Semiparametric Estimation Methods for Nonlinear Panel Data Models and Mismeasured Dependent Variables

by

Jason Abrevaya

A. B., Harvard University (1992)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Economics

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 1996

Author ....
Department of Economics
May 18, 1996

Certified by . ...
Jerry Hausman
Professor of Economics
Thesis Supervisor

Certified by . ...
Whitney Newey
Professor of Economics
Thesis Supervisor

Accepted by ....
Richard Eckaus
Chairman, Departmental Committee on Graduate Students

Semiparametric Estimation Methods for Nonlinear Panel Data Models and Mismeasured Dependent Variables

by

Jason Abrevaya

Submitted to the Department of Economics
on May 18, 1996, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Economics

Abstract

This thesis consists of three papers on semiparametric estimation in various econometric models. Panel models are the focus of the first two chapters, and mismeasured dependent variables are the focus of the final chapter.

The first chapter considers a panel version of the linear transformation model, in which the dependent variable is subject to an unknown, strictly monotonic transformation. Examples of the model include the multiple-spell proportional hazard model and dependent-variable transformation models (e.g., the Box-Cox model). Two alternative estimators are shown to be consistent and asymptotically normal.

The second chapter demonstrates that the maximum score estimator can be used to consistently estimate the parameters of a general linear index panel model. The model places no restrictions on the fixed effects, requires only weak nonparametric assumptions on the error term, and allows for general forms of censoring and truncation. A smoothed version of the estimator is shown to be asymptotically normal (with a convergence rate approaching $n^{-1/2}$).

The third chapter, joint with Jerry Hausman, considers mismeasurement of the dependent variable in a general linear index model. The monotone rank estimator is shown to be consistent in the presence of any mismeasurement process that obeys a simple stochastic-dominance condition. The proportional hazard model, with an application to unemployment spells, is studied in detail.

Thesis Supervisor: Jerry Hausman
Title: Professor of Economics

Thesis Supervisor: Whitney Newey
Title: Professor of Economics

Acknowledgments

The National Science Foundation provided financial support while this dissertation was in progress. Jon Gruber provided the data extracts used in the first and third chapters.

I owe a debt of gratitude to my thesis advisors, Jerry Hausman and Whitney Newey. I thank Jerry for stimulating my interest in econometrics and for teaching me how to do research. Our frequent meetings were the most valuable part of my graduate school education. I can only hope that his intuition and his knack for tackling important problems has rubbed off on me. I thank Whitney for telling me to abandon the computer simulations and to prove the results that were in front of my face. I also thank him for always knowing the answers to my questions and for leaving at least a few econometrics papers unwritten so that the rest of us have something to do.

My fellow students have shaped my life and my ideas for the past four years. I thank Matt Barmack and Jeff Kubik for helping to make our office into a playground refuge and for "trucking" with me rain or shine. I also thank Leora Friedberg, Harrison Hong, Catherine Wolfram, and the rest of my classmates for their continued friendship since we arrived at M.I.T. Neal Rappaport made life bearable during my first two years at M.I.T.; the very definition of reliability, he was always there for a cup of coffee, a set of tennis, or a day at the movies.

Special thanks to Lou Masur for convincing me that academia was the place for me (and for acknowledging me in his book).

My greatest gratitude is reserved for Allison Kramer, my partner in life, who sacrificed everything to come to Boston. I thank her for being there through all the ups and downs, for convincing me that I could finish, and, most of all, for making it all worthwhile.

Contents

Introduction

This thesis consists of three papers on semiparametric estimation of econometric models. The work builds upon that of Manski (1987) and Han (1987), both of which introduce estimators that use only the ordinal information contained in the dependent variables. These techniques are shown to be useful in the context of panel data models and in the presence of mismeasured dependent variables.

The first two chapters introduce consistent estimators for two different fixed-effects frameworks. Much has been written about the difficulties of consistently estimating the parameters of fixed-effects panel data models. The standard first-differencing trick which eliminates the fixed effect from a linear model extends to only certain nonlinear models, including the conditional logit model for binary data, the Poisson model for count data, and certain parametric models for duration data. Each of these models share an exponential form which allows for cancellation of the fixed effect akin to first differencing in the linear panel model. Semiparametric methods, which do not require any parametric assumptions on the error term, exist for consistent estimation of the binary choice model (Manski (1987)) and the linear censored and truncated models (Honoré (1992)).

The first chapter considers a panel version of the linear transformation model, in which the dependent variable is $h(y_t)$ for an unspecified, strictly increasing h. Examples of the model include the multiple-spell proportional hazard model and dependent-variable transformation models (e.g., the Box-Cox model). No restrictions are placed on the fixed effect. Two alternative estimators, a "change" estimator and a "leapfrog" estimator, are proposed. The "change" estimator is derived by noting that the model can be transformed into a semiparametric binary response model. The "leapfrog" estimator, which also allows h to vary over time, is a variation on the *maximum rank correlation estimator* of Han (1987). The proposed estimators are shown to be $\sqrt{n}$–consistent and asymptotically normal. Specification testing is discussed, using a general covariance result for such estimators. Monte Carlo evidence is reported, and an empirical application to the consumption smoothing effects of unemployment benefits is considered.

The second chapter demonstrates that the *maximum score estimator* of Manski (1987) can be used to consistently estimate the parameters of a general linear index panel model. The model is one in which the dependent variable y_t is a monotonic function of the linear index $x_t\beta$. The model places no restrictions on the fixed effects, requires only weak nonparametric assumptions on the error term, and allows for general forms of censoring and truncation. A smoothed version of the estimator has a convergence rate between $n^{-2/5}$ and $n^{-1/2}$.

The third chapter, joint with Jerry Hausman, 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 (1992) is shown to be consistent in the presence of any mismeasurement process that obeys a simple stochastic-dominance condition. 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).

These three papers have the common theme that throwing away information about the dependent variables can actually help matters in certain situations. In the first two chapters, the ordinal information contained in the dependent variables is used to develop consistent semiparametric estimators for models for which no consistent estimators were previously known. In the third chapter, estimators that use the actual values of the dependent variable are shown to be inconsistent when the dependent variable is mismeasured; however, an estimator that uses only the ordinal information of the dependent variables remains consistent for a wide range of possible mismeasurement processes, all of which share the property that the ordinal information is correct "on average."

The aforementioned estimators are somewhat computationally intensive since the underlying objective functions are not smooth and alternative optimization techniques need to be applied. With the rapid increase in computing speed of recent years, however, estimators like those discussed in this thesis are sure to see more widespread use in empirical work.

References

Cavanagh, C. and R. P. Sherman (1992): "Rank Estimators for Monotone Index Models," Bellcore Economics Discussion Paper no. 84.

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

Honoré, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60: 533–565.

Manski, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica* 55: 357–362.

Chapter 1

Estimation of the Linear Transformation Panel Model

1.1 Introduction

The *linear transformation model* has received much attention in the statistics and econometrics literature. Most research has focused on the non-panel model

$$h(y_i) = x_i\beta + \epsilon_i \quad (i = 1, \ldots, n) \tag{1.1}$$

where h is a strictly increasing function.[1]

When h is left unspecified, semiparametric techniques can be used to estimate β up to scale. The methods of Cavanagh and Sherman (1992), Han (1987b), Ichimura (1993), or Powell et. al. (1989) all yield $\sqrt{n}$–consistent estimators of β. With an estimate of β and an i.i.d. assumption on ϵ, one can even nonparametrically estimate h and the distribution of ϵ using the techniques of Horowitz (1996) or Ye and Duan (1995).

This paper considers the panel version of (1.1):

$$h(y_{it}) = x_{it}\beta + \gamma_t + \alpha_i + \epsilon_{it} \quad (i = 1, \ldots, n;\ t = 1, \ldots, T), \tag{1.2}$$

where α is a fixed effect and γ is a time effect.[2] The focus of this paper is on estimation of β, the coefficients on the time-varying covariates. Since h is left unspecified, the goal is to estimate β up to scale. Time-invariant covariates are not included explicitly since they can be thought of as being part of the fixed effect. The semiparametric methods developed in this paper will only identify the coefficients on the time-varying covariates. We focus on the two-period model ($T = 2$), with extension to larger T discussed in Section 1.6. Our notation is $\Delta x \equiv x_2 - x_1$, and similarly for other variables.

The aforementioned estimators of β in the non-panel setting are not applicable

[1]Of course, h could be strictly decreasing. In that case, define $\tilde{h} \equiv -h$ so that $\tilde{h}$ satisfies (1.1) with $-\beta$ as the parameter vector.

[2]A location normalization $h(0) = c_0$ is required for identification of the time effect relative to $x\beta$.

to the panel model due to the possible correlation between the fixed effect α and the regressors. The first-differencing method used to partial out the fixed effect in the linear panel model can not be extended to our more general setting. First-differencing kills the fixed effect but leaves $h(y_2)-h(y_1)$ as the dependent variable. For unknown h, the value of the dependent variable is not known, making standard within estimation impossible.

Transforming the dependent variable incorrectly and performing within estimation yields inconsistent estimates. If h is the true transformation of the dependent variable but g is the strictly increasing transformation used by the econometrician, we have

$$\begin{aligned} g(y_2)-g(y_1) &= g(h^{-1}(x_2\beta+\gamma_2+\alpha+\epsilon_2)) - g(h^{-1}(x_1\beta+\gamma_1+\alpha+\epsilon_1)) \\ &= \frac{d\, g(h^{-1}(z))}{dz}(\Delta x\beta+\Delta\gamma+\Delta\epsilon), \end{aligned}$$

where z is some value between $x_1\beta+\gamma_1+\alpha+\epsilon_1$ and $x_2\beta+\gamma_2+\alpha+\epsilon_2$ (applying the mean-value theorem). When g and h are not scalar multiples of each other, $\frac{d\, g(h^{-1}(z))}{dz}$ is nonconstant, resulting in inconsistent estimates of β due to the correlation between z and Δx. The degree of misspecification of g will determine the extent of the problem.

To avoid the problems with the within estimator, we consider semiparametric techniques for estimating β up to scale. In Section 1.2, we discuss some interesting examples of the linear transformation panel model. In Sections 1.3 and 1.4 of this paper, two estimators of β in (1.2) are proposed, and $\sqrt{n}$-consistency and asymptotic normality of the estimators are proven. The estimators are nonparametric with respect to the distribution of the error term ϵ and place no restrictions on the fixed effect α. The "change" estimator of Section 1.3 uses within-individual, across-time comparisons of dependent variables (looking at Δy_i). The "leapfrog" estimator of Section 1.4 instead uses across-individual, within-time comparisons of dependent variables (comparing y_{it} against y_{jt} for $i \neq j$). Section 1.5 reports the results from a Monte Carlo experiment using the estimators proposed. Section 1.6 presents a general covariance theorem as well as applications of the theorem, including several specification tests. Section 1.7 considers an empirical application that examines the consumption-smoothing effects of unemployment insurance. Section 1.8 concludes. The Appendix discusses computational issues and contains proofs of the theorems.

1.2 Applications

Before discussing the estimators, we motivate the paper with a few examples that satisfy (1.2). The common feature in the examples is that leaving h unspecified in (1.2) allows for greater flexibility and robustness in the estimation.

1.2.1 Multiple-Spell Proportional Hazard Models

Most of the literature on proportional hazards models, in both the single- and multiple-spell models, uses a random-effects approach to model heterogeneity. The obvious drawback of this approach is that the heterogeneity is assumed to be independent of the covariates. For multiple-spell models, panel data allows the possibility of using a fixed-effects approach where heterogeneity can be correlated with the covariates.

Examples of multiple-spell models which have been studied in the fixed-effects framework are consumer purchasing behavior (Gönül and Srinivasan (1993)) and child mortality (Olsen and Wolpin (1983)). The existing fixed-effects estimation techniques require some specification of the baseline hazard. Chamberlain (1985), for instance, considers marginal and conditional likelihood methods for eliminating the fixed effect in models with Weibull, lognormal, or gamma specifications. Gönül and Srinivasan (1993) propose a concentrated-likelihood method which can be used regardless of the functional form of the baseline hazard, but the chosen baseline enters into the concentrated-likelihood function. The estimators of this paper effectively eliminate the fixed effect without having to make any assumptions about the baseline hazard.

The proportional hazards model, incorporating fixed effects, specifies the hazard function

$$f(\tau) = \lambda(\tau)\exp(x_{it}\beta + \gamma_t + \alpha_i), \tag{1.3}$$

where $\lambda(\cdot)$ is the baseline hazard function and t indexes spells (rather than calendar time). The x_{it} differ over spells but remain constant within a given spell.

If y_{it} is the duration of spell t for individual i, equation (1.3) yields

$$h(y_{it}) = x_{it}\beta + \gamma_t + \alpha_i + \epsilon_{it} \tag{1.4}$$

where ϵ has the extreme density (i.e., $g(\epsilon) = \exp(\epsilon - \exp(\epsilon))$) and

$$h(y) = \ln \int_0^y \lambda(\tau)d\tau.$$

Since this model fits into the linear transformation panel framework, β can be estimated consistently (up to scale) without having to specify the baseline hazard function which determines h. Using the results of Section 1.6, one could do a specification test of a specific functional form for the baseline hazard.

There are a few drawbacks to the semiparametric approach. First, since the baseline is not parametrically specified, duration dependence can not be inferred. (The two-stage estimation technique discussed below can be used if one is willing to specify a parametric family for the hazard function after estimating β.) Second, much of the interesting duration data are censored in some way. The estimators of this paper do not apply when y is censored. Extension of the estimators to the case of truncated and censored y is being pursued in a separate paper.

1.2.2 Dependent-Variable Transformation Models

Even when economic models are able to make qualitative predictions, they do not always suggest the functional forms that should be used to test the predictions in empirical work. In an effort to allow the data to guide the choice of functional form, parametrized transformations of variables have been considered where the transformation parameters are estimated along with the other parameters of interest. In this paper, we focus on transformation of the dependent variable and allow for an unknown strictly monotonic transformation. The proposed estimators can estimate β without having to jointly estimate the transformation parameters.

The most widely used parametrized transformation is the Box-Cox transformation:

$$h(y,\lambda)=\begin{cases}(y^\lambda-1)/\lambda & \text{if } \lambda>0\\ \log y & \text{if } \lambda=0\end{cases} \tag{1.5}$$

with λ to be estimated along with the other parameters.

Maximum likelihood estimation (MLE) can be used to estimate the Box-Cox panel model. The estimation, which requires a parametric assumption on ϵ, is slightly more complicated than in the non-panel case (see Abrevaya (1996)). One can also first-difference the model and use the non-linear instrumental variables (NLIV) technique of Amemiya and Powell (1981), which does not require any parametric assumptions on the error term. First-differencing (with λ constant over the two periods) gives

$$\frac{y_2^\lambda-y_1^\lambda}{\lambda}=\Delta x\beta+\Delta\gamma+\Delta\epsilon. \tag{1.6}$$

The NLIV technique requires use of enough moment conditions (stemming from independence of the x's and their interactions from the error disturbances) to estimate β, $\Delta\gamma$, and λ.

More generally, MLE or NLIV can be used to estimate any model where the strictly monotonic h is parametrized by λ (possibly a vector), and (1.2) is rewritten

$$h(y_t,\lambda)=x_t\beta+\gamma_t+\alpha+\epsilon_t. \tag{1.7}$$

For each λ, $h(\cdot,\lambda)$ is a strictly increasing transformation. For λ to be identified, we need

$$\nexists\lambda_1,\lambda_2(\lambda_1\neq\lambda_2) \text{ and } a,b\in\mathcal{R} \text{ s.t. } h(y,\lambda_1)=ah(y,\lambda_2)+b\ \forall y\in\mathcal{R}.$$

First-differencing yields

$$h(y_2,\lambda)-h(y_1,\lambda)=\Delta x\beta+\Delta\gamma+\Delta\epsilon. \tag{1.8}$$

Bickel and Doksum (1981) have shown that the joint estimation of β and λ in the Box-Cox model is sensitive to correct specification of the model. That is, if the estimate of λ is incorrect, it can have large effects on the estimate of β. Also, it might be the case that none of the transformations from the chosen parametrized family is a correct specification of the model, in which case the estimate of β will be inconsistent. We can avoid these problems by estimating β consistently in a first stage using the "change" or "leapfrog" estimator, neither of which uses any information about λ. Then, in a second stage, a consistent estimate of λ (if (1.7) holds) can be obtained using a first-stage estimate $\hat{\beta}$.

For the second stage, we propose NLIV using interactions of the x's as instruments.[3] Since $\hat{\beta}$ only estimates β up to scale and doesn't estimate an intercept term, we will need to estimate scale and intercept coefficients in addition to λ. Plugging $\hat{\beta}$ into (1.8) yields

$$h(y_2, \lambda) - h(y_1, \lambda) = a + b(\Delta x \hat{\beta}) + u,$$

where u is independent of the x's under either of the assumptions on the error term needed for consistency of the "change" and "leapfrog" estimators. The independence gives moment conditions that can be used for NLIV estimation. For instance, we have $E[u' x_t] = E[u' \Delta x] = 0$, as well as moment conditions using any interactions of the x's as instruments. As long as enough moment conditions are used, we can estimate (a, b, λ) consistently. One needs to correct the standard errors for these estimates appropriately, taking into account the use of the estimate $\hat{\beta}$ rather than the true β.

1.2.3 Repeated Ratings Data

Consider a situation where n objects are rated by some entity over several time periods indexed by t. For instance, colleges are rated by students and administrative officials in national publications each year, products are rated by consumers, employees are given performance ratings by their employers, students are rated by testing agencies, etc.[4]

Let R_{it} denote the rating given to object i in year t. We hope to explain variation in ratings with time-varying characteristics of the objects. There's no reason to expect that the ratings themselves are a linear function of intrinsic quality. For instance, the quality difference between the top-rated college and the 20th-rated college is probably much larger than the quality difference between the 100th- and 120th-rated colleges. Rather than trying to specify the exact nature of the nonlinearity, we say that perceived quality is an unspecified, strictly increasing function of rating in each period,

$$Q_{it} = h_t(R_{it}).$$

Then, the linear regression of quality on time-varying covariates and a fixed effect becomes

$$h_t(R_{it}) = x_{it}\beta + \gamma_t + \alpha_i + \epsilon_{it}. \tag{1.9}$$

The fixed effect α can pick up any time-invariant covariates as well as any reputation

[3] Han (1987a) offers a semiparametric alternative which can be generalized to the panel setting to estimate λ. Given $\hat{\beta}$, the estimator $\hat{\lambda}$ maximizes (over ℓ)

$$\sum_{i \neq j} 1(\Delta x_i \hat{\beta} > \Delta x_j \hat{\beta}) 1(h(y_{i2}, \ell) - h(y_{i1}, \ell) > h(y_{j2}, \ell) - h(y_{j1}, \ell)).$$

[4] The key here is that the ratings of the objects are comparable, which is why we stress that they are rated by the same entity. If different entities rate different objects, it's impossible to compare the ratings. This problem is analogous to trying to compare different consumers' utilities in microeconomics.

effects that might exist. Notice that (1.9) is a linear transformation panel model where the transformation is indexed by time. The "leapfrog" estimator of Section 1.4 will be shown to be consistent in such a situation.

1.3 "Change" Estimation

The linear transformation panel model can be estimated using within-individual information after a transformation to a binary response model. If $\Delta\epsilon$ is i.i.d., note that

$$\begin{aligned}\Pr(\Delta y > 0|\Delta x) &= \Pr(h(y_2) > h(y_1)|\Delta x) \\ &= \Pr(\Delta x\beta + \Delta\gamma + \Delta\epsilon > 0) \\ &= F(\Delta x\beta + \Delta\gamma)\end{aligned}$$

where F is the c.d.f. of $-\Delta\epsilon$. Putting an i subscript on $h(\cdot)$ does not affect this derivation. As long as h is the same for a given observational unit at each time period, the "change" estimator will be consistent. In contrast, the "leapfrog" estimator of the next section will be consistent as long as h is the same in a given time period for each observational unit.

Defining $d \equiv 1(\Delta y > 0)$, one can estimate a binary response model with dependent variable d, independent variables Δx, and an unspecified increasing $F : \mathcal{R} \to [0,1]$. If one is willing to assume the form of F, an efficient estimate (within the class of estimators using binary data for the dependent variable) can be obtained using MLE. When F is assumed to be unknown, several semiparametric $\sqrt{n}$-consistent estimators can be used, including those developed by Han (1987b), Ichimura (1993), and Klein and Spady (1992).[5]

We discuss Han's MRC estimator because of its computational ease and its similarity to the "leapfrog" estimator proposed in the next section. Specifically, let $\hat{\beta}_c$ be the estimator that maximizes the objective function

$$S_c(b) = \sum_{i \neq j} 1(\Delta x_i b > \Delta x_j b) d_i \tag{1.10}$$

over the set $\{b \in \mathcal{R}^k | b_k = 1\}$. A normalization is needed here since $1(\Delta x_i b > \Delta x_j b)$ is not affected by the scale of b (i.e., $1(\Delta x_i b > \Delta x_j b) = 1(\Delta x_i(cb) > \Delta x_j(cb))$ for $c > 0$). The MRC estimator also does not identify $\Delta\gamma$ since $1(\Delta x_i b > \Delta x_j b)$ is not affected by addition of a constant (i.e., $1(\Delta x_i b > \Delta x_j b) = 1(\Delta x_i b + c > \Delta x_j b + c)$ for all c).

[5]To test a specified distribution for F, one can perform a Hausman (1978) test of the parametric estimator versus a semiparametric estimator. The former is efficient under correct specification but inconsistent under misspecification.

$\hat{\beta}_c$ can be interpreted as a *rank estimator* since it maximizes

$$\sum_i \text{Rank}(\Delta x_i b) d_i, \tag{1.11}$$

where the Rank($\cdot$) function is defined as follows:[6]

$$\Delta x_{i_1} b < \Delta x_{i_2} b < \ldots < \Delta x_{i_n} b \Longrightarrow \text{Rank}(\Delta x_{i_m} b) = m. \tag{1.12}$$

For consistency and asymptotic normality of $\hat{\beta}_c$, we need several assumptions. Throughout the paper, we follow closely the notation used by Sherman (1993) and Cavanagh and Sherman (1992).

Assumption 1 *β is an interior element of the parameter space $\mathcal{B}$, a compact subset of $\{b \in \mathcal{R}^k | b_k = 1\}$.*

Assumption 2 *Δx is an i.i.d. k-dimensional random variable s.t.:*
(i) The support of Δx is not contained in a proper linear subspace of $\mathcal{R}^k$.
(ii) The k'th component of Δx has everywhere positive Lebesgue density, conditional on the other components.

