

Under the further assumptions that h_η is continuous and η has positive density everywhere along the real line, there exists a function $\tilde{h}$ such that

$$\Pr(h(t,\eta) \leq v) = \Pr(\eta \leq \tilde{h}(v,t)). \tag{17}$$

Combining (16) and (17) gives

$$F_{y|y^*} = G(\tilde{h}(v,t)), \tag{18}$$

where G is the c.d.f. of η.

3 Parametric Estimation

In this section, the focus will be on parametrization of the measurement model and maximum likelihood estimation of the parametrized model. The basic assumption needed for this approach is that the mismeasurement can be modeled in terms of a finite-dimensional parameter. At the true value of the parameter, the underlying conditional c.d.f. $F_{y|y^*}$ will be modeled correctly; for all other values, the c.d.f. will be modeled incorrectly. Technicalities concerning identification will be avoided for the most part, though some issues will be considered in the examples below.

Example 1 continued

Under a parametric assumption on the c.d.f. of ϵ, Hausman et. al. (1998) discuss maximum likelihood estimation of $(\beta_o, \alpha_0, \alpha_1)$ in the simple model of misclassification discussed in the previous section. If the mismeasurement has been modeled incorrectly (e.g., if the misclassification probabilities depend on the level of y^* rather than just its sign), the estimates will be inconsistent; a more complicated likelihood function would need to be specified.

Example 2 continued

The probabilities for the observed dependent variable are given by

$$\Pr(y = t|x) = \sum_{s=1}^{K} \alpha_{s,t} \Pr(y^* \in S_s), \tag{19}$$

where $y^* = x\beta_o + \epsilon$. Assume the cutoff points for the sets $S_1, S_2, \ldots, S_K$ are known by the researcher (e.g., $S_1 = \{v : v \leq 10\}, S_2 = \{v : 10 < v \leq 20\}, S_3 = \{v : v > 20\}$). Denote the cutoff points by $-\infty = c_0 < c_1 < \cdots < c_{K-1} < c_K = \infty$ so that $S_s = (c_{s-1}, c_s]$. In situations where the cutoff points are unknown (e.g., an ordered qualitative variable having values "poor," "good," and "excellent"), they can be jointly estimated in the likelihood function.

In order to form a likelihood function, a parametrization of the distribution of ϵ is needed, so that $\Pr(y^* \in S_s)$ can be written in terms of estimable parameters. We assume a parametrized distribution for H, the c.d.f. of $-\epsilon$. For simplicity, assume that H is the c.d.f. of a normal random variable having a standard deviation of σ.[3] Then, the likelihood function is

$$\ln \mathcal{L}(b, \hat{\sigma}, \{a_{s,t}\}) = \sum_i \sum_{t=1}^{K} 1(y_i = t) \ln \left\{ \sum_{s=1}^{K} a_{s,t} [H(x_i b - c_s, \hat{\sigma}) - H(x_i b - c_{s-1}, \hat{\sigma})] \right\}, \quad (20)$$

subject to the constraints $\sum_t \alpha_{s,t} = 1$ for each s. If no other restrictions are placed on the misclassification probabilities, there are $K(K-1)$ parameters to be estimated in addition to β_o and σ. For large K, this approach is cumbersome and will result in inefficient estimates. Depending on prior knowledge about the misclassification, though, one might be willing to impose further restrictions on the misclassification probabilities. An example would be that the observable variable is at worst misclassified into an adjacent cell (i.e., $a_{s,t} = 0$ if $|s-t| > 1$), in which case only $2(K-1)$ additional parameters are estimated.

As in the binary choice case, consistency depends on the correct specification of the misclassification process. If the misclassification depends on the level of y^* and not just the subset S_s to which it belongs, maximization of the above likelihood function will yield inconsistent estimates.

Example 3 continued

Recall that G denotes the c.d.f. of η and define $f_{y^*|x}$ as the conditional density of y^*. Then, the c.d.f. of the observable dependent variable y can be written as

$$\begin{aligned} \Pr(y \leq v|x) &= \int F_{y|y^*}(v|t) f_{y^*|x}(t|x)\, dt \\ &= \int G(\tilde{h}(v,t)) f_{y^*|x}(t|x)\, dt, \end{aligned} \quad (21)$$

[3] The standard deviation of the error can be estimated here since the cutoff points are known, identifying β_o and σ. In the binary choice case, only the ratio β_o/σ is identified.

where the second equality follows from (18). To form a parametrized likelihood function, we need to parametrize both G and $f_{y^*|x}$, which requires parametric assumptions on the distributions of η and ϵ, respectively. The parametrization of $f_{y^*|x}$ also requires knowledge of the function g from (1). For simplicity, assume that the parametric assumptions are fully described by the parameters σ_η and σ_ϵ. The further assumptions needed are that h (and therefore $\tilde{h}$) is known (e.g., it is known whether the mismeasurement is additive or multiplicative), $\tilde{h}$ is differentiable with respect to its second argument, and G is differentiable.

Then, differentiation of (21) yields the likelihood function

$$\ln \mathcal{L}(b, \hat{\sigma}_\eta, \hat{\sigma}_\epsilon) = \sum_i \ln \left\{ \int \tilde{h}_2(y_i, t) G'(\tilde{h}(y_i, t), \hat{\sigma}_\eta) f_{y^*|x}(t|x_i, \hat{\sigma}_\epsilon)\, dt \right\}. \tag{22}$$

Equation (22) is similar to likelihoods used to capture heterogeneity, in which mixing distributions are used to model random effects. Maximization of (22) usually requires numerical integration. In general, consistency of the estimates of β_o requires that h, G, and $f_{y^*|x}$ are correctly specified.

4 Semiparametric Estimation

In this section, we discuss semiparametric estimation in the presence of mismeasured dependent variables. The approach described is extremely useful when the researcher suspects mismeasurement but lacks any additional prior information for forming a reliable model of mismeasurement. Even if the researcher is confident of the underlying mismeasurement process, the semiparametric approach can, at the very least, serve as a useful specification check of the model.

The section is organized as follows. First, we discuss the monotone rank estimator (MRE) developed by Cavanagh and Sherman (1998). Second, we describe an intuitive sufficient condition for the consistency of the estimator in the presence of mismeasured dependent variables. The key insight is that the consistency of the semiparametric estimator does not require a model of the measurement error. Third, we interpret the sufficient condition in the context of our examples.

4.1 The Monotone Rank Estimator

The MRE, as defined by Cavanagh and Sherman (1998), is the vector $\hat{\beta}^{\text{MRE}}$ that maximizes the objective function

$$S^{\text{MRE}}(b) = \sum_i M(y_i) \cdot \text{Rank}(x_i b), \tag{23}$$

over the set $\mathcal{B} \equiv \{b \in \mathcal{R}^d : |b_d| = 1\}$, where $M : \{y_1, \ldots, y_n\} \to \mathcal{R}$ is some increasing function (i.e., $y_i > y_j \implies M(y_i) > M(y_j)$). There are d covariates contained in x, which does not include a constant (since rankings are not affected by location). b_d denotes the d-th component of the vector b. The Rank$(\cdot)$ function is defined as follows:[4]

$$x_{i_1} b < x_{i_2} b < \cdots < x_{i_n} b \implies \text{Rank}(x_{i_m} b) = m. \tag{24}$$

Since the ranking of $x_i b$ is unaffected by the scale of b (i.e., $\text{Rank}(x_i b) = \text{Rank}(x_i(cb))$ for $c > 0$), β_o is only identified up-to-scale using the MRE and a normalization ($|b_d| = 1$) is required.

The key condition needed for consistency of $\hat{\beta}^{\text{MRE}}$ is

$$H(z) = \text{E}[M(y)|x\beta_o = z] \text{ increasing in } z. \tag{25}$$

By "increasing in z" (here and in the following conditions), we mean that $H(z)$ is a nonconstant, increasing function of z (see Assumption (A1) in the appendix). The monotonicity condition (25) says that, on average, higher $x\beta_o$ are associated with higher y. This "correlation" is maximized by the objective function (23).

4.2 Sufficient Condition for Consistency

The monotonicity condition is satisfied for the latent variable y^* in the model given by (1). That is,

$$(1) \implies H^*(z) = \text{E}[M(y^*)|x\beta_o = z] \text{ increasing in } z. \tag{26}$$

To have the monotonicity condition hold for the observed dependent variable y when there is mismeasurement, it suffices to have

$$\text{E}[M(y)|y^* = t] \text{ increasing in } t. \tag{27}$$

[4] It is innocuous to consider strict inequalities here due to a continuity assumption on x needed for consistency; see the proof of Theorem 1 in the appendix.

A sufficient condition for (27), and thus for (25), is that the distribution of y for a higher y^* first-order stochastically dominates the distribution of y for a lower y^*. This result is analogous to the result in microeconomics that a portfolio having returns which first-order stochastically dominate the returns of another portfolio results in higher expected utility. The "returns" here correspond to the distribution of y (conditional on y^*) and the "utility function" corresponds to the increasing function $M(\cdot)$.

We now state the basic consistency theorem. The underlying assumptions and proof are in the appendix.[5]

Theorem 1 *Under suitable assumptions (in appendix), $\hat{\beta}^{\text{MRE}}$ (for any choice of increasing M) is an asymptotically normal, $\sqrt{n}$-consistent estimate of β_o in (1) if*

$$(i) \qquad t_1 > t_2 \implies F_{y|y^*}(v|t_1) \leq F_{y|y^*}(v|t_2)\ \forall v \tag{28}$$

$$(ii) \qquad \exists \tilde{t}\ s.t.\ t_1 > \tilde{t} > t_2 \implies \exists v\ s.t.\ F_{y|y^*}(v|t_1) < F_{y|y^*}(v|t_2). \tag{29}$$

Condition (i) corresponds to first-order stochastic dominance in the *weak* sense for $t_1 > t_2$. Condition (i) combined with condition (ii) corresponds to first-order stochastic dominance in the *strong* sense for $t_1 > \tilde{t} > t_2$. The asymptotic distribution for $\hat{\beta}^{\text{MRE}}$ is derived in Cavanagh and Sherman (1998).

The usefulness of this theorem is that the stochastic-dominance conditions have an intuitive interpretation when mismeasurement of the dependent variable is a potential problem. The question that the researcher needs to ask herself is, "Are observational units with larger 'true' values for their dependent variable more likely to *report* larger values than observational units with smaller 'true' values?" For the application discussed later in this paper, that of unemployment duration, we expect that the answer to this question is "yes."

4.3 The Examples Revisited

In this section, we discuss the stochastic-dominance conditions of Theorem 1 in the context of the examples that were introduced in Section 2.

[5] One of the assumptions, common in the semiparametric literature, is that one of the covariates is continuous (conditional on the others). The MRE can still be used when all covariates are discrete (unlike the estimator of Powell et. al. (1989)); see Cavanagh and Sherman (1998) for the properties of $\hat{\beta}^{\text{MRE}}$ in this situation.

