ESTIMATING DERIVATIVES IN NONSEPARABLE MODELS WITH LIMITED DEPENDENT VARIABLES

By

Joseph G. Altonji, Hidehiko Ichimura and Taisuke Otsu

July 2008
Revised May 2011

COWLES FOUNDATION DISCUSSION PAPER NO. 1668R

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.econ.yale.edu/

Estimating Derivatives in Nonseparable Models with Limited Dependent Variables

Joseph G. Altonji
Yale University and NBER

Hidehiko Ichimura
University of Tokyo

Taisuke Otsu
Yale University

March 1996
This Draft: April 2011[1]

[1]Our research was supported by the National Science Foundation under SBR-9512009, the Institute for Policy Research, Northwestern University, the Economic Growth Center and the Cowles Foundation, and Yale University (Altonji), JSPS Basic Research (B) 18330040 (Ichimura), and the National Science Foundation under SES-0720961 (Otsu).

Abstract

We present a simple way to estimate the effects of changes in a vector of observable variables X on a limited dependent variable Y when Y is a general nonseparable function of X and unobservables, and X is independent of the unobservables. We treat models in which Y is censored from above, below, or both. The basic idea is to first estimate the derivative of the conditional mean of Y given X at x with respect to x on the uncensored sample without correcting for the effect of x on the censored population. We then correct the derivative for the effects of the selection bias. We discuss nonparametric and semiparametric estimators for the derivative. We also discuss the cases of discrete regressors and of endogenous regressors in both cross section and panel data contexts.

1 Introduction

Many problems in economics involve dependent variables that are censored in some way. For example, one may wish to know how consumers who demand a positive amount of a good respond to changes in prices, income, or age. Furthermore, many of the restrictions placed on demand by consumer theory apply only to consumers who are not at corner solutions and must be tested using the uncensored observations. For example, one might wish to estimate a compensated price effect. In the factor demand literature, the problem of zero inputs often arises. For these reasons, a vast empirical literature has used the Tobit or generalized Tobit models to study the effects of a set of independent variables X on a censored dependent variable Y.

Unfortunately, almost all the literature on censored regression relies heavily on the assumptions of additive separability and/or monotonicity in the error term U.[1] In contrast, nonseparability and nonmonotonicity are likely to be the rule rather than exception in the choice problems based on constrained optimization that characterize much of economics. In a world of heterogeneous consumers, the demand function $Y = M(X, U)$ is unlikely to be additively separable in the observed price, income, and preference variables X and the unobserved preference variables U, especially given that preferences, prices, and endowments interact through budget and time constraints. Monotonicity in the vector U is also unlikely. Unobserved heterogeneity across firms in technology and efficiency will also enter factor input demand functions in a nonseparable way unless one artificially restricts the form that unobserved heterogeneity can take, as is often done by simply tacking an error term onto the demand model.[2]

Altonji, Hayashi and Kotlikoff's (1997) study of altruism based models of money transfers from parents to children is a concrete example of nonseparability and censoring in a consumer demand context. In their application, $Y = M(X_p, X_c, X_2, U)$ if $M(X_p, X_c, X_2, U)$ is positive and is 0 otherwise, where X_p and X_c are the endowments of the parents and child and X_2 and U are vectors of observed and unobserved preferences of the parents and child. They point that nonseparability of M is a generic property of transfer equations that are based on a consumer choice framework with interdependent preferences. Furthermore, a key theoretical prediction of altruism models of transfers, $\partial M(X_p, X_c, X_2, U)/\partial X_p - \partial M(X_p, X_c, X_2, U)/\partial X_c = 1$, applies only if $M(X_p, X_c, X_2, U) > 0$, so one must account for both censoring and nonseparability if one wishes to test it.[3]

In this paper, we present a simple way to estimate the effects of changes in a vector of observable regressors X on a censored dependent variable Y when Y is a general nonseparable function of X and unobservables U, and X is independent of U. The general model we consider includes models of the form $Y = M(X, U)$ if $L(X) < M(X, U)$ and $Y = C_L$ otherwise, where $M(X, U)$ is a differentiable function with respect to X indexed by U, $L(X)$ is an unknown function of X, and $Y = C_L$ indicates that Y is censored from below. The parameter

[1] See, e.g., Chay and Powell (2001) for a survey.

[2] If the dependent variable is not censored or truncated, one can estimate average derivatives of nonseparable regression models using the methods of Stoker (1986) and Powell, Stock and Stoker (1989) among others.

[3] Similar issues arise in the public finance literature concerning the extent to which public transfers crowd out charitable giving, and whether targeted grants (such as food stamps) affect consumption patterns differently from income.

of interest is the average derivative $\beta(x) = E[\nabla M(X,U)|X = x, L(X) < M(X,U)]$, where $\nabla M(X,U)$ is the partial derivative of $M(x,u)$ with respect to x evaluated at $(x,u) = (X,U)$. Note that in the ordinary linear censored regression model with an additive error (i.e., the Tobit model), $\beta(x)$ is constant and coincides with the slope coefficients of the regressors.

Our estimation strategy is simple. The basic idea is (i) estimate the derivative of $\Psi(x) = E[M(X,U)|X = x, L(X) < M(X,U)]$ with respect to x *without* correcting for the influence of x on the composition of the uncensored population (i.e., selection bias) and then (ii) correct the partial derivative $\nabla\Psi(x)$ for the effects of the selection bias. It turns out that the correction term has a simple structure which only depends on $\Psi(x)$, $L(x)$, and on the level and derivative of $\Pr\{L(X) < M(X,U)|X = x\}$, the probability that Y is uncensored given $X = x$. We consider models in which Y is censored from both above and below but do not address the case in which the boundaries are stochastic conditional on X.

The paper continues in Section 2, where we provide a brief literature review. In Section 3 we present a canonical nonseparable censored dependent variable model. We then show that $\beta(x)$ is identified from knowledge of certain estimable functions of x. Starting from the expression for $\beta(x)$ that underlies our identification result, Section 4 discusses nonparametric and semiparametric estimation of $\beta(x)$. Section 5 briefly discusses extensions to the case of discrete regressors and to the case of endogenous regressors in both cross section and panel data contexts. In Section 6 we provide some encouraging Monte Carlo evidence on the performance of our estimators.

2 Previous Literature

Some early efforts on estimation of parameters in nonseparable models are found in Han (1987), Matzkin (1991), and Powell (1991). One of the difficulties in nonseparable models is to define an estimable parameter of interest. Han (1987) considered estimation of β in models where $Y = M(X'\beta, U)$, Matzkin (1991) considered estimation of m in models where $Y = M(m(X), U)$, and Powell (1991) considered estimation of β in models where $Y = M(X, \beta, U)$. All models assume that U is a scalar and that M is nondecreasing in U. Han (1987) and Matzkin (1991) allow the function M to be unknown and Powell (1991) assumes it to be known. As the above authors discuss, these models generalize many limited dependent variable models, some hazard models, and some transformation models.

Since the early drafts of our paper were circulated, a few papers on nonparametric estimation of features of censored dependent variable models have appeared. Lewbel and Linton (2002) consider the additively separable model

$$M(X,U) = m(X) + U, \quad L(X) = c, \; H(X) = \infty,$$

where the constant c is known. Note that in their model $\beta(x) = \nabla m(x)$. Under the additive error model they show that $\nabla m(x)$ is the derivative of $E[I\{Y > c\}(Y - c)|X = x]$ with respect to x divided by the conditional probability that Y is uncensored given $X = x$. We show that this result holds much more generally when we replace $\nabla m(x)$ with $\beta(x)$. We do not require an additive error structure and allow for censoring from above and below and for censoring points that depend generally on X.

Chen, Dahl and Kahn (2005) provide an estimator for $m(x)$ based on conditional quantiles in a model similar to Lewbel and Linton's. They assume $M(X,U) = m(X) + \sigma(X)U$, where U is a scalar and independent of X and $\sigma(X)$ is strictly positive. Their approach breaks down if monotonicity in U is dropped or the second additive error term appears in the model m. They do not consider estimation of $\beta(x)$ or the case in which $L(X)$ depends on X. In contrast, we place almost no restrictions on $M(X,U)$ but only consider estimation of $\beta(x)$.