Assumption $\mathbf{E}_c$ *$\Delta\epsilon$ is an i.i.d. random variable.*

A final assumption requires further notation. Let $z \equiv (y_1, y_2, \Delta x)$ denote an observation from the set $\mathcal{S} \subseteq \mathcal{R} \times \mathcal{R} \times \mathcal{R}^k$. For each $z \in \mathcal{S}$ and $b \in \mathcal{B}$, define

$$d(z) \equiv 1(\Delta y > 0) \tag{1.13}$$

and

$$\tau_c(z,b) \equiv \mathrm{E}_Z \left[\{d(z) > d(Z)\}\{\Delta x b > \Delta X b\} + \{d(Z) > d(z)\}\{\Delta X b > \Delta x b\} \right] \tag{1.14}$$

where $Z \equiv (Y_1, Y_2, \Delta X)$. Write ∇_m for the m'th partial derivative operator applied to the first $k-1$ components of b, and

$$|\nabla_m| \sigma(b) \equiv \sum_{i_1,\ldots,i_m} \left| \frac{\partial^m}{\partial b_{i_1} \cdots \partial b_{i_m}} \sigma(b) \right|. \tag{1.15}$$

Finally, the symbol $\|\cdot\|$ denotes the matrix norm: $\|(a_{ij})\| \equiv \left(\sum_{i,j} a_{ij}^2\right)^{1/2}$.

We say that "*$\tau(z,\cdot)$ is Taylor-expandable*" if, for $\mathcal{N}$ a neighborhood of β,
(i) For each $z \in \mathcal{S}$, all mixed second partial derivatives of $\tau(z,\cdot)$ exist on $\mathcal{N}$.
(ii) There is an integrable function $\Gamma(z)$ such that for all $z \in \mathcal{S}$ and $b \in \mathcal{N}$,

$$\|\nabla_2 \tau(z,b) - \nabla_2 \tau(z,\beta)\| \leq \Gamma(z)|b|.$$

[6]It's innocuous to consider strict inequalities here due to a continuity assumption on Δx needed for consistency (see Assumption 2 (ii) below).

(iii) $\mathrm{E}|\nabla_1\tau(\cdot,\beta)|^2 < \infty$.
(iv) $\mathrm{E}|\nabla_2|\tau(\cdot,\beta) < \infty$.
(v) The $(d-1)\times(d-1)$ *matrix* $\mathrm{E}\nabla_2\tau(\cdot,\beta)$ *is negative definite.*

Then, we have

Theorem 1 *If Assumptions 1, 2, and* E_c *hold and* $\tau_c(z,\cdot)$ *is Taylor-expandable, then*

$$\sqrt{n}(\hat{\beta}_c - \beta) \xrightarrow{d} \begin{pmatrix} W_c \\ 0 \end{pmatrix}$$

where $W_c \sim N(0, V_c^{-1}\Delta_c V_c^{-1})$, *with* $\Delta_c = \mathrm{E}\nabla_1\tau_c(\cdot,\beta)[\nabla_1\tau_c(\cdot,\beta)]'$ *and* $2V_c = \mathrm{E}\nabla_2\tau_c(\cdot,\beta)$.

Both V_c and Δ_c can be written as functions of the model's primitives. Let $g_o(\cdot)$ denote the marginal density of $\Delta x\beta$, let $\Delta\tilde{x}$ denote the first $k-1$ components of Δx, and define $\Delta\tilde{x}_o \equiv \mathrm{E}(\Delta\tilde{x}|\Delta x\beta)$. Then, we have

Theorem 2 *If* F *and* g_o *are differentiable and* $\mathrm{E}|\Delta x|^2 < \infty$, *then*

$$\Delta_c = \mathrm{E}(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)^2 F(\Delta x\beta)[1 - F(\Delta x\beta)] \tag{1.16}$$

$$\textit{and} \quad 2V_c = -\mathrm{E}(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)F'(\Delta x\beta)g_o(\Delta x\beta). \tag{1.17}$$

The "change" estimator suffers from some drawbacks. First, the i.i.d. assumption on $\Delta\epsilon$ needed for consistency may be considered too restrictive for certain applications. One way to weaken this restriction is to use the maximum score estimator (MSE) of Manski (1975) instead of the MRC estimator. The MSE requires only a median restriction on $\Delta\epsilon$ and allows for heteroskedasticity across i. The estimator $\hat{\beta}_{MSE}$ maximizes

$$\sum_i \mathrm{sgn}(\Delta x_i b) d_i, \tag{1.18}$$

where the sign function $\mathrm{sgn}(v) \equiv [1(v > 0) - 1(v < 0)]$. Unfortunately, the MSE is not $\sqrt{n}$-consistent. To quickly see that the MSE is less efficient than the MRC estimator, recall the interpretation of $\hat{\beta}_c$ as a rank estimator in (1.11). The Rank($\cdot$) function of the MRC estimator takes full advantage of F's monotonicity. At the true value β, MSE only uses the fact that $d = 1$ is more likely for a positive $\Delta x\beta$ than a negative $\Delta x\beta$, whereas MRC uses the monotonicity of F to compare across $\Delta x\beta$'s of all levels. The MSE does identify the coefficient on a constant term in Δx (and thus estimates $\Delta\gamma$) since the sign of $\Delta x\beta$ is affected by addition or subtraction of a constant whereas the rankings of $\Delta x\beta$ are not.

A second drawback is that the "change" estimator, using either MRC or MSE, throws away a lot of information about the dependent variables by using only the sign of Δy in the objective function. There is additional information contained in the levels of the y_t's which should be utilized for more efficient estimation.

Moreover, the efficiency of the "change" estimator depends on the importance of the time trend in the dependent variable. The estimator is most useful in situations where Δy is positive for many observational units and negative for many others. In the extreme case that Δy is the same sign for all observational units, the parameters will not be identified at all. One can simply look at the percentage of positive Δy before doing any estimation to get a sense of how well the "change" estimator can be expected to perform; a value near 50% is ideal.

Finally, the consistency of the "change" estimator hinges critically on the fact that h does not change over time. If h were to be a function of t (e.g., a Box-Cox model with different transformation parameters λ_1 and λ_2 in the two periods), a different estimator would have to be used.

The "leapfrog" estimator of the next section is proposed as an attractive alternative to the "change" estimator since it addresses each of the aforementioned weaknesses of "change" estimation.

1.4 "Leapfrog" Estimation

In this section, we develop a "leapfrog" estimator that works in the presence of significant time trends, allows h to vary over time, and allows for forms of heteroskedasticity across observational units. The basic idea behind the estimator is to compare dependent variables of a given time period across different observational units. We have

$$h_t(y_{it}) - h_t(y_{jt}) = (x_{it} - x_{jt})\beta + (\alpha_i - \alpha_j) + (\epsilon_{it} - \epsilon_{jt}). \tag{1.19}$$

Note that h now has a time subscript, allowing for a different strictly monotonic transformation h_t in each time period. Equation (1.19) immediately yields

$$y_{it} > y_{jt} \Longleftrightarrow (x_{it} - x_{jt})\beta + (\alpha_i - \alpha_j) + (\epsilon_{it} - \epsilon_{jt}) > 0. \tag{1.20}$$

Also, note that

$$\Delta x_i\beta > \Delta x_j\beta \Longrightarrow (x_{i2} - x_{j2})\beta + (\alpha_i - \alpha_j) > (x_{i1} - x_{j1})\beta + (\alpha_i - \alpha_j). \tag{1.21}$$

If $(\epsilon_{i1} - \epsilon_{j1})$ and $(\epsilon_{i2} - \epsilon_{j2})$ have the same marginal distribution, combining (1.20) and (1.21) yields

$$\Delta x_i\beta > \Delta x_j\beta \Longrightarrow \Pr(y_{i2} > y_{j2}|\Delta x_i, \Delta x_j) > \Pr(y_{i1} > y_{j1}|\Delta x_i, \Delta x_j). \tag{1.22}$$

Assumption E_c is thus replaced with

Assumption E_ℓ *For all i and j, $(\epsilon_{i1} - \epsilon_{j1})$ and $(\epsilon_{i2} - \epsilon_{j2})$ have the same marginal distribution, conditional on $(x_{i1}, x_{i2}, x_{j1}, x_{j2})$.*

This assumption is rather weak. For instance, if errors are stationary across time for each individual but heteroskedastic across individuals, Assumption E_ℓ is satisfied.

Condition (1.22) suggests maximizing the objective function

$$S(b) = \sum_{i \neq j} 1(\Delta x_i b > \Delta x_j b)[1(y_{i2} > y_{j2}) - 1(y_{i1} > y_{j1})]. \tag{1.23}$$

We rewrite (1.23) to yield the "leapfrog" objective function:

$$\begin{aligned} S(b) &= \textstyle\sum_{i\neq j} 1(\Delta x_i b > \Delta x_j b)[1(y_{i2} > y_{j2}) + 1(y_{i1} < y_{j1}) - 1] \\ &= \textstyle\sum_{i\neq j} 1(\Delta x_i b > \Delta x_j b)[1(y_{i2} > y_{j2}, y_{i1} < y_{j1}) + 1(\text{sgn}(y_{i1} - y_{j1}) \neq \text{sgn}(y_{i2} - y_{j2}))] \\ &= S_\ell(b) + \textstyle\sum_{i\neq j} 1(\Delta x_i b > \Delta x_j b) 1(\text{sgn}(y_{i1} - y_{j1}) \neq \text{sgn}(y_{i2} - y_{j2})), \end{aligned}$$

where

$$S_\ell(b) = \sum_{i\neq j} 1(\Delta x_i b > \Delta x_j b) 1(y_{i1} < y_{j1}, y_{i2} > y_{j2}). \tag{1.24}$$

Note that the same estimate will maximize both (1.23) and (1.24) since

$$\begin{aligned} S(b) - S_\ell(b) &= \sum_{i\neq j} 1(\Delta x_i b > \Delta x_j b)[1(\text{sgn}(y_{i1} - y_{j1}) \neq \text{sgn}(y_{i2} - y_{j2}))] \\ &= \frac{1}{2}\sum_{i\neq j} 1(\text{sgn}(y_{i1} - y_{j1}) \neq \text{sgn}(y_{i2} - y_{j2})) \end{aligned}$$

is not a function of b.

Let $\hat{\beta}_\ell$ be the estimator that maximizes (1.24) over the set $\{b \in \mathcal{R}^k | b_k = 1\}$. We call $\hat{\beta}_\ell$ the "leapfrog" estimator since the objective function rewards leapfrogging by those observational units having higher $\Delta x\beta$. Leapfrogging by i over j means that i has a smaller dependent variable than j at $t = 1$ but a larger dependent variable at $t = 2$.

Since $\hat{\beta}_\ell$ depends on leapfrogging behavior, one can calculate the percentage of observation-pairs that "leapfrog" each other to get a sense of how well the "leapfrog" estimator will perform. In the extreme case that there is no leapfrogging whatsoever in the sample, the objective function $S_\ell(b)$ will be equal to zero for all b.

Using the notation of the previous section, for each $z \in \mathcal{S}$ and $b \in \mathcal{B}$, define

$$H(y_1, y_2, v) \equiv \Pr_Z[y_1 < Y_1, y_2 > Y_2 | \Delta X\beta = v] - \Pr_Z[y_1 > Y_1, y_2 < Y_2 | \Delta X\beta = v]$$

and

$$\tau_\ell(z, b) \equiv \mathrm{E}_Z[\{y_1 < Y_1, y_2 > Y_2\}\{\Delta xb > \Delta Xb\} + \{Y_1 < y_1, Y_2 > y_2\}\{\Delta Xb > \Delta xb\}]$$

where $Z \equiv (Y_1, Y_2, \Delta X)$.

The asymptotic normality theorems are analogous to those for $\hat{\beta}_c$.

Theorem 3 *If Assumptions 1, 2, and E_ℓ hold and $\tau_\ell(z, \cdot)$ is Taylor-expandable, then*

$$\sqrt{n}(\hat{\beta}_\ell - \beta) \xrightarrow{d} \begin{pmatrix} W_\ell \\ 0 \end{pmatrix}$$

where $W_\ell \sim N(0, V_\ell^{-1}\Delta_\ell V_\ell^{-1})$, with $\Delta_\ell = \mathrm{E}\nabla_1\tau_\ell(\cdot, \beta)[\nabla_1\tau_\ell(\cdot, \beta)]'$ and $2V_\ell = \mathrm{E}\nabla_2\tau_\ell(\cdot, \beta)$.

Theorem 4 *If g_o is differentiable, H is differentiable with respect to its third argu-*

ment, and $\mathrm{E}|\Delta x|^2 < \infty$*, then*

$$\Delta_\ell = \mathrm{E}(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)^2 H(y_1, y_2, \Delta x\beta)^2 \tag{1.25}$$

$$\text{and} \quad 2V_\ell = \mathrm{E}(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)H_3(y_1, y_2, \Delta x\beta)g_o(\Delta x\beta). \tag{1.26}$$

1.5 Monte Carlo Results

In this section, we report the results of Monte Carlo simulations using the proposed estimators. For comparison, we study the least-squares within estimator to see the effect of misspecification. In particular, we look first at a correctly specified linear model and then at models which are slight deviations from the linear model. We consider the power transformation family studied by Bickel and Doksum (1981), which allows for negative y values unlike the Box-Cox family:

$$h(y, \lambda) = \frac{|y|^\lambda \mathrm{sgn}(y) - 1}{\lambda} \quad \text{for positive } \lambda.$$

We consider three values for λ (0.90, 0.95, and 1), where $\lambda = 1$ corresponds to the linear model. The associated transformations are shown in Figure 1.1 for the range of y used in the simulations. As the figure shows, we are considering very mild misspecification in the Monte Carlo design. More serious misspecifications will result in greater bias than reported in this section.

We consider the following specification for the Monte Carlo design:

$$\begin{aligned} h(y_{i1}) &= 1 + x_{1i1} - 2x_{2i1} + 4x_{3i1} + \alpha_i + \epsilon_{i1} \\ h(y_{i2}) &= \phantom{1 +{}} x_{1i2} - 2x_{2i2} + 4x_{3i2} + \alpha_i + \epsilon_{i2}, \end{aligned}$$

where x_{1i1} and x_{1i2} are lognormally distributed, x_{2i1} and x_{2i2} are uniformly distributed, x_{3i1} and x_{3i2} are dummy variables (equal to one with probability 0.3), ϵ_{i1} and ϵ_{i2} are normally distributed, and the fixed effect is

$$\alpha_i = 0.5(x_{2i1} + x_{2i2}) + \eta_i,$$

where η_i is normally distributed. Aside from the fixed effect, there is no additional correlation between any of the random variables. The true coefficient vector for this design, including an intercept term, is $\beta = (\beta_0, \beta_1, \beta_2, \beta_3) = (1, 1, -2, 4)$.

For each of the three λ values, four estimators were used to estimate β: (1) the OLS within estimator; (2) the maximum score estimator; (3) the "change" estimator, $\hat{\beta}_c$; and (4) the "leapfrog" estimator, $\hat{\beta}_\ell$. Samples of 100 and 200 were drawn according to the design described above. The results for 250 simulations for each sample size and each λ value are given in Table 1.1. For comparability, the ratios of the coefficients (to β_3) are shown. As explained in Sections 1.3 and 1.4, the "change" and "leapfrog" estimators do not estimate the intercept β_0.

Figure 1.1:

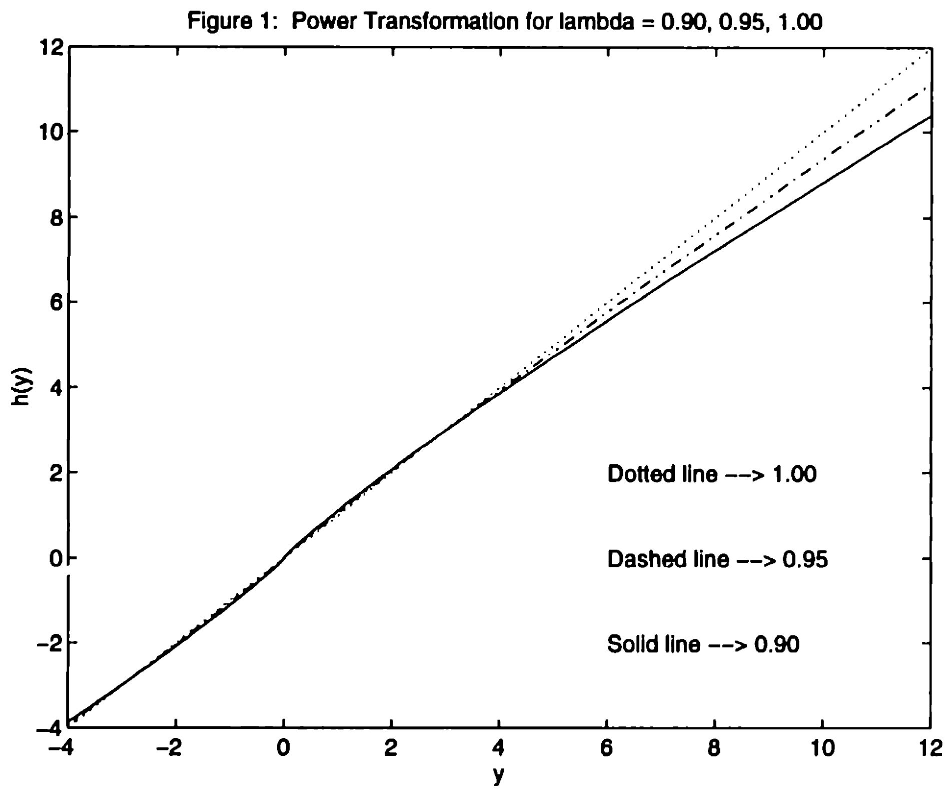


Figure 1: Power Transformation for lambda = 0.90, 0.95, 1.00
12
10
8
6
4
2
0
-2
-4
h(y)
-4
-2
0
2
4
6
8
10
12
y
Dotted line --> 1.00
Dashed line --> 0.95
Solid line --> 0.90

Table 1.1: Monte Carlo results

	$\lambda = 1$ (linear)			$\lambda = 0.95$			$\lambda = 0.90$		
	β_0/β_3	β_1/β_3	β_2/β_3	β_0/β_3	β_1/β_3	β_2/β_3	β_0/β_3	β_1/β_3	β_2/β_3
True	0.2500	0.2500	-0.5000	0.2500	0.2500	-0.5000	0.2500	0.2500	-0.5000
$n = 100$									
OLS	0.2521	0.2495	-0.5045	0.2453	0.2401	-0.5041	0.2476	0.2293	-0.4905
	(0.0383)	(0.0203)	(0.1003)	(0.0386)	(0.0196)	(0.1002)	(0.0411)	(0.0193)	(0.0948)
MSE	0.2600	0.2650	-0.5210	0.2609	0.2623	-0.5236	0.2541	0.2569	-0.5096
	(0.1017)	(0.1049)	(0.1958)	(0.0991)	(0.0865)	(0.2104)	(0.1049)	(0.0823)	(0.1935)
"Change"	——	0.2480	-0.4956	——	0.2558	-0.5122	——	0.2515	-0.5012
		(0.0655)	(0.2031)		(0.0663)	(0.1971)		(0.0648)	(0.2127)
"Leapfrog"	——	0.2471	-0.5001	——	0.2533	-0.4979	——	0.2538	-0.5000
		(0.0409)	(0.1524)		(0.0392)	(0.1440)		(0.0426)	(0.1444)
$n = 200$									
OLS	0.2511	0.2505	-0.4993	0.2506	0.2392	-0.5042	0.2449	0.2286	-0.4899
	(0.0268)	(0.0143)	(0.0660)	(0.0294)	(0.0135)	(0.0646)	(0.0263)	(0.0145)	(0.0632)
MSE	0.2418	0.2540	-0.5012	0.2567	0.2571	-0.5121	0.2500	0.2568	-0.4910
	(0.0795)	(0.0583)	(0.1664)	(0.0838)	(0.0656)	(0.1584)	(0.0786)	(0.0616)	(0.1699)
"Change"	——	0.2504	-0.4945	——	0.2473	-0.5112	——	0.2475	-0.4959
		(0.0443)	(0.1416)		(0.0415)	(0.1286)		(0.0420)	(0.1462)
"Leapfrog"	——	0.2492	-0.4964	——	0.2474	-0.5029	——	0.2519	-0.4988
		(0.0269)	(0.0935)		(0.0252)	(0.0867)		(0.0287)	(0.0971)

The estimates are sample averages over 250 simulations. Sample standard errors are reported in parentheses.

Not surprisingly, OLS is the most efficient estimator when the model is truly linear. The standard errors for the "change" and "leapfrog" estimators are between 150% and 300% larger. As the true model departs from linearity, the inconsistency of OLS becomes apparent. The OLS estimate of β_1/β_3 is around 0.24 for $\lambda = 0.95$ (4% off from 0.25) and around 0.23 for $\lambda = 0.90$ (8% off from 0.25); the OLS estimate of β_2/β_3 for $\lambda = 0.90$ is about 2% off. The "change" and "leapfrog" estimators are fairly similar across the λ values, with the "leapfrog" estimator slightly outperforming the "change" estimator in terms of accuracy (most notably for $\lambda = 0.95$). The MSE doesn't appear to perform too well overall; even for $\lambda = 0.90$, the MSE is not much better than the inconsistent OLS. Presumably the smoothed MSE with bias correction (see Horowitz (1992) and Chapter 2 of this thesis) would perform better.

1.6 Covariance Theorem

In order to develop specification tests involving the proposed estimators, we need to be able to determine covariances between estimators which maximize objective functions like those in (1.10) and (1.24). In this section, we state a general theorem to serve this purpose. The theorem will also be used to estimate β when $T > 2$ by efficiently combining estimates which each use only information from two time periods. We state the theorem and then discuss the applications. A new piece of notation, P_n, is the empirical measure that places mass $1/n$ on each z_i.