Example 1 continued

From the conditional c.d.f.'s in (9) and (10), the stochastic-dominance conditions require $(1-\alpha_0) > \alpha_1$ or, equivalently, $(\alpha_0 + \alpha_1) < 1$. If $(\alpha_0 + \alpha_1) > 1$, the responses are so badly misreported that the MRE would actually estimate $-\beta_o$ rather than β_o.

Unlike the parametric approach of Section 3, the MRE remains consistent if the misclassification probabilities are functions of the level of y^*. With the function $\alpha(t)$ given by (11), the stochastic-dominance conditions are satisfied if $\alpha(t)$ is weakly increasing everywhere and strictly increasing along some region having positive probability.

Example 2 continued

In this setting, the stochastic-dominance conditions have discretized representations. Condition (i) is equivalent to

$$s_1 > s_2 \Longrightarrow \sum_{i=k}^{K} \alpha_{s_1,i} \geq \sum_{i=k}^{K} \alpha_{s_2,i} \quad \forall k \in \{1, \ldots, K\} \tag{30}$$

and condition (ii) is equivalent to

$$\exists s_1 > s_2 \text{ s.t. } \sum_{i=k}^{K} \alpha_{s_1,i} > \sum_{i=k}^{K} \alpha_{s_2,i} \text{ for some } k \in \{1, \ldots, K\}. \tag{31}$$

Looking at the transition matrix A defined in (13), the first condition means that the elements of the first column must be weakly decreasing as you go down row-by-row, the sum of the elements of the first two columns must be weakly decreasing as you go down row-by-row, and so on. Alternatively, the elements of the K-th column must be weakly increasing as you go down row-by-row, the sum of the elements of the last two columns must be weakly increasing as you go down row-by-row, and so on. The second condition has a similar interpretation.

As in Example 1, the MRE will be robust to situations in which the misclassification probabilities are functions of the level of y^*. Conditions analogous to those above can be derived rather easily.

Example 3 continued

The model of mismeasurement given by (15) satisfies the stochastic-dominance conditions. To see this, write the conditional c.d.f. as

$$\begin{aligned} F_{y|y^*}(v|t) &= \Pr(h(t,\eta) \leq v) \\ &= \Pr(\eta \leq \tilde{h}(v,t)) \\ &= \int_{-\infty}^{\tilde{h}(v,t)} dG(u), \end{aligned}$$

where G is the c.d.f. of η. Differentiating with respect to t yields

$$\begin{aligned} \partial F_{y|y^*}(v|t)/\partial t &= g(\tilde{h}(v,t))\frac{\partial \tilde{h}(v,t)}{\partial t} \\ &= -g(\tilde{h}(v,t))\frac{h_1(t,\tilde{h}(v,t))}{h_2(t,\tilde{h}(v,t))} < 0, \end{aligned}$$

where $g(s) = dG(s)/ds$ and h_j is the derivative of h with respect to the j-th argument. Thus, conditions (i) and (ii) hold.

Unlike the parametric approach, there is no need to specify the function h or the distribution G. As long as h has positive partial derivatives and η is i.i.d., the MRE will be consistent. This result is rather strong considering the wide range of mismeasurement models described by (15).

5 Covariate-Dependent Measurement Error

In this section, we modify the MRE to handle measurement error in the dependent variable that is not independent of the covariates. We limit our attention to dependence upon a single covariate, x_1.[6] We consider two cases below.

The first case covers discrete x_1, where the stochastic-dominance conditions of Theorem 1 hold for each subgroup of observations having the same value for x_1 but not necessarily across different values of x_1. For instance, if measurement error differs systematically for union workers and non-union workers, then the conditions may not hold for the whole sample but will hold for the subsample of union workers and the subsample of non-union workers.

[6]Extension to more covariates is straightforward, but there is a large loss in efficiency if the measurement error is allowed to be a function of too many covariates. If the measurement error truly is a function of nearly all the covariates, there's little hope of identifying β semiparametrically.

The second case covers continuous x_1, where some cutoff S exists such that those observations having $x_1 < S$ and those having $x_1 > S$ satisfy the stochastic model, but the whole sample doesn't necessarily satisfy it.

5.1 Dependence upon a Discrete Covariate

Let the range of x_1 be $\{1, \ldots, K\}$. Then, the basic idea is to use the MRE on each subgroup for which the stochastic-dominance conditions apply. Since all the observations in a subgroup have the same value for x_1, we lose identification of β_1. Note that $x\beta_o = x_1\beta_1 + x_{-1}\beta_{-1}$, where "$-1$" indicates all components but the first. Within a subgroup, $x_1\beta_1$ is the same for all the observations and has no effect on the rankings of $x\beta_o$ within the subgroup. The focus, then, is to estimate β_{-1} up to scale.

The MRE can be used to estimate β_{-1} consistently within each subgroup. For each $j \in \{1, \ldots, K\}$, let $\hat{\beta}^j_{-1}$ maximize the objective function

$$T_j(b_{-1}) = \sum_i 1(x_{i1} = j) M_j(y_i) \text{Rank}_j(x_{i,-1}b_{-1}), \tag{32}$$

over the set $\mathcal{B}_{-1} \equiv \{b_{-1} \in \mathcal{R}^{d-1} : |b_{-1,d-1}| = 1\}$, where the subscript j indicates that the function applies to the observations within the subgroup defined by $x_1 = j$. Then, we can take a linear combination of the subgroup estimates to yield a consistent estimate $\hat{\beta}_{-1}$ for the whole sample. For instance,

$$\hat{\beta}_{-1} = \frac{1}{n} \sum_{j=1}^{K} n_j \hat{\beta}^j_{-1}. \tag{33}$$

Since the asymptotic distribution of each $\hat{\beta}^j_{-1}$ is known, the asymptotic distribution of $\hat{\beta}_{-1}$ follows simply.[7]

Having estimated β_{-1}, one can do a specification test of this model for measurement error against the alternative of covariate-independent measurement error. The latter allows for consistent estimation of β_o using MRE on the whole sample. The covariance of these estimators can be derived using results from Abrevaya (1999), allowing for a χ^2-test of their difference.[8]

[7] The key for the asymptotic argument is that each $n_j \to \infty$ as $n \to \infty$.

[8] Rejection based on the χ^2-test statistic may be caused by something other than the behavior of the measurement error. For instance, the same estimators and specification test apply when x_1-dependent heteroskedasticity is suspected (violation of equation (1)). That is, the observations within the subgroups partitioned by x_1 are homoskedastic but possibly heteroskedastic across subgroups.

5.2 Dependence upon a Continuous Covariate with Cutoff

The basic idea here is the same as for the discrete case. The difference is that continuity of x_1 retains the identification of β_1, since even within subgroups $x_1\beta_1$ will differ across observations due to the continuity. We consider a single cutoff point S, so that the observations are split into two subgroups, defined by $x_1 < S$ and $x_1 > S$. Then, let the estimators $\hat{\beta}_1$ and $\hat{\beta}_2$ maximize (over the set $\mathcal{B}$) the objective functions

$$T_1(b) = \sum_i [1(x_{i1} < S)M_1(y_i)\text{Rank}_1(x_i b)] \tag{34}$$

$$T_2(b) = \sum_i [1(x_{i1} > S)M_2(y_i)\text{Rank}_2(x_i b)] \tag{35}$$

respectively, where the subscript indicates the subgroup to which the function applies. Then, a consistent estimate $\hat{\beta}$ is

$$\hat{\beta} = \frac{1}{n}\left(n_1\hat{\beta}_1 + (n - n_1)\hat{\beta}_2\right), \tag{36}$$

where n_1 is the number of observations having $x_1 < S$.

A specification test can be constructed here in the same manner as above. Also, multiple cutoff points can be handled by defining additional subgroups appropriately.

6 Application to Duration Models

In this section, we consider dependent variable mismeasurement in the context of duration models. Beginning with a brief review of the proportional hazard model, we discuss the potential problem of unspecified heterogeneity and its equivalence to mismeasured durations for a certain class of models. We consider several parametric and semiparametric estimation techniques, including the Cox partial likelihood and the Han-Hausman-Meyer flexible MLE. We demonstrate inconsistency of these estimators when durations are mismeasured. The MRE remains consistent when the mismeasurement follows the form discussed in the previous section. Our results are illustrated in Monte Carlo simulations under different specifications. Finally, we estimate an unemployment duration model using data from the Survey of Income and Program Participation (SIPP) to see the effects of mismeasured durations. Few studies have taken into account mismeasurement of unemployment spells in estimating a duration model. One exception is Romeo (1995), which explicitly models the measurement error (using cross-validation data) and forms a parametric likelihood incorporating the errors-in-variables and a flexible hazard specification. Since the MRE doesn't require an explicit model of the measurement error, the consistency of the coefficient parameters using our approach does not depend on correct specification of the errors-in-variables. Also, the likelihood approach of Romeo (1995) is quite cumbersome since the likelihood is complicated and requires numerical integration.

6.1 The Proportional Hazard Model

We briefly review the proportional hazard model, which has been used extensively in empirical analysis of duration data in economics; for a more complete treatment, see Kalbfleisch and Prentice (1980) or Lancaster (1990, pp. 245ff).

We consider a standard proportional hazard model with exponential index, where the hazard function is $h(\tau) = h_o(\tau)e^{x\beta_o}$, with $h_o(\cdot)$ called the "baseline hazard function." The "integrated baseline hazard function" is $H_o(\tau) = \int_{-\infty}^{t} h_o(s)ds$. The spell duration t satisfies

$$-\ln H_o(t) = x\beta_o + \epsilon, \tag{37}$$

where $-\epsilon$ follows an extreme value distribution (with density function $f(u) = e^u \exp(-e^u)$). When the baseline hazard is strictly positive, the integrated baseline hazard is strictly increasing with well-defined inverse. We can then write the duration t as a closed-form function of $x\beta_o$:

$$t = H_o^{-1}(\exp(-x\beta_o - \epsilon)). \tag{38}$$

Negating (38) puts the proportional hazard model into the latent variable context of equation (1):

$$y^* \equiv -t = -H_o^{-1}(\exp(-x\beta_o - \epsilon)), \tag{39}$$

so that

$$g(x\beta_o, \epsilon) = -H_o^{-1}(\exp(-x\beta_o - \epsilon)) \tag{40}$$

has $g_1 > 0 \; \forall \epsilon$.