We wish to emphasize that interest in the conditional average effect $\beta(x)$ depends on the question at hand and the source of the bounds $L(x)$ and $H(x)$. If $L(x)$ and $H(x)$ are due to data problems, (such as lower or upper bounds on reported values in a survey of business income) rather than a natural part of the model (such as nonnegative consumer expenditures) one is likely to be more interested in $E[\nabla M(X,U)|X=x]$, the unconditional average effect of x, than in $\beta(x)$.[4] If M takes the form of $m(X)+U$, as in Lewbel and Linton for example, then $\beta(x)$ is both the conditional and the unconditional effect. Our estimator leaves $M(X,U)$ and the distribution of U essentially unrestricted. Note that

$$E[Y|X] = E[Y|X, \mathit{uncensored}]\Pr\{\mathit{uncensored}|X\} + E[Y|X, \mathit{censored}]\Pr\{\mathit{censored}|X\}.$$

Since there are no data on Y when censored, $E[Y|X, \mathit{censored}]$ cannot be identified without assuming separability or imposing restrictions on both $M(X,U)$ and the distribution of U. Therefore, $E[Y|X]$ cannot be identified without a further assumption. One could combine our estimator of the conditional average effect with an estimator of the average effect for the censored cases that requires stronger assumptions.

Over the past decade there has been an explosion of research on nonseparable models with particular attention to models with endogenous regressors.[5] This literature is concerned with estimation of the partial effects of X on Y as well as with estimation of the structural function $M(X,U)$ and the distribution function of U given X, which we do not address. In the nonseparable simultaneous equation literature, monotonicity in a scalar valued U plays a key role in the identification of $M(x,u)$ at given (x,u), but it may not be a reasonable assumption for consumer expenditure problems or for choice problems based on constrained optimization in general. We do not assume monotonicity in U or that U is a scalar.

3 The Model and Identification of $\beta(x)$

We first introduce the model and parameter of interest. Let $X \in \mathbb{R}^k$ be a $k \times 1$ random vector of observables, and $M(X,U)$ be a random function of X, where the unobservable random object U indexes a class of differentiable functions from $\mathbb{R}^k$ to $\mathbb{R}$. The random object U does not need to be a scalar random variable or a finite dimensional random vector. In

[4]As noted in the introduction $\beta(x)$ can be used to test theoretical restrictions that apply to all the uncensored cases. The parameter $\beta(x)$ would permit one to test these restrictions even in the case of censoring due to survey reporting limits, in the case when $H(x)$ is a rationing limit on consumer expenditures that is a known or unknown function of x, and in the case where a tax or subsidy leads to a discontinuity in $\nabla M(X,U)$ when $M(X,U) > H(X)$. (In the latter case, the researcher could impose censoring on Y to avoid the discontinuity.)

[5]See, e.g., Altonji and Matzkin (2005), Blundell and Powell (2003), Chernozhukov, Imbens and Newey (2007), Chesher (2003, 2005), Florens *et. al.* (2008), Hoderlein and Mammen (2007), Imbens and Newey (2009), and Matzkin (2007).

our model, $M(X,U)$ is a latent variable. Instead we observe Y:

$$Y = \begin{cases} M(X,U) & \text{if } L(X) < M(X,U) < H(X), \\ C_L & \text{if } M(X,U) \leq L(X), \\ C_H & \text{if } H(X) \leq M(X,U), \end{cases} \tag{1}$$

where $L(X)$ and $H(X)$ are scalar valued functions of X and C_L and C_H are constants that indicate whether Y is censored from below or above, respectively. Our notation allows for the possibility that the functions $M(X,U)$, $L(X)$, and $H(X)$ do not depend on all of the elements of X. The linear censored regression model (i.e., the Tobit model) is a special case of (1) in which U is a scalar random variable, $M(X,U) = X'\beta + U$, $L(X) = 0$, and $H(X) = \infty$. For notational convenience we introduce three indicator random variables: $I_M(X) = I\{L(X) < M(X,U) < H(X)\}$, $I_L(X) = I\{M(X,U) \leq L(X)\}$, and $I_H(X) = I\{H(X) \leq M(X,U)\}$, where $I\{A\} = 1$ if the event A occurs and 0 otherwise, and the argument U is suppressed to simplify the notation.

Let ∇ denote the partial derivative with respect to x. The parameter of interest, $\beta(x)$, is the average derivative of $M(X,U)$ with respect to X given that $X = x$ and Y is not censored:

$$\beta(x) = E[\nabla M(X,U)|X = x, I_M(X) = 1]. \tag{2}$$

Note that in the Tobit model mentioned above, $\beta(x)$ corresponds to the constant slope parameter β.

We now discuss identification of the parameter of interest $\beta(x)$. For the sake of exposition only we momentarily assume that U is a scalar with the Lebesgue density $d\mu$ and that $M(x,u)$ is continuous and monotonic with respect to u for each x. If U and X are independent, the parameter of interest $\beta(x)$ is written as

$$\beta(x) = \int_{u_L(x)}^{u_H(x)} \nabla M(x,u) d\mu(u)/G_M(x), \tag{3}$$

where $u_L(x)$ and $u_H(x)$ solve $M(x,u) = L(x)$ and $M(x,u) = H(x)$, respectively, and $G_M(x) = \Pr\{I_M(X) = 1|X = x\}$. Denote

$$\Psi(x) = E[M(X,U)|X = x, I_M(X) = 1] = \int_{u_L(x)}^{u_H(x)} M(x,u) d\mu(u)/G_M(x).$$

Let us examine the relationship between the derivative of $\Psi(x)$ and $\beta(x)$. The Leibniz integral rule implies

$$\begin{aligned} \nabla[\Psi(x)G_M(x)] = & \int_{u_L(x)}^{u_H(x)} \nabla M(x,u) d\mu(u) \\ & + M(x,u_H(x)) d\mu(u_H(x)) \nabla u_H(x) - M(x,u_L(x)) d\mu(u_L(x)) \nabla u_L(x). \end{aligned} \tag{4}$$

Note that $M(x,u_H(x)) = H(x)$ and $M(x,u_L(x)) = L(x)$. Let $G_H(x) = \Pr\{I_H(X) = 1|X = x\}$ and $G_L(x) = \Pr\{I_L(X) = 1|X = x\}$. Then $\nabla G_H(x) = -d\mu(u_H(x))\nabla u_H(x)$ and

$\nabla G_L(x) = d\mu(u_L(x))\nabla u_L(x)$. Therefore, using these facts and noting $\nabla[\Psi(x)G_M(x)] = G_M(x)\nabla\Psi(x) + \Psi(x)\nabla G_M(x)$, $\beta(x)$ can be written as

$$\beta(x) = \nabla\Psi(x) + [\Psi(x)\nabla G_M(x) + H(x)\nabla G_H(x) + L(x)\nabla G_L(x)]/G_M(x). \tag{5}$$

The second term in (5) corrects for the fact that x affects selection of the population for which Y is observed. Given $X = x$, the correction term can be identified from (i) $\Psi(x)\nabla G_M(x)$, the product of the conditional mean of Y given that Y is uncensored and the derivative of the probability that Y is uncensored, (ii) $H(x)\nabla G_H(x)$, the product of the upper bound $H(x)$ and the derivative of the probability that Y is censored from above, and (iii) $L(x)\nabla G_L(x)$, the product of the lower bound $L(x)$ and the derivative of the probability that Y is censored from below. All components are normalized by $G_M(x)$, the probability that Y is uncensored.[6]

We now consider the general case where U need not be a scalar and continuous and $M(X,U)$ need not be monotonic and continuous in U. In particular, we impose the following assumptions. Let $\mathcal{N}_x$ be a neighborhood of x.

Assumption. *Assume that*

1. *U and X are independent,*

2. *$L(x)$ and $H(x)$ are continuous at x and satisfy $L(x') < H(x')$ for all $x' \in \mathcal{N}_x$,*

3. *$G_L(x)$, $G_M(x)$, and $G_H(x)$ are differentiable at x and $G_M(x) > 0$,*

4. *$M(x',U)$ is continuously differentiable a.s. at each $x' \in \mathcal{N}_x$, and there exists a function B such that for any $x' \in \mathcal{N}_x$, $|\nabla M(x',U)| \le B(U)$ a.s., and $E[B(U)] < \infty$,*

5. $\Pr\{M(X,U) = L(X)|X = x\} = \Pr\{M(X,U) = H(X)|X = x\} = 0$.

The first assumption is stronger than the usual conditional mean independence assumption $E[U|X] = 0$ in a regression framework. However, the maximum likelihood estimator for the Tobit model requires U to be normal and independent of X. In Section 5, we discuss the case of endogenous regressors. The second assumption reflects the definition of L and H as the lower and upper bounds. The fourth assumption is standard and guarantees that one may change the order of differentiation and integration. The rest of the assumptions are natural given that we wish to identify some aspects of derivatives. Here we implicitly assume that all elements of X are continuous.