Theorem 5 *Let $\Gamma_{1n}(\cdot)$ and $\Gamma_{2n}(\cdot)$ be objective functions which are sample analogues of $\Gamma_1(\cdot)$ and $\Gamma_2(\cdot)$. Both $\Gamma_1(\cdot)$ and $\Gamma_2(\cdot)$ are maximized at $\theta_o = \mathbf{0}$, whereas the sample analogues are maximized at θ_{1n} and θ_{2n}, respectively. Suppose θ_{1n} and θ_{2n} are $\sqrt{n}$-consistent estimates of $\mathbf{0}$, an interior point of Θ. Suppose that, uniformly over $O_p(1/\sqrt{n})$ neighborhoods of $\mathbf{0}$,*

$$\Gamma_{1n}(\theta) = \frac{1}{2}\theta' V_1 \theta + \frac{1}{\sqrt{n}}\theta' W_{1n} + o_p(1/n)$$

$$\Gamma_{2n}(\theta) = \frac{1}{2}\theta' V_2 \theta + \frac{1}{\sqrt{n}}\theta' W_{2n} + o_p(1/n)$$

where V_1 and V_2 are negative definite matrices and

$$W_{1n} = \sqrt{n} P_n \nabla_1 \tau_1(\cdot, \mathbf{0}), \quad W_{2n} = \sqrt{n} P_n \nabla_1 \tau_2(\cdot, \mathbf{0}),$$

where both $\tau_1(z, \cdot)$ and $\tau_2(z, \cdot)$ are Taylor-expandable.[7]

Then,

$$\sqrt{n}\begin{pmatrix} \theta_{1n} \\ \theta_{2n} \end{pmatrix} \xrightarrow{d} N\left(0, \begin{pmatrix} V_1^{-1}\Delta_1 V_1^{-1} & V_1^{-1}\Delta_{12} V_2^{-1} \\ V_2^{-1}\Delta_{12}' V_1^{-1} & V_2^{-1}\Delta_2 V_2^{-1} \end{pmatrix}\right),$$

[7]Note that θ is not a normalized parameter. For instance, it can be thought of as the first $k-1$ components of β from the previous sections. As a result, ∇_m is the m'th partial derivative applied to *all* components of θ. Likewise, Taylor-expandability is defined in this way.

where

$$\begin{aligned}\Delta_1 &= E\nabla_1\tau_1(\cdot,\mathbf{0})[\nabla_1\tau_1(\cdot,\mathbf{0})]' \\ \Delta_2 &= E\nabla_1\tau_2(\cdot,\mathbf{0})[\nabla_1\tau_2(\cdot,\mathbf{0})]' \\ \textit{and}\ \Delta_{12} &= E\nabla_1\tau_1(\cdot,\mathbf{0})[\nabla_1\tau_2(\cdot,\mathbf{0})]'.\end{aligned}$$

To make Theorem 5 useful in practice, we can give a specific expression for Δ_{12} (which can be estimated using kernel techniques) when the estimators maximize objective functions of the form $\sum_{i\neq j} 1(w_i b > w_j b)1(\cdots)$. To be precise, we state

Theorem 6 *Let θ_{1n} and θ_{2n} be estimates satisfying the conditions of Theorem 5. Suppose that θ_{1n} and θ_{2n} are the first $k-1$ components of $\hat{\beta}_1$ and $\hat{\beta}_2$ (elements of $\mathcal{B}$), respectively. Suppose also that*

$$\tau_1(z_1,b) = \int_{W_1 b < w_1 b} S(\mathbf{y_1}, W_1\beta)G_1(dW_1) + \int \rho(\mathbf{y_1}, W_1\beta)G_1(dW_1)$$

$$\tau_2(z_2,b) = \int_{W_2 b < w_2 b} T(\mathbf{y_2}, W_2\beta)G_2(dW_2) + \int \nu(\mathbf{y_2}, W_2\beta)G_2(dW_2)$$

where $z_i = (\mathbf{y_i}, w_i)$, $Z_i = (\mathbf{Y_i}, W_i)$, and $G_i(\cdot)$ is the distribution of w_i for $i = 1,2$.
If $E|w_1' w_2| < \infty$, then

$$\Delta_{12} = E(\tilde{w}_1 - \tilde{w}_{1o})'(\tilde{w}_2 - \tilde{w}_{2o})g_1(w_1\beta)g_2(w_2\beta)S(\mathbf{y_1}, w_1\beta)T(\mathbf{y_2}, w_2\beta),$$

where $\tilde{w}_i$ denotes the first $k-1$ components of w_i, $\tilde{w}_{io} \equiv E(\tilde{w}_i|w_i\beta)$, and $g_i(\cdot)$ is the marginal density of $w_i\beta$ for $i = 1,2$.

The estimators we've considered have $w_1 = w_2 = \Delta x$ and $\mathbf{y_1} = \mathbf{y_2} = (y_1, y_2)$, but Theorem 6 is stated more generally so that it can be applied for different w_i and $\mathbf{y_i}$. Two such applications, testing for fixed effects and estimating the model for $T > 2$, will be discussed.

The main use of these covariance results is to perform specification tests and to combine estimates for greater efficiency. If under the null hypothesis, both $\hat{\beta}_1$ and $\hat{\beta}_2$ are consistent estimates of β, but under the alternative hypothesis, only $\hat{\beta}_2$ is a consistent estimate of β, we can perform a χ^2-test since

$$(\hat{\beta}_1^{-k} - \hat{\beta}_2^{-k})'\mathrm{Var}(\hat{\beta}_1^{-k} - \hat{\beta}_2^{-k})^{-1}(\hat{\beta}_1^{-k} - \hat{\beta}_2^{-k}) \overset{H_o}{\approx} \chi^2_{k-1}, \tag{1.27}$$

where the superscript "$-k$" indicates only the first $k-1$ components. The variance term has the form

$$\begin{aligned}\mathrm{Var}(\hat{\beta}_1^{-k} - \hat{\beta}_2^{-k}) &= \mathrm{Var}(\hat{\beta}_1^{-k}) + \mathrm{Var}(\hat{\beta}_2^{-k}) - \mathrm{Cov}(\hat{\beta}_1^{-k}, \hat{\beta}_2^{-k}) - \mathrm{Cov}(\hat{\beta}_1^{-k}, \hat{\beta}_2^{-k})' \\ &= V_1^{-1}\Delta_1 V_1^{-1} + V_2^{-1}\Delta_2 V_2^{-1} - V_1^{-1}\Delta_{12}V_2^{-1} - V_2^{-1}\Delta_{12}'V_1^{-1}.\end{aligned}$$

Theorems 5 and 6 allow us to estimate $\mathrm{Var}(\hat{\beta}_1^{-k} - \hat{\beta}_2^{-k})$ (using consistent estimates $\widehat{V}_1$, $\widehat{V}_2$, $\widehat{\Delta}_1$, $\widehat{\Delta}_2$, and $\widehat{\Delta}_{12}$) and then perform an asymptotic χ^2-test.

To improve efficiency of two consistent estimates $\hat{\beta}_1$ and $\hat{\beta}_2$, we can take a linear combination of the two, taking into account their correlation.[8] In particular, for a $(k-1) \times (k-1)$ matrix A, consider

$$\hat{\beta}_A^{-k} = A\hat{\beta}_1^{-k} + (I - A)\hat{\beta}_2^{-k}.$$

The optimal weighting matrix A^* is given by

$$\begin{aligned} A^* &= (\mathrm{Var}(\hat{\beta}_1^{-k}) + \mathrm{Var}(\hat{\beta}_2^{-k}) - \mathrm{Cov}(\hat{\beta}_1^{-k}, \hat{\beta}_2^{-k}) - \mathrm{Cov}(\hat{\beta}_1^{-k}, \hat{\beta}_2^{-k})')^{-1}(\mathrm{Var}(\hat{\beta}_2^{-k}) - \mathrm{Cov}(\hat{\beta}_1^{-k}, \hat{\beta}_2^{-k})) \\ &= (V_1^{-1}\Delta_1 V_1^{-1} + V_2^{-1}\Delta_2 V_2^{-1} - V_1^{-1}\Delta_{12}V_2^{-1} - V_2^{-1}\Delta_{12}'V_1^{-1})^{-1}(V_2^{-1}\Delta_2 V_2^{-1} - V_1^{-1}\Delta_{12}V_2^{-1}). \end{aligned}$$

Using Theorems 5 and 6, we can consistently estimate A^* to achieve the asymptotically efficient linear combination of $\hat{\beta}_1$ and $\hat{\beta}_2$.

Testing/combining β_c and β_ℓ: If $\hat{\beta}_1 \equiv \hat{\beta}_c$ and $\hat{\beta}_2 \equiv \hat{\beta}_\ell$, we can apply the results from above. The χ^2-test of their difference could test the validity of Assumptions E_c and E_ℓ. A linear combination of $\hat{\beta}_c$ and $\hat{\beta}_\ell$ would be appropriate if both Assumptions E_c and E_ℓ hold (i.e., if both $\hat{\beta}_c$ and $\hat{\beta}_\ell$ are consistent). If the conditions of Theorem 6 hold,

$$\Delta_{12} = E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)^2 1(\Delta y > 0)H(y_1, y_2, \Delta x\beta).$$

Testing a specified h: Both $\hat{\beta}_c$ and $\hat{\beta}_\ell$ are consistent when h is unknown as long as the appropriate assumptions hold. If h is known, we can estimate β more efficiently by using the exact form of h. Specifically, let $\hat{\beta}_1$ be the MRC estimate using $h(y_2) - h(y_1)$ as the dependent variable and Δx as the independent variables. Notice that

$$\begin{aligned} & \Pr(h(y_{i2}) - h(y_{i1}) > h(y_{j2}) - h(y_{j1})) > \Pr(h(y_{i2}) - h(y_{i1}) < h(y_{j2}) - h(y_{j1})) \\ \Longleftrightarrow\ & \Pr(h(y_{i2}) - h(y_{i1}) > h(y_{j2}) - h(y_{j1})) > 1/2 \end{aligned}$$

and

$$\begin{aligned} \Pr(h(y_{i2}) - h(y_{i1}) &> h(y_{j2}) - h(y_{j1})) \\ &= \Pr((\Delta x_i - \Delta x_j)\beta + (\Delta\epsilon_i - \Delta\epsilon_j) > 0) \qquad (1.28) \\ &= \Pr((\Delta x_i - \Delta x_j)\beta + (\epsilon_{i2} - \epsilon_{j2}) - (\epsilon_{i1} - \epsilon_{j1}) > 0). \qquad (1.29) \end{aligned}$$

The equality in (1.28) implies that $\hat{\beta}_1$ is consistent under Assumption E_c, while the equality in (1.29) implies that $\hat{\beta}_1$ is consistent under Assumption E_ℓ. Thus, if h is

[8] An alternative way of combining the information used by two estimators is to do so explicitly in an objective function. For instance, if the assumptions for both the "change" and "leapfrog" estimators are true, then a single estimator which maximizes $(S_c(b) + S_\ell(b))$ would be consistent. This approach would require determining the covariance matrix of the new estimator, whereas the approach described in this section merely uses the covariance matrices derived in the previous sections.

correctly specified, $\hat{\beta}_1$ will be consistent under either assumption. If Assumption E_c holds, one can perform a specification test of $\hat{\beta}_1$ against $\hat{\beta}_c$ to test if h is correctly specified. Similarly, if Assumption E_ℓ holds, a test of $\hat{\beta}_1$ against $\hat{\beta}_\ell$ tests the specification of h.

Define

$$\begin{aligned} S^h(y_1, y_2, v) &\equiv \Pr_Z[h(y_2) - h(y_1) > h(Y_2) - h(Y_1) | \Delta X\beta = v] \\ &- \Pr_Z[h(y_2) - h(y_1) < h(Y_2) - h(Y_1) | \Delta X\beta = v]. \end{aligned}$$

Then, for $\hat{\beta}_2 \equiv \hat{\beta}_c$,

$$\Delta_{12} = E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)^2 1(\Delta y > 0)S^h(y_1, y_2, \Delta x\beta).$$

and for $\hat{\beta}_2 \equiv \hat{\beta}_\ell$,

$$\Delta_{12} = E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)^2 H(y_1, y_2, \Delta x\beta)S^h(y_1, y_2, \Delta x\beta).$$

Testing for correlated fixed effects: In order to test whether α is correlated with x, one can do a specification test of either $\hat{\beta}_c$ or $\hat{\beta}_\ell$ against a single-period MRC estimator $\hat{\beta}_t^{MRC}$ (i.e., MRC using dependent variable y_t and independent variables x_t for a given t). The single-period MRC estimator is consistent if the condition

$$x_{it}\beta > x_{jt}\beta \Longrightarrow \Pr(y_{it} > y_{jt}) > 1/2$$

holds. We have

$$\begin{aligned} \Pr(y_{it} > y_{jt}) &= \Pr(x_{it}\beta + \alpha_i + \epsilon_{it} > x_{jt}\beta + \alpha_j + \epsilon_{jt}) \\ &= \Pr((x_{it} - x_{jt})\beta > (\alpha_i + \epsilon_{it}) - (\alpha_j + \epsilon_{jt})), \end{aligned}$$

which yields the consistency condition if α is independent of x and, for all i and j, the median of $(\alpha_i + \epsilon_{it}) - (\alpha_j + \epsilon_{jt})$ is zero. Notice that this condition is weaker than the restriction made by Han (1987b) since we are considering only a subset of the models to which MRC applies. Whereas an i.i.d. assumption is needed for MRC to be consistent for more general models, the restriction here allows for forms of heteroskedasticity when dealing with the linear transformation model. For instance, if $(\alpha_i + \epsilon_{it})$ is symmetric for each i but not necessarily identically distributed, the median condition holds.

If Assumption E_c (or Assumption E_ℓ) holds, then $\hat{\beta}_c$ (or $\hat{\beta}_\ell$) will be consistent even if there is correlation between the fixed effect and the covariates. If the median restriction holds for a given t but there is correlation between the fixed effect and the covariates, the estimate $\hat{\beta}_t^{MRC}$ described above will be inconsistent. Then, a χ^2-test of the difference between $\hat{\beta}_c$ (or $\hat{\beta}_\ell$) and $\hat{\beta}_t^{MRC}$ serves as a specification test.

Let $\hat{\beta}_2 \equiv \hat{\beta}_t^{MRC}$. Define

$$T(y_t, v) \equiv \Pr_Z[y_t > Y_t | x_t\beta = v] - \Pr_Z[y_t < Y_t | x_t\beta = v].$$

Then, for $\hat{\beta}_1 \equiv \hat{\beta}_\ell$,

$$\Delta_{12} = E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\tilde{x}_t - \tilde{x}_{to})H(y_1, y_2, \Delta x\beta)T(y_t, x_t\beta),$$

and for $\hat{\beta}_1 \equiv \hat{\beta}_c$,

$$\Delta_{12} = E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\tilde{x}_t - \tilde{x}_{to})1(\Delta y > 0)T(y_t, x_t\beta).$$

Estimation when $T > 2$: For simplicity, we consider $T = 3$ and the "leapfrog" estimator. Extension to $T > 3$ and the "change" estimator follows trivially. The idea is to consistently estimate β with the "leapfrog" estimator in periods 1 and 2, yielding $\hat{\beta}_\ell^{12}$, and in periods 1 and 3, yielding $\hat{\beta}_\ell^{13}$. To improve efficiency, we can take a linear combination of $\hat{\beta}_\ell^{12}$ and $\hat{\beta}_\ell^{13}$.[9] For these two estimates,

$$\Delta_{12} = E(\Delta\tilde{x}^{12} - \Delta\tilde{x}_o^{12})'(\Delta\tilde{x}^{13} - \Delta\tilde{x}_o^{13})g_o^{12}(\Delta x^{12}\beta)g_o^{13}(\Delta x^{13}\beta)H^{12}(y_1, y_2, \Delta x^{12}\beta)H^{13}(y_1, y_3, \Delta x^{13}\beta),$$

where the superscripts indicate the time-period pair and $g_o^{st}(\cdot)$ and $H^{st}(\cdot)$ are defined analogously to Sections 1.3 and 1.4.

1.7 An Empirical Example

Panel data has been used extensively in the study of consumption patterns in macroeconomics and public finance. The standard approach is to assume some form for the utility function so that an Euler equation can be derived and estimated. Hall and Mishkin (1982), in their paper on the sensitivity of consumption to transitory income, were the first to use panel data at the micro level to study changes in consumption over time. Many papers since Hall and Mishkin (1982) have studied the effect of certain covariates on changes in consumption. Two examples are Hausman and Paquette (1987), which looks at the effect of involuntary unemployment, and Zeldes (1989), which looks at the effect of liquidity constraints.

Food consumption is the usual measure of consumption used in empirical work. For a utility-maximization model to have implications about food consumption, one needs to assume separability of the utility function. Following Hausman and Paquette (1987), we specify the utility function for individual i at time t as

$$U_{it}(C_{1it}, C_{2it}, L_{it}) = f_1(C_{1it})e^{W_{it}+bL_{it}} + f_2(C_{2it}, L_{it}), \tag{1.30}$$

where C_1 is food consumption, C_2 is non-food consumption, L is leisure, and the multiplicative term reflects the taste for food consumption relative to non-food consumption. So that the utility-maximization problem yields a closed-form solution, it is

[9] For $T = 3$, one can also use $\hat{\beta}_\ell^{23}$ in the combination. Note that this is different from the linear panel model, where one of the first-differencing estimates would be redundant (since $y_3 - y_2 = (y_3 - y_1) - (y_2 - y_1)$).

generally assumed that $f_1(\cdot)$ exhibits either constant absolute risk aversion (CARA),

$$f_1(C_1) = e^{-\rho C},$$

or constant relative risk aversion (CRRA),

$$f_1(C_1) = \frac{C^{1-\rho}}{1-\rho}.$$

The CARA formulation has an implication about the expected change in the level of consumption,

$$E(C_{t+1} - C_t) = r_{t+1} + x_{t+1}\beta, \tag{1.31}$$

where r reflects the effect of the interest rate and the x's are changes in covariates which affect consumption. The CRRA formulation, on the other hand, has an implication about the expected growth of consumption,

$$E(\ln(C_{t+1}) - \ln(C_t)) = r_{t+1} + x_{t+1}\beta. \tag{1.32}$$

Either (1.31) or (1.32) can be estimated using OLS within regression so that any fixed effects will be wiped out. In this section, we will estimate a consumption-change model using the linear transformation panel model. We try this approach for two reasons. The first reason is that different utility functions give different implications about how consumption changes, as can be seen from (1.31) and (1.32). Allowing for an unspecified transformation of C in the estimation offers flexibility so that there is less concern that the results are being driven by reliance on a specific functional form for the utility function.[10] The second reason is that consumption is well-known to be poorly reported (see Zeldes (1989) for a discussion and references). There is reason to expect that the rank-type estimators developed in this paper are more robust to such mismeasurement of the dependent variable than the least-squares estimator is. In Chapter 3 of this thesis, we find this to be the case in non-panel models; the panel analogue is currently being pursued in a separate paper.

We use our estimation techniques on data from a study by Gruber (1994) on the consumption smoothing benefits of unemployment insurance (UI). Using data from the Panel Study of Income Dynamics (PSID) and information on UI benefits across states and over time, Gruber finds that the fall in food consumption for the unemployed is three times smaller than it would be in the absence of UI. The equation estimated (for individuals becoming unemployed at time t) is

$$\ln(C_{i,t+1}) - \ln(C_{i,t}) = x_i\beta + \gamma UI_i + \epsilon_i, \tag{1.33}$$

[10]Even though the "change" and "leapfrog" estimators offer more flexibility, it's not clear what the implications for the utility function are if the estimates differ from the within estimates. The problem is that it's quite difficult to find closed-form solutions for the utlity-maximization problem, so even a parametrization of $h(C)$ doesn't necessarily translate into something readily interpretable about $f_1(C)$.

where C is food consumption, UI is the "replacement rate" (ratio of benefits to past wages) for which an individual is eligible, and x is a vector of individual characteristics. The parameter of interest here is γ since positive γ is evidence of the consumption smoothing effect of UI eligibility. We relax the formulation in (1.33) by substituting an unknown h for the log function:

$$h(C_{i,t+1}) - h(C_{i,t}) = x_i\beta + \gamma UI_i + \epsilon_i. \tag{1.34}$$

The sample, drawn from between 1968 and 1987, consists of 1,605 household heads who become unemployed. The UI variable is constructed from the wage data and state-specific information using a simulation program described in Gruber (1994). The covariates used in x include dummies for race, sex, and marital status in addition to age, number of kids, wage, education, and the growth in "food needs." Even though many of these variables are time-invariant, we include them in the analysis to compare with the results from Gruber (1994) and to allow for different impacts upon consumption across different demographic characteristics. "Food needs," as constructed in Zeldes (1989), is a weighted average of food costs for the family (as estimated by the Department of Agriculture for different age categories), adjusted for the size of the family (due to economies of scale).

Table 1.2: Estimation Results for Consumption Equation (normalized)

	(1) OLS (logs)	(2) MRC (logs)	(3) $\hat{\beta}_c$	(4) $\hat{\beta}_\ell$	(5) MSE
White	0.4943	0.3813	0.2968	0.3636	1.0688
	(0.2867)	(0.3372)	(0.4162)	(0.3884)	
Black	0.3655	0.3203	0.3488	0.2972	1.3096
	(0.2635)	(0.3400)	(0.4072)	(0.3707)	
Female	0.3348	0.2111	0.1512	0.1356	0.6739
	(0.1924)	(0.1813)	(0.1385)	(0.1493)	
Married	0.1111	-0.0393	-0.0768	-0.0941	0.1301
	(0.1504)	(0.0385)	(0.1127)	(0.0973)	
Age	-0.0060	-0.0134	-0.0169	-0.0104	-0.0193
	(0.0052)	(0.0080)	(0.0129)	(0.0103)	
# Kids	0.0749	0.0937	0.0554	0.1068	0.2259
	(0.0488)	(0.0697)	(0.0746)	(0.1310)	
After-Tax Wage (x 10^3)	*-0.0007*	*-0.0018*	*-0.0017*	*-0.0022*	*-0.0011*
	(0.0015)	(0.0031)	(0.0039)	(0.0048)	
Education	0.0249	0.0162	0.0188	0.0144	0.0405
	(0.0206)	(0.0162)	(0.0346)	(0.0252)	
UI Replacement Rate	1.2543	1.4105	1.3116	1.5172	4.2989
	(0.5799)	(0.9229)	(0.9722)	(0.8786)	
Growth in Food Needs	1.0000	1.0000	1.0000	1.0000	1.0000
	——	——	——	——	——
# obs	1605	1605	1605	1605	1605
χ^2 test vs. OLS (d.f.=12)	——	8.26	19.70	15.42	——
(p-value)		(0.765)	(0.073)	(0.219)	

The estimation results using the within estimator and semiparametric alternatives are reported in Table 1.2.[11] So that we may compare the results, we have normalized all of the coefficients by dividing by the coefficient on the growth in food needs. This variable is a natural choice since we can see the effect of consumption on the other covariates relative to actual consumption needs. Time variables (entering linearly, quadratically, and cubically) are included for the each of the estimations to capture any discount-rate effects. A constant term is also included for the least squares and maximum score estimators. These coefficients are not reported in the interest of saving space.