6.1.1 Unobserved Heterogeneity

We can introduce unobserved heterogeneity into the proportional hazard model by specifying the hazard function as

$$h(\tau) = h_o(\tau)e^{x\beta_o + u}, \tag{41}$$

where the heterogeneity term u is independent of x. This model satisfies (1) with

$$g\left(x\beta_o, \begin{pmatrix} \epsilon \\ u \end{pmatrix}\right) = -H_o^{-1}(\exp(-x\beta_o - u - \epsilon)). \tag{42}$$

6.1.2 Parametric Estimation: Weibull model

The most often used parametrization of the proportional hazard model is the Weibull model where the baseline hazard is specified as

$$h_o(\tau) = \alpha\tau^{\alpha-1}. \tag{43}$$

For the Weibull model, the integrated baseline hazard is $H_o(\tau) = \tau^\alpha$, so that equation (38) simplifies to

$$-\alpha \ln t = x\beta_o + \epsilon. \tag{44}$$

This simple model can be estimated using either OLS or MLE. The latter is generally used if there is right-censoring (the censoring point is observed rather than a completed duration), with the likelihood function modified appropriately.

When unobserved heterogeneity is included, the Weibull model becomes

$$-\alpha \ln t = x\beta_o + u + \epsilon. \tag{45}$$

Lancaster (1985) notes that this heterogeneity can arise from multiplicative measurement error in the dependent variable. If $\tilde{t} = e^\eta t$ is the observed (mismeasured) duration and the true duration t satisfies (44), then

$$\begin{aligned} -\alpha \ln \tilde{t} &= -\alpha \ln(e^\eta t) \\ &= -\alpha \ln t - \alpha\eta \\ &= x\beta + (\epsilon - \alpha\eta), \end{aligned}$$

so that the observed duration $\tilde{t}$ can be thought of as arising from a Weibull model with heterogeneity, as in equation (45). Without censoring, least-squares regression of $-\ln t$ on x yields consistent estimates of (β_o/α) if η is independent of x. The mean of η can be non-zero since it will be absorbed in the constant term of the regression. If the value of α is assumed (e.g., $\alpha = 1$ is the "unit exponential model") and the assumed value is correct, then the OLS estimate of β_o is consistent. Usually, though, we are interested in estimating α. In the model without heterogeneity, one can estimate the variance of (ϵ/α) using the residuals from the regression of $-\ln t$ on x. An estimate of α can be imputed since ϵ is known to have an extreme-value distribution. Using the same method when durations are mismeasured but the mismeasurement is ignored, the residuals are used to estimate the variance of $(\epsilon/\alpha - \eta)$. Since the variance of $(\epsilon/\alpha - \eta)$ is larger than the variance of (ϵ/α), the imputed estimate of α will be too low.[9] The resulting estimate of β_o, then, will be biased toward zero even though the estimates of the ratios of the coefficients are consistent.

Lancaster (1985) reaches the same conclusions looking at MLE estimation of β_o for the Weibull model without censoring. The results for consistency of the parameter ratios using OLS or MLE are specific to the Weibull model with uncensored data and i.i.d. measurement errors across observations. This point is important since most applied duration work has moved away from Weibull-type specifications (which restrict the baseline hazard to be monotonic) in order to allow for more flexible hazard specifications. Unlike the MLE, the MRE will yield consistent ratios for β_o in the presence of censoring and more general measurement error (as in Example 3 of the previous sections) in a proportional hazard model with arbitrary baseline hazard.

6.1.3 Semiparametric Estimation

We discuss two approaches, the Cox partial likelihood and the Han-Hausman-Meyer flexible MLE, that estimate proportional hazard models without parametrizing the baseline hazard. Both approaches have the virtue of flexibility, but both are inconsistent in the presence of mismeasured durations.

The Cox (1972) partial-likelihood approach estimates β_o without specifying the baseline

[9] We assume η is also independent of ϵ.

hazard. The estimation uses only information about the ordering of the durations and maximizes the partial likelihood function

$$\ln \mathcal{L}(b) = \sum_i \left[x_i b - \ln \sum_{j \in R(i)} e^{x_j b} \right], \tag{46}$$

where $R(i)$, the "risk set" of observation i, contains all observations that survive at least until time t_i:

$$R(i) = \{j | t_j \geq t_i\}. \tag{47}$$

The estimator works in the presence of right censoring but does not handle ties (equal durations) in a natural way.

In a Monte Carlo study, Ridder and Verbakel (1983) show that neglected heterogeneity results in inconsistent partial likelihood estimates that are attenuated toward zero. Our results for mismeasured durations are similar, and the intuitive reason for inconsistency is straightforward: mismeasurement causes durations to be ordered incorrectly. As a result, the risk sets used in the partial likelihood function are wrong. We can write the first-order conditions resulting from (46):

$$\frac{1}{n} \sum_i x_i = \frac{1}{n} \sum_i \left[\frac{\sum_{j \in R(i)} x_j e^{x_j b}}{\sum_{j \in R(i)} e^{x_j b}} \right]. \tag{48}$$

Letting $n \to \infty$ and evaluating at the true parameter β yields

$$\mathrm{E}[x_i] = \mathrm{E} \left[\frac{\sum_{j \in R(i)} x_j e^{x_j \beta}}{\sum_{j \in R(i)} e^{x_j \beta}} \right]. \tag{49}$$

The problem is that we observe incorrect risk sets $\widetilde{R}(i)$ rather than $R(i)$, and in general, we'll have

$$\mathrm{E} \left[\frac{\sum_{j \in R(i)} x_j e^{x_j \beta}}{\sum_{j \in R(i)} e^{x_j \beta}} \right] \neq \mathrm{E} \left[\frac{\sum_{j \in \widetilde{R}(i)} x_j e^{x_j \beta}}{\sum_{j \in \widetilde{R}(i)} e^{x_j \beta}} \right] \tag{50}$$

so that the MLE estimate does not correspond to the first-order condition given by (49).

Another approach, developed by Han and Hausman (1990) and Meyer (1990), has similar difficulties when durations are mismeasured. Unlike partial likelihood estimation, the Han-Hausman-Meyer (HHM hereafter) approach handles ties in a natural way and also extends easily to unobserved heterogeneity. The basic idea of the HHM estimator is to group the observed durations into K intervals $\{(-\infty, c_1], (c_1, c_2], \ldots, (c_{K-1}, \infty)\}$, so that we observe

$$d_{ik} = \begin{cases} 1 & t_i \in [c_{k-1}, c_k) \\ 0 & \text{otherwise} \end{cases} \tag{51}$$

and maximize the likelihood function

$$\ln \mathcal{L}(b, \{\delta_k\}) = \sum_i \sum_k d_{ik} \ln \left[F(\delta_k + x_i b) - F(\delta_{k-1} + x_i b) \right], \tag{52}$$

where F is the extreme value c.d.f. The δ_k's are jointly estimated, with their true value being $\ln H_o(c_k)$ for each k. The result is a step function estimate of the baseline hazard along with the estimates of β_o. For the extension to unobserved gamma heterogeneity, see Han and Hausman (1990).

Mismeasurement in the HHM framework causes durations to be classified in incorrect intervals. This misclassification results in misspecification of the likelihood function and inconsistent estimation. The extent of the problem will depend on the form of the mismeasurement (i.e., how often the observed durations fall into the wrong interval).

6.1.4 MRE Estimation

Since the proportional hazard model with unobserved heterogeneity is a special case of the general model in (1), the MRE will consistently estimate β_o up-to-scale when the mismeasurement satisfies the conditions of Theorem 1. It is important to note that "random" right-censoring (i.e., the distribution of the censoring time is independent of the covariates) will also not affect the consistency of the MRE. With right-censoring, (39) becomes

$$y^* = \min(-H_o^{-1}(\exp(-x\beta_o - \epsilon)), C^*), \tag{53}$$

where the censoring time C^* is independent of x. This specification still satisfies (1), with

$$g\left(x\beta_o, \begin{pmatrix} \epsilon \\ u \\ C^* \end{pmatrix}\right) = \min(-H_o^{-1}(\exp(-x\beta_o - u - \epsilon)), C^*). \tag{54}$$

6.2 Monte Carlo results

For the Monte Carlo experiments in this section, we sample covariate values from the unemployment-spell dataset used in the following section. The "observed" dependent variables are created from specifications of the underlying hazard function and the mismeasurement process. Three covariates are considered:[10]

$$
\begin{aligned}
x_1 &= \text{KIDS} \\
x_2 &= \text{MARRIED} \\
x_3 &= \ln(\text{BENEFIT}/\text{WAGE}),
\end{aligned}
$$

corresponding to the number of children, an indicator variable for marriage status, and the natural log of the replacement ratio (benefit divided by wage). The three chosen covariates are discrete, binary, and continuous, respectively. See Table 5 for descriptive statistics on these variables.

In our Monte Carlo designs, we consider two different baseline hazard specifications for the proportional hazard model. The first is the exponential baseline hazard, which is constant (equivalent to a Weibull hazard with $\alpha = 1$). The second is a log-logistic baseline hazard, which takes the general form

$$h_o(\tau) = \frac{\gamma\alpha\tau^{\alpha-1}}{1+\tau^{\alpha}\gamma}, \tag{55}$$

and has integrated hazard function $H_o(\tau) = \ln(1 + \gamma\tau^{\alpha})$. Our simulations consider the log-logistic hazard with $\alpha = 3$ and $\gamma = 2$, which is non-monotonic (first increasing and then decreasing); see Kiefer (1988) for more details.

For each simulation, a given number of observations (either 500 or 1000) are drawn at random from the SIPP extract in order to yield the three covariates x_1, x_2, and x_3. Then, a "true" duration (denoted t) is generated according to (38), using one of the two baseline hazards specified above. The true value of β_o is 0.25 for the coefficient of x_1, 1.00 for the coefficient of x_2, and -0.50 for the coefficient of x_3.[11] The true exponential durations have mean 22.4 and standard deviation 31.4, and the true log-logistic durations have mean 7.2 and standard deviation 3.2.

For the observed duration (denoted $\tilde{t}$), we consider four different models of measurement error:

[10] Only three covariates were used so that many simulations for the various estimators could be conducted.

[11] The location parameter was set to 4 for the exponential hazard and 8 for the log-logistic hazard.

Model 1: $\tilde{t} = t$ (no measurement error)

Model 2: $\tilde{t} = e^{\eta} t$, where $\eta \sim \mathrm{N}(-0.5, 1)$

Model 3: $\tilde{t} = \eta t$, where $\eta = \begin{cases} 1 & \text{with probability } 1/2 \\ 1.25 & \text{with probability } 1/4 \\ 0.75 & \text{with probability } 1/4 \end{cases}$

Model 4: $\tilde{t} = t + |\eta|$, where $\eta \sim \mathrm{N}(0, 4)$

In each of the models, the resulting observed duration $\tilde{t}$ is positive. Models 2 and 3 are multiplicative measurement-error models; in both models, the expected value of the scaling factor is equal to one. Model 4 is an additive measurement-error model.