Under these assumptions, we can show that the derivative formula in (5) still holds true and obtain the main theorem.

Theorem 3.1. *Under Assumptions 1-5, the expression for $\beta(x)$ in (5) holds true.*

The proof is contained in the appendix. We emphasize that this theorem applies to any random object U and that the region of integration need not be rectangular. In particular, U may be a vector and $M(X,U)$ need not be monotone in U. When $L(x) = -\infty$, the term $L(x)\nabla G_L(x)$ does not appear in (5) and when $H(x) = \infty$, the term $H(x)\nabla G_H(x)$ does not appear. In the case of fixed censoring from below at zero (i.e., $L(x) = 0$ and $H(x) = \infty$), the formula is $\beta(x) = \nabla\Psi(x) + \Psi(x)\nabla G_M(x)/G_M(x)$.

[6] Since identification of $\beta(x)$ by (5) requires knowledge of conditional probabilities $G_M(x)$, $G_L(x)$, and $G_H(x)$ or their derivatives, our identification strategy does not apply to truncated dependent variables.

Remark 3.1. [Comparison with control function approach] Consider the standard Tobit model for simplicity. The conventional control function approach (e.g., Heckman, 1976) is: (i) obtain the conditional mean function $E[Y|X = x, Y > 0] = x'\beta + Q(x)$ parametrically or semiparametrically, where $Q(x) = E[U|X = x, Y > 0]$, and then (ii) estimate β and $Q(x)$ jointly. In contrast, our approach is: (i) estimate $\nabla E[Y|X = x, Y > 0] = \beta + \nabla Q(x)$, and then (ii) estimate the correction term $\nabla Q(x)$ to estimate β. More generally, $\beta(x)$ is given by (5), where the last three terms on the right hand side correspond to the correction terms for sample selection. We emphasize that our approach can handle a general nonadditive random object including a random function.

Remark 3.2. [Reduction to one dimensional error] As pointed out by a referee, one could reduce the model (1) with the random object U to one with a scalar error term, say $\tilde{U}$, by setting $\tilde{U} = F(M(X,U)|X)$, where $F(\cdot|X)$ is the conditional distribution function of $M(X,U)$ given X, and defining $\tilde{M}(X,\tilde{U})$ by $\tilde{M}(X,\tilde{U}) = F^{-1}(\tilde{U}|X)$. If $F(\cdot|X)$ is strictly increasing, $\tilde{U}$ is uniformly distributed on $(0,1)$. Clearly $\tilde{M}(X,\tilde{U}) = M(X,U)$, as one may verify by substituting the definition of $\tilde{U}$ in the right-hand side of the definition of $\tilde{M}$, and the estimand $\beta(x)$ remains the same whether we write the model in terms of M and U or $\tilde{M}$ and $\tilde{U}$. Lemma 4.1 in the next section provides another model which implies observationally equivalent $\beta(x)$. In general, there is a potentially large class of models which can yield observationally equivalent $\beta(x)$. Although we could present the above identification result in terms of $\tilde{U}$ and $\tilde{M}$, we prefer the presentation in terms of U and M for the following reasons. First, we regard U and M as primitive objects which have direct economic interpretations, such as unobserved preferences for U and demand functions for M. The transformed objects $\tilde{U}$ and $\tilde{M}$ are rather artificial. Therefore, the conditions of the theorem are more intuitive and easier for an applied researcher to verify when presented in terms of U and M. Second, the representation by $\tilde{U}$ and $\tilde{M}$ does not facilitate the proof of Theorem 3.1. The above intuitive argument only leads to a simple proof if $F(\cdot|X)$ is *strictly* monotonic. This excludes non-strict monotonic M with respect to U, discrete U, and U with non-contiguous support. When U is a general random object such as a utility function, it may require some additional conditions. In these general cases, $\tilde{U}$ is not uniformly distributed but is a mixed distribution of continuous and discrete points. With a mixed distribution of $\tilde{U}$, the domain of integration with respect to $\tilde{U}$ is hard to characterize and the intuitive argument in (3)-(5) under monotonicity and a continuous distribution on a contiguous support cannot be applied directly. One-dimensionality alone does not facilitate the proof, and having to impose some auxiliary assumptions to prove identification might mask the fact that only weak conditions are required on the primitives of the model.

4 Estimation

We can nonparametrically estimate $\beta(x)$ by plugging nonparametric estimators for the unknown functions $\Psi(x)$, $\nabla\Psi(x)$, $G_M(x)$, $\nabla G_M(x)$, $\nabla G_L(x)$, $\nabla G_H(x)$, $L(x)$, and $H(x)$ into the identification formula (5). In Altonji, Ichimura and Otsu (2008) (hereafter, AIO (2008)), we suggest: (i) estimate the conditional mean and derivative functions $\Psi(x)$, $\nabla\Psi(x)$, $G_M(x)$, $\nabla G_M(x)$, $\nabla G_L(x)$, and $\nabla G_H(x)$ by local polynomial regression (see, e.g, Fan and Gijbels, 1996), and (ii) estimate the boundary functions $L(x)$ and $H(x)$ by local polynomial extreme quantile regression (Chernozhukov, 1998, and Ichimura, Otsu and Altonji, 2008), where the quantile point drifts to zero (to estimate $L(x)$) or one (to estimate $H(x)$) at a certain rate as the sample size increases.[7] AIO (2008) show that the nonparametric estimator $\hat{\beta}(x)$ is consistent and is asymptotically normal at the $\sqrt{nh_n^{k+2}}$-rate with the bandwidth h_n for the local polynomial regression.

To avoid the curse of dimensionality, the slower convergence rate of the nonparametric estimator $\hat{\beta}(x)$ for larger k, AIO (2008) also propose an average derivative estimator over a compact subset $\bar{\mathbb{X}}$ of the support of X, that is $\hat{\beta} = n^{-1}\sum_{i=1}^{n} I\{X_i \in \bar{\mathbb{X}}\}\hat{\beta}(X_i)/(n^{-1}\sum_{i=1}^{n} I\{X_i \in \bar{\mathbb{X}}\})$. They show that this estimator is asymptotically normal at the $\sqrt{n}$-rate. An alternative parameter is the average derivative over $\bar{\mathbb{X}}$ conditional on $I_M(X_i) = 1$, which may be estimated as $\hat{\beta}^* = n^{-1}\sum_{i=1}^{n} I\{X_i \in \bar{\mathbb{X}}, I_M(X_i) = 1\}\hat{\beta}(X_i)/(n^{-1}\sum_{i=1}^{n} I\{X_i \in \bar{\mathbb{X}}, I_M(X_i) = 1\})$.

Another way to circumvent the curse of dimensionality is to impose a priori parametric restriction about potential functional forms on $\Psi(x)$, $G_H(x)$, $G_L(x)$, $L(x)$, and $H(x)$ without specifying the distribution of U. The following lemma identifies the conditions where the parametric specification can provide a consistent estimate of $\beta(x)$.

Assumption. *Assume that*

2'. *$L(x)$ and $H(x)$ are continuously differentiable at all x and there exists $\varepsilon > 0$ such that $L(x) + \varepsilon < \Psi(x) < H(x) - \varepsilon$ for all x,*

3'. *$G_L(x)$, $G_M(x)$, and $G_H(x)$ are continuously differentiable at all x and there exist p_1, p_2, p_L, and p_H such that $0 < p_L < G_L(x) < p_1 < p_2 < 1 - G_H(x) < p_H < 1$ for all x,*

4'. *$M(x, U)$ is continuously differentiable a.s. at all x, and there exists a function B such that for any x, $|\nabla M(x, U)| \leq B(U)$ a.s., and $E[B(U)] < \infty$,*

5'. $\Pr\{M(X, U) = L(X)|X = x\} = \Pr\{M(X, U) = H(X)|X = x\} = 0$ *for all x.*

Lemma 4.1. *Suppose Y is generated from (1) and Assumptions 1 and 2'-5' hold. Then there exist a function $\tilde{M}(x, \tilde{u}) = \tilde{M}_0(x) + \tilde{M}_1(x)\tilde{u}_1 + \tilde{M}_2(x)\tilde{u}_2$ with $\tilde{M}_1(x), \tilde{M}_2(x) > 0$ for all x, and random variables $\tilde{U} = (\tilde{U}_1, \tilde{U}_2)$ which are independent of X such that $E[M(X, U)|X = x, I_M(X) = 1] = E[\tilde{M}(X, \tilde{U})|X = x, I_M(X) = 1]$ and $E[\nabla M(X, U)|X = x, I_M(X) = 1] = E[\nabla\tilde{M}(X, \tilde{U})|X = x, I_M(X) = 1]$ for all x.*

Assumptions 2'-5' are global counterparts (over all x) of Assumptions 2-5, respectively. Note that these assumptions are very mild so that we can adopt various parametric functional forms for $\Psi(x)$, $G_H(x)$, $G_L(x)$, $L(x)$, and $H(x)$ to estimate consistently $\beta(x)$. On the other hand, it should be noted that those parametric specifications do not necessarily provide a consistent estimator for the observed joint distribution (Y, X). Also it is remarkable that we do not need to consider more general forms of $\tilde{M}(X, \tilde{U})$ than the one specified in this lemma. The reason is that the parameter of interest in our analysis is the conditional mean of the derivative $\nabla M(X, U)$ rather than the whole function $M(X, U)$ or the distribution of U.[8] For example, suppose that we parametrize $L(x; \theta_L)$, $H(x; \theta_H)$, $G_L(x; \theta_L, \theta_H, \theta_R)$,

[7] For example, in STATA our estimator can be implemented using the *lpoly* and *qreg* packages for local polynomial regression and quantile regression, respectively.