The first two columns estimate the model in (1.33), restricting $h(\cdot)$ to be $\ln(\cdot)$, by least squares and maximum rank correlation, respectively. Column (1) simply replicates the results from Gruber (1994). Under the assumption that ϵ is i.i.d., the χ^2 test between OLS and MRC does not reject that ϵ is distributed normally (in which case OLS is efficient). The final three columns of Table 1.2 estimate the more flexible model of (1.34). The maximum score estimates are reported without standard errors. Under the assumption of i.i.d. normal errors, a Hausman (1978) specification test of OLS versus an alternative is appropriate. The test against the "change" estimator of column (3) rejects with a p-value of 7.3%, while the test against the "leapfrog" estimator has a p-value of 21.9%. Even with this moderate size data set, there is some evidence of misspecification of the change-in-logs model. The bottom line in terms of the effect of unemployment insurance, however, remains unchanged. The coefficient on the replacement rate is positive for all of the estimators and significant at the 10% level for the "leapfrog" estimator. The "leapfrog" estimate is about 20% higher than the OLS estimate.

Our semiparametric estimates suggest a greater relative consumption-smoothing effect of UI than the within estimates. Gruber (1994) finds that a difference of 30–40% in the consumption-smoothing effect of UI can have large implications for the optimal replacement rate.

[11]The covariance matrices for columns (2)–(4) were computed using kernel estimation of the appropriate primitives. A window width of $\hat{\sigma}n^{-1/5}$ was used, where $\hat{\sigma}$ is the empirical standard error of the estimated index. The numerical derivatives needed to estimate the V's were somewhat sensitive to the window choice for the gradient, but bootstrapping the derivatives resulted in more stable estimates.

Table 1.3: Optimal Replacement Rates

ρ	$\hat{\gamma}$ from Gruber (1994)	$\hat{\gamma}/0.7$	$0.7\times\hat{\gamma}$
1	0	0	0
1.5	0	0.091	0
2	0.035	0.286	0
2.5	0.202	0.403	0
3	0.314	0.481	0.076
3.5	0.394	0.537	0.190
4	0.453	0.579	0.275

Source: Gruber (1994), Table 6.

In Table 1.3, we replicate a table from Gruber (1994) that gives optimal replacement rate calculations for different values of the relative risk aversion parameter. The first column of replacement rates are imputed from Gruber's estimate γ of the consumption-smoothing effect. The next two columns show the optimal replacement rates imputed from $\gamma/0.7$ (larger effect) and $0.7 \times \gamma$ (smaller effect). Since our estimate is only relative to the food-needs coefficient, our estimates are only suggestive of the implications for the optimal replacement rate calculation.

1.8 Conclusions and Extensions

The estimators proposed in this paper represent a first step in the semiparametric approach to the linear transformation panel model. The most obvious open question is, given estimates of β, how can we nonparametrically estimate the transformation function h? One possible approach is to use series estimation of h via polynomial and spline approximations. It would be preferable to have a more efficient estimation strategy as is the case for the non-panel case, where Horowitz (1996) and Ye and Duan (1995) obtain $\sqrt{n}$-consistent estimates of h.

Another possible extension of this work is to allow more flexibility with respect to the independent variables. There is a bit of a dichotomy in the model since the dependent variable is subject to a completely unspecified transformation whereas the functional forms of the independent variables are assumed to be known. One way of dealing with this dichotomy is to allow for parametrized transformations of the independent variables as well. For instance, we could subject each x to a Box-Cox transformation, where the transformation parameters would be estimated along with β. Our conjecture is that the "change" and "leapfrog" estimators can be modified in a natural way to accommodate this flexibility. As a simple example, consider a panel model having only a single independent variable subject to a power transformation,

$$h(y_{it}) = x_{it}^{\lambda} + \gamma_t + \alpha_i + \epsilon_{it}. \tag{1.35}$$

The suitable objective functions for "change" and "leapfrog" estimation of the transformation parameter λ would be

$$\sum_{i \neq j} 1(x_{i2}^{\ell} - x_{i1}^{\ell} > x_{j2}^{\ell} - x_{j1}^{\ell}) 1(\Delta y_i > 0)$$

and

$$\sum_{i \neq j} 1(x_{i2}^{\ell} - x_{i1}^{\ell} > x_{j2}^{\ell} - x_{j1}^{\ell}) 1(y_{i1} < y_{j1}, y_{i2} > y_{j2}).$$

Allowing for multiple transformed independent variables would result in similar objective functions. The normalization of the parameter vector (β, λ) needed for identification would apply to β, the coefficients on the additive terms of the right-hand side of the model.[12]

[12]This same idea could be used to extend the non-panel MRC estimator, an extension which has

Finally, the issue of efficiency has been mostly ignored in this paper. Since efficiency is being sacrificed for flexibility, it's important to try and re-weight estimators to "squeeze" as much efficiency out of them as possible. For "change" estimation, Cavanagh and Sherman (1993) have shown that a one-step correction of the MRC estimator achieves the semiparametric efficiency bound. Perhaps there is an analogous result for "leapfrog" estimation. Even a naive re-weighting strategy, though, would improve matters. The intuition for re-weighting is that observation-pairs (i, j) with larger $|\Delta x_i\beta - \Delta x_j\beta|$ should be weighted more heavily in the objective function since the direction of leapfrogging for these pairs should be more predictable than for pairs with lower $|\Delta x_i\beta - \Delta x_j\beta|$. A Monte Carlo investigation of the efficiency gains from using such re-weighting would be interesting.

Appendix A: Computational Issues

Estimation of $\hat{\beta}_c$ or $\hat{\beta}_\ell$ requires maximization of a non-smooth objective function, making standard optimization techniques ineffective. Instead, we suggest the use of the Nelder-Mead simplex method since it avoids gradient techniques.[13] The Nelder-Mead method will find local optima on occasion, so one needs to iterate the method several times to check that the optimum is global. This problem is more prevalent when the parameter space being searched has higher dimension.

Although the objective functions $S_c(\cdot)$ and $S_\ell(\cdot)$ have a similar form, their computational characteristics differ for large n. From (1.11), computing $S_c(b)$ requires ranking (sorting) the observations by $\Delta x\beta$ and taking a dot product. The fastest sorting algorithms require $O(n \log n)$ operations, so computation of $S_c(b)$ also requires $O(n \log n)$ operations. Unfortunately, no sorting technique can be used to compute $S_\ell(b)$, meaning that computation of $S_\ell(b)$ requires $O(n^2)$ operations. To speed up computation of $S_\ell(b)$, store an array of pairs (i, j) for which i leapfrogs j. Each time the objective function is calculated, simply look up the value of $1(y_{i1} < y_{j1}, y_{i2} > y_{j2})$ rather than re-checking for leapfrogging among the $n(n-1)/2$ observation-pairs.

Kernel methods are not necessary for finding $\hat{\beta}_c$ or $\hat{\beta}_\ell$, but they need to be used to obtain consistent estimates for any of covariance matrices from Sections 1.3–1.6. For instance, Δ_c, V_c, Δ_ℓ, and V_ℓ are consistently estimated by

$$\widehat{\Delta}_c = \frac{1}{n}\sum_{i=1}^{n}(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})'(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})\widehat{g_o}(\Delta x_i\hat{\beta}_c)^2\widehat{F}(\Delta x_i\hat{\beta}_c)[1 - \widehat{F}(\Delta x_i\hat{\beta}_c)], \quad (1.36)$$

$$\widehat{V}_c = -\frac{1}{2n}\sum_{i=1}^{n}(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})'(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})\widehat{F}'(\Delta x_i\hat{\beta}_c)\widehat{g_o}(\Delta x_i\hat{\beta}_c), \quad (1.37)$$

$$\widehat{\Delta}_\ell = \frac{1}{n}\sum_{i=1}^{n}(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})'(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})\widehat{g_o}(\Delta x_i\hat{\beta}_\ell)^2\widehat{H}(y_{i1}, y_{i2}, \Delta x_i\hat{\beta}_\ell)^2, \quad (1.38)$$

---

not been considered in the literature.

[13]MATLAB has a built-in function for the simplex method. Numerical Recipes also has a program for implementing the simplex method. A GAUSS version, along with code for the objective functions, is available from the author upon request.

$$\widehat{V}_\ell = \frac{1}{2n}\sum_{i=1}^{n}(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})'(\Delta\tilde{x}_i - \widehat{\Delta\tilde{x}_{io}})\widehat{H}_3(y_{i1}, y_{i2}, \Delta x_i\hat{\beta}_\ell)\widehat{g_o}(\Delta x_i\hat{\beta}_\ell), \tag{1.39}$$

where $\hat{\beta}_c$ and $\hat{\beta}_\ell$ are the "change" and "leapfrog" estimates, respectively, and $\widehat{\Delta\tilde{x}_o}$, $\widehat{g_o}(\cdot)$, $\widehat{F}(\cdot)$, and $\widehat{H}(\cdot)$ denote kernel estimates of their underlying primitives.

An alternative to the kernel approach is to employ bootstrapping to estimate V_c, Δ_c, V_ℓ, and Δ_ℓ. This alternative is time-intensive, especially when used in conjunction with "leapfrog" estimation, but computers have become fast enough to make bootstrap estimation feasible.

Appendix B: Proofs of Theorems

Proof of Theorem 1: See Theorem 4 of Sherman (1993).

Proof of Theorem 2: See Section 6 of Sherman (1993).

Proof of Theorem 3: Since $\tau_\ell(\cdot,\cdot)$ satisfies the same technical assumptions as $\tau_c(\cdot,\cdot)$, the proof is nearly identical to Theorem 4 of Sherman (1993). Here, we consider the class of functions

$$\mathcal{F} = \{f(\cdot,\cdot,b) : b \in \mathcal{B}\},$$

where, for each $(z_1, z_2) \in \mathcal{S} \times \mathcal{S}$ (for $z_i \equiv (y_{i1}, y_{i2}, \Delta x_i)$) and $b \in \mathcal{B}$,

$$f(z_1, z_2, b) = 1(y_{12} > y_{22}, y_{11} < y_{21}) \cdot 1(\Delta x_1 b > \Delta x_2 b).$$

To apply Theorem 4 of Sherman (1993), we need only check that $\mathcal{F}$ is Euclidean for the constant envelope 1. The details are similar to Section 5 of Sherman (1993).

Consider $t, \gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22} \in \mathcal{R}$ and $\delta_1, \delta_2 \in \mathcal{R}^k$. For each $(z_1, z_2) \in \mathcal{S} \times \mathcal{S}$, define

$$g(z_1, z_2, t; \gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \delta_1, \delta_2) = \gamma t + \gamma_{11}y_{11} + \gamma_{12}y_{12} + \gamma_{21}y_{21} + \gamma_{22}y_{22} + \Delta x_1\delta_1 + \Delta x_2\delta_2$$

and

$$\mathcal{G} = \{g(\cdot,\cdot,\cdot : \gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22}, \delta_1, \delta_2) : \gamma, \gamma_{11}, \gamma_{12}, \gamma_{21}, \gamma_{22} \in \mathcal{R}, \delta_1, \delta_2 \in \mathcal{R}^k\}.$$

Then, $\mathcal{G}$ is a $(2k+5)$-dimensional vector space of real-valued functions on $\mathcal{S} \times \mathcal{S} \times \mathcal{R}$.

Consider the set of subgraphs of functions belonging to $\mathcal{F}$. For each $b \in \mathcal{B}$,

$$\begin{aligned}
\text{subgraph}(f(\cdot,\cdot,b)) &= \{(z_1, z_2, t) : 0 < t < f(z_1, z_2, b)\} \\
&= \{y_{12} - y_{22} > 0\}\{y_{11} - y_{21} < 0\}\{\Delta x_1 b - \Delta x_2 b > 0\}\{t \geq 1\}^c\{t > 0\} \\
&= \{g_1 > 0\}\{g_2 > 0\}\{g_3 > 0\}\{g_4 \geq 1\}^c\{g_5 > 0\}
\end{aligned}$$

where $g_1, g_2, g_3, g_4, g_5 \in \mathcal{G}$. By Lemma 2.4 of Pakes and Pollard (1989), we know that the class of sets of the form $\{g \geq c\}$ and $\{g > c\}$, with $g \in \mathcal{G}$ and $c \in \mathcal{R}$, is a VC class (or a "polynomial class"). So, the subgraph of $f(\cdot,\cdot,b)$ is the intersection of five sets, four belonging to a VC class and the fifth being the complement of a set belonging to a VC class. Then, $\{\text{subgraph}(f) : f \in \mathcal{F}\}$ is a VC class of sets (Pakes and Pollard (1989), Lemma 2.5), which implies $\mathcal{F}$ is Euclidean for every envelope (Pakes and Pollard (1989), Lemma 2.12).

Proof of Theorem 4: The results of Section 6 of Sherman (1993) apply. Note that

$$\tau_\ell(z,b) = \int_{\Delta Xb<\Delta xb} H(y_1,y_2,\Delta Xb)G(d\Delta X) + \int \rho(y_1,y_2,\Delta X\beta)G(d\Delta X),$$

where $G(\cdot)$ is the distribution of Δx and $\rho(y_1,y_2,v) = Pr_Z[y_1 > Y_1, y_2 < Y_2|\Delta X\beta = v]$. The second integral does not depend on b, so $\nabla_1\tau_\ell(z,\beta) = (\Delta\tilde{x} - \Delta\tilde{x}_o)g_o(\Delta x\beta)H(y_1,y_2,\Delta x\beta)$. Taking an outer product and taking expectations gives Δ_ℓ. Also,

$$\begin{aligned}
2V_\ell &= E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)\left.\frac{\partial}{\partial v}[H(y_1,y_2,v)g_o(v)]\right|_{v=\Delta x\beta} \\
&= E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)[H_3(y_1,y_2,\Delta x\beta)g_o(\Delta x\beta) + H(y_1,y_2,\Delta x\beta)g_o'(\Delta x\beta)] \\
&= E(\Delta\tilde{x} - \Delta\tilde{x}_o)'(\Delta\tilde{x} - \Delta\tilde{x}_o)H_3(y_1,y_2,\Delta x\beta)g_o(\Delta x\beta).
\end{aligned}$$

Proof of Theorem 5: Let

$$\Gamma_n\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} \equiv \Gamma_{1n}(\theta_1) + \Gamma_{2n}(\theta_2).$$

By definition, $\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}$ maximizes $\Gamma_n(\cdot)$. Also, $\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$ maximizes the sample analogue of $\Gamma_n(\cdot)$ since expectation is a linear operator.

Uniformly over $O_p(1/\sqrt{n})$ neighborhoods of $\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}$, we have

$$\begin{aligned}
\Gamma_n\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} &= \frac{1}{2}\theta_1'V_1\theta_1 + \frac{1}{\sqrt{n}}\theta_1'W_{1n} + \frac{1}{2}\theta_2'V_2\theta_2 + \frac{1}{\sqrt{n}}\theta_2'W_{2n} + o_p(1/n) \\
&= \frac{1}{2}\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}'\begin{pmatrix}V_1 & 0\\ 0 & V_2\end{pmatrix}\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix} + \frac{1}{\sqrt{n}}\begin{pmatrix}\theta_1\\ \theta_2\end{pmatrix}'\begin{pmatrix}W_{1n}\\ W_{2n}\end{pmatrix} + o_p(1/n).
\end{aligned}$$

Since $E\nabla_1\tau_1(\cdot,\mathbf{0}) = E\nabla_1\tau_2(\cdot,\mathbf{0}) = \mathbf{0}$ and $\tau_1(z,\cdot)$ and $\tau_2(z,\cdot)$ are Taylor-expandable, we have

$$\begin{pmatrix}W_{1n}\\ W_{2n}\end{pmatrix} \xrightarrow{d} N\left(0,\begin{pmatrix}\Delta_1 & \Delta_{12}\\ \Delta_{12}' & \Delta_2\end{pmatrix}\right).$$

Also, $\begin{pmatrix}V_1 & 0\\ 0 & V_2\end{pmatrix}$ is negative definite since V_1 and V_2 are. Applying Theorem 2 of Sherman (1993) and multiplying out gives the desired result.

Proof of Theorem 6: Similar to proof of Theorem 4.

References

ABREVAYA, J. (1996): "Maximum Likelihood Estimation of the Box-Cox Panel Model," M.I.T. Department of Economics, mimeo.

AMEMIYA, T. AND J. POWELL (1981): "A Comparison of the Box-Cox Maximum Likelihood Estimator and the Nonlinear Two-Stage Least-Squares Esimator," *Journal of Econometrics* 17: 351–381.

BICKEL, P. J. AND K. A. DOKSUM (1981): "An Analysis of Transformations Revisited," *Journal of the American Statistical Association* 76: 296–311.

CAVANAGH, C. AND R. P. SHERMAN (1992): "Rank Estimators for Monotone Index Models," Bellcore Economics Discussion Paper no. 84.

CHAMBERLAIN, G. (1985): "Heterogeneity, Omitted Variable Bias, and Duration Dependence," in *Longitudinal Analysis of Labor Market Data*, eds. J. J. Heckman and B. Singer, New York: Cambridge University Press, pp. 3–38.

GÖNÜL, F. AND K. SRINIVISAN (1993): "Consumer Purchase Behavior in a Frequently Bought Product Category: Estimation Issues and Managerial Insights from a Hazard Function Model with Heterogeneity," *Journal of the American Statistical Association* 88: 1219–1227.

GRUBER, J. (1994): "The Consumption Smoothing Benefits of Unemployment Insurance," NBER Working Paper Series, no. 4750.

HALL, R. H. AND F. S. MISHKIN (1982): "The Sensitivity of Consumption to Transitory Income: Estimates from Panel Data on Households," *Econometrica* 50: 461–481.

HAN, A. K. (1987a): "A Non-Parametric Analysis of Transformations," *Journal of Econometrics* 35: 191–209.

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

HAUSMAN, J. A. (1978): "Specification Tests in Econometrics," *Econometrica* 46: 1251–1271.

HAUSMAN, J. A. AND L. PAQUETTE (1987): "Involuntary Early Retirement and Consumption," in *Work, Health, and Income Among the Elderly*, ed. by G. Burtless, Washington, D. C.: Brookings Institution.

HOROWITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60: 505–531.

HOROWITZ, J. L. (1996): "Semiparametric Estimation of a Regression Model with an Unknown Transformation of the Dependent Variable," *Econometrica* 64: 103–137.

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

KLEIN, R. W. AND R. H. SPADY (1993): "An Efficient Semiparametric Estimator for Binary Response Models," *Econometrica* 61: 387–421.

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3: 205–228.

MANSKI, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica* 55: 357–362.

OLSEN, R. J. AND K. I. WOLPIN (1983): "The Impact of Exogenous Child Mortality on Fertility: A Waiting Time Regression with Dynamic Regressors," *Econometrica* 51: 731–751.

PAKES, A. AND D. POLLARD (1989): "Simulation and the Asymptotics of Optimization Estimators," *Econometrica* 57: 1027–1057.

POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989): "Semiparametric Esimation of Index Coefficients," *Econometrica* 57: 474–523.

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

YE, J. AND N. DUAN (1995): "Nonparametric $n^{-1/2}$-Consistent Estimation for the General Transformation Model," University of Chicago GSB, mimeo.

ZELDES, S. P. (1989): "Consumption and Liquidity Constraints: An Empirical Investigation," *Journal of Political Economy* 97: 305–346.

Chapter 2

Maximum Score Estimation of Linear Index Panel Models

2.1 Introduction

Much has been written about the difficulties in consistently estimating the parameters of nonlinear fixed-effects panel data models. The standard first-differencing trick which eliminates the fixed effect from a linear model extends to only certain nonlinear models, including the conditional logit model for binary data, the Poisson model for count data, and certain parametric models for duration data. Each of these models share an exponential form which allows for cancellation of the fixed effect akin to first differencing in the linear panel model. Semiparametric methods, which do not require any parametric assumptions on the error term, exist for consistent estimation of the binary choice model (Manski (1987)) and the linear censored and truncated models (Honoré (1992)). The basic insight of this paper is that Manski's maximum score estimator can be used to consistently estimate (up to scale) the parameters of interest in a wide class of nonlinear panel data models.

We consider a general linear index panel model of the following form:

$$y_t^* = g(x_t\beta, \boldsymbol{\alpha}) + \epsilon_t, \tag{2.1}$$

$$y_t = d(y_t^*) \tag{2.2}$$

where x_t is a k-dimensional vector of explanatory variables, $\boldsymbol{\alpha}$ is a vector of fixed effects, g is unknown and strictly increasing in the index (i.e., $g_1 > 0\ \forall\boldsymbol{\alpha}$), and $d : \mathcal{R} \to \mathcal{R}$ is a known (nondegenerate) weakly increasing function. Without loss of generality, we consider the case of two time periods ($t = 1, 2$).[1]

The linear model, binary and ordered choice models, proportional hazard model, and censored model (linear or nonlinear) are all special cases of this general model.

[1]Extension to $T > 2$ is straightforward. Maximum score estimation can be done for each possible pair of time periods. The pairwise estimates can then be combined to form a more efficient estimate. Unbalanced panels present no additional difficulties since only pairwise estimates are used. See Charlier et. al. (1995) for more details.

The model doesn't require that the fixed effects enter additively, nor does it require that y_t^* be a linear function of $x_t\beta$. For instance, maximum score estimation will still be consistent even when the true model has complicated interactions with the fixed effects, as in

$$y_t^* = \alpha_1(x_t\beta)^{\alpha_2} + \alpha_3 + \epsilon_t,$$

where $\alpha_1 > 0$ and $\alpha_2 > 0$ (to ensure $g_1 > 0$). Not surprisingly, efficiency must be sacrificed in order to gain this generality.

The transformation d describes the form in which the dependent variable is observed. When d is a strictly increasing function, the data are *uncensored* — that is, we directly observe the latent variable y_t^* (or, equivalently for our purposes, some strictly increasing transformation of the y_t^*). When y_t^* is not always directly observable, we say the data are *censored.* In this terminology, the binary choice model is a censored model having $d(v) = 1(v > 0)$. The traditional censored model has $d(v) = v \cdot 1(v > 0)$.