Tables 1 and 2 report the exponential simulation results for sample sizes of 500 and 1000. Tables 3 and 4 report the log-logistic simulation results for sample sizes of 500 and 1000. For each estimator and design, the mean and standard deviation of the estimates from 100 simulations are reported. Results are reported for the three estimators discussed in Section 6.1.3 (the Cox partial likelihood estimator, the HHM estimator, and the HHM estimator that allows for gamma heterogeneity).[12] The MRE estimation results (using $M(\cdot) = \mathrm{Rank}(\cdot)$) are in the final column of each table. Since the MRE estimates β_o up-to-scale, the estimates from each simulation were normalized so that the estimate vector had length one (i.e., $\hat{\beta}_1^2 + \hat{\beta}_2^2 + \hat{\beta}_3^2 = 1$). Normalized estimation results are also reported for the Cox and HHM estimators to allow comparison. The true value for β_o is $\{0.25, 1.00, -0.50\}$, and the true value for the normalized β_o is $\{0.2182, 0.8729, -0.4364\}$.

For Model 1 (no measurement error) in each table, each of the estimators performs quite well. The Cox and HHM estimators estimate both the scale and ratios of β_o appropriately. As expected, the standard deviations of the estimators increases from left to right (Cox having the smallest, MRE having the largest). When mismeasurement is introduced, the Cox and HHM estimates of β_o become biased as expected. For the exponential hazard, Model 4 does not represent serious mismeasurement since the additive error is small compared to

[12]Five bins were used for both HHM estimators.

Table 1: Exponential Hazard Simulations ($n = 500$)

($\beta_o = \{0.25, 1.00, -0.50\}$, normalized $\beta_o = \{0.2182, 0.8729, -0.4364\}$)

		Cox		HHM		HHM (gamma)		MRE
		Estimates	Normalized	Estimates	Normalized	Estimates	Normalized	Normalized
Model 1	x_1	0.2488	0.2155	0.2471	0.2129	0.2569	0.2120	0.2129
		(0.0474)	(0.0497)	(0.0560)	(0.0575)	(0.0613)	(0.0570)	(0.0790)
	x_2	1.0150	0.8683	1.0278	0.8710	1.0730	0.8705	0.8671
		(0.1122)	(0.0482)	(0.1318)	(0.0538)	(0.1643)	(0.0542)	(0.0667)
	x_3	-0.5043	-0.4337	-0.4992	-0.4259	-0.5229	-0.4274	-0.4206
		(0.0958)	(0.0831)	(0.1083)	(0.0929)	(0.1225)	(0.0931)	(0.1138)
Model 2	x_1	0.1778	0.2484	0.1794	0.2330	0.2319	0.2320	0.2010
		(0.0501)	(0.0831)	(0.0532)	(0.0808)	(0.0745)	(0.0771)	(0.1143)
	x_2	0.6153	0.8355	0.6711	0.8467	0.8733	0.8500	0.8630
		(0.1170)	(0.0885)	(0.1212)	(0.0847)	(0.1980)	(0.0813)	(0.1090)
	x_3	-0.3339	-0.4539	-0.3485	-0.4407	-0.4469	-0.4361	-0.4413
		(0.1149)	(0.1407)	(0.1217)	(0.1457)	(0.1698)	(0.1458)	(0.1718)
Model 3	x_1	0.2626	0.2448	0.2634	0.2430	0.2833	0.2422	0.2037
		(0.0468)	(0.0543)	(0.0494)	(0.0558)	(0.0599)	(0.0565)	(0.0819)
	x_2	0.9477	0.8653	0.9536	0.8599	1.0362	0.8609	0.8669
		(0.1019)	(0.0439)	(0.1111)	(0.0545)	(0.1566)	(0.0548)	(0.0740)
	x_3	-0.4576	-0.4144	-0.4726	-0.4212	-0.5023	-0.4192	-0.4296
		(0.1119)	(0.0845)	(0.1352)	(0.1011)	(0.1529)	(0.1024)	(0.1231)
Model 4	x_1	0.2250	0.2522	0.2528	0.2240	0.2668	0.2223	0.2199
		(0.0441)	(0.0480)	(0.0461)	(0.0523)	(0.0553)	(0.0533)	(0.0762)
	x_2	1.0045	0.8726	1.0010	0.8707	1.0522	0.8725	0.8699
		(0.1280)	(0.0469)	(0.1385)	(0.0540)	(0.1545)	(0.0527)	(0.0690)
	x_3	-0.4822	-0.4202	-0.4790	-0.4182	-0.4982	-0.4142	-0.4188
		(0.1144)	(0.0907)	(0.1307)	(0.1062)	(0.1403)	(0.1054)	(0.1241)

Table 2: Exponential Hazard Simulations ($n = 1000$)

$(\beta_o = \{0.25, 1.00, -0.50\},$ normalized $\beta_o = \{0.2182, 0.8729, -0.4364\})$

		Cox		HHM		HHM (gamma)		MRE
		Estimates	Normalized	Estimates	Normalized	Estimates	Normalized	Normalized
Model 1	x_1	0.2581	0.2277	0.2547	0.2237	0.2649	0.2260	0.2231
		(0.0321)	(0.0333)	(0.0378)	(0.0379)	(0.0425)	(0.0376)	(0.0528)
	x_2	0.9962	0.8738	0.9991	0.8722	1.0269	0.8703	0.8713
		(0.0885)	(0.0397)	(0.0983)	(0.0446)	(0.1241)	(0.0435)	(0.0509)
	x_3	-0.4789	-0.4214	-0.4852	-0.4248	-0.5017	-0.4279	-0.4314
		(0.0774)	(0.0686)	(0.0885)	(0.0765)	(0.0835)	(0.0733)	(0.0919)
Model 2	x_1	0.1739	0.2437	0.1847	0.2434	0.2446	0.2383	0.2145
		(0.0409)	(0.0718)	(0.0424)	(0.0702)	(0.0526)	(0.0568)	(0.0735)
	x_2	0.6059	0.8295	0.6542	0.8444	0.8838	0.8461	0.8815
		(0.0907)	(0.0613)	(0.0854)	(0.0471)	(0.1513)	(0.0476)	(0.0652)
	x_3	-0.3514	-0.4865	-0.3568	-0.4639	-0.4792	-0.4644	-0.4300
		(0.0558)	(0.0875)	(0.0575)	(0.0764)	(0.0876)	(0.0812)	(0.1155)
Model 3	x_1	0.2578	0.2431	0.2587	0.2403	0.2749	0.2405	0.2067
		(0.0335)	(0.0382)	(0.0320)	(0.0357)	(0.0410)	(0.0361)	(0.0566)
	x_2	0.9373	0.8726	0.9534	0.8736	1.0146	0.8731	0.8702
		(0.0837)	(0.0348)	(0.0834)	(0.0350)	(0.0976)	(0.0362)	(0.0555)
	x_3	-0.4364	-0.4064	-0.4432	-0.4058	-0.4759	-0.4065	-0.4267
		(0.0712)	(0.0673)	(0.0755)	(0.0689)	(0.0968)	(0.0701)	(0.0934)
Model 4	x_1	0.2440	0.2106	0.2429	0.2070	0.2533	0.2067	0.2150
		(0.0233)	(0.0281)	(0.0309)	(0.0336)	(0.0374)	(0.0343)	(0.0664)
	x_2	1.0193	0.8736	1.0399	0.8794	1.0858	0.8798	0.8794
		(0.0836)	(0.0278)	(0.0965)	(0.0350)	(0.1126)	(0.0341)	(0.0511)
	x_3	-0.5051	-0.4337	-0.4971	-0.4211	-0.5193	-0.4208	-0.4334
		(0.0690)	(0.0549)	(0.0819)	(0.0658)	(0.0942)	(0.0644)	(0.0876)

Table 3: Log-Logistic Hazard Simulations ($n = 500$)

($\beta_o = \{0.25, 1.00, -0.50\}$, normalized $\beta_o = \{0.2182, 0.8729, -0.4364\}$)

		Cox		HHM		HHM (gamma)		MRE
		Estimates	Normalized	Estimates	Normalized	Estimates	Normalized	Normalized
Model 1	x_1	0.2576	0.2267	0.2584	0.2296	0.2706	0.2290	0.2241
		(0.0545)	(0.0542)	(0.0568)	(0.0573)	(0.0637)	(0.0572)	(0.0728)
	x_2	0.9938	0.8654	0.9836	0.8630	1.0302	0.8628	0.8647
		(0.1031)	(0.0468)	(0.1202)	(0.0556)	(0.1401)	(0.0560)	(0.0637)
	x_3	-0.4978	-0.4325	-0.4917	-0.4313	-0.5147	-0.4317	-0.4278
		(0.1163)	(0.0879)	(0.1279)	(0.1014)	(0.1342)	(0.1029)	(0.1293)
Model 2	x_1	0.0901	0.2375	0.0874	0.2147	0.1256	0.2141	0.2153
		(0.0440)	(0.1554)	(0.0460)	(0.1427)	(0.0861)	(0.1384)	(0.1557)
	x_2	0.3010	0.8433	0.3339	0.8545	0.6011	0.8650	0.8543
		(0.0980)	(0.1884)	(0.1062)	(0.1865)	(0.2371)	(0.1901)	(0.1779)
	x_3	-0.1368	-0.4056	-0.1456	-0.3845	-0.2827	-0.4420	-0.4176
		(0.1092)	(0.2745)	(0.1135)	(0.2634)	(0.2231)	(0.2652)	(0.2479)
Model 3	x_1	0.1880	0.2205	0.1907	0.2281	0.2425	0.2252	0.2268
		(0.0494)	(0.0652)	(0.0506)	(0.0706)	(0.0687)	(0.0742)	(0.0966)
	x_2	0.7375	0.8530	0.7221	0.8479	0.9540	0.8527	0.8562
		(0.0973)	(0.0676)	(0.1055)	(0.0718)	(0.2212)	(0.0684)	(0.0890)
	x_3	-0.3870	-0.4473	-0.3835	-0.4496	-0.4954	-0.4442	-0.4338
		(0.1174)	(0.1225)	(0.1259)	(0.1310)	(0.1679)	(0.1227)	(0.1478)
Model 4	x_1	0.1987	0.1962	0.1987	0.1929	0.2173	0.1938	0.2183
		(0.0538)	(0.0567)	(0.0558)	(0.0599)	(0.0701)	(0.0596)	(0.0757)
	x_2	0.8902	0.8686	0.9165	0.8754	1.0327	0.8747	0.8632
		(0.1096)	(0.0582)	(0.1280)	(0.0643)	(0.1630)	(0.0633)	(0.0705)
	x_3	-0.4615	-0.4520	-0.4519	-0.4352	-0.5002	-0.4565	-0.4267
		(0.1066)	(0.0999)	(0.1189)	(0.1124)	(0.1447)	(0.1117)	(0.1295)