[8] The point made in Remark 3.2 also applies here. Although Lemma 4.1 is useful to assist the search for parametric functional forms, it is more convenient and general to present the identification result in terms of primitive objects U and M, as in Theorem 3.1

$G_H(x;\theta_L,\theta_H,\theta_R)$, and $\Psi(x;\theta_L,\theta_H,\theta_R)$ to satisfy the conditions in this lemma, where the parameters θ_L, θ_H, and θ_R do not overlap. We can first estimate θ_L and θ_H by the extreme quantile regression estimators $\hat{\theta}_L$ and $\hat{\theta}_H$ (Chernozhukov, 2005), and then estimate θ_R by maximizing the criterion function:

$$\begin{aligned}\ell(\theta_R) \quad = \quad & \sum_{i=1}^{n}[I_L(X_i)\log G_L(X_i;\hat{\theta}_L,\hat{\theta}_H,\theta_R) + I_M(X_i)\log G_M(X_i;\hat{\theta}_L,\hat{\theta}_H,\theta_R) \\ & + I_H(X_i)\log G_H(X_i;\hat{\theta}_L,\hat{\theta}_H,\theta_R)] - \sum_{i=1}^{n}(Y_i - \Psi(X_i;\hat{\theta}_L,\hat{\theta}_H,\theta_R))^2 I_M(X_i).\end{aligned}$$

If $\sqrt{n}(\hat{\theta}_L - \theta_L) = o_p(1)$ and $\sqrt{n}(\hat{\theta}_H - \theta_H) = o_p(1)$, standard conditions guarantee the asymptotic normality of the semiparametric estimator for $\beta(x)$ at the $\sqrt{n}$-rate.

5 Extensions

In this section we briefly discuss some extensions of our approach.[9]

5.1 Endogenous Regressors in a Cross Section

One may use a control function approach to handle correlation between X and U. Assume that we have observables W which are not independent of X. Write X as $X = \varphi(W) + V$, where $\varphi(W)$ is defined so that $E[V|W] = 0$ a.s. Assume

$$U \perp W|V. \tag{6}$$

This assumption is strong, but will be hard to avoid unless one is willing to impose additional restrictions on $M(X,U)$, such as monotonicity in a scalar valued function of U. Let $d\mu(\cdot|\cdot)$ be the generic notation for conditional densities. The object $\beta(x)$ in (2) can be written as

$$\beta(x) = \int_v \int_{u\in\{u:I_M(x)=1\}} \{\nabla M(x,u)d\mu(u|x,v)/G_M(x,v)\}d\mu(v|x), \tag{7}$$

where $G_M(x,v) = \Pr\{I_M(X) = 1|X = x, V = v\}$. $G_L(x,v)$ and $G_H(x,v)$ are defined similarly. Let $\Psi(x,v) = E[Y|X = x, V = v, I_M(X) = 1]$. By (6), $\Psi(x,v)$ is written as

$$\Psi(x,v) = \int_{u\in\{u:I_M(x)=1\}} M(x,u)d\mu(u|\varphi(w),v)/G_M(x,v) = \int_{u\in\{u:I_M(x)=1\}} M(x,u)d\mu(u|v)/G_M(x,v).$$

[9] In AIO (2008), we consider measurement error in the outcome. Suppose that $H(x) = \infty$, $L(x)$ is some known constant, and instead of Y and $I_M(X)$, we observe $Y^* = I_R I_M(X)(e_1 Y + e_2)$ and $I_M^* = I_R I_M(X)$, respectively, where the indicator I_R for reporting is 1 with probability p and is 0 with probability $1-p$ that is independent of (X, U, e_1, e_2), the multiplicative measurement error e_1 is a positive random variable with mean μ that is independent of (X, U, I_R), and the additive measurement error e_2 is a random variable with mean 0 that is independent of (X, U, I_R). It is easy to show that if one uses Y^* instead of Y to estimate $\beta(x)$ in (5), then the probability limit of the estimator of $\beta(x)$ is $\mu\beta(x)$. If $L(x)$ has to be estimated, this form of measurement error will lead to bias even for the case of $\mu = 1$. More general forms of measurement error, such as correlation of I_R with (X, U, e_1, e_2), will lead to bias even if $\mu = 1$ and $L(x)$ is known.

Differentiating $\Psi(x,v)$ with respect to x holding v fixed leads to

$$\begin{aligned}\nabla\Psi(x,v) &= \int_{u\in\{u:I_M(x)=1\}} \nabla M(x,u)d\mu(u|v)/G_M(x,v) \\ &\quad -\{H(x)\nabla G_H(x,v)+L(x)\nabla G_L(x,v)+\Psi(x,v)\nabla G_M(x,v)\}/G_M(x,v).\end{aligned}$$

Rearrangement of the above equation establishes $\beta(x)=\int\beta(x,v)d\mu(v|x)$ where

(8)

$$\beta(x,v)=\nabla\Psi(x,v)+\{\Psi(x,v)\nabla G_M(x,v)+H(x)\nabla G_H(x,v)+L(x)\nabla G_L(x,v)\}/G_M(x,v).$$

Taking v as known, the functions on the right hand side of (8) can be estimated using the nonparametric and semiparametric approaches discussed in Section 4. Thus we can estimate $\beta(x,v)$ for each (x,v). The conditional density $d\mu(v|x)$ can be estimated by using the residual $\hat{V}$ of nonparametric regression of X on W as a proxy for V. Integration over v for the product of the estimators of $\beta(x,v)$ and $d\mu(v|x)$ yields an estimator of $\beta(x)$. Also, $\beta(x)=E[\beta(X,V)|X=x]$ can be estimated by nonparametric regression of $\beta(X,\hat{V})$ on X.

Our treatment of endogeneity is closely related to a number of estimation procedures in the literature in which an estimated control variable is used, particularly Smith and Blundell (1986) and Rivers and Vuong (1988) in the context of the Tobit and probit models. Because of nonseparability between X and U, one must use (7) to "undo" the effects of conditioning on V. Blundell and Powell (2004) and Altonji and Matzkin (2001) use the same idea in settings that differ from ours. Chesher (2003) and Imbens and Newey (2009) consider the case in which $X=g(Z,V)$, g is monotone in scalar V, M takes the form of $M(X,V,U)$, and M is monotone in scalar U. See also Matzkin (2003). Following their approach, one can recover V from the cumulative distribution function of X given Z and proceed as outlined above if Z and (V,U) are independent.[10] We suspect that the specification of $M(X,U)$ and estimation method used in Florens *et. al.* (2008) could be used here as well.

A number of papers in the literature discuss estimation in nonseparable models with endogenous variables when a control variable Z that is excluded from X is observed directly and has the property $U\perp X|Z$. If one has such a variable, then one can estimate $\beta(x)$ as $\int\beta(x,z)d\mu(z|x)$, where one obtains $\beta(x,z)$ by replacing v with z, V with Z, and $d\mu(v|x)$ with the conditional density $d\mu(z|x)$ of Z given $X=x$ in the equations leading to $\beta(x,v)$. The problem with this strategy, of course, is that it may be hard to think of applications in which an appropriate Z variable is directly available.