We say that the data are *truncated* when y_1 and y_2 are observed if and only if $y_1^*, y_2^* \in \mathcal{T} \subset \mathcal{R}$. The set $(\mathcal{R} - \mathcal{T})$ is the region where truncation occurs. In the traditional truncated model (as studied by Honoré (1992)), we have $\mathcal{T} = [0, \infty)$ and $(\mathcal{R} - \mathcal{T}) = (-\infty, 0)$. Our classification of truncated data allows for more general forms of truncation.[2]

In Section 2.2, we define the maximum score estimator. In Section 2.3, we show that the maximum score estimator applied to our model is consistent for uncensored data, censored data, and truncated data. The three cases require slightly different assumptions on the distribution of $\Delta\epsilon$. Uncensored data require only a median restriction. Censored and truncated data require a symmetry restriction, with truncated data also requiring a unimodality restriction. In Section 2.4, we extend the smoothing technique of Horowitz (1992), allowing for a smoothed version of the maximum score estimator having convergence rate between $n^{-2/5}$ and $n^{-1/2}$.

2.2 The Maximum Score Estimator

The estimator is a form of the maximum score estimator (MSE) developed by Manski (1975, 1987) for the binary choice model in non-panel and panel settings. The objective function is

$$S_n(b) \equiv \frac{1}{n}\sum_{i=1}^{n} \text{sgn}(\Delta y) 1(\Delta x b > 0), \tag{2.3}$$

where n is the number of observational units and the sign function is given by

$$\text{sgn}(v) = 1(v > 0) - 1(v < 0).$$

The only information from the dependent variables which is used is whether the observable dependent variable has gone up or gone down. Observations for which

[2]We require that the form of censoring and/or truncation remains the same through time. This restriction is not substantive, though, since the econometrician can censor and/or truncate the data herself so that the function $d(\cdot)$ and/or the region $\mathcal{T}$ does not change over time.

there has been no change in the observable dependent variable have $\text{sgn}(\Delta y) = 0$ and do not contribute to the objective function.

The MSE is

$$\hat{\beta} \equiv \underset{b:|b_1|=1}{\arg\max}\ S_n(b).$$

The coefficient vector is normalized since scale does not affect the sign of the index (i.e., $S_n(b) = S_n(kb)$ for $k > 0$). When there is any form of censoring, the MSE is a conditional estimator since only observational units having non-zero Δy contribute to the objective function. The correction for this conditioning is discussed below in Section 2.4. When there is no censoring, all the observational units are used in the estimation since $\{\Delta y = 0\}$ is a zero-probability event. The covariates Δx should include a constant term, which corresponds to a time trend (i.e., the difference between the constant term at $t = 2$ and the constant term at $t = 1$).

2.3 Consistency of the MSE

The assumptions needed for strong consistency are basically those used by Manski (1987), except for the slightly stronger restrictions needed on the error distribution.

The following sampling assumption covers both the non-truncated case ($\mathcal{T} = \mathcal{R}$) and the truncated case ($\mathcal{T} \subset \mathcal{R}$).

Assumption 1 (Sampling) *An i.i.d. sample of n observational units is generated according to (1) conditional on $y_1^*, y_2^* \in \mathcal{T} \subseteq \mathcal{R}$. The observables are $\{(y_{it}, x_{it}) : i = 1, \ldots, n; t = 1, 2\}$.*

To guarantee identification of the parameter vector, we need the following assumption on the regressors.

Assumption 2 (Continuous Regressors with Full Rank) *(a) The support of Δx is not contained in any proper linear subspace of $\mathcal{R}^k$; (b) $\beta^1 \neq 0$ and for almost every $\Delta\tilde{x} \equiv (\Delta x^2, \ldots, \Delta x^k)$, the distribution of Δx^1 conditional on $\Delta\tilde{x}$ has everywhere positive density with respect to Lebesgue measure.*

Assumption 2(a) is the usual full-rank condition. Assumption 2(b), standard in semiparametric estimation, ensures that $\Delta x b$ has everywhere positive density when $b^1 \neq 0$.

Our normalization convention ($|\beta^1| = 1$) requires a compactness assumption.[3]

Assumption 3 (Compact Parameter Space) $|\beta^1| = 1$, *and $\tilde{\beta} \equiv (\beta^2, \ldots, \beta^k)'$ is contained in a compact subset $\tilde{B}$ of $\mathcal{R}^{k-1}$.*

[3]The normalization made by Manski (1987) is $\|\beta\| = 1$, which directly implies that β lies in a compact subset of $\mathcal{R}^k$.

To ensure that $\Pr[\Delta y > 0|x_1, x_2, \boldsymbol{\alpha}]$ and $\Pr[\Delta y < 0|x_1, x_2, \boldsymbol{\alpha}]$ are both positive, we assume that the errors have infinite support.

Assumption 4 (Continuous Errors) *Conditional on $(x_1, x_2, \boldsymbol{\alpha})$, ϵ_1 and ϵ_2 are continuously distributed on $\mathcal{R}$.*

The following alternative assumptions on $\Delta\epsilon$ correspond, respectively, to uncensored data, censored data, and truncated data. Assumptions 5.2 and 5.3 are the same assumptions as those made by Honoré (1992) for the linear censored and truncated panel models, respectively.

Assumption 5.1 (Zero Median) *The distribution of $\Delta\epsilon$ conditional on $(x_1, x_2, \boldsymbol{\alpha})$ has median zero.*

Assumption 5.2 (Symmetry) *The distribution of $\Delta\epsilon$ conditional on $(x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2)$ is symmetric around zero.*

Assumption 5.3 (Symmetry and Unimodality) *The distribution of $\Delta\epsilon$ conditional on $(x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2)$ is symmetric around zero and strictly unimodal.*

Similar to Powell (1986), we say that a continuous distribution is *strictly unimodal* if it achieves a unique maximum (which must be at zero to also satisfy symmetry) and is monotone on either side of the maximum.

The following lemma (Lemma 1 from Honoré (1992)) gives sufficient conditions for Assumptions 5.2 and 5.3 to hold.

Lemma 1 *If, conditional on $(x_1, x_2, \boldsymbol{\alpha})$, ϵ_1 and ϵ_2 are independent, and identically and continuously distributed, then Assumption 5.2 holds. If the (common) marginal density of ϵ_1 and ϵ_2 is also strictly log-concave, then Assumption 5.3 holds.*

The conditions of Lemma 1 allow for heteroskedasticity across observational units, but require homoskedasticity over time for each observational unit. As Honoré (1992) points out, some form of serial dependence of the error terms can be captured in an additive fixed effect (e.g., if ϵ_1 and ϵ_2 are jointly normal with equal variance and arbitrary positive correlation). Also, many of the usual distributions used in economics, including the normal and logistic distributions, satisfy the log-concavity sufficiency condition for Assumption 5.3.[4]

The key condition driving consistency of the MSE is

$$\begin{aligned} \Delta x\beta > 0 &\iff \Pr(\Delta y > 0|x_1, x_2) > \Pr(\Delta y < 0|x_1, x_2) \\ \Delta x\beta = 0 &\iff \Pr(\Delta y > 0|x_1, x_2) = \Pr(\Delta y < 0|x_1, x_2) \\ \Delta x\beta < 0 &\iff \Pr(\Delta y > 0|x_1, x_2) < \Pr(\Delta y < 0|x_1, x_2), \end{aligned} \tag{2.4}$$

[4] The uniform distribution is also log-concave, but doesn't satisfy the continuity of Assumption 4.

which can be written more concisely as

$$\text{sgn}(\text{E}[\text{sgn}(\Delta y)|x_1, x_2]) = \text{sgn}(\Delta x\beta)$$

or

$$\text{Med}[\text{sgn}(\Delta y)|x_1, x_2, \Delta y \neq 0] = \text{sgn}(\Delta x\beta).$$

The following lemma formally states that the aforementioned assumptions are sufficient for (2.4) to hold.

Lemma 2 *Assume Assumptions 1-4 hold. Then, (2.4) holds when either (a) the data are uncensored and Assumption 5.1 holds, (b) the data are censored and Assumption 5.2 holds, or (c) the data are truncated (and possibly censored) and Assumption 5.3 holds.*

The proof of Lemma 2 is given in the Appendix. The uncensored case is trivial. The censored and truncated cases can be seen graphically. The idea is to consider any realization of $(x_1, x_2, \boldsymbol{\alpha})$ and $(\epsilon_1 + \epsilon_2)$. This realization is associated with some value of $y_1^* + y_2^*$. On a graph with y_1^* and y_2^* on the axes, $y_1^* + y_2^* = K$ is a line along which we can determine the relative likelihood of Δy being positive or negative.

We look first at the case of censored data. Figures 2.1 and 2.2 show the binary choice model and the traditional censored model, respectively. A particular $y_1^* + y_2^* = K$ line is shown in Figure 2.1 as well. For censored data, the region of positive Δy is simply a mirror image of the region of negative Δy through the 45°-line passing through the origin. The formula for the line through the origin is $\Delta\epsilon = g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha})$. The $\Delta\epsilon = 0$ locus is parallel to this line and will be below and to the right of it for positive $\Delta x\beta$ (above and to the left for negative $\Delta x\beta$).

Figure 2.3 shows, for positive and negative $\Delta x\beta$, the cross-section of a censored data graph along a $y_1^* + y_2^* = K$ line (e.g., the cross-section along the line pictured in Figure 2.1). Following Assumption 5.2 for the censored case, a symmetric distribution for $\Delta\epsilon$ is drawn. The figure illustrates that positive Δy is more likely than negative Δy when $\Delta x\beta$ is positive, and vice versa when $\Delta x\beta$ is negative. For positive $\Delta x\beta$, the area underneath the right tail (labeled "$\Delta y > 0$" in the figure) is larger than the area underneath the left tail (labeled "$\Delta y < 0$" in the figure). The opposite is true for negative $\Delta x\beta$.

The case of truncated data is similar. Figure 2.4 shows the traditional truncated model. Again, the positive Δy and negative Δy regions are mirror images of each other. Figure 2.5 shows the cross-section along a $y_1^* + y_2^* = K$ line. The dotted portion of the axes in Figure 2.5 shows the region where $\Delta\epsilon$ can not possibly lie (due to truncation). Notice that, unlike the censored case, the positive Δy and negative Δy regions need not extend indefinitely. This feature necessitates the additional assumption of unimodality for truncated data. Figure 2.6 shows a situation where (2.4) fails for a symmetric, but non-unimodal, $\Delta\epsilon$ distribution. In Figure 2.6, the area under the density in the $\Delta y < 0$ region is larger than the area under the density in the $\Delta y > 0$ region even though $\Delta x\beta$ is positive.

Figure 2.1: Binary Choice Model

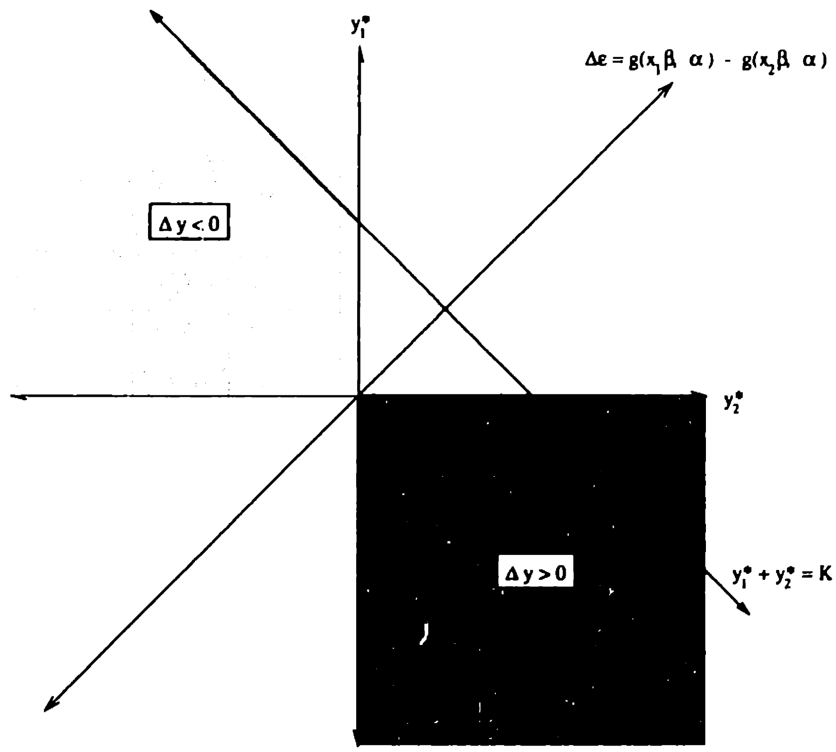


Figure 2.2: Censored Model

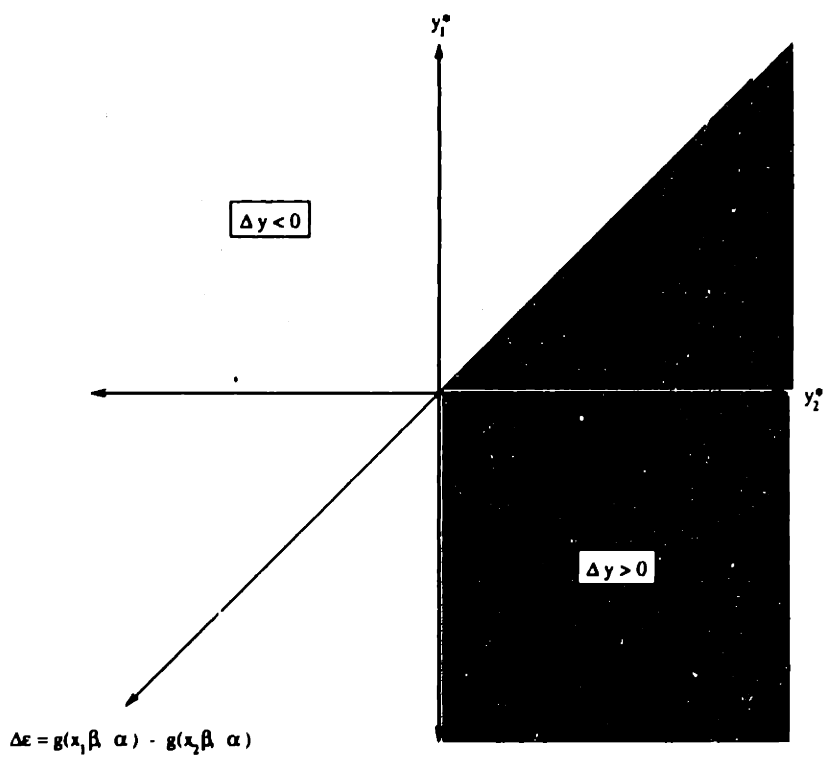


Figure 2.3: Cross-Section for Censored Data

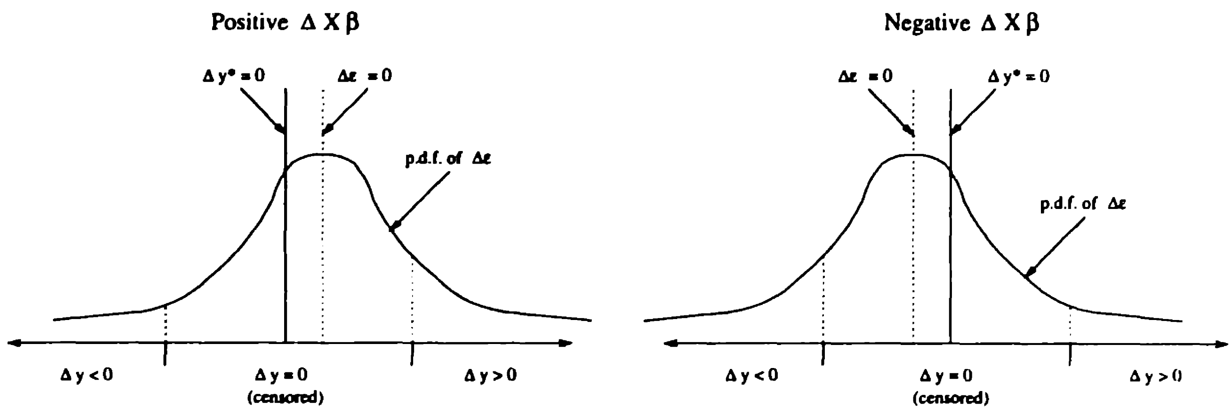


Figure 2.4: Truncated Model

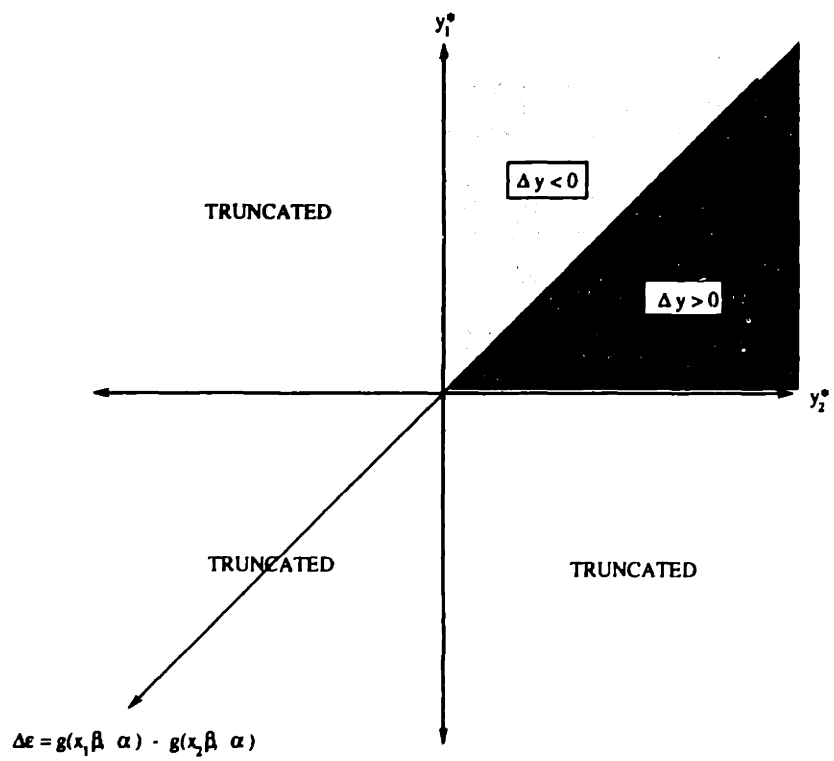


Figure 2.5: Cross-Section for Truncated Data

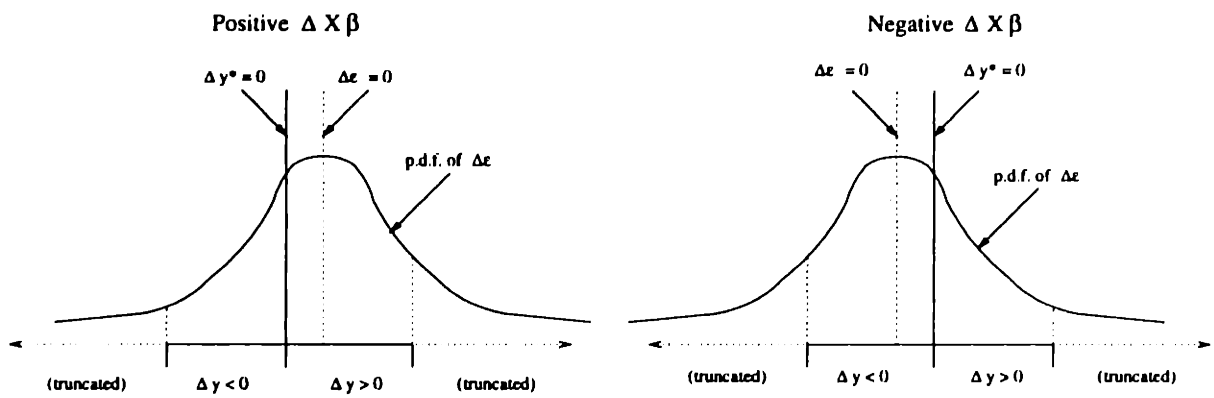
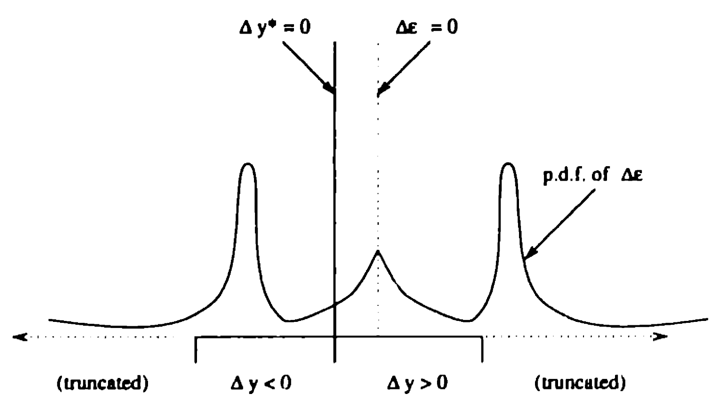
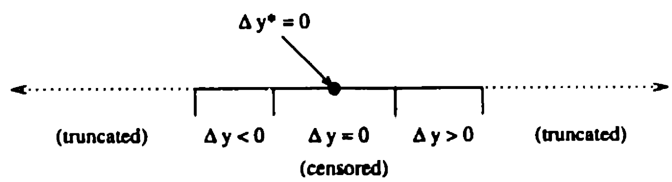


Figure 2.6: Symmetric, Non-Unimodal Distribution for Truncated Data



This graphical intuition still works when the data are both censored and truncated. Figure 2.7 shows a line representing a possible cross-section for such a model. Unimodality is still needed to avoid a situation like Figure 2.6.

Figure 2.7: Cross-Section for Censored, Truncated Data



Lemma 2 directly yields consistency.

Theorem 1 (Consistency of the MSE) *If Assumptions 1–4 and the appropriate error assumption (Assumption 5.1, 5.2, or 5.3 as in Lemma 2) hold, and*

$$\hat{\beta}_n \equiv \underset{b:|b_1|=1,\, \tilde{b}\in\tilde{B}}{\arg\max}\ S_n(b),$$

where $\tilde{b} \equiv (b^2, \ldots, b^k)'$*, then* $\lim_{n\to\infty} \hat{\beta}_n = \beta$ *almost surely.*

Finally, we note the difference between the assumptions on the error distribution made here and the stationarity assumption of Manski (1987). Stationarity of the errors across time for a given individual implies

$$\text{sgn}(\text{E}[\Delta y|x_1, x_2]) = \text{sgn}(\Delta x\beta).$$

In the binary choice model, this condition is equivalent to the one used here since

$$\Delta y = \text{sgn}(\Delta y) \Longrightarrow \text{Med}[\text{sgn}(\Delta y)|x_1, x_2] = \text{sgn}(\text{E}[\Delta y|x_1, x_2]).$$

In general, though, the conditions need not be equivalent.