Table 4: Log-Logistic Hazard Simulations ($n = 1000$)

($\beta_o = \{0.25, 1.00, -0.50\}$, normalized $\beta_o = \{0.2182, 0.8729, -0.4364\}$)

		Cox		HHM		HHM (gamma)		MRE
		Estimates	Normalized	Estimates	Normalized	Estimates	Normalized	Normalized
Model 1	x_1	0.2514	0.2218	0.2534	0.2234	0.2610	0.2232	0.2196
		(0.0316)	(0.0332)	(0.0336)	(0.0350)	(0.0362)	(0.0353)	(0.0502)
	x_2	0.9905	0.8683	0.9979	0.8740	1.0298	0.8740	0.8749
		(0.0889)	(0.0439)	(0.0837)	(0.0394)	(0.1153)	(0.0397)	(0.0474)
	x_3	-0.4917	-0.4324	-0.4789	-0.4205	-0.4926	-0.4205	-0.4229
		(0.0951)	(0.0853)	(0.0935)	(0.0837)	(0.0940)	(0.0844)	(0.1021)
Model 2	x_1	0.0811	0.2488	0.0839	0.2251	0.1384	0.2196	0.2308
		(0.0318)	(0.1222)	(0.0326)	(0.1025)	(0.0646)	(0.1040)	(0.1283)
	x_2	0.2923	0.8328	0.3389	0.8579	0.5981	0.8606	0.8700
		(0.0860)	(0.1403)	(0.0942)	(0.1345)	(0.2276)	(0.1334)	(0.1507)
	x_3	-0.1323	-0.3999	-0.1407	-0.3745	-0.2395	-0.3693	-0.4229
		(0.0738)	(0.2315)	(0.0735)	(0.2179)	(0.1496)	(0.2218)	(0.2281)
Model 3	x_1	0.1940	0.2306	0.1991	0.2374	0.2575	0.2323	0.2239
		(0.0375)	(0.0566)	(0.0384)	(0.0595)	(0.0568)	(0.0554)	(0.0663)
	x_2	0.7396	0.8641	0.7340	0.8598	0.9654	0.8628	0.8740
		(0.0848)	(0.0420)	(0.0907)	(0.0560)	(0.1210)	(0.0521)	(0.0643)
	x_3	-0.3727	-0.4349	-0.3713	-0.4329	-0.4853	-0.4316	-0.4289
		(0.0833)	(0.0809)	(0.1050)	(0.1048)	(0.1399)	(0.1007)	(0.1242)
Model 4	x_1	0.2258	0.2248	0.2282	0.2216	0.2491	0.2198	0.2192
		(0.0348)	(0.0370)	(0.0408)	(0.0461)	(0.0477)	(0.0454)	(0.0554)
	x_2	0.8829	0.8754	0.9141	0.8808	1.0076	0.8816	0.8713
		(0.0652)	(0.0347)	(0.0775)	(0.0359)	(0.1153)	(0.0357)	(0.0488)
	x_3	-0.4225	-0.4192	-0.4224	-0.4077	-0.4670	-0.4069	-0.4284
		(0.0761)	(0.0713)	(0.0841)	(0.0770)	(0.1116)	(0.0771)	(0.1122)

the magnitude of the spells. For Models 2 and 3, the exponential estimates are attenuated toward zero. The HHM gamma estimator is less biased toward zero since it allows for heterogeneity, but since the mismeasurement is not a gamma distribution the HHM gamma estimator remains biased (especially for Model 2). For the log-logistic simulations, the Cox and HHM estimators are biased for Models 2–4 with the HHM gamma estimator being the least biased.

The MRE seems to be relatively insensitive to the form of mismeasurement. Across the four tables, the MRE estimates for the normalized parameter vector seem unbiased. As expected, the standard deviations for the MRE estimates increase with the severity of the measurement error. The other three estimators are biased for the normalized parameter in certain instances. In Table 2, the Cox and HHM estimators for Model 2 yield coefficient estimates on x_3 which are too negative and coefficient estimates on x_2 which are too small; the bias on these estimates are around 5–10%. The bias actually seems to have gotten worse (in comparison to Table 1) with more observations. For Model 3 with exponential hazard, the estimates on x_1 and x_3 are biased upwards. For Model 4 with exponential hazard, the estimators do well since the mismeasurement is not too severe. In Table 4, the log-logistic Cox and HHM estimates in Model 2 for the normalized parameter are quite biased (especially the coefficient on x_3). There are also slight biases evident in Models 3 and 4.

These simulations suggest that the Cox and HHM estimates of β_o will be biased in the presence of mismeasurement and that the Cox and HHM estimates of β_o up-to-scale can also be biased. In the next section, we consider an empirical example with a much larger sample size than the simulations considered here. Since the bias of the Cox and HHM estimators will not disappear with an increase in sample size, the MRE seems appropriate for estimation of the parameter vector up-to-scale.

6.3 Mismeasured Duration Data in the SIPP

Several studies have examined the extent of measurement error in reporting of unemployment durations, particularly in the Current Population Survey (CPS) and the Panel Study of Income Dynamics (PSID). While we use the Survey of Income and Program Participation (SIPP), the same stylized facts should apply. We highlight a few of the regularities that have been found:

Reporting errors are widespread. Poterba and Summers (1984), using Reinterview Surveys for the CPS, compare month-to-month questionnaires and find that 37% of unemployed workers overstated unemployment duration (i.e., their estimate in a given month was more than five weeks larger than their estimate in the preceding month). This percentage counts only those responses which are *inconsistent*, which is a lower bound on the percentage of responses which are incorrect. Mathiowetz and Duncan (1988), using a validation study of the PSID, find that the average absolute difference between interview response and company records for reporting of unemployment hours was 45 hours (per year) in 1981 and 52 hours (per year) in 1982. A more disturbing finding by Mathiowetz and Duncan (1988) is that many unemployment spells, particularly those lasting less than three months, are not reported at all in the PSID.

Longer spells have more reporting errors. The evidence supports the conventional wisdom that people have trouble accurately recalling events which occurred long ago. Bowers and Horvath (1984) find that only 8–20% of workers who are unemployed for over a year give consistent responses in the CPS. Poterba and Summers (1984) have a similar finding.

Responses tend to be focal. Since people don't always keep detailed records of their unemployment spells, they tend to give "focal responses" when questioned about their unemployment duration. For instance, people are more likely to say that they were unemployed for two months rather than seven weeks or nine weeks. Sider (1985) finds that modes in the PSID data occur at durations corresponding to monthly, quarterly, half-yearly, and yearly points. One explanation that has been given to account for certain spikes is that unemployment benefits usually run out after 26 weeks (half a year) or 39 weeks (three-quarters of a year) so that many people go back to work at these times. This explanation can only account for a fraction of the focal responses since spikes appear at other regular intervals and the same 26-week and 39-week spikes are also seen among workers who have not yet completed their unemployment spells.

Demographic variables do not explain the mismeasurement. There has been no evidence that individual characteristics have an effect on the likelihood of reporting error. Three of the previously discussed studies (Bowers and Horvath (1984), Mathiowetz and Duncan (1988), and Poterba and Summers (1984)) regress some function of reporting error on demographic variables including age, education, race, and sex. In each instance, the coefficients on the demographic variables are insignificant.[13] Factors which are significant in explaining reporting errors include the length of the unemployment spell, the time between the spell and the interview, and the reason for unemployment (layoff, temporary layoff, voluntary leave, etc.).

Studies of measurement error in the SIPP have focused on whether or not people correctly report participation in government transfer programs. The SIPP is a longitudinal panel study that interviews people eight times at four-month intervals and collects monthly data on earnings, participation in government transfer programs, assets and liabilities, and employment history. Marquis and Moore (1990) match responses in the SIPP against federal and state administrative records to determine the extent of reporting errors. They find that reporting error for participation is quite small (about 1.5% for unemployment insurance participation). The reporting error for change in participation is also small (about 0.6% for unemployment insurance participation). An interesting finding is that people are twice as likely to report change in participation "on seam" as they are to report change in participation "off seam." ("On seam" means that the change in participation occurs in two adjacent months that fall in different interview periods.) This "seam bias" is akin to the focal response errors discussed above. People tend to over-report participation change "on seam" since it is a focal response to say that the change has occurred just recently rather than recalling when in the last four months it actually occurred.

Unlike Marquis and Moore (1990), our primary concern is with the mismeasurement of unemployment durations and not the mismeasurement of participation in the UI program. Our sample consists of 15,103 males between the ages of 21 and 55 who experience an

[13]The sole exception is that Poterba and Summers (1984) find that teenage women tend to underreport their duration increment. Our analysis of the SIPP does not include teenage workers.

unemployment spell between 1986 and 1992 and are eligible for UI benefits.[14] In our sample, 4,205 (27.8%) receive UI benefits at some point during their spell. There are 2,237 (14.8%) people whose unemployment spell is right-censored, meaning that the spell was ongoing when the interviewee left the SIPP.

Table 5 reports summary statistics for the full sample and subsample of UI recipients and non-recipients. Standard deviations for the non-indicator variables are given in parentheses. The uncensored spells of UI recipients last an average of 6.61 weeks longer than the uncensored spells of non-recipients. A higher percentage of UI recipients are married (67%) than are non-recipients (57%). Previous weekly wage and, in turn, benefit eligibility is higher for those receiving UI. Many of the other characteristics are similar across UI recipients and non-recipients.

To highlight the focal response phenomenon in the data, we show several histograms of unemployment duration. Figures 1 and 2 graph durations for all right-censored spells and all uncensored spells, respectively. The x-axis is labeled at four-month intervals, corresponding to the time between successive interviews. The spikes for the right-censored sample are quite noticeable. The spikes for the uncensored sample are also present, but they are less noticeable due to the large number of spells that last fewer than four months.

Figures 3 and 4 give more detailed histograms for the uncensored spells, focusing on the subsample unemployed between 25 and 75 weeks. Figure 3 graphs durations for UI recipients, and Figure 4 graphs durations for non-recipients. The four-month spikes (at 35, 52, and 69 weeks) are evident. In Figure 3, there are also spikes at 26 and 39 weeks since benefits generally elapse at those times. The histograms unfortunately don't tell us much about the overall mismeasurement of unemployment durations; they just serve to highlight the extent of focal responses. The MRE, though, handles mismeasurement beyond focal responses as long as the mismeasurement satisfies the stochastic-dominance condition of Section 4.

We estimate a proportional hazard model using the previously discussed estimation techniques. In Table 6, we report coefficient estimates obtained from Weibull MLE, Cox partial likelihood, HHM MLE, and HHM MLE with gamma heterogeneity.[15]

[14] For those with multiple spells, we consider only the first spell.

[15] For the HHM estimates, monthly bins were used for estimation of the underlying baseline hazard function. The estimates were not very sensitive to alternative bin sizes.