5.2 Endogenous Regressors in a Panel

Suppose that one has panel data observations Y_{it}, X_{it}, and $I_{Mit}=I\{Y_{it}\text{ is uncensored}\}$, where i is a group indicator and t is a time indicator $(t=1,\ldots,T)$. It may be possible to construct a suitable control variable Z from the panel data on X_{it}. Following Altonji and Matzkin (2001, 2005), if the conditional distribution of U_{it} is *exchangeable* in $(X_{i1},\ldots,X_{iT})$,

[10]These papers and others discussed by Blundell and Powell (2003), Chesher (2007), and Matzkin (2007) focus on estimation of $M(x,u)$ and $\nabla M(x,u)$ for given (x,u) as well as the distribution of U. Identifying these objects is much more demanding than identifying an average derivative such as $\beta(x)$ so it is not surprising that stronger assumptions are required. Note that $\beta(x)$ is what Altonji and Matzkin (2005) call a local average response.

then symmetric functions of $(X_{i1}, \ldots, X_{iT})$, such as the group mean of X_{it} for each i, might be a suitable choice for Z_i such that $X_{it} \perp U_{it} \mid Z_i$. See Altonji and Matzkin (2001, 2005) for the details.[11] [12]

In some applications within group variation in U_{it} may be related to X_{it} conditional on Z_i. Following the lines of the papers above, the estimated control variable approach in Section 5.1 can be extended by writing $X_{it} = \varphi(W_{it}, Z_i) + V_{it}$ with $E[V_{it}|W_{it}, Z_i] = 0$ a.s. and assuming $U_{it} \perp W_{it}|Z_i, V_{it}$. In this case, the parameter of interest can be written as $\beta(x) = \int_{z,v} \beta(x, z, v) d\mu(z, v|x_{it} = x)$, where $\beta(x, z, v)$ is defined by replacing v with (z, v), V with (Z, V), and $d\mu(v|x)$ with the conditional density $d\mu(z, v|x)$ of (Z, V) given $X = x$ in (8).

The panel data version of our estimator complements Honoré's (1992) trimmed LAD estimator, which permits one to estimate θ in censored and truncated regression models when $M(X_{it}, U_{it}) = X_{it}\theta + U_{it}$. His estimator is based on differencing the panel observations in clever ways and is quite distinct from our approach. See Arellano and Honoré (2001, Section 7) for additional discussion and references.

5.3 Discrete Regressors

Our identification strategy may be generalized to the case where the regressor vector X contains not only continuous regressors X_C but also discrete ones X_D. Let $\beta_C(x_C, x_D)$ denote the vector of average derivatives of $M(X, U)$ with respect to X_C given $I_M(X) = 1$, $X_C = x_C$, and $X_D = x_D$. It would be straightforward to extend our methods above to allow estimation of $\beta_C(x_C, x_D)$. However, estimation of the effect of X_D raises issues of identification. For simplicity, assume X_D is a scalar binary variable and $L(X) = 0$ and $H(X) = \infty$. There are a number of ways we can define parameters of interest. For example, we can consider identification of

$$\beta_D^{01}(x_C, x_D) = E[I_M(X_C, 1)M(X_C, 1, U) - M(X_C, 0, U)|I_M(X_C, 0) = 1, X_C = x_C].$$

This is the effect of a shift in X_D from 0 to 1 on the average value of Y chosen by those for whom $I_M(x_C, 0) = 1$ (initially uncensored).[13] In our setup, the object $\beta_D^{01}(x_C, x_D)$ is not identified in general. In AIO (2008), we assume that $M(x_C, 0, u') < M(x_C, 0, u'')$ if and only if $M(x_C, 1, u') < M(x_C, 1, u'')$ (see Heckman, Smith and Clements, 1997), and obtain the following estimable bounds for $\beta_D^{01}(x_C, x_D)$:

$$\begin{aligned} & \Psi(x_C, 1) \max\{G_M(x_C, 1) + G_M(x_C, 0) - 1, 0\}/G_M(x_C, 0) - \Psi(x_C, 0) \\ \leq\ & \beta_D^{01}(x_C, x_D) \\ \leq\ & \Psi(x_C, 1)G_M(x_C, 1)/G_M(x_C, 0) - \Psi(x_C, 0). \end{aligned}$$

[11] Exchangeability alone does not restrict the symmetric functions sufficiently to permit one to identify the functions in $\beta(x, z)$ nonparametrically. Consequently, some restrictions on the functions in $\beta(x, z)$ (e.g. linear index restrictions) would be needed.

[12] To identify the average derivative at some t, we need panel data for X_{it} to construct Z_i but cross section data at t are sufficient for Y_{it} and I_{Mit}.

[13] The alternative parameter $E[M(X_C, 1, U) - M(X_C, 0, U)|I_M(X_C, 1) = 1, I_M(X_C, 0) = 1, X_C = x_C]$ can be analyzed analogously.

6 A Monte Carlo Investigation

We now evaluate the finite sample performance of our nonparametric and semiparametric estimators for average derivatives. In Table 1, we report the results of Monte Carlo experiments based on the model

$$\begin{aligned} Y &= \max\{M(X,U), L(X)\}, \\ M(X,U) &= \alpha_0 + \alpha_1 X + \alpha_2 X U_1 + U_2,\ L(X) = a_0 + a_1 X, \end{aligned}$$

where $U = (U_1, U_2)' \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right)$ and $X \sim \text{Uniform}[0,4]$. We consider three cases for the parameter values,

$$\begin{aligned} \text{Case 1} &: (\alpha_0, \alpha_1, \alpha_2, a_0, a_1) = (1, -0.5, 1, 0, 0), \\ \text{Case 2} &: (\alpha_0, \alpha_1, \alpha_2, a_0, a_1) = (1, -0.5, 1, 0, 0.5), \\ \text{Case 3} &: (\alpha_0, \alpha_1, \alpha_2, a_0, a_1) = (1, -0.5, 0, 0, 0). \end{aligned}$$

In Cases 1 and 3, the censoring point $L(X) = 0$ is treated as known. Case 2 requires estimation of the boundary function $L(X)$. In Case 3, the function $M(X,U)$ is linear and separable. So, $\beta(x)$ is constant and the conventional Tobit is the maximum likelihood estimator. The column headings report the values of x at which $\beta(x)$ is evaluated. The column labelled "Avg. β" reports results for the averaged estimator $\hat{\beta} = \sum_{i=1}^n I\{X_i \in \bar{\mathbb{X}}\}\hat{\beta}(X_i)/\sum_{i=1}^n I\{X_i \in \bar{\mathbb{X}}\}$ with $\bar{\mathbb{X}} = [0.5, 3.5]$. The column labelled "Avg. β^*" reports results for $\hat{\beta}^* = \sum_{i=1}^n I\{X_i \in \bar{\mathbb{X}}, I_M(X_i) = 1\}\hat{\beta}(X_i)/\sum_{i=1}^n I\{X_i \in \bar{\mathbb{X}}, I_M(X_i) = 1\}$ and thus weights $\hat{\beta}(X_i)$ by the distribution of X for the uncensored cases. The rows labeled "True Value" reports the true values of $\beta(x)$ when x is 0.5, 1, 1.5, 2, 2.5, 3, and 3.5, and the true values of $E[\beta(X_i)|X_i \in \bar{\mathbb{X}}]$, and $E[\beta(X_i)|X_i \in \bar{\mathbb{X}}, I_M(X_i) = 1]$. Note that $\beta(x)$ varies substantially with x in both Cases 1 and 2, and the variation is due entirely to selection. The rows labelled "AIO-NP" report the results for a nonparametric estimator, the rows labelled "AIO-SP" report the results for a semiparametric estimator, and the rows labelled "Tobit" report the results for the Tobit estimator. The rows labelled "Unadjusted" report the nonparametric estimator of $\nabla\Psi(x)$, the first term of (5). The first rows for each panel of the estimators report the means of the estimates across Monte Carlo replications. The rows labelled with "sd" report the standard deviations of the estimates across the replications. The rows labelled "se" report the means of the asymptotic standard error estimates, and the rows labelled "90%" are the coverages rates of the 90% asymptotic confidence intervals. The sample size is 2,000 and the number of replications is 5,000.

For AIO-NP, we estimate the functions Ψ, $\nabla\Psi$, G_M, and ∇G_M by local second-order polynomial regressions with the Epanechnikov kernel and the rule of thumb bandwidth in Fan and Gijbels (1996). In Case 2, the boundary function $L(x)$ is estimated by local linear quantile regression at the first percentile with the uniform kernel and the bandwidth at 0.5. For AIO-SP, we specify $\Psi(x;\theta_1)$ to be a fourth order polynomial in x plus a constant term and estimate θ_1 by OLS. We do not impose the restriction that the estimated values of $\Psi(x;\theta_1)$ are greater than 0 for all x. For the conditional probability $G_M(x;\theta_2)$, we specify $G_M(x;\theta_2) = \Phi(P(x;\theta_2))$ where $\Phi(\cdot)$ is the standard normal distribution function and $P(x;\theta_2)$ is a fourth order polynomial in x plus a constant. We estimate θ_2 by the maximum likelihood. In Case 2, $L(x)$ is estimated by the linear quantile regression at the first percentile.[14] The Tobit estimation assumes that the analyst does not know the form of $M(X, U)$ and approximates it with a fourth-order polynomial with an additively separable normal error term. In Case 2, the Tobit is estimated assuming the true boundary function $L(x)$ is known.