2.4 Smoothed MSE and Asymptotic Normality

The MSE's objective function is a step function and difficult to analyze using traditional asymptotic methods. Kim and Pollard (1990) show that the binary choice MSE converges at a rate of $n^{-1/3}$ to a random variable that maximizes a certain Gaussian process. Unfortunately, the properties of the limiting distribution are largely unknown, making inference impossible. Horowitz (1992) has developed a smoothed (differentiable) version of the MSE for the non-panel binary choice model. Under certain distributional assumptions, the smoothed MSE is asymptotically normal with a convergence rate that is at least $n^{-2/5}$ and can be made arbitrarily close to $n^{-1/2}$, depending on the strength of the smoothness assumptions. Kyriazidou (1994) and Charlier et. al. (1995) have extended this smoothing method to the panel version of the binary choice model. With hardly any modification, the smoothing method also extends to the general MSE of this paper.

We follow closely the notation of Horowitz (1992) in this section. First, we let $K : \mathcal{R} \to \mathcal{R}$ be a continuous function such that

(K1) $|K(v)| < M$ for some finite M and all $v \in \mathcal{R}$
(K2) $\lim_{v\to-\infty} K(v) = 0$ and $\lim_{v\to\infty} K(v) = 1$

Let $\{\sigma_n : \sigma_n > 0, n = 1, 2, \ldots\}$ be a sequence satisfying $\sigma_n \to 0$. Define the smoothed objective function as

$$S_n(b; \sigma_n) \equiv \frac{1}{n}\sum_{i=1}^{n} \mathrm{sgn}(\Delta y)K(\Delta x b/\sigma_n). \tag{2.5}$$

The smoothed MSE

$$\hat{\beta}^s \equiv \underset{b:|b_1|=1}{\arg\max}\ S_n(b; \sigma_n).$$

is consistent under the same assumptions as those in Section 2.2.

Theorem 2 (Consistency of Smoothed MSE) *If Assumptions 1-4 and the appropriate error assumption (Assumption 5.1, 5.2, or 5.3 as in Lemma 2) hold, and*

$$\hat{\beta}^s_n \equiv \underset{b:|b_1|=1,\ \tilde{b}\in\tilde{B}}{\arg\max}\ S_n(b; \sigma_n),$$

where $\tilde{b} \equiv (b^2, \ldots, b^k)'$, *then* $\lim_{n\to\infty} \hat{\beta}^s_n = \beta$ *almost surely.*

Analyzing the asymptotic normality of the smoothed MSE requires some additional notation and several technical assumptions. The only differences from Horowitz (1992) are that we consider first differences and condition on the event $\{\Delta y \neq 0\}$. To start, define

$$T_n(b; \sigma_n) \equiv \partial S_n(b; \sigma_n)/\partial \tilde{b} \tag{2.6}$$

and

$$Q_n(b; \sigma_n) \equiv \partial S_n^2(b; \sigma_n)/\partial \tilde{b}\partial \tilde{b}'. \tag{2.7}$$

Let $z \equiv \Delta x\beta$. By Assumptions 2 and 4, the distribution of z conditional on $\Delta\tilde{x}$ and on the event $\{\Delta y \neq 0\}$ has everywhere positive density with respect to Lebesgue measure for almost every $\Delta\tilde{x}$. Let $p(z|\Delta\tilde{x}, \Delta y \neq 0)$ denote this conditional density, and for each positive integer i, define

$$p^{(i)}(z|\Delta\tilde{x}, \Delta y \neq 0) \equiv \partial^i p(z|\Delta\tilde{x}, \Delta y \neq 0)/\partial z^i$$

whenever the derivative exists, and, for notational purposes, define $p^{(0)}(z|\Delta\tilde{x}, \Delta y \neq$

$0) \equiv p(z|\Delta\tilde{x}, \Delta y \neq 0)$. Let $F(\cdot|z, \Delta\tilde{x}, \Delta y \neq 0)$ denote the c.d.f. of $\Delta\epsilon$ conditional on z, $\Delta\tilde{x}$, and the event $\{\Delta y \neq 0\}$. For each positive integer i, define

$$F^{(i)}(-z|z, \Delta\tilde{x}, \Delta y \neq 0) \equiv \partial^i F(-z|z, \Delta\tilde{x}, \Delta y \neq 0)/\partial z^i$$

whenever the derivative exists. Define the scalar constants α_A and α_D by

$$\alpha_A \equiv \int_{-\infty}^{\infty} v^h K'(v)dv$$

and

$$\alpha_D \equiv \int_{-\infty}^{\infty} [K'(v)]^2 dv$$

whenever these quantities exist. For each integer $h \geq 2$, define the $(k-1) \times 1$ vector A and the $(k-1) \times (k-1)$ matrices D and Q by

$$A \equiv -2\alpha_A \sum_{i=1}^{h} \{[i!(h-i)!]^{-1} \mathrm{E}[F^{(i)}(0|0, \Delta\tilde{x}, \Delta y \neq 0) p^{(h-i)}(0|\Delta\tilde{x}, \Delta y \neq 0)\Delta\tilde{x}']\} \Pr(\Delta y \neq 0),$$

$$D \equiv \alpha_D \mathrm{E}[\Delta\tilde{x}'\Delta\tilde{x} p(0|\Delta\tilde{x}, \Delta y \neq 0)] \Pr(\Delta y \neq 0),$$

and

$$Q \equiv 2\mathrm{E}[\Delta\tilde{x}'\Delta\tilde{x} F^{(1)}(0|0, \Delta\tilde{x}, \Delta y \neq 0) p(0|\Delta\tilde{x}, \Delta y \neq 0)] \Pr(\Delta y \neq 0),$$

whenever these quantities exist.

The following assumptions are needed for the asymptotic results.

Assumption 6 *The components of $\Delta\tilde{x}$ and of the matrices $\Delta\tilde{x}'\Delta\tilde{x}$ and $\Delta\tilde{x}'\Delta\tilde{x}\Delta\tilde{x}'\Delta\tilde{x}$ have finite first absolute moments.*

Assumption 7 $(\log n)/(n\sigma_n^4) \to 0$ *as* $n \to \infty$.

Assumption 8 *(a) K is twice differentiable everywhere, $|K'(\cdot)|$ and $|K''(\cdot)|$ are uniformly bounded, and each of the following integrals over $(-\infty, \infty)$ is finite: $\int [K'(v)]^4 dv$, $\int [K''(v)]^2 dv$, $\int |v^2 K''(v)| dv$; (b) For some integer $h \geq 2$ and each integer i $(1 \leq i \leq h)$, $\int |v^i K'(v)| dv < \infty$, and*

$$\int_{-\infty}^{\infty} v^i K'(v)dv = \begin{cases} 0 & \text{if } i < h \\ r\,(\text{nonzero}) & \text{if } i = h; \end{cases} \tag{2.8}$$

(c) For any integer i between 0 and h, any $\delta > 0$, and any sequence $\{\sigma_n\}$ converging to 0,

$$\lim_{n\to\infty} \sigma_n^{i-h} \int_{|\sigma_n v|>\delta} |v^i K'(v)| dv = 0 \tag{2.9}$$

and

$$\lim_{n\to\infty} \sigma_n^{-1} \int_{|\sigma_n v|>\delta} |K''(v)| dv = 0. \tag{2.10}$$

Assumption 9 *For each integer i such that $1 \leq i \leq h-1$, all z in a neighborhood of 0, almost every $\Delta\tilde{x}$, and some $M < \infty$, $p^{(i)}(z|\Delta\tilde{x}, \Delta y \neq 0)$ exists and is a continuous function of z satisfying $|p^{(i)}(z|\Delta\tilde{x}, \Delta y \neq 0)| < M$. In addition, $|p(z|\Delta\tilde{x}, \Delta y \neq 0)| < M$ for all z and almost every $\Delta\tilde{x}$.*

Assumption 10 *For each integer i such that $1 \leq i \leq h$, all z in a neighborhood of 0, almost every $\Delta\tilde{x}$, and some $M < \infty$, $F^{(i)}(-z|z, \Delta\tilde{x}, \Delta y \neq 0)$ exists and is a continuous function of z satisfying $|F^{(i)}(-z|z, \Delta\tilde{x}, \Delta y \neq 0)| < M$. (This assumption is satisfied if $[\partial^{i+j} F(\Delta\epsilon|z, \Delta\tilde{x}, \Delta y \neq 0)/\partial\Delta\epsilon^i \partial w^j]_{\Delta\epsilon=-z}$ is bounded and continuous in a neighborhood of $z = 0$ for almost every $\Delta\tilde{x}$ whenever $i + j \leq h$.)*

Assumption 11 *$\tilde{\beta}$ is an interior point of $\tilde{B}$.*

Assumption 12 *The matrix Q is negative definite.*

The only additional condition needed to apply the results of Horowitz (1992) is

$$F(0|0, \Delta\tilde{x}, \Delta y \neq 0) = 1/2. \tag{2.11}$$

Horowitz (1992) makes the explicit assumption that the median of the error term is zero. This assumption is equivalent to $F(0|0, \tilde{x}) = 1/2$, the analogous condition for the non-panel binary choice model. Here, (2.11) holds if the appropriate assumptions needed for Lemma 2 hold:

$$\begin{aligned} & \mathrm{E}[\mathrm{sgn}(\Delta y)|z = 0, \Delta\tilde{x}] = 0 \text{ by Lemma 2} \\ \Longrightarrow\ & \mathrm{E}[\mathrm{sgn}(\Delta y)|z = 0, \Delta\tilde{x}, \Delta y \neq 0] = 0 \\ \Longrightarrow\ & \mathrm{E}[2 \cdot 1(\Delta\epsilon \leq g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha})) - 1|z = 0, \Delta\tilde{x}, \Delta y \neq 0] = 0 \\ \Longrightarrow\ & \mathrm{E}[2 \cdot 1(\Delta\epsilon \leq 0) - 1|z = 0, \Delta\tilde{x}, \Delta y \neq 0] = 0 \\ \Longrightarrow\ & F(0|0, \Delta\tilde{x}, \Delta y \neq 0) = 1/2 \end{aligned}$$

The main results are embodied in the following two theorems. Theorem 3 gives the asymptotic distribution of the smoothed MSE. Theorem 4 shows how A, D, and Q can be consistently estimated from the observed data.

Theorem 3 *If Assumptions 1-4, 6-12, and the appropriate version of Assumption 5 hold for some $h \geq 2$, and $\{\hat{\beta}_n^s\}$ is a sequence of solutions to the maximization of $S_n(b; \sigma_n)$, then:*
(a) If $n\sigma_n^{2h+1} \to \infty$ as $n \to \infty$, $\sigma_n^{-h}(\hat{\tilde{\beta}}_n^s - \tilde{\beta}) \xrightarrow{p} -Q^{-1}A$.
(b) If $n\sigma_n^{2h+1}$ has a finite limit λ as $n \to \infty$,

$$\sqrt{n\sigma_n}(\hat{\tilde{\beta}}_n^s - \tilde{\beta}) \xrightarrow{d} N(-\sqrt{\lambda}Q^{-1}A, Q^{-1}DQ^{-1}).$$

(c) Let $\sigma_n = (\lambda/n)^{2h+1}$ with $0 < \lambda < \infty$, Ω be any nonstochastic, positive semidefinite matrix such that $A'Q^{-1}\Omega Q^{-1}A \neq 0$, E_A denote the expectation with respect to the

asymptotic distribution of $n^{h/(2h+1)}(\tilde{\hat{\beta}}_n^s - \tilde{\beta})$, *and mean square error equal to* $E_A(\tilde{\hat{\beta}}_n^s - \tilde{\beta})'\Omega(\tilde{\hat{\beta}}_n^s - \tilde{\beta})$. *Then, mean square error is minimized by setting*

$$\lambda = \lambda^* \equiv [\text{trace}(Q^{-1}\Omega Q^{-1}D)]/(2hA'Q^{-1}\Omega Q^{-1}A),$$

in which case

$$n^{h/(2h+1)}(\tilde{\hat{\beta}}_n^s - \tilde{\beta}) \xrightarrow{d} \mathrm{N}(-(\lambda^*)^{h/(2h+1)}Q^{-1}A, (\lambda^*)^{-1/(2h+1)}Q^{-1}DQ^{-1}).$$

Theorem 4 *Let* $\hat{\beta}_n^s$ *be a consistent smoothed MSE based on* $\sigma_n \propto n^{-1/(2h+1)}$. *For* $b \in \{-1,1\} \times \tilde{B}$ *and* $j = 1, \ldots, n$, *define*

$$t_j(b,\sigma) = sgn(\Delta y)(\Delta\tilde{x}_j/\sigma)K'(x_j b/\sigma).$$

Let $\sigma_n^* \propto n^{-\delta/(2h+1)}$, *where* $0 < \delta < 1$. *Let* $n_o = \sum_{j=1}^n 1(\Delta y \neq 0)$. *Then,*
(a) $\hat{A}_n \equiv \frac{n_o}{n}(\sigma_n^*)^{-h}T_n(\hat{\beta}_n^s, \sigma_n^*) \xrightarrow{p} A$;
(b) $\hat{D}_n \equiv \frac{n_o}{n}(\sigma_n/n)\sum_{j=1}^n t_j(\hat{\beta}_n^s, \sigma_n)t_j(\hat{\beta}_n^s, \sigma_n)' \xrightarrow{p} D$; *and,*
(c) $\hat{Q}_n \equiv \frac{n_o}{n}Q_n(\hat{\beta}_n^s, \sigma_n) \xrightarrow{p} Q$.

Multiplying by n_o/n achieves the necessary conditioning correction in Theorem 4 since n_o/n is an estimate of $\Pr(\Delta y \neq 0)$. (Alternatively, one could do the asymptotics in terms of n_o and rewrite Theorem 3 accordingly.)

Horowitz (1992) discusses how to choose the bandwidth for estimation and how to correct the small-sample bias. The same techniques can be used here as well.

Appendix

Proof of Lemma 1: See Lemma 1 of Honoré (1992).

Proof of Lemma 2: We consider each of the three cases separately.

(a) d is strictly increasing, which implies

$$\begin{aligned} \text{sgn}(\Delta y) &= \text{sgn}(d(g(x_2\beta,\alpha)+\epsilon_2) - d(g(x_1\beta,\alpha)+\epsilon_1)) \\ &= \text{sgn}(g(x_2\beta,\alpha) - g(x_1\beta,\alpha) + \Delta\epsilon). \end{aligned}$$

Then, positive $\Delta x\beta$ implies $g(x_2\beta,\alpha) - g(x_1\beta,\alpha) > 0$ for any α, which means $\Pr(\Delta y > 0|x_1,x_2) > 1/2 > \Pr(\Delta y < 0|x_1,x_2)$ since $\text{Med}(\Delta\epsilon|x_1,x_2,\alpha) = 0$. Similar for negative $\Delta x\beta$.

(b) Consider any realization of (x_1,x_2,α) and $(\epsilon_1+\epsilon_2)$. Let $L \equiv \frac{1}{2}[(g(x_1\beta,\alpha)+\epsilon_1) + (g(x_2\beta,\alpha)+\epsilon_2)]$. Since d is nondegenerate, there exists at least one place on the real line where d is nonconstant (i.e., $\exists v$ s.t. $d(v) > d(v-\delta)\ \forall\delta > 0$). Pick the closest such point to L:

$$v^* \equiv \arg\inf_{v\in V}|L - v|,$$

where $V \equiv \{v : d(v) > d(v-\delta) \; \forall \delta > 0\}$. Let $c^* \equiv 2|L - v^*|$. Then,

$$\Pr(\Delta y > 0 | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2) = \Pr(\Delta y^* > c^* | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2)$$

and

$$\Pr(\Delta y < 0 | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2) = \Pr(\Delta y^* < -c^* | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2).$$

Since $\Delta y^* = g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha}) + \Delta\epsilon$, we have

$$\begin{aligned} &\Pr(\Delta y > 0 | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2) = \\ &\quad \Pr(\Delta\epsilon > c^* - [g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha})] | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2) \end{aligned}$$

and

$$\begin{aligned} &\Pr(\Delta y < 0 | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2) = \\ &\quad \Pr(-\Delta\epsilon > c^* - [g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha})] | x_1, x_2, \boldsymbol{\alpha}, \epsilon_1 + \epsilon_2). \end{aligned}$$

Then, since g is strictly increasing in its first argument, the desired result follows from symmetry of $\Delta\epsilon$ conditional on $(x_1, x_2, \boldsymbol{\alpha})$ and $(\epsilon_1 + \epsilon_2)$. (3) holds since the result is true for any such realization, allowing us to condition only on x_1 and x_2.

(c) Consider any realization of $(x_1, x_2, \boldsymbol{\alpha})$ and $(\epsilon_1 + \epsilon_2)$, and define L as above. We first map the non-truncated region $\mathcal{T}$ into a set $\mathcal{P}$ of "possible" values for Δy^* given L. Defining $h(v) = \sqrt{2}(v - L)$ and $j(v) = -h(v)$, we have $\mathcal{P} = h(\mathcal{T}) \cup j(\mathcal{T})$ by some simple geometry. Define $\mathcal{P}^+ \equiv \{v \in \mathcal{P} : v > 0\}$ and $\mathcal{P}^- \equiv \{v \in \mathcal{P} : v < 0\}$. Then

$$\Pr(\Delta y > 0 | x_1, x_2, \boldsymbol{\alpha}) = \frac{\Pr(\Delta y^* \in \mathcal{P}^+ | x_1, x_2, \boldsymbol{\alpha})}{\Pr(\Delta y^* \in \mathcal{P} | x_1, x_2, \boldsymbol{\alpha})}$$

and

$$\Pr(\Delta y < 0 | x_1, x_2, \boldsymbol{\alpha}) = \frac{\Pr(\Delta y^* \in \mathcal{P}^- | x_1, x_2, \boldsymbol{\alpha})}{\Pr(\Delta y^* \in \mathcal{P} | x_1, x_2, \boldsymbol{\alpha})}.$$

Let $q(\cdot)$ denote the p.d.f. of Δy^* conditional on $(x_1, x_2, \boldsymbol{\alpha})$ and $(\epsilon_1 + \epsilon_2)$. We show that for any v in $\mathcal{P}^+$ (meaning $-v$ is in $\mathcal{P}^-$), we have $q(v) > q(-v)$ for positive $\Delta x\beta$. Note that $\Delta y^* = v \iff \Delta\epsilon = g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha}) + v$. Let $f(\cdot)$ denote the p.d.f. of $\Delta\epsilon$. Consider positive $\Delta x\beta$. If $v > g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha})$, we have

$$\begin{aligned} q(v) &= f(g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha}) + v) \\ &= f(g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha}) - v) \text{ by symmetry} \\ &> f(g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha}) - v) \text{ by unimodality} \\ &= q(-v). \end{aligned}$$

If $v < g(x_2\beta, \boldsymbol{\alpha}) - g(x_1\beta, \boldsymbol{\alpha})$, we have

$$\begin{aligned} q(v) &= f(g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha}) + v) \\ &> f(g(x_1\beta, \boldsymbol{\alpha}) - g(x_2\beta, \boldsymbol{\alpha}) - v) \text{ by unimodality} \\ &= q(-v). \end{aligned}$$

So, $q(v) > q(-v)$ for all v in $\mathcal{P}^+$. Since $\Pr(\Delta y^* \in \mathcal{S}) = \int_{v \in \mathcal{S}} q(v)dv$, we have

$$\Pr(\Delta y^* \in \mathcal{P}^+ | x_1, x_2, \alpha, \epsilon_1 + \epsilon_2) > \Pr(\Delta y^* \in \mathcal{P}^- | x_1, x_2, \alpha, \epsilon_1 + \epsilon_2)$$
$$\Longrightarrow \Pr(\Delta y > 0 | x_1, x_2, \alpha, \epsilon_1 + \epsilon_2) > \Pr(\Delta y < 0 | x_1, x_2, \alpha, \epsilon_1 + \epsilon_2)$$

Again, (3) holds since the result is true for any realization of (x_1, x_2, α) and $(\epsilon_1 + \epsilon_2)$, allowing us to condition only on x_1 and x_2.

Proof of Theorem 1: See Lemma 1 of Manski (1987). Replace $(z_1 - z_o)$ by $\text{sgn}(\Delta y)$ to make the proof applicable.

Proof of Theorem 2: Using weights in the objective function has no effect on consistency. See Theorem 1 of Horowitz (1992) for details.

Proof of Theorems 3 and 4: See Theorems 2 and 3 of Horowitz (1992).

References

CHARLIER, E., B. MELENBERG, AND A. H. O. VAN SOEST (1995): "A Smoothed Maximum Score Estimator for the Binary Choice Panel Data Model with an Application to Labour Force Participation," *Statistica Neerlandica* 49: 324–342.

HONORÉ, B. (1992): "Trimmed LAD and Least Squares Estimation of Truncated and Censored Regression Models with Fixed Effects," *Econometrica* 60: 533–565.

HOROWITZ, J. L. (1992): "A Smoothed Maximum Score Estimator for the Binary Response Model," *Econometrica* 60: 505–531.

KIM, J., AND D. POLLARD (1990): "Cube Root Asymptotics," *Annals of Statistics* 18: 191–219.

KYRIAZIDOU, K. (1994): "Semiparametric Estimation of a Sample Selection Panel Model," University of Chicago, mimeo.

MANSKI, C. F. (1975): "Maximum Score Estimation of the Stochastic Utility Model of Choice," *Journal of Econometrics* 3: 205–228.

MANSKI, C. F. (1987): "Semiparametric Analysis of Random Effects Linear Models from Binary Panel Data," *Econometrica* 55: 357–362.

POWELL, J. (1986): "Symmetrically Trimmed Least Squares Estimation for Tobit Models," *Econometrica* 54: 1435–1460.

Chapter 3

Estimation in the Presence of Mismeasured Dependent Variables (joint with Jerry Hausman)

3.1 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 and Scott-Morton (1994) 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 OLS).

This paper provides a treatment of mismeasured dependent variables in a more general 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* developed by Cavanagh and Sherman (1992), is proposed as an attractive alternative to parametric estimation. 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 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 higher latent dependent variable *first-order stochastically dominates* the random variable associated with a lower latent dependent variable.

The paper is organized as follows. Section 3.2 describes the general model of interest, formalizes the mismeasurement process, and introduces three illustrative examples. Section 3.3 describes maximum likelihood estimation methods in the presence of mismeasurement, focusing on the examples from Section 3.2. Section 3.4 introduces the semiparametric approach, formalizes the sufficient condition for consistency, and interprets the condition in the context of the examples. Section 3.5 extends the semiparametric approach to situations in which the measurement error in the dependent variable is dependent upon covariates. Section 3.6 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 3.7 concludes.