Table 5: Summary Statistics for SIPP Sample

(means and standard deviations (in parentheses))

	All	UI	Non-UI
Number of spells	15,103	4,205	10,898
Censored spells	2,237	667	1,570
Uncensored spell length (in weeks)	10.92 (13.28)	15.71 (15.89)	9.10 (11.65)
Censored spell length (in weeks)	37.05 (31.15)	40.19 (29.42)	35.72 (31.77)
Age	34.30 (9.67)	35.76 (9.34)	33.74 (9.74)
HS grad/no college	0.37	0.40	0.36
Some college	0.25	0.24	0.26
College grad	0.18	0.17	0.18
# children	0.66 (1.04)	0.72 (1.06)	0.64 (1.03)
White	0.86	0.88	0.85
Married	0.60	0.67	0.57
Prev. weekly wage	389.4 (264.8)	441.2 (269.5)	369.4 (260.2)
Weekly benefit (eligible)	164.3 (65.8)	183.0 (63.9)	157.1 (65.1)

The variable names are fairly self-explanatory. The demographic variables are all dummy variables with the exception of KIDS, which is the number of children. The age dummies (AGE31TO40, AGE41TO50, AGE51UP) are all zero for men between the ages of 21 and 30. The education dummies (HSDGRAD, SOMECOLL, COLLGRAD) are all zero for high school dropouts. As a result, the coefficient estimates on these dummies should be interpreted as comparisons to the excluded groups. The variables WAGE and BENEFIT correspond to the pre-unemployment weekly wage and weekly unemployment benefit eligibility, respectively.[16] We allow for the possibility that UI recipients and non-recipients behave differently by having separate coefficients on ln(WAGE) and ln(BENEFIT) for the two groups. In the table, the variable UI indicates that the worker received UI benefits at some point during his unemployment spell. One would expect UI recipients to be more sensitive to benefit levels than non-recipients. We have also allowed the coefficients on ln(WAGE) and ln(BENEFIT) to differ from each other. Many empirical studies have focused on the effect of the "replacement rate" (the ratio of benefits to earnings) on unemployment spells, implicitly constraining the two coefficients to be of equal magnitude and opposite sign.[17] The approach taken here allows for a test of the replacement-rate approach (for both recipients and non-recipients).

A positive (negative) coefficient indicates that the associated variable causes longer (shorter) unemployment spells. The results are quite similar across the columns of Table 6. None of the signs on the coefficients are too surprising. The estimates indicate that the following groups (all other things being equal) have longer unemployment spells: older workers, workers with fewer children, single workers, high school dropouts, and non-white workers. The effects of previous wage and level of unemployment benefits also have the predicted signs. Those with higher previous wage (and, thus, higher opportunity cost of remaining unemployed) have shorter spells. Those with higher UI benefits have longer spells.

The Weibull estimate of α (the variable parametrizing the hazard in (44)) is 0.5979 (with a standard error of 0.0068), indicating a decreasing baseline hazard. There is evidence of

[16]Eligibility rather than actual receipt is used to deal with selection issues and possible misreporting. The benefit eligibility was calculated using a UI simulator created by Jon Gruber which takes into account the regional UI laws pertaining to the given worker and the reported quarterly wages before unemployment. See Gruber (1997) for details.

[17]See Atkinson and Micklewright (1991) for a survey.

Table 6: Duration Results for SIPP Sample

(standard errors in parentheses)

	Weibull	Cox	HHM	HHM het
AGE31TO40	0.1761	0.1406	0.1968	0.1781
	(0.0226)	(0.0226)	(0.0222)	(0.0405)
AGE41TO50	0.2711	0.2234	0.3208	0.3306
	(0.0267)	(0.0268)	(0.0260)	(0.0475)
AGE51UP	0.5978	0.4879	0.6456	0.6462
	(0.0381)	(0.0379)	(0.0347)	(0.0617)
HSGRAD	-0.0892	-0.0694	-0.0809	-0.1222
	(0.0237)	(0.0252)	(0.0252)	(0.0438)
SOMECOLL	-0.0909	-0.0736	-0.1184	-0.1157
	(0.0289)	(0.0271)	(0.0271)	(0.0475)
COLLGRAD	-0.0320	-0.0126	-0.0106	0.0807
	(0.0262)	(0.0303)	(0.0309)	(0.0536)
KIDS	-0.0294	-0.0277	-0.0416	-0.0566
	(0.0086)	(0.0094)	(0.0089)	(0.0174)
WHITE	-0.3523	-0.3112	-0.4606	-0.5817
	(0.0272)	(0.0264)	(0.0288)	(0.0441)
MARRIED	-0.2725	-0.2458	-0.3105	-0.4352
	(0.0219)	(0.0218)	(0.0217)	(0.0385)
ln(WAGE)×UI	-0.1573	-0.1383	-0.3967	-0.1657
	(0.0330)	(0.0423)	(0.0413)	(0.0771)
ln(BENEFIT)×UI	0.1318	0.1396	0.1330	0.1700
	(0.0413)	(0.0519)	(0.0516)	(0.0943)
ln(WAGE)×(1-UI)	-0.2200	-0.2100	-0.3585	-0.4108
	(0.0251)	(0.0257)	(0.0237)	(0.0468)
ln(BENEFIT)×(1-UI)	0.1134	0.1406	0.0104	0.2573
	(0.0355)	(0.0363)	(0.0351)	(0.0660)
CONSTANT	3.1034			
	(0.0993)			
α	0.5979			
	(0.0068)			
θ				0.8703
				(0.0123)

heterogeneity from the HHM estimates in the final column of Table 6. The unobserved heterogeneity is assumed to be a gamma distribution with mean 1 and variance $1/\theta$. Using the delta method, the estimate of θ yields a variance estimate of 1.1490 (with standard error of 0.0187) which is significantly different from zero.

Due to the presence of mismeasured durations and the evidence of heterogeneity, the MRE seems like an appropriate estimator for this data. In Table 7, we report the results from estimation of the proportional hazard model using the MRE with $M(y) = y$, the identity function.[18] As a comparison, we also list the results from the Weibull, Cox, HHM, and HHM with gamma heterogeneity. Since the MRE only identifies the parameters up to scale, all of the coefficient estimates have been rescaled so that each estimate vector has length one. The standard errors and 95% confidence intervals for the MRE were constructed using bootstrap estimates.[19] The standard errors for the other estimates were derived using the delta method.

The first important point about the MRE results concerns their precision. Semiparametric estimation always involves a tradeoff between precision and flexibility. Oftentimes, allowing for too much flexibility of the model results in estimates which are too imprecise to be meaningful in practice. In our application, though, the estimates remain statistically significant. Almost all of the demographic variables retain the predicted signs, and the 95% confidence intervals imply statistical significance since they do not contain zero (except for COLLGRAD).

The MRE coefficient estimates of the demographic variables are generally in agreement with the estimates of the other techniques. The striking difference between the MRE results and the other estimates is the effect of UI benefit levels on unemployment duration. The benefit coefficients for both UI recipients and non-recipients are not significantly different from zero. For a given wage, the variation in benefit levels seems to have little effect on the

[18] The results are very similar using other choices for $M(\cdot)$, such as Rank$(\cdot)$ and $\ln(\cdot)$.

[19] Cavanagh and Sherman (1998) provide formulas which can be used in conjunction with kernel techniques to compute consistent estimates of the standard errors. In this application, however, these estimates were sensitive to the choice of kernel windows. As a result, 200 bootstrap replications were used to construct standard errors and confidence intervals. Cavanagh and Sherman (1998) find that the bootstrap is quite accurate for estimating standard errors in a large sample. There have been no definitive studies on its use in small samples.

Table 7: Normalized Results for SIPP Sample

	Weibull	Cox	HHM	HHM het	MRE
AGE31TO40	0.1991	0.1849	0.1801	0.1497	0.1741
	(0.0280)	(0.0381)	(0.0179)	(0.0278)	(0.0285)
					[0.1099, 0.2245]
AGE41TO50	0.3066	0.2937	0.2937	0.2779	0.2974
	(0.0307)	(0.0425)	(0.0192)	(0.0299)	(0.0305)
					[0.2395, 0.3480]
AGE51UP	0.6762	0.6413	0.5911	0.5431	0.7394
	(0.0272)	(0.0427)	(0.0166)	(0.0262)	(0.0319)
					[0.6740, 0.7914]
HSGRAD	-0.1008	-0.0913	-0.0741	-0.1027	-0.1167
	(0.0305)	(0.0437)	(0.0213)	(0.0312)	(0.0329)
					[-0.1860, -0.0458]
SOMECOLL	-0.1028	-0.0967	-0.1084	-0.0973	-0.0930
	(0.0369)	(0.0468)	(0.0227)	(0.0336)	(0.0309)
					[-0.1511, -0.0269]
COLLGRAD	-0.0362	-0.0166	-0.0097	0.0678	-0.0603
	(0.0334)	(0.0524)	(0.0259)	(0.0378)	(0.0374)
					[-0.1346, 0.0102]
KIDS	-0.0332	-0.0364	-0.0381	-0.0476	-0.0207
	(0.0112)	(0.0165)	(0.0076)	(0.0126)	(0.0095)
					[-0.0363, -0.0003]
WHITE	-0.3985	-0.4092	-0.4216	-0.4889	-0.3658
	(0.0339)	(0.0435)	(0.0223)	(0.0288)	(0.0388)
					[-0.4448, -0.2995]
MARRIED	-0.3082	-0.3231	-0.2843	-0.3658	-0.2813
	(0.0247)	(0.0334)	(0.0164)	(0.0233)	(0.0332)
					[-0.3593, -0.2239]
ln(WAGE)×UI	-0.1779	-0.1818	-0.3632	-0.1392	-0.1947
	(0.0415)	(0.0713)	(0.0309)	(0.0533)	(0.0564)
					[-0.2441, -0.0245]
ln(BENEFIT)×UI	0.1491	0.1836	0.1217	0.1428	0.0894
	(0.0515)	(0.0859)	(0.0411)	(0.0640)	(0.0666)
					[-0.1276, 0.1436]
ln(WAGE)×(1-UI)	-0.2489	-0.2761	-0.3282	-0.3452	-0.2220
	(0.0308)	(0.0426)	(0.0197)	(0.0317)	(0.0361)
					[-0.2993, -0.1617]
ln(BENEFIT)×(1-UI)	0.1282	0.1848	0.0095	0.2162	0.0044
	(0.0440)	(0.0597)	(0.0294)	(0.0420)	(0.0454)
					[-0.0929, 0.0909]

Standard errors are in parentheses. The MRE s.e.'s are standard errors of the bootstrap estimates. The 95% confidence intervals for the MRE estimates are shown in brackets.

length of unemployment. This is not to say that benefits have no effect on unemployment duration; benefit eligibility, after all, is a function of previous wage. The HHM estimates for the benefit coefficients are the only ones that fall within the 95% confidence interval of the MRE estimates. The only other estimator that allows for heterogeneity (HHM allowing for gamma heterogeneity) has significantly positive estimates for both the recipients' benefit coefficient and the non-recipients' benefit coefficient, which is quite different from the conclusion from the MRE estimates.