In Case 1 Tobit substantially underestimates $\beta(x)$ for all values of x. The unadjusted estimator substantially overestimates $\beta(x)$. Thus, the effect of the correction term (i.e., the second term in (5)) is not negligible. The bias in AIO-NP is very small for all values of x, but (surprisingly) is bit larger for Avg. β and Avg β^*. The standard errors of AIO-NP are close to the Monte Carlo standard deviations of the estimates. Coverage rates are close the nominal level of 90%. AIO-SP also exhibits little bias. Not surprisingly, it has a smaller standard deviation than AIO-NP. The standard errors of AIO-SP are close to the Monte Carlo standard deviations and coverage rates are close to 90%.

In Case 2, we need to estimate the boundary function $L(x)$. In spite of this additional complication, the results are similar to Case 1. Both AIO-NP and AIO-SP track $\beta(x)$ closely, while Tobit with knowledge of $L(x)$ and the unadjusted regression estimator are both substantially biased. The standard errors for AIO-NP are overstated and coverage rates are above the nominal value.

In Case 3, Tobit is the maximum likelihood estimator for $\beta(x)$, which equals -0.5 for all values of x. This case is useful for evaluating the efficiency loss of our estimators compared to Tobit. The standard deviations of AIO-NP are about three times larger than those of Tobit for the average derivative and typically about 1/3 larger at specific values of x. The efficiency loss of AIO-SP compared to Tobit is small.

In AIO (2008), we report results for additional parameter values and for designs in which $U_1 = U_2$. The results are generally consistent with those we report here. However, we have found cases in which the fourth-order polynomial for AIO-SP does not provide an adequate approximation over the whole range of x, leading to significant bias for some values of $\beta(x)$. For example, in Case 1 when X is uniform over $[-4, 4]$, AIO-SP underestimates $\beta(x)$ at $x = 2$ by about .21, while AIO-NP tracks $\beta(x)$ closely. One can, of course, alter the bias/variance properties of AIO-NP by changing the bandwidth. Similarly, one can alter the functional forms for AIO-SP. The best choice depends on sample sizes and prior information about functional forms.

Overall, the Monte Carlo results are very encouraging.

References

[1] Altonji, J. G., Hayashi, F. and L. J. Kotlikoff (1997) "Parental altruism and inter vivos transfers: theory and evidence," *Journal of Political Economy*, 105, 1121-1166.

[2] Altonji, J. G. and R. L. Matzkin (2001) "Panel data estimators for nonseparable models with endogenous regressors," Working paper.

[14] In results not reported, both the nonparametric and semiparametric quantile regression estimators for $L(x)$ perform very well in terms of bias and standard deviations.

[3] Altonji, J. G., Ichimura, H. and T. Otsu (2008) "Estimating derivatives in nonseparable models with limited dependent variables," Cowles Foundation Discussion Paper #1668.

[4] Altonji, J. G. and R. L. Matzkin (2005) "Cross section and panel data estimators for nonseparable models with endogenous regressors," *Econometrica*, 73, 1053-1102.

[5] Arellano, M. and B. E. Honoré (2001) "Panel data models: some recent developments," in Heckman, J. J. and E. Learner (eds.), *Handbook of Econometrics*, vol. 5, 3229-3296, Elsevier, Amsterdam.

[6] Blundell, R. and J. L. Powell (2003) "Endogeneity in nonparametric and semiparametric regression models," in Dewatripont, M., Hansen, L. P. and S. J. Turnovsky (eds.) *Advances in Economics and Econometrics: Theory and Applications: Eighth World Congress*, vol. II, Cambridge University Press.

[7] Blundell, R. and J. L. Powell (2004) "Endogeneity in semiparametric binary response models," *Review of Economic Studies*, 71, 655-679.

[8] Chay, K. Y. and J. L. Powell (2001) "Semiparametric censored regression models," *Journal of Economic Perspectives*, 15, 29-42.

[9] Chen, S., Dahl, G. B. and S. Khan (2005) "Nonparametric identification and estimation of a censored location-scale regression model," *Journal of the American Statistical Association*, 100, 212-221.

[10] Chernozhukov, V. (1998) "Nonparametric extreme regression quantiles," Working paper.

[11] Chernozhukov, V. (2005) "Extremal quantile regression," *Annals of Statistics*, 33, 806-839.

[12] Chernozhukov, V., Imbens, G. W. and W. K. Newey (2007) "Instrumental variable estimation of nonseparable models," *Journal of Econometrics*, 139, 4-14.

[13] Chesher, A. (2003) "Local identification in nonseparable models," *Econometrica*, 71, 1405-1441.

[14] Chesher, A. (2005) "Nonparametric identification under discrete variation," *Econometrica*, 73, 1525-1550.

[15] Chesher, A. (2007) "Identification of nonadditive structural functions," in Blundell, R., Persson, T. and W. K. Newey (eds.) *Advances in Economics and Econometrics: Theory and Applications: Ninth World Congress*, vol. III, Cambridge University Press.

[16] Fan, J. and I. Gijbels (1996) "*Local Polynomial Modelling and Its Applications*," Chapman & Hall/CRC.

[17] Florens J. P., Heckman, J. J., Meghir, C. and E. Vytlacil (2008) "Identification of treatment effects using control functions in models with continuous, endogenous treatment and heterogeneous effects," *Econometrica*, 76, 1191-1206.

[18] Han, A. K. (1987) "Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator," *Journal of Econometrics*, 35, 303-316.

[19] Heckman, J. J. (1976) "The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models," *Annals of Economic and Social Measurement*, 5, 475-492.

[20] Heckman, J. J., Smith, J. and N. Clements (1997) "Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts," *Review of Economic Studies*, 64, 487-535.

[21] Hoderlein, S. and E. Mammen (2007) "Identification of marginal effects in nonseparable models without monotonicity," *Econometrica*, 75, 1513-1518.

[22] Honoré, B. E. (1992) "Trimmed LAD and least squares estimation of truncated and censored regression models with fixed effects," *Econometrica*, 60, 533-565.

[23] Imbens, G. W. and W. K. Newey (2009) "Identification and estimation of triangular simultaneous equations models without additivity," *Econometrica*, 77, 1481-1512.

[24] Ichimura, H., Otsu, T. and J. G. Altonji (2008) "Nonparametric intermediate order regression quantiles," Working paper.

[25] Lewbel, A. and O. Linton (2002) "Nonparametric censored and truncated regression," *Econometrica*, 70, 765-779.

[26] Matzkin, R. L. (1991) "A nonparametric maximum rank correlation estimator," in Barnett, W. A., Powell, J. L. and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

[27] Matzkin, R. L. (2003) "Nonparametric estimation of nonadditive random functions," *Econometrica*, 71, 1339-1375.

[28] Matzkin, R. L. (2007) "Nonparametric identification," in Engle, R. F. and D. L. McFadden (eds.), *Handbook of Econometrics*, vol. 6, part 2, 5307-5368, Elsevier, Amsterdam.

[29] Powell, J. L. (1991) "Estimation of monotonic regression models under quantile restrictions," in Barnett, W. A., Powell, J. L. and G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, Cambridge University Press.

[30] Powell, J. L., Stock, J. H. and T. M. Stoker (1989) "Semiparametric estimation of index coefficients," *Econometrica*, 57, 1403-1430.

[31] Rivers, D. and Q. Vuong (1988) "Limited information estimators and exogeneity tests for simultaneous probit models," *Journal of Econometrics*, 39, 347-366.

[32] Smith, R. J. and R. W. Blundell (1986) "An exogeneity test for a simultaneous equation Tobit model with an application to labor supply," *Econometrica*, 54, 679-685.

[33] Stoker, T. M. (1986) "Consistent estimation of scaled coefficients," *Econometrica*, 54, 1461-1481.