3.2 The Model

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

$$y^* = g(x\beta_o, \epsilon), \ \epsilon \text{ i.i.d.}, \tag{3.1}$$

where g is an unknown function with strictly positive partial derivatives everywhere.[1] The model given by (3.1) is quite general. For instance, it includes models with nonlinearity on the left-hand-side,

$$f(y^*) = x\beta_o + \epsilon, \tag{3.2}$$

and models with nonlinearity on the right-hand-side,

$$y^* = f(x\beta_o) + \epsilon, \tag{3.3}$$

where f is strictly increasing. Both of these models have ϵ entering additively, though that is not a restriction made in (3.1).

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)

[1] Additional disturbances may be included in (3.1). As an example, such a specification would allow for unobserved heterogeneity, as will be discussed in more detail in Section 3.6.

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

$$F_{y|y^*}(v|t) = \Pr(y \leq v|y^* = t). \tag{3.4}$$

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

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

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

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

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 0-1 case. The third example considers the case of a continuous dependent variable. The duration model application considered in Section 3.6 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 and Scott-Morton (1993), assume 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 α_0 and α_1 are

$$\alpha_0 \equiv \Pr(y = 1|y^* < 0) \tag{3.5}$$
$$\alpha_1 \equiv \Pr(y = 0|y^* > 0). \tag{3.6}$$

In the traditional binary response model, $\alpha_0 = \alpha_1 = 0$ since zeros are never misreported as ones and vice-versa.

With misreporting, the conditional c.d.f. $F_{y|y^*}$ is

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 0 \\ 1 - \alpha_0 & \text{if } v \in [0,1) \quad \text{if } t < 0 \\ 1 & \text{if } v \geq 1 \end{cases} \tag{3.7}$$

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 0 \\ \alpha_1 & \text{if } v \in [0,1) \\ 1 & \text{if } v \geq 1 \end{cases} \quad \text{if } t > 0. \tag{3.8}$$

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

$$\alpha(t) \equiv \Pr(y = 1|y^* = t) \text{ for } t \in \mathcal{R}. \tag{3.9}$$

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^* 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

$$\alpha_{s,t} = \Pr(y = t|y^* \in S_s). \tag{3.10}$$

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

$$A = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \cdots & \cdots & \alpha_{1,K} \\ \alpha_{2,1} & \alpha_{2,2} & \cdots & \cdots & \alpha_{2,K} \\ \vdots & & \ddots & & \vdots \\ \vdots & & & \ddots & \vdots \\ \alpha_{K,1} & \cdots & \cdots & \alpha_{K,K-1} & \alpha_{K,K} \end{bmatrix} \tag{3.11}$$

with the elements of each row adding up to one.

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

$$F_{y|y^*}(v|t) = \begin{cases} 0 & \text{if } v < 1 \\ \sum_{i=1}^{\lfloor v \rfloor} \alpha_{s,i} & \text{if } v \in [1,K] \\ 1 & \text{if } v > K \end{cases} \quad \text{for } t \in S_s. \tag{3.12}$$

As in Example 1, a more complicated model could allow the misclassification probabilities to depend on the actual leve 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,

$$y = h(y^*, \eta), \tag{3.13}$$

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

$$F_{y|y^*}(v|t) = \Pr(h(t, \eta) \leq v). \tag{3.14}$$

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

$$\Pr(h(t, \eta) \leq v) = \Pr(\eta \leq \tilde{h}(v, t)). \tag{3.15}$$

Combining (3.14) and (3.15) gives

$$F_{y|y^*}(v|t) = G(\tilde{h}(v, t)), \tag{3.16}$$

where G is the c.d.f. of η.

3.3 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

Consider the binary response model of the previous section. If H is the c.d.f. of $-\epsilon$,

$$\Pr(y = 1) \quad = \quad (1 - \alpha_1)H(x\beta_o) + \alpha_0(1 - H(x\beta_o)) \tag{3.17}$$

$$= \alpha_0 + (1 - \alpha_0 - \alpha_1)H(x\beta_o).$$

When $\alpha_0 = \alpha_1 = 0$ (no misclassification), equation (3.17) collapses to $\Pr(y = 1) = H(x\beta_o)$.

Parametric estimation of the binary response model proceeds by assuming the form of F (usual normal or logistic) and specifying a likelihood function. In the absence of misclassification, the log-likelihood is

$$\ln \mathcal{L}(b) = \sum_i \{y_i \ln H(x_i b) + (1 - y_i) \ln [1 - H(x_i b)]\}. \tag{3.18}$$

Taking misclassification into account, the log-likelihood becomes

$$\ln \mathcal{L}(b) = \textstyle\sum_i \{y_i \ln [\alpha_0 + (1 - \alpha_0 - \alpha_1)H(x_i b)] + (1 - y_i) \ln [(1 - \alpha_0) - (1 - \alpha_0 - \alpha_1)H(x_i b)]\}. \tag{3.19}$$

If misclassification exists but (3.18) is maximized instead of (3.19), the resulting estimate of β_o will be biased and inconsistent. Maximization of (3.19) yields a consistent estimate of β_o if α_0 and α_1 are known. If the misclassification probabilities are unknown, they can be estimated using

$$\ln \mathcal{L}(b, a_0, a_1) = \textstyle\sum_i \{y_i \ln [a_0 + (1 - a_0 - a_1)H(x_i b)] + (1 - y_i) \ln [(1 - a_0) - (1 - a_0 - a_1)H(x_i b)]\}. \tag{3.20}$$

If the mismeasurement has been modeled correctly, maximization of (3.20) yields consistent estimates of α_0, α_1, and β_o. If the mismeasurement has been modeled incorrectly (e.g., if the misclassification probabilities also depend on the level of y^* rather than just the sign of y^*), maximization of (3.20) will not yield consistent estimates. A more complicated likelihood function would need to be specified.

Example 2 continued

The parametric approach for mismeasured discrete dependent variables is similar to the approach for Example 1. The probability for the observed dependent variables is

$$\begin{aligned} \Pr(y = t) &= \sum_{s=1}^{K} \Pr(y = t | y^* \in S_s) \Pr(y^* \in S_s) \\ &= \sum_{s=1}^{K} \alpha_{s,t} \Pr(y^* \in S_s). \end{aligned} \tag{3.21}$$

We assume that the cutoff points for the sets $S_1, S_2, \ldots, S_K$ are known by the researcher (e.g., $S_1 = \{v : v \leq 10\}$, $S_2 = \{v : 10 < v \leq 20\}$, and $S_3 = \{v : v > 20\}$). Denote the cutoff points by $-\infty = c_0, c_1, \ldots, c_{K-1}, c_K = \infty$ 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.[2] Then, the likelihood function is

$$\ln \mathcal{L}(b, \hat{\sigma}, \{a_{s,t}\}) = \sum_i \sum_{t=1}^{K} 1(y_i = t) \ln \left\{ \sum_{s=1}^{K} a_{s,t} \left[H(x_i b - c_s) - H(x_i b - c_{s-1}) \right] \right\}, \quad (3.22)$$

subject to the constraints $\sum_t a_{s,t} = 1$ 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

$$\begin{aligned} \Pr(y \le v|x) &= \int F_{y|y^*}(v|t) f_{y^*|x}(t|x)\, dt \\ &= \int G(\tilde{h}(v,t)) f_{y^*|x}(t|x)\, dt \quad \text{by (3.16).} \end{aligned} \quad (3.23)$$

To form a parametrized likelihood function, we need to parametrize both G and $f_{y^*|x}$, which boils down to making parametric assumptions on the distributions of η and ϵ, respectively. The parametrization of $f_{y^*|x}$ also requires knowledge of the function g from (3.1). 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 (3.23) yields the likelihood function

$$\ln \mathcal{L}(b, \hat{\sigma}_\eta, \hat{\sigma}_\epsilon) = \sum_i \ln \left\{ \int \tilde{h}_2(y_i, t) G'(\tilde{h}(y_i, t)) f_{y^*|x}(t|x_i)\, dt \right\}. \quad (3.24)$$

Equation (3.24) is similar to likelihoods used to capture heterogeneity, in which mix-

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

ing distributions are used to model random effects. Maximization of (3.24) requires numerical integration.

In general, consistency of the estimates of β_o requires that h, G, and $f_{y^*|x}$ are correctly specified.

3.4 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 developed by Cavanagh and Sherman (1992). 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.

3.4.1 The Monotone Rank Estimator

The MRE, as defined by Cavanagh and Sherman (1992), is the estimator $\hat{\beta}^{\mathrm{MRE}}$ that maximizes the objective function

$$S^{\mathrm{MRE}}(b) = \sum_i M(y_i)\mathrm{Rank}(x_i b), \tag{3.25}$$

over the set $\mathcal{B} = \{b \in \mathcal{R}^d : |b_d| = 1\}$, where $M : \{y_1, \ldots, y_n\} \longrightarrow \mathcal{R}$ is some increasing function (i.e., $y_i > y_j \Longrightarrow M(y_i) > M(y_j)$). There are d covariates contained in x, which does not include a constant. The Rank$(\cdot)$ function is defined as follows:[3]

$$x_{i_1} b < x_{i_2} b < \ldots < x_{i_n} b \Longrightarrow \mathrm{Rank}(x_{i_m} b) = m.$$

Since the ranking of $x_i b$ is unaffected by the scale of b (i.e., $\mathrm{Rank}(x_i b) = \mathrm{Rank}(x_i(cb))$ 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 $\hat{\beta}^{\mathrm{MRE}}$ is

$$H(z) = \mathrm{E}[M(y)|x\beta_o = z] \text{ increasing in } z. \tag{3.26}$$

The monotonicity condition (3.26) says that, on average, higher $x\beta_o$ are associated with higher y. This "correlation" is maximized by the objective function (3.25).

[3]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.

3.4.2 Sufficient Condition for Consistency

The monotonicity condition is satisfied for the latent variable y^* in the model given by (3.1). That is,

$$(3.1) \implies H^*(z) = \mathrm{E}[M(y^*)|x\beta_o = z] \text{ increasing in } z. \tag{3.27}$$

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

$$\mathrm{E}[M(y)|y^* = t] \text{ increasing in } t. \tag{3.28}$$

A sufficient condition for (3.28), and thus for (3.26), 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 additional technical assumptions and proof are in the appendix.

Theorem 7 *Under standard assumptions needed for consistency of semiparametric estimators, $\hat{\beta}^{\mathrm{MRE}}$ (for any choice of increasing M) is an asymptotically normal, $\sqrt{n}$-consistent estimate of β_o in the model described by (3.1) if*

$$(i) \quad t_1 > t_2 \implies F_{y|y^*}(v|t_1) \leq F_{y|y^*}(v|t_2)\ \forall v \tag{3.29}$$

$$(ii) \quad \exists \tilde{t}\ s.t.\ t_1 > \tilde{t} > t_2 \implies \exists v\ s.t.\ F_{y|y^*}(v|t_1) < F_{y|y^*}(v|t_2). \tag{3.30}$$

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 $\hat{\beta}^{\mathrm{MRE}}$ is derived in Cavanagh and Sherman (1992).

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."

3.4.3 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 3.2.

Example 1 continued

From the conditional c.d.f.'s in (3.7) and (3.8), 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 3.3, the MRE estimator remains consistent if the misclassification probabilities are functions of the level of y^*. With the function $\alpha(t)$ given by (3.9), the stochastic-dominance conditions of Theorem 1 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) from Theorem 1 is equivalent to

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

and condition (ii) is equivalent to

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

Looking at the transition matrix A defined in (3.11), 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 (3.13) satisfies the stochastic-dominance conditions. To see this, write the conditional c.d.f. as

$$\begin{aligned} F_{y|y^*}(v|t) &= \Pr(h(t, \eta) \leq v) \\ &= \int 1(h(t, u) \leq v) \, dG(u), \end{aligned}$$

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

$$\partial F_{y|y^*}(v|t)/\partial t = \int h_{y^*}(t,u) 1(h(t,u) \leq v)\, dG(u),$$

which is positive for all v since h_{y^*} is positive. 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 (3.13).

3.5 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.[4] 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.

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 conditions, but the whole sample doesn't necessarily satisfy them.

3.5.1 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

$$T_j(b_{-1}) = \sum_i 1(x_{i1} = j) M_j(y_i) \mathrm{Rank}_j(x_{i,-1} b_{-1}), \tag{3.31}$$

over the set $\mathcal{B}_{-1} = \{b_{-1} \in \mathcal{R}^{d-1} : |b_{-1,d-1}| = 1\}$, where the subscript j indicates that the function applies to the observations within the subgroup defined by $x_1 = j$.

[4]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 is little hope of identifying β_o semiparametrically.

Then, we can take a linear combination of the subgroup estimates to yield a consistent estimator $\hat{\beta}_{-1}$ for the whole sample. For instance,

$$\hat{\beta}_{-1} = \frac{1}{n}\sum_{j=1}^{K} n_j \hat{\beta}^j_{-1}. \tag{3.32}$$

Since the asymptotic distribution of each $\hat{\beta}^j_{-1}$ is known, the asymptotic distribution of $\hat{\beta}_{-1}$ follows simply.[5]

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 Chapter 1 of this thesis, allowing for a χ^2-test of their difference.[6]

3.5.2 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

$$T_1(b) = \sum_i [1(x_{i1} < S) M_1(y_i) \text{Rank}_1(x_i b)] \tag{3.33}$$

$$T_2(b) = \sum_i [1(x_{i1} > S) M_2(y_i) \text{Rank}_2(x_i b)] \tag{3.34}$$

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

$$\hat{\beta} = \frac{1}{n}\left(n_1\hat{\beta}_1 + (n - n_1)\hat{\beta}_2\right), \tag{3.35}$$

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.

---

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

[6]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 (3.1)). That is, the observations within the subgroups partitioned by x_1 are homoskedastic but possibly heteroskedastic across subgroups.

3.6 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.

3.6.1 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).

We consider a standard proportional hazard model with exponential index, where the hazard function is

$$h(t) = h_o(t)e^{x\beta_o},$$

with $h_o(\cdot)$ called the "baseline hazard function." Then, the "integrated baseline hazard function" $H_o(t) = \int_{-\infty}^{t} h_o(\tau)d\tau$ satisfies

$$-\ln H_o(t) = x\beta_o + \epsilon,$$

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

$$t = H_o^{-1}(\exp(-x\beta_o - \epsilon)). \tag{3.36}$$

Negating (3.36) puts the proportional hazard model into the latent variable context of equation (3.1):

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

so that

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

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

Unobserved Heterogeneity

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

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

where the heterogeneity term u is independent of x. This model satisfies (3.1) with

$$g\left(x\beta_o, (u+\epsilon)\right) = -H_o^{-1}(\exp(-x\beta_o - (u+\epsilon))).$$

Parametric Estimation: Weibull model

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

$$h_o(t) = \alpha t^{\alpha-1}.$$

For the Weibull model, the integrated baseline hazard is $H_o(t) = t^\alpha$, so that equation (3.36) simplifies to

$$-\alpha \ln t = x\beta_o + \epsilon. \tag{3.37}$$

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) or if the right-hand-side variables change over time.

When unobserved heterogeneity is included, the Weibull model becomes

$$-\alpha \ln t = x\beta_o + u + \epsilon. \tag{3.38}$$

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 (3.37), then

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

so that the observed duration $\tilde{t}$ can be thought of as arising from a Weibull model with heterogeneity, as in equation (3.38). 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.[7] 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.[8] 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.

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

$$\ln \mathcal{L}(b) = \sum_i \left[x_i b - \ln \sum_{j \in R(i)} e^{x_j b} \right], \tag{3.39}$$

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

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

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. 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.

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

[8] Lancaster (1985) shows that the ratios remain consistent when $\eta \sim N(x\delta, \sigma_\eta^2)$ so that dependence on x is allowed if the heterogeneity is normally distributed.

We can write the first-order conditions resulting from (3.39):

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

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

$$\mathrm{E}[x_i] = \mathrm{E}\left[\frac{\sum_{j\in R(i)} x_j e^{x_j \beta_o}}{\sum_{j\in R(i)} e^{x_j \beta_o}}\right]. \tag{3.40}$$

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

$$\mathrm{E}\left[\frac{\sum_{j\in R(i)} x_j e^{x_j \beta_o}}{\sum_{j\in R(i)} e^{x_j \beta_o}}\right] \neq \mathrm{E}\left[\frac{\sum_{j\in \tilde{R}(i)} x_j e^{x_j \beta_o}}{\sum_{j\in \tilde{R}(i)} e^{x_j \beta_o}}\right]$$

so that the MLE estimate does not correspond to the first-order condition given by (3.40).

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 $\{(-\infty, t_1], (t_1, t_2], \ldots, (t_{K-1}, \infty)\}$, so that we observe

$$d_{ik} = \begin{cases} 1 & t_i \in [t_{k-1}, t_k) \\ 0 & \text{otherwise} \end{cases}$$

and maximize the likelihood function

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

where F is the extreme value c.d.f. The δ_k's are jointly estimated, with their true value being $\ln H_o(t_k)$ 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).

3.6.2 Monte Carlo results

We consider two Monte Carlo designs for the proportional hazard model to compare the performance of the Weibull MLE, Cox partial likelihood, HHM flexible MLE, and the MRE. The first design specifies a Weibull model with $\alpha = 1$, yielding a unit exponential baseline hazard. The second design specifies a non-monotonic baseline hazard, having

$$H_o^{-1}(v) = v - \kappa \sin(\gamma v)$$

and

$$h_o(t) = \left[1 - \gamma\kappa\cos(\gamma H_o(t))\right]^{-1}.$$

$|\gamma\kappa| < 1$ ensures a positive non-monotonic baseline hazard; in our simulations, we use $\gamma = 4$ and $\kappa = 0.2$.

To model mismeasured durations, we introduce multiplicative lognormal noise with unit mean. Observed durations are generated as $\tilde{t} = e^{\eta}t$, where $\eta \sim \mathrm{N}(-\sigma_\eta^2/2, \sigma_\eta^2)$. (As discussed in Section 3.6.1, multiplicative lognormal noise is equivalent to unobserved heterogeneity when the model is Weibull.) Each simulation uses two covariates, created by drawing (z_1, z_2) from a bivariate normal distribution with correlation of -0.5 truncated at 2.5 standard deviations and then setting $x_1 = e^{z_1}$ and $x_2 = e^{z_2}$.

For each experiment, we generate 10,000 observations and run 100 simulations for Weibull MLE, Cox partial likelihood, and MRE and 25 simulations for HHM flexible MLE. For the MRE estimates, we use the Rank($\cdot$) function for $M(\cdot)$ in the objective function (3.25). We use 12 intervals for the HHM estimates. The results are reported in Tables 3.1 and 3.2, where the true coefficient vector is $\beta_o = (1, -1)'$ and $\sigma_\eta^2 = 0$ (no measurement error) or $\sigma_\eta^2 = 1$ (measurement error). Since the MRE estimates only the ratio of the coefficients, we have scaled the MRE estimates to have vector length equal to one.

Table 3.1 gives results for the unit exponential baseline hazard. Without mismeasurement, all of the estimators are consistent as expected. When noise is added to the dependent variable, the coefficient estimates of the non-MRE estimators are attenuated toward zero. The attenuation for the Weibull MLE, partial likelihood, and standard HHM estimates is around 30%. We also report results for the HHM estimator which allows for unobserved gamma heterogeneity. The estimates $(-0.8828, 0.8887)$ have attenuation of just over 11%. As explained in Section 3.6.1, the ratios of the estimates for the Weibull MLE are consistent even though the absolute values are attenuated toward zero. The ratios of the estimates diverge from the true value of -1 for the Cox partial likelihood and regular HHM estimator. The extension of the HHM MLE to allow for gamma heterogeneity does well in estimating the ratio as well.

Table 3.1: Unit Exponential Baseline Hazard

	$\sigma_\eta^2 = 0$ (no measurement error)			$\sigma_\eta^2 = 1$ (measurement error)		
	β_1	β_2	$\|\beta_2/\beta_1\|$	β_1	β_2	$\|\beta_2/\beta_1\|$
True	-1.0000	1.0000	1.00	-1.0000	1.0000	1.00
Weibull	-0.9979	0.9991	1.00	-0.6846	0.6862	1.00
	(0.0096)	(0.0101)	(0.014)	(0.0089)	(0.0090)	(0.018)
Cox	-0.9983	0.9990	1.00	-0.6353	0.7222	1.14
	(0.0132)	(0.0120)	(0.018)	(0.0132)	(0.0106)	(0.029)
HHM	-0.9962	0.9991	1.00	-0.6714	0.7098	1.06
	(0.0156)	(0.0164)	(0.023)	(0.0116)	(0.0157)	(0.030)
HHM (gamma)	——	——	——	-0.8828	0.8887	1.01
				(0.0217)	(0.0200)	(0.034)
MRE	-0.7059	0.7081	1.00	-0.7052	0.7086	1.00
	(0.0118)	(0.0118)	(0.024)	(0.0153)	(0.0152)	(0.031)

The estimates are sample averages over 100 simulations (25 simulations for Han-Hausman) for 10,000 observations. Sample s.e.'s are in parentheses.

Table 3.2 gives results for the non-monotonic baseline hazard. The Weibull MLE is biased even without mismeasurement since the model is misspecified. The mean of the Weibull MLE estimates is $(-0.8138, 0.9069)$. The Cox partial likelihood and HHM estimates are consistent since they allow for a flexible baseline hazard. When there is mismeasurement, however, all of the esimates are again attenuated toward zero. Since the measurement error no longer reduces to additive heterogeneity as it did in the first design, none of the non-MRE estimators yield consistent estimates of the coefficient ratios. The HHM estimator allowing for gamma heterogeneity has $|\beta_2/\beta_1|$ of 1.05, which is biased by 5%. The MRE estimates are again consistent for the ratios of the coefficients and, as in Table 3.1, the sample standard errors are higher due to the additional noise present in the Monte Carlo design.