The estimates can also be used to test the applicability of the replacement-rate model. To simplify matters, we introduce notation for the four parameters of interest:

$$\begin{aligned}
\beta_{w,UI} &= \text{coefficient on } \ln(WAGE) \times UI \\
\beta_{b,UI} &= \text{coefficient on } \ln(BENEFIT) \times UI \\
\beta_{w,no\,UI} &= \text{coefficient on } \ln(WAGE) \times (1 - UI) \\
\beta_{b,no\,UI} &= \text{coefficient on } \ln(BENEFIT) \times (1 - UI)
\end{aligned}$$

The hypothesis underlying the replacement-rate model for recipients is

$$H_1 : \beta_{w,UI} = -\beta_{b,UI}. \tag{56}$$

The hypothesis underlying the replacement-rate model for non-recipients is

$$H_2 : \beta_{w,no\,UI} = -\beta_{b,no\,UI}. \tag{57}$$

The p-values associated with tests of these two hypotheses using the non-MRE estimates is shown in Table 8. All four techniques reject the replacement-rate model for non-recipients. Only the HHM estimates provide evidence against the replacement-rate model for recipients. The HHM estimates accounting for heterogeneity have a p-value of 0.928, which provides support for the replacement-rate model. As for the MRE estimates, confidence intervals were constructed using the 200 bootstrap replications. The 95% confidence interval for $\beta_{w,UI}/\beta_{b,UI}$ was $(-0.68, 2.26)$, and the 95% confidence interval for $\beta_{w,no\,UI}/\beta_{b,no\,UI}$ was $(-0.35, 0.54)$. Only two of the 200 values for $\beta_{w,UI}/\beta_{b,UI}$ were below -1, and none of the 200 values for $\beta_{w,no\,UI}/\beta_{b,no\,UI}$ were below -1. This finding calls into question the applicability of the replacement-rate model for both recipients and non-recipients, in contrast to the results from HHM with heterogeneity.

Table 8: p-values for Replacement-Rate Tests

	Weibull	Cox	HHM	HHM het
$H_1 : \beta_{w,UI} = -\beta_{b,UI}$	0.203	0.946	0.000	0.928
$H_2 : \beta_{w,no\,UI} = -\beta_{b,no\,UI}$	0.000	0.001	0.000	0.000

7 Conclusion

This paper has proposed semiparametric estimation in the presence of mismeasured dependent variables in a general linear index model. The stochastic-dominance condition of Section 4 is a strong result in that it applies to many forms of mismeasurement and is easy for the researcher to interpret. In addition, use of the MRE doesn't require any prior model of the mismeasurement.

This work was motivated by the fact that unemployment duration data is known to be poorly measured. The proportional hazard model used to analyze such data fits nicely into the general framework in which semiparametric estimation remains consistent. The results of Section 6 show that the semiparametric approach has different implications for the effect of previous wages and unemployment benefits on the length of unemployment spells.

Though the MRE is quite robust to mismeasurement of the dependent variable in the proportional hazard model, it does have the drawback of only estimating β_o up-to-scale. The MRE estimates indicate the relative impact of covariates on y^*. In order to say anything about either the conditional expectation of y^* given x or the baseline hazard function, the researcher would need to explicitly model the mismeasurement process. In the binary-choice model, Hausman et. al. (1998) use the MRE to estimate β_o up-to-scale and then derive conditional expectations and marginal effects from the mismeasurement model described in Section 2. Future research might consider specific mismeasurement models for unemployment-spell data and determine which quantities of interest can be estimated in conjunction with the semiparametric approach discussed in this paper.

Appendix

Proof of Theorem 1: The assumptions from Cavanagh and Sherman (1998) are:

(A0) $\mathrm{E}[M(y)|x]$ depends on x only through $x\beta_o$.

(A1) $\mathrm{E}[M(y)|x]$ is a nonconstant, increasing function of $x\beta_o$.

(A2) The support of x is not contained in any proper linear subspace of $\mathcal{R}^d$.

(A3) The d'th component of x has an everywhere positive Lebesgue density, conditional on the other components.

(A4) The parameter space $\mathcal{B}$ is a compact subset of $\{b \in \mathcal{R}^d : |b_d| = 1\}$.

(A5) $\mathrm{E}[M(y)^2] < \infty$.

We make one additional technical assumption:

(iii) $\lim_{z\to\infty} g(z,\epsilon) = \infty$ and $\lim_{z\to-\infty} g(z,\epsilon) = -\infty\ \forall\epsilon$.

We assume (A2)–(A5) and show that conditions (i), (ii), and (iii) imply (A0) and (A1). (A0) is trivial since the model for the latent variable depends on x only through $x\beta_o$ and the mismeasurement is independent of x.

Let $H(\cdot)$ be the c.d.f. of $-\epsilon$. Write $\mathrm{E}[M(y)|x\beta_o = z]$ as

$$\begin{aligned}
\mathrm{E}[M(y)|x\beta_o = z] &= \int \mathrm{E}[M(y)|x\beta_o = z, \epsilon = -u]dH(u) \\
&= \int \mathrm{E}[M(y)|y^* = g(z,-u)]dH(u) \\
&= \int\int M(y)dF_{y|y^*}(y|g(z,-u))dH(u).
\end{aligned}$$

Since M is an increasing function, first-order stochastic dominance implies that

$$\int M(y)dF_{y|y^*}(y|t)$$

is increasing in t; then,

$$\int M(y)dF_{y|y^*}(y|g(z,-u))$$

is increasing in z for any u, and so the integral with respect to $dH(u)$ is an increasing function of z. Assumption (A3) and conditions (ii)–(iii) ensure that the function is nonconstant. Thus, (A1) holds and the MRE is consistent.

Additional assumptions are needed for asymptotic normality; see Cavanagh and Sherman (1998).

References

Abrevaya, J. (1999): "Leapfrog Estimation of a Fixed-Effects Model with Unknown Transformation of the Dependent Variable," *Journal of Econometrics*, forthcoming.

Atkinson, A. B. and J. Micklewright (1991): "Unemployment Compensation and Labor Market Transitions: A Critical Review," *Journal of Economic Literature* 29: 1679–1727.

Bowers, N. and F. W. Horvath (1984): "Keeping Time: An Analysis of Errors in the Measurement of Unemployment Duration," *Journal of Business and Economic Statistics* 2: 140–149.

Cavanagh, C. and R. P. Sherman (1998): "Rank Estimators for Monotonic Index Models," *Journal of Econometrics* 84: 351–381.

Cox, D. R. (1972): "Regression Models and Life Tables (with discussion)," *Journal of the Royal Statistical Society* B, 34: 187–220.

Gruber, J. (1997): "The Consumption Smoothing Benefits of Unemployment Insurance," *American Economic Review* 87: 192–205.

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

Han, A. K. and J. A. Hausman (1990): "Flexible Parametric Estimation of Duration and Competing Risk Models," *Journal of Applied Econometrics* 5: 1–28.

Hausman, J. A., J. Abrevaya, and F. M. Scott-Morton (1998): "Misclassification of a Dependent Variable in a Discrete-Response Setting," *Journal of Econometrics* 87: 239–269.

Ichimura, H. (1993): "Semiparametric Least Squares (SLS) and Weighted SLS Estimation of Single-Index Models," *Journal of Econometrics* 58: 71–120.

Kalbfleisch, J. D. and R. L. Prentice (1980): *The Statistical Analysis of Failure Time Data*, New York: John Wiley and Sons.

Kiefer, N. M. (1988): "Economic Duration Data and Hazard Functions," *Journal of Economic Literature* 26: 646–679.

Lancaster, T. (1985): "Generalised Residuals and Heterogeneous Duration Models with Applications to the Weibull Model," *Journal of Econometrics* 28: 155–169.

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

Marquis, K. M. and J. C. Moore (1990): "Measurement Errors in the Survey of Income and Program Participation (SIPP) Program Reports," Bureau of the Census, mimeo.

Mathiowetz, N. A. and G. J. Duncan (1988): "Out of Work, Out of Mind: Response Errors in Retrospective Reports of Unemployment," *Journal of Business and Economic Statistics* 6: 221–229.

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

Poterba, J. M. and L. H. Summers (1984): "Response Variation in the CPS: Caveats for the Unemployment Analyst," *Monthly Labor Review* 107 (March): 37–43.

Poterba, J. M. and L. H. Summers (1995): "Unemployment Benefits and Labor Market Transitions: A Multinomial Logit Model with Errors in Classification," *Review of Economics and Statistics* 77: 207–216.

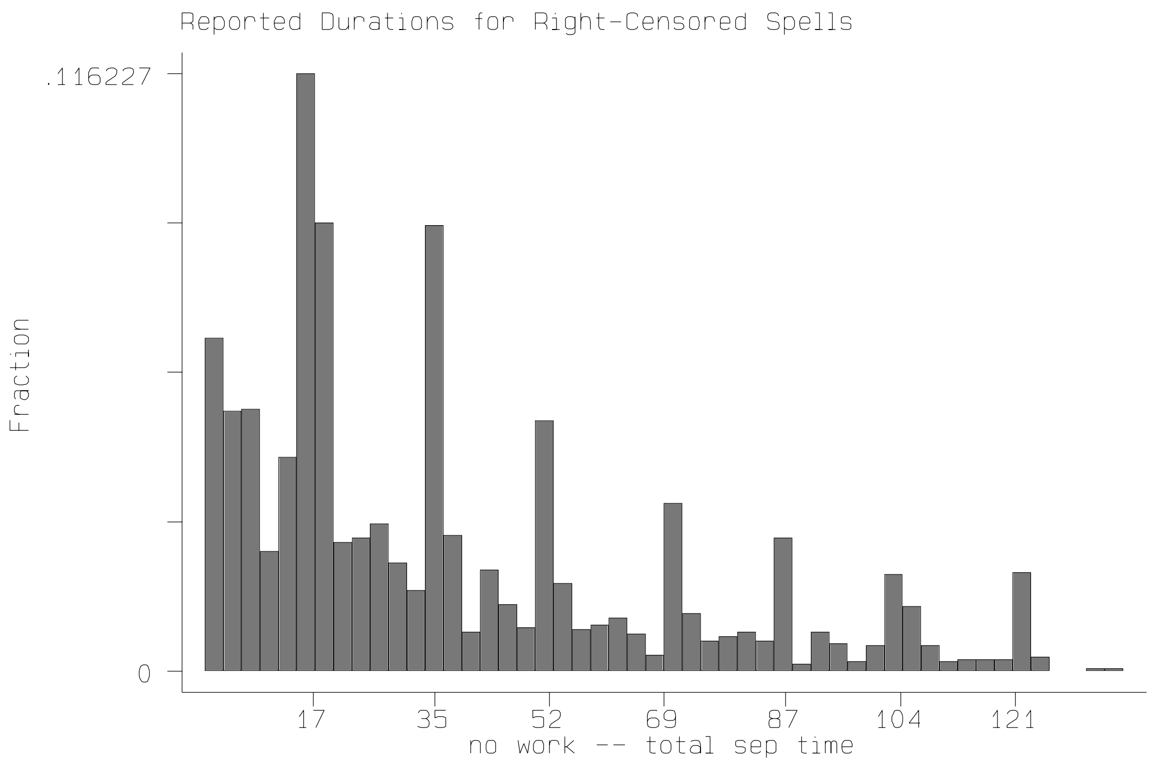
Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric Estimation of Index Coefficients," *Econometrica* 57: 474–523.