A Appendix

A.1 Proof of Theorem 3.1

It is sufficient to prove the derivative formula (4) for ∇_1, the partial derivative with respect to the first element of x, i.e.,

(9)
$$\nabla_1 \int M(x,u) I_M(x) d\mu(u) = \int \nabla_1 M(x,u) I_M(x) d\mu(u) - H(x)\nabla_1 G_H(x) - L(x)\nabla_1 G_L(x).$$

The left hand side of (9) is written as

$$\begin{aligned}
& \lim_{\varepsilon\to 0} \left[\int M(x+\varepsilon\mathbf{e_1}, u) I_M(x+\varepsilon\mathbf{e_1}) d\mu(u) - \int M(x,u) I_M(x) d\mu(u) \right] /\varepsilon \\
= & \lim_{\varepsilon\to 0} \int \left[M(x+\varepsilon\mathbf{e_1}, u) - M(x,u) \right] I_M(x+\varepsilon\mathbf{e_1}) d\mu(u)/\varepsilon \\
& + \lim_{\varepsilon\to 0} \int M(x,u) \left[I_M(x+\varepsilon\mathbf{e_1}) - I_M(x) \right] d\mu(u)/\varepsilon \ = T_1 + T_2,
\end{aligned}$$

where $\mathbf{e_1} = \pm(1,0,\ldots,0)$. Assumptions 2, 4, and 5 imply $\lim_{\varepsilon\to 0} I_M(x+\varepsilon\mathbf{e_1}) = I_M(x)$ a.s. Thus, Assumption 4 and the Lebesgue dominated convergence theorem imply that $T_1 = \int \nabla_1 M(x,u) I_M(x) d\mu(u)$. We now consider T_2. By the definition of I_M and Assumption 2,

$$\begin{aligned}
I_M(x+\varepsilon\mathbf{e_1}) - I_M(x) = & \left[I\{L(x+\varepsilon\mathbf{e_1}) < M(x+\varepsilon\mathbf{e_1}, U)\} + I\{M(x+\varepsilon\mathbf{e_1}, U) < H(x+\varepsilon\mathbf{e_1})\} \right] \\
& - \left[I\{L(x) < M(x,U)\} + I\{M(x,U) < H(x)\} \right],
\end{aligned}$$

a.s. for all ε sufficiently close to zero. So, T_2 can be written as

$$\begin{aligned}
T_2 = & \lim_{\varepsilon\to 0} \int M(x,u) \left[I\{L(x+\varepsilon\mathbf{e_1}) < M(x+\varepsilon\mathbf{e_1}, u)\} - I\{L(x) < M(x,u)\} \right] d\mu(u)/\varepsilon \\
& + \lim_{\varepsilon\to 0} \int M(x,u) \left[I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} - I\{M(x,u) < H(x)\} \right] d\mu(u)/\varepsilon.
\end{aligned}$$

Since $I\{L(x+\varepsilon\mathbf{e_1}) < M(x+\varepsilon\mathbf{e_1}, u)\} = 1 - I\{M(x+\varepsilon\mathbf{e_1}, u) \leq L(x+\varepsilon\mathbf{e_1})\}$ for all ε sufficiently close to zero, the following lemma completes the proof.

Lemma A.1. *Under Assumptions 1-5,*

(10)
$$\begin{aligned}
& \lim_{\varepsilon\to 0} \int M(x,u) \left[I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} - I\{M(x,u) < H(x)\} \right] d\mu(u)/\varepsilon \\
& = -H(x)\nabla_1 G_H(x).
\end{aligned}$$

Proof. It is sufficient to show that both an upper bound and a lower bound of the left hand side of (10) converge to the right hand side as $\varepsilon \to 0$. The left hand side of (10) equals

$$\begin{aligned}
& \lim_{\varepsilon\to 0} \int M(x,u) I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) \geq H(x)\} d\mu(u)/\varepsilon \\
& - \lim_{\varepsilon\to 0} \int M(x,u) I\{M(x+\varepsilon\mathbf{e_1}, u) \geq H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) < H(x)\} d\mu(u)/\varepsilon. \qquad (11)
\end{aligned}$$

Since the argument is analogous, we only show the result for an upper bound.

For any small $\varepsilon > 0$ that satisfies the neighborhood condition in Assumption 4, by the mean value theorem there exists $0 < \tilde{\varepsilon} < \varepsilon$ such that $M(x+\varepsilon\mathbf{e_1}, U) = M(x, U) + \nabla M(x + \tilde{\varepsilon}\mathbf{e_1}, U)\varepsilon$ a.s. Thus, by Assumption 4

$$\begin{aligned} M(x+\varepsilon\mathbf{e_1}, U) &\leq M(x,U) + \sup_{0<\tilde{\varepsilon}<\varepsilon} \nabla M(x+\tilde{\varepsilon}\mathbf{e_1}, U)\varepsilon \leq M(x,U) + \sup_{x' \in N(x,\varepsilon)} \nabla M(x', U)\varepsilon \\ &\leq M(x,U) + B(U)\varepsilon, \end{aligned}$$

a.s., where $N(x, \varepsilon)$ is a neighborhood around x with radius ε. Analogously by replacing the supremum with the infimum, we can show that $M(x + \varepsilon e_1, U) \geq M(x, U) - B(U)\varepsilon$ a.s.

By these inequalities, (11) can be bounded from above by

$$\begin{aligned} &\lim_{\varepsilon\to 0} \int H(x+\varepsilon\mathbf{e_1}) I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) \geq H(x)\} d\mu(u)/\varepsilon \\ &+ \lim_{\varepsilon\to 0} \int B(u) I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) \geq H(x)\} d\mu(u) \\ &- \lim_{\varepsilon\to 0} \int H(x+\varepsilon\mathbf{e_1}) I\{M(x+\varepsilon\mathbf{e_1}, u) \geq H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) < H(x)\} d\mu(u)/\varepsilon \\ &+ \lim_{\varepsilon\to 0} \int B(u) I\{M(x+\varepsilon\mathbf{e_1}, u) \geq H(x+\varepsilon\mathbf{e_1})\} I\{M(x,u) < H(x)\} d\mu(u). \end{aligned}$$

By Assumptions 2, 4, and 5, the Lebesgue dominated convergence theorem implies that the second term and the fourth term converge to zero. The first term and the third term can be rewritten as

$$\lim_{\varepsilon\to 0} H(x+\varepsilon\mathbf{e_1}) \int \left[I\{M(x+\varepsilon\mathbf{e_1}, u) < H(x+\varepsilon\mathbf{e_1})\} - I\{M(x,u) < H(x)\} \right] d\mu(u)/\varepsilon$$

which is the right hand side of (10) under Assumptions 2 and 3. The conclusion is obtained.

A.2 Proof of Lemma 4.1

The basic idea of the proof is as follows. First, independently from x, we pick any strictly increasing distribution functions F_1 and F_2 with continuous densities f_1 and f_2 such that

$$\sup_x |H(x) - L(x)| \{F_j^{-1}(p_H) - F_j^{-1}(p_L)\} \max_{\tilde{u}_j \in [F_j^{-1}(p_L), F_j^{-1}(p_H)]} f_j(\tilde{u}_j) < 2\varepsilon(p_2 - p_1), \tag{12}$$

for $j = 1, 2$, where p_1, p_2, p_L, and p_H satisfy Assumption 3' and ε satisfies Assumption 2'. Since $\sup_x |H(x) - L(x)|$ is bounded by a constant from Assumption 3', it is possible to choose such F_1 and F_2. We set the joint density of $\tilde{U}$ as $f_{\tilde{U}}(\tilde{u}_1, \tilde{u}_2) = f_1(\tilde{u}_1) f_2(\tilde{u}_2)$. Second, we pick any point x. Third, for the given x, we show the existence of $(\tilde{M}_0(x), \tilde{M}_1(x), \tilde{M}_2(x))$ satisfying the equivalence $\Psi(x) = E[M(X,U)|X = x, I_M(X) = 1] = E[\tilde{M}(X,\tilde{U})|X = x, I_M(X) = 1]$. Fourth, observe that we can apply this argument for any x with the same F_1 and F_2 above to show the equivalence on $\Psi(x)$ for all x (note: F_1 and F_2 do not depend on x by definition). Finally, showing that $\tilde{M}(x, \tilde{u})$ is differentiable in x (for almost every $\tilde{u}$) implies $E[\nabla M(X,U)|X = x, I_M(X) = 1] = E[\nabla \tilde{M}(X, \tilde{U})|X = x, I_M(X) = 1]$. Hereafter, we show the third and final steps.