Table 3.2: Non-Monotonic Baseline Hazard

	$\sigma_\eta^2 = 0$ (no measurement error)			$\sigma_\eta^2 = 1$ (measurement error)		
	β_1	β_2	$\|\beta_2/\beta_1\|$	β_1	β_2	$\|\beta_2/\beta_1\|$
True	-1.0000	1.0000	1.00	-1.0000	1.0000	1.00
Weibull	-0.8138	0.9069	1.11	-0.6352	0.7120	1.12
	(0.0072)	(0.0108)	(0.017)	(0.0076)	(0.0092)	(0.020)
Cox	-0.9994	0.9995	1.00	-0.6599	0.7773	1.18
	(0.0131)	(0.0123)	(0.018)	(0.0116)	(0.0111)	(0.027)
HHM	-1.0003	0.9908	0.99	-0.7021	0.7830	1.12
	(0.0152)	(0.0132)	(0.020)	(0.0119)	(0.0124)	(0.026)
HHM	——	——	——	-0.9070	0.9546	1.05
(gamma)				(0.0228)	(0.0228)	(0.036)
MRE	-0.7073	0.7067	1.00	-0.7089	0.7049	0.99
	(0.0130)	(0.0131)	(0.026)	(0.0155)	(0.0155)	(0.031)

The estimates are sample averages over 100 simulations (25 simulations for Han-Hausman) for 10,000 observations. Sample s.e.'s are in parentheses.

3.6.3 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 aforementioned 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.[9] 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 adjacents 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 in the SIPP 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 unemployment spell between 1986 and 1992 and are eligible for UI benefits.[10] 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.

[9]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.

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

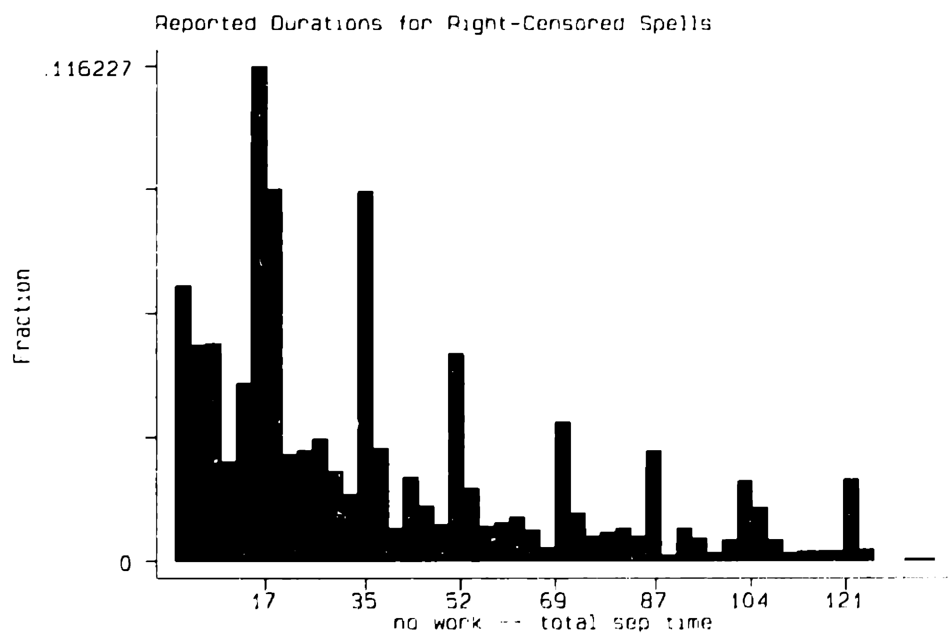
Table 3.3: Summary Statistics for SIPP Sample

	All	UI	Non-UI
Number of spells	15,103	4,205	10,898
Censored spells (percentage)	2,237 (14.8%)	667 (15.9%)	1,570 (14.4%)
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)

Table 3.3 reports summary statistics for the full sample and the subsamples of UI recipients and non-recipients. 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 3.1 and 3.2 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.

Figure 3.1:



Figures 3.3 and 3.4 give more detailed histograms for the uncensored spells, focusing on the subsample unemployed between 25 and 75 weeks. Figure 3.3 graphs durations for UI recipients, and Figure 3.4 graphs durations for non-recipients. The four-month spikes (at 35, 52, and 69 weeks) are evident. In Figure 3.3, 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 3.4.

Figure 3.2:

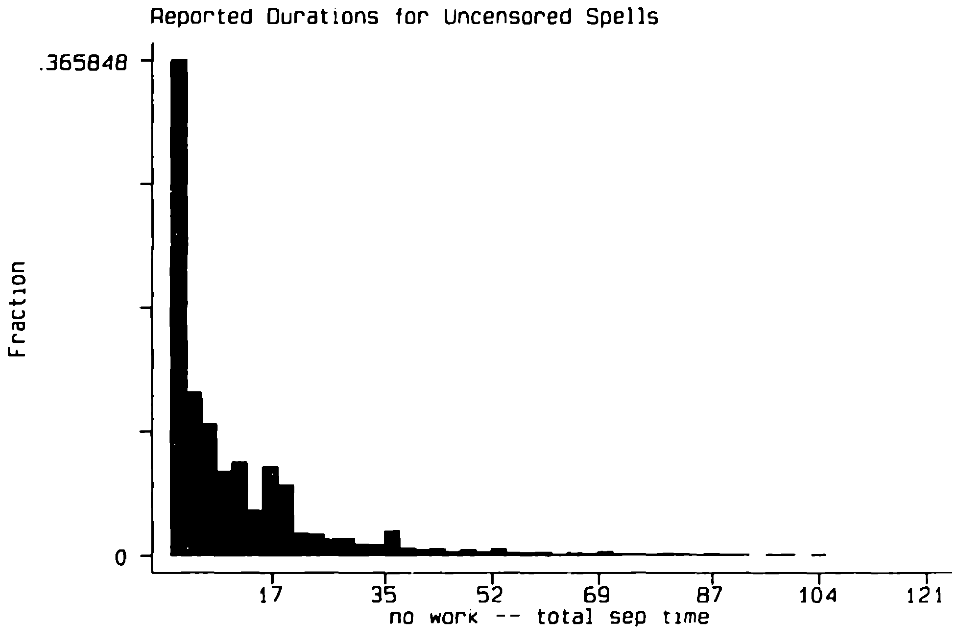


Figure 3.3:

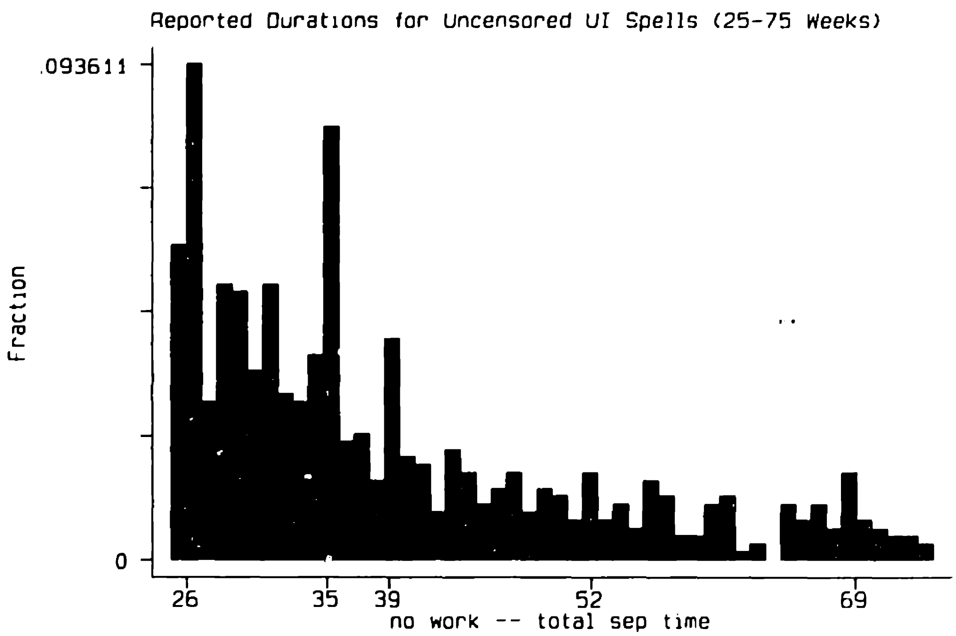
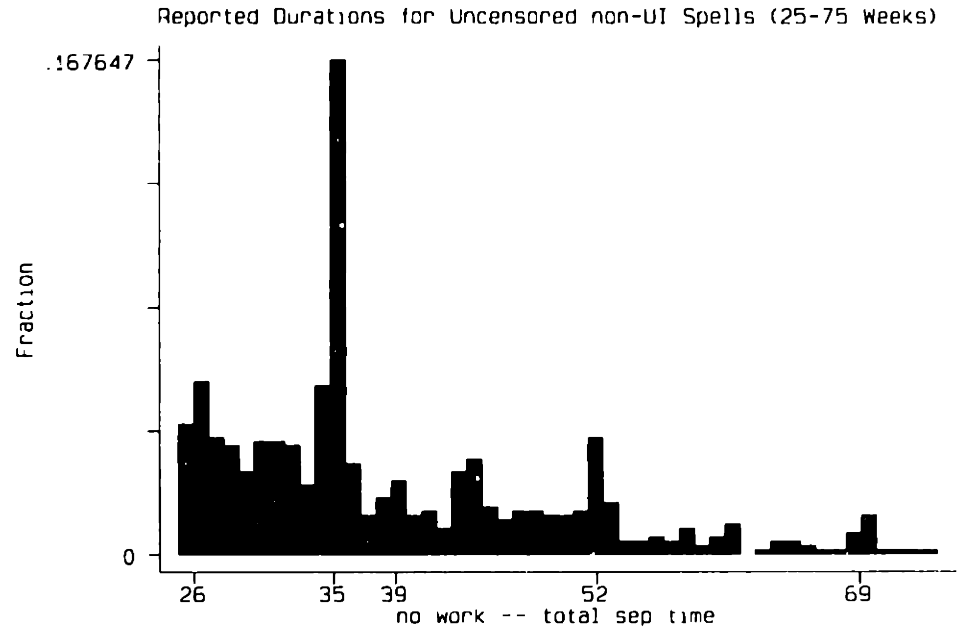


Figure 3.4:



We estimate a proportional hazard model using the aforementioned estimation techniques. In Table 3.4, we report coefficient estimates obtained from Weibull MLE, Cox partial likelihood, HHM MLE, and HHM MLE with gamma heterogeneity.[11]

[11]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 3.4: Duration Results for SIPP Sample

	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)

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 age of 21 and 30. The education dummies (HSGRAD, 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.[12] 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.

A positive (negative) coefficient indicates that the associated variable causes longer (shorter) unemployment spells. The results are quite similar across the columns of Table 3.4. 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 (3.37)) is 0.5979 (with a standard error of 0.0068), indicating a decreasing baseline hazard. There is evidence of heterogeneity from the HHM estimates in the final column of Table 3.4. 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 is the appropriate estimator for this data. In Table 3.5, we report the results from estimation of the proportional hazard model using the MRE with $M(y) = y$, the identity function.[13]

[12] 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 UI laws pertaining to the given worker and the reported quarterly wages before unemployment.

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

Table 3.5: 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 are shown in brackets.

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.[14] 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 sign and the 95% confidence intervals imply statistical significance since they do not contain zero (except for college graduates).

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 has little effect on the 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 HHM allowing for gamma heterogeneity, though, has significantly positive estimates, with the coefficient for UI recipients at the boundary of the MRE confidence interval.

While consistency of the MRE estimates does not require a model of the unemployment spell mismeasurement, estimation of the underlying baseline hazard (or, equivalently, the underlying survival function) will require some specification of the mismeasurement. To estimate the survival function, we use a variant of the Kaplan-Meier procedure that takes the linear index $x\beta_o$ into account by weighting observations by an appropriate function of $x_i\hat{\beta}$. Since consistency of this procedure requires consistency of $\hat{\beta}$, the MRE estimates are used. Figures 3.5 and 3.6 are the estimated survival function and baseline hazard function at the mean index value. These figures tell much the same story as the histograms in Figures 3.1–3.4. The survival function has a noticeable drop at both four months and eight months, and the hazard function has spikes at four month intervals. These spikes are certainly special to the SIPP and do not represent a general feature about unemployment in the United States. Any inferences based on these estimates would be misleading.

As a first step in dealing with the mismeasurement, we pool durations into 5-week groupings (1 to 5 weeks, 6 to 10 weeks, 11 to 15 weeks, etc.) and apply Kaplan-Meier. The Kaplan-Meier estimates will be consistent if durations are mismeasured, but

[14] Cavanagh and Sherman (1992) 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, we are far more confident in the bootstrap results.

Figure 3.5:

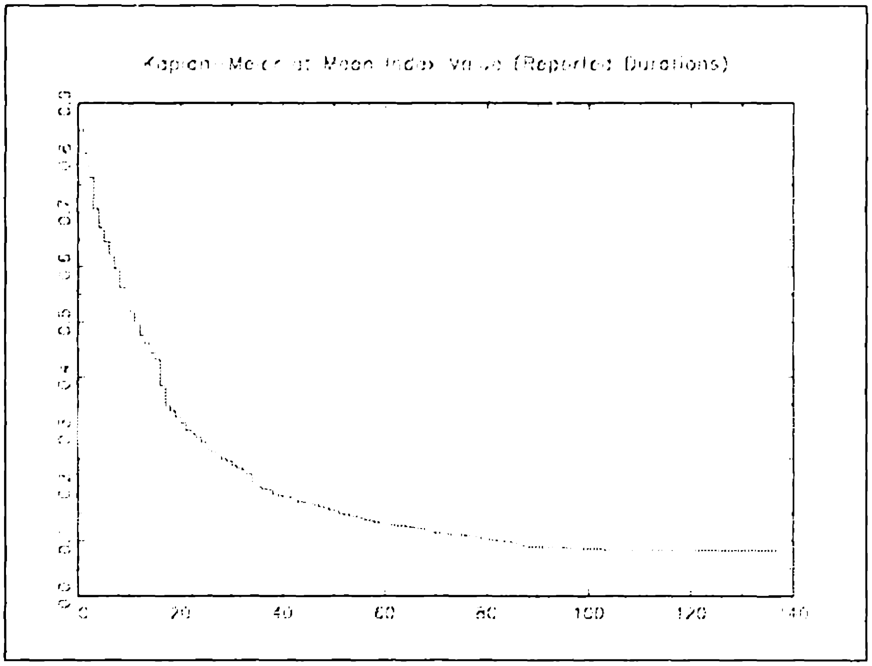
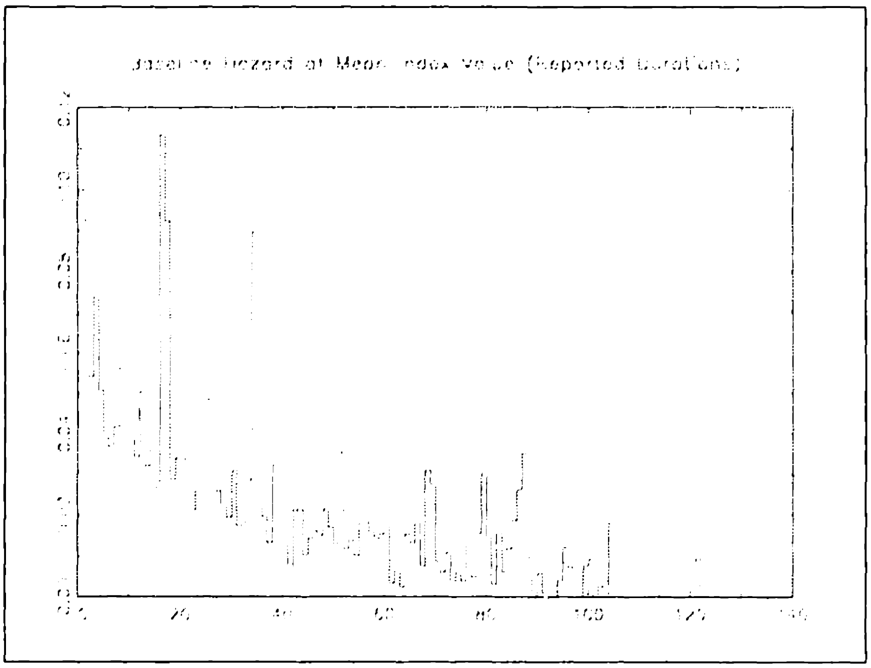
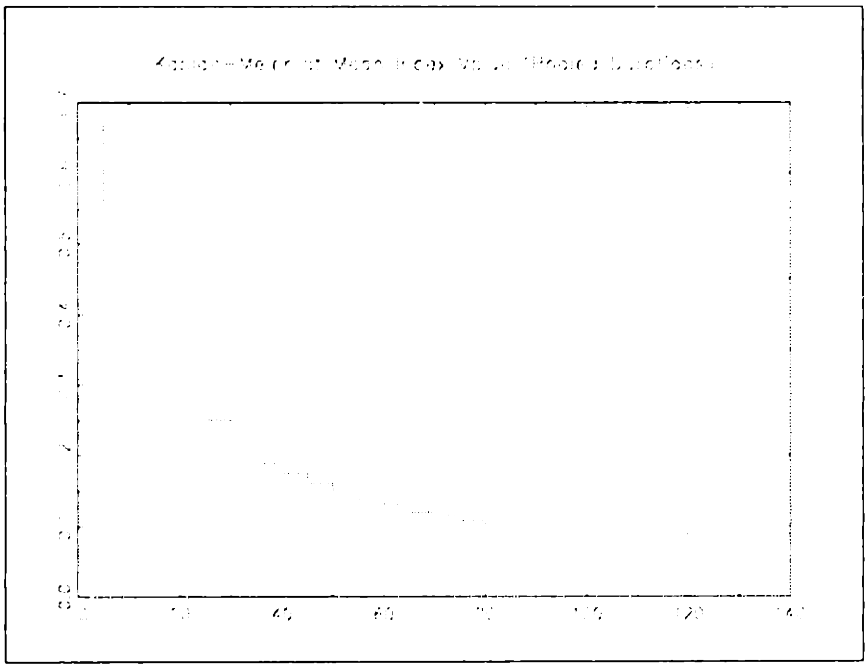


Figure 3.6:



always fall into the correct five-week interval. Of course, this hypothesis is not likely to be true; if it were, the HHM technique would result in consistent estimates of β_o and the MRE would not be needed at all. As shown in Figures 3.7 and 3.8, though, even this simplistic view of the mismeasurement smooths out the survival and hazard functions. The pooling of durations here is akin to fixed-window kernel smoothing.

Figure 3.7:

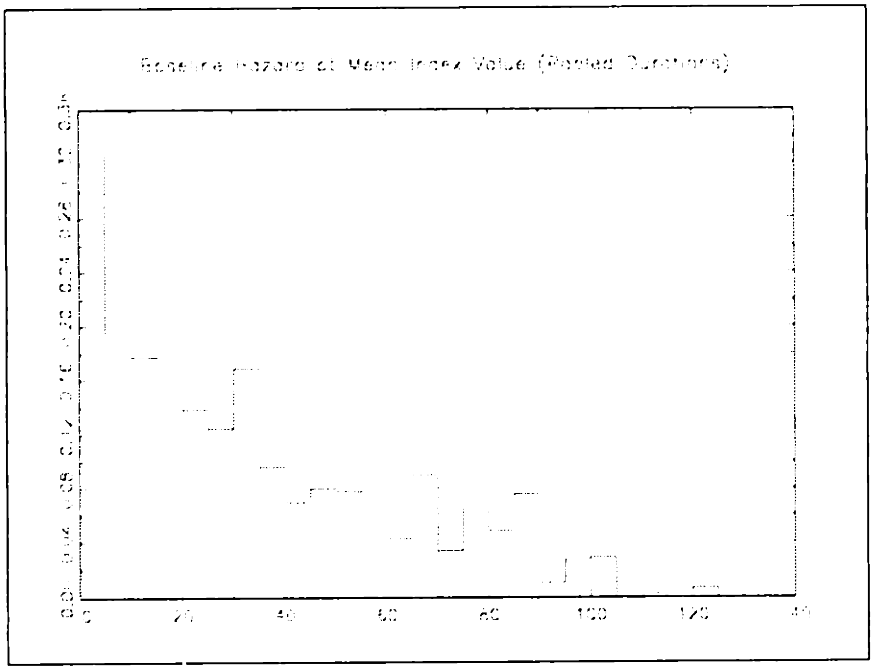


A more realistic model of the mismeasurement would directly consider the focal response phenomenon. For instance, an observed duration of four months is likely mismeasured and one can specify a probability distribution of the true duration (e.g., equal weights placed on all durations between three and five months). Once a complete description of the mismeasurement is formed, the Kaplan-Meier estimator can be applied to a new sample which consists of several replications of the original sample, where the durations depend on the observed duration and the specified model of mismeasurement. Several different models can be considered and compared to gauge the sensitivity of the results to alternative assumptions. Pursuit of this topic is beyond the scope of this paper but is worthy of future research.

3.7 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 3.4 is a strong result in that it applies to many forms of mismeasure-

Figure 3.8:



ment 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 mismeasured. 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 3.6 show that the semiparametric approach has different implications for the effect of previous wages and unemployment benefits on the length of unemployment spells. Section 3.6 also suggests ways in which the underlying hazard function may be estimated through modeling of the mismeasurement process.

Appendix

Proof of Theorem 1:

The consistency proof in Sherman and Cavanagh (1992) requires the following assumptions:

(A1) The support of x is not contained in any proper linear subspace of $\mathcal{R}^d$.
(A2) The d'th component of x has everywhere positive Lebesgue density, conditional on the other components.
(A3) The parameter space $\mathcal{B}$ is a compact subset of $\{b \in \mathcal{R}^d : b_d = 1\}$.

(A4) The function $H(z) = \mathrm{E}[M(y)|x\beta_o = z]$ is increasing.
(A5) The random variables $M(y)$ and $x\beta_o$ have nonzero correlation.
(A6) $\mathrm{E}[M(y)^2] < \infty$.

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 (A1)–(A3) and (A6) and show that conditions (i), (ii), and (iii) imply (A4) and (A5).

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

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

Then, $\partial g(z,-u)/\partial z > 0$ and condition (i) yield

$$\frac{\partial \mathrm{E}[M(y)|x\beta_o = z]}{\partial z} \geq 0.$$

The continuity assumption on x (assumption (A2)), assumption (ii), and condition (iii) ensure that the inequality will be strict for some z. Thus, (A4) and (A5) hold, and the MRE is consistent.

Additional assumptions are needed for asymptotic normality of $\hat{\beta}^{\mathrm{MRE}}$ (see Cavanagh and Sherman (1992)).

References

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 (1992): "Rank Estimators for Monotone Index Models," Bellcore Economics Discussion Paper no. 84.

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

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. AND F. M. SCOTT-MORTON (1994): "Misclassification of a Dependent Variable in a Discrete Response Setting," M.I.T. Department of Economics, mimeo.

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

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.

MANSKI, C. F. (1985): "Semiparametric Analysis of Discrete Response: Asymptotic Properties of the Maximum Score Estimator," *Journal of Econometrics* 27: 313–333.

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.

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.

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