Ridder, G. and W. Verbakel (1983): "On the Estimation of the Proportional Hazard Model in the Presence of Heterogeneity," University of Amsterdam, mimeo.

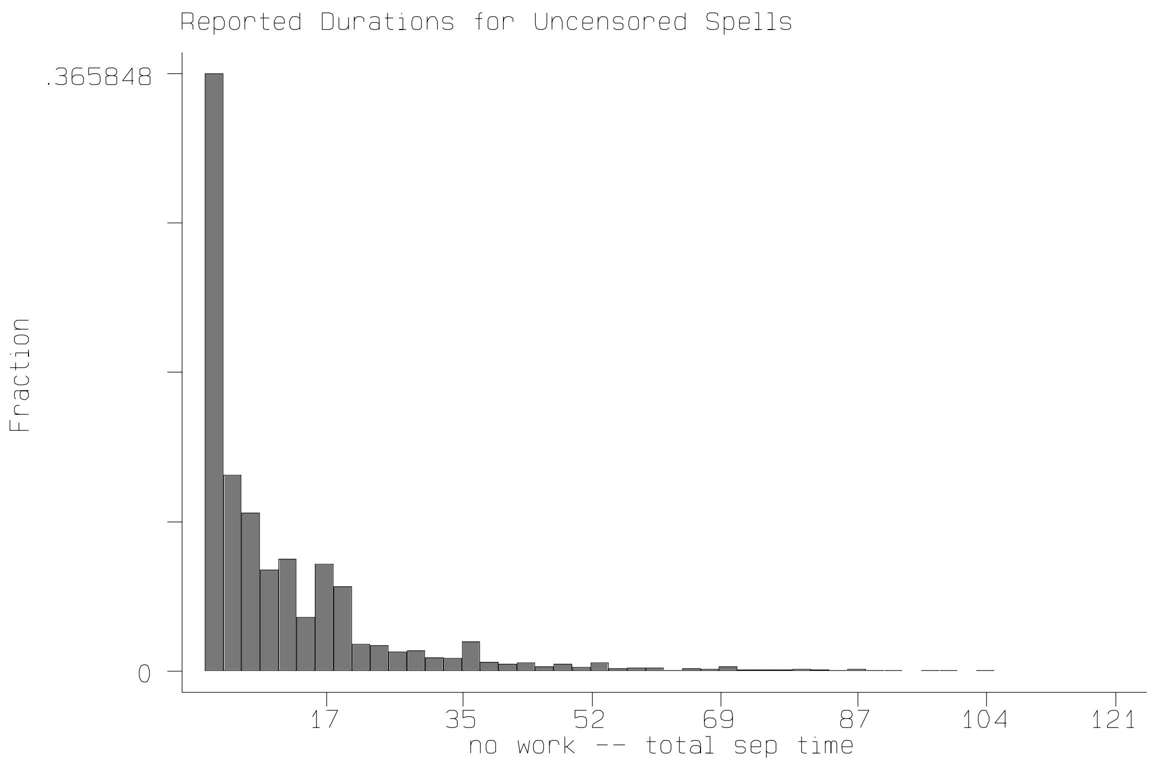
Romeo, C. J. (1995): "Errors-in-Variables Modeling of the Unemployment Duration Data in a Four Month Panel of CPS Microdata," Rutgers University, mimeo.

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

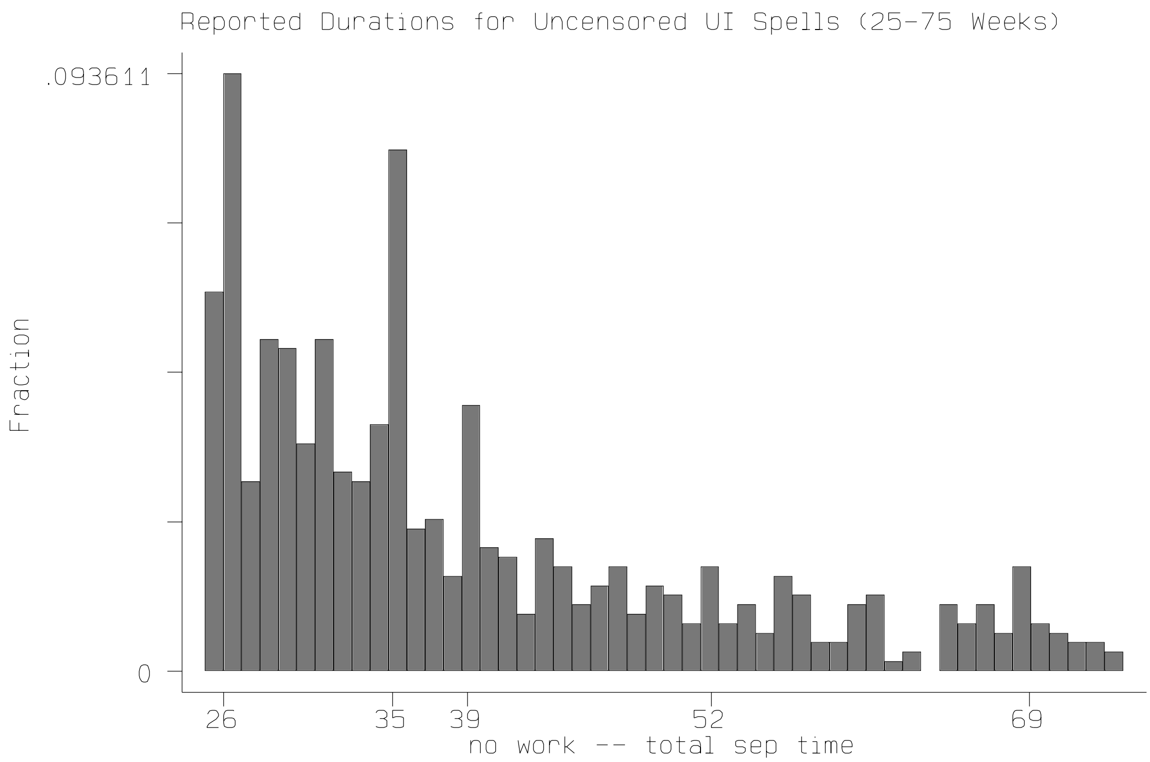
Sider, H. (1985): "Unemployment Duration and Incidence: 1968–82," *American Economic Review* 75: 461–472.

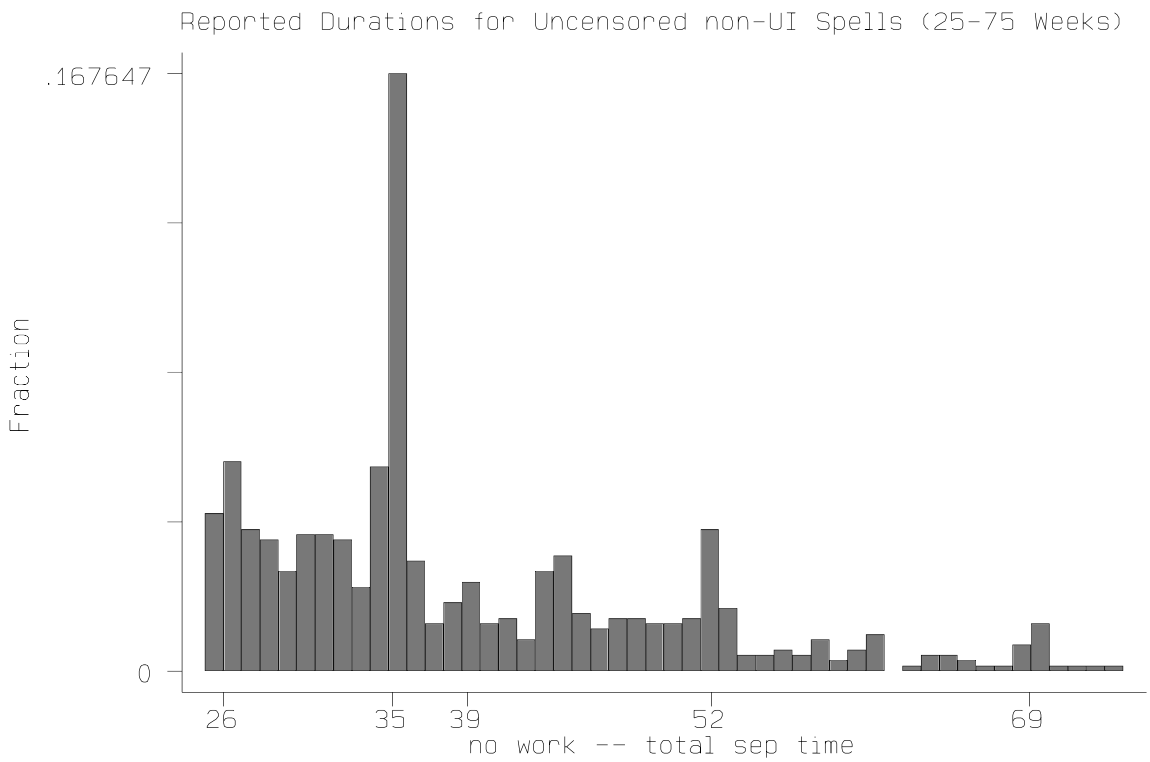


Reported Durations for Right-Censored Spells
Fraction
.116227
0
17
35
52
69
87
104
121
no work -- total sep time



Reported Durations for Uncensored Spells
.365848
0
Fraction
17
35
52
69
87
104
121
no work -- total sep time


Reported Durations for Uncensored UI Spells (25-75 Weeks)
.093611
0
Fraction
26
35
39
52
69
no work -- total sep time


Reported Durations for Uncensored non-UI Spells (25-75 Weeks)
.167647
0
Fraction
26
35
39
52
69
no work -- total sep time