Proof of the third step. For given x, we want to find $(\tilde{M}_0(x), \tilde{M}_1(x), \tilde{M}_2(x))$ such that $\tilde{M}(x,\tilde{u}) = \tilde{M}_0(x) + \tilde{M}_1(x)\tilde{u}_1 + \tilde{M}_2(x)\tilde{u}_2$, $G_L(x) = \Pr\{\tilde{M}(X,\tilde{U}) \le L(X)|X = x\}$, $G_H(x) = \Pr\{\tilde{M}(X,\tilde{U}) \ge H(X)|X = x\}$, and $\Psi(x) = E[\tilde{M}(X,\tilde{U})|I_M(X) = 1, X = x]$. For notational convenience, we hereafter drop the arguments x from functions and suppress tilde, denoting $(\tilde{M}_0(x), \tilde{M}_1(x), \tilde{M}_2(x))$ as (M_0, M_1, M_2) and $(\tilde{u}_1, \tilde{u}_2)$ as (u_1, u_2). Note that

(13)
$$G_L = \int_{-\infty}^{\infty} f_1(u_1) F_2\left(\frac{L - M_0 - M_1 u_1}{M_2}\right) du_1 = \int_{-\infty}^{\infty} f_2(u_2) F_1\left(\frac{L - M_0 - M_2 u_2}{M_1}\right) du_2,$$

(14)
$$1 - G_H = \int_{-\infty}^{\infty} f_1(u_1) F_2\left(\frac{H - M_0 - M_1 u_1}{M_2}\right) du_1 = \int_{-\infty}^{\infty} f_2(u_2) F_1\left(\frac{H - M_0 - M_2 u_2}{M_1}\right) du_2,$$

$$\begin{aligned} \Psi G_M &= M_0 G_M + M_1 \int_{-\infty}^{\infty} u_1 f_1(u_1) \left[F_2\left(\frac{H - M_0 - M_1 u_1}{M_2}\right) - F_2\left(\frac{L - M_0 - M_1 u_1}{M_2}\right)\right] du_1 \\ &\quad + M_2 \int_{-\infty}^{\infty} u_2 f_2(u_2) \left[F_1\left(\frac{H - M_0 - M_2 u_2}{M_1}\right) - F_1\left(\frac{L - M_0 - M_2 u_2}{M_1}\right)\right] du_2. \qquad (15) \end{aligned}$$

Reparameterize so that $\lambda = M_1/M_2$. By holding λ constant, we can find $M_0^*(\lambda)$ and $M_2^*(\lambda)$ that solve (13) and (14) with respect to M_0 and M_2, respectively. Let l_λ and h_λ denote the solutions to $G_L = \int_{-\infty}^{\infty} f_1(u_1) F_2(l_\lambda - \lambda u_1) du_1$ and $1 - G_H = \int_{-\infty}^{\infty} f_1(u_1) F_2(h_\lambda - \lambda u_1) du_1$, respectively. Then by the definitions, $M_0^*(\lambda)$ and $M_2^*(\lambda)$ are written as $M_0^*(\lambda) = \frac{h_\lambda L - l_\lambda H}{h_\lambda - l_\lambda}$ and $M_2^*(\lambda) = \frac{H-L}{h_\lambda - l_\lambda}$. By substituting these solutions, the right hand side of the expression for ΨG_M above can be regarded as a function of λ (denote the function by $m(\lambda)$). Thus, for the conclusion it is sufficient to check the existence of $\lambda^* > 0$ that solves $\Psi G_M = m(\lambda)$. Note that $m(\lambda)$ is continuous in λ because of the continuity of F_1 and F_2. Thus, by the intermediate value theorem and Assumption 2', the existence of λ^* can be verified by showing

$$\lim_{\lambda \to 0} m(\lambda) < (L + \varepsilon) G_M, \quad \lim_{\lambda \to \infty} m(\lambda) > (H - \varepsilon) G_M, \qquad (16)$$

for some $\varepsilon > 0$ satisfying Assumption 2'.

We now show the first statement of (16). Note that $h_\lambda \to h_0$ and $l_\lambda \to l_0$ as $\lambda \to 0$, where h_0 and l_0 solve $F_2(h_0) = 1 - G_H$ and $F_2(l_0) = G_L$, respectively, and that $m(\lambda) \to LG_M + \frac{H-L}{h_0 - l_0} \int_{l_0}^{h_0} (u - l_0) f_2(u) du$ as $\lambda \to 0$. Since $G_M > p_2 - p_1$ by Assumption 3', the requirement (12) on F_2 implies the first statement of (16). Similarly, since $m(\lambda) \to HG_M - \frac{H-L}{h_\infty - l_\infty} \int_{l_\infty}^{h_\infty} (h_\infty - u) f_1(u) du$ as $\lambda \to \infty$ (where h_∞ and l_∞ solve $F_1(h_\infty) = 1 - G_H$ and $F_1(l_\infty) = G_L$, respectively), the requirement (12) on F_1 implies the second statement of (16). This completes the proof of the third step.

Proof of the final step. Since $(\tilde{M}_0(x), \tilde{M}_1(x), \tilde{M}_2(x))$ satisfies (13)-(15) for all x and Assumptions 2'-4' guarantee the differentiability of $(\tilde{M}_0(x), \tilde{M}_1(x), \tilde{M}_2(x))$, it follows that $\tilde{M}(x, \tilde{u})$ is differentiable in x for almost every $\tilde{u}$.

Table 1:

Case 1: $M(X,U) = 1 - 0.5X + XU_1 + U_2$, $L(X) = 0$, percentage uncensored: 54.7%									
	Evaluation Point of $\beta(x)$							**Avg.**β	**Avg.**β^*
	0.5	**1**	**1.5**	**2**	**2.5**	**3**	**3.5**		
True Value	**-0.310**	**-0.085**	**0.092**	**0.214**	**0.297**	**0.355**	**0.398**	**0.156**	**0.120**
AIO-NP	-0.298	-0.097	0.063	0.189	0.270	0.352	0.409	0.0985	0.0986
sd	0.164	0.134	0.182	0.217	0.286	0.332	0.638	0.184	0.184
se	0.206	0.146	0.178	0.220	0.270	0.327	0.665		
90%	0.959	0.925	0.899	0.904	0.892	0.902	0.909		
AIO-SP	-0.295	-0.078	0.075	0.200	0.305	0.374	0.391	0.158	0.123
sd	0.146	0.141	0.135	0.140	0.222	0.239	0.521	0.065	0.057
se	0.146	0.143	0.137	0.141	0.223	0.240	0.525		
90%	0.899	0.903	0.902	0.904	0.906	0.904	0.900		
Tobit	-0.626	-0.426	-0.241	-0.081	0.041	0.116	0.131	-0.137	-0.178
sd	0.182	0.157	0.140	0.120	0.186	0.176	0.375	0.048	0.049
Unadjusted	0.050	0.300	0.447	0.522	0.562	0.592	0.608	0.411	0.411
sd	0.134	0.106	0.139	0.163	0.212	0.244	0.466	0.128	0.129
Case 2: $M(X,U) = 1 - 0.5X + XU_1 + U_2$, $L(X) = 0.5X$, percentage uncensored: 40.2%									
True Value	**-0.260**	**0.064**	**0.318**	**0.486**	**0.599**	**0.676**	**0.732**	**0.401**	**0.312**
AIO-NP	-0.224	0.043	0.272	0.458	0.572	0.653	0.752	0.403	0.360
sd	0.167	0.154	0.206	0.257	0.326	0.384	0.743	0.261	0.269
se	0.270	0.190	0.225	0.279	0.352	0.442	0.909		
90%	0.991	0.955	0.929	0.925	0.937	0.942	0.953		
AIO-SP	-0.234	0.078	0.314	0.492	0.615	0.680	0.709	0.408	0.323
sd	0.154	0.151	0.142	0.160	0.248	0.267	0.609	0.074	0.060
se	0.153	0.152	0.145	0.159	0.248	0.269	0.607		
90%	0.893	0.904	0.903	0.902	0.902	0.893	0.896		
Tobit	-0.709	-0.357	-0.052	0.193	0.370	0.466	0.473	0.090	-0.023
sd	0.183	0.159	0.139	0.122	0.186	0.176	0.376	0.045	0.051
Unadjusted	0.351	0.662	0.830	0.913	0.951	0.967	0.991	0.775	0.775
sd	0.134	0.116	0.154	0.186	0.238	0.277	0.524	0.142	0.144
Case 3: $M(X,U) = 1.0 - 0.5X + U_2$, $L(X) = 0$, percentage uncensored: 50.0%									
True Value	**-0.5**	**-0.5**	**-0.5**	**-0.5**	**-0.5**	**-0.5**	**-0.5**	**-0.5**	**-0.5**
AIO-NP	-0.500	-0.499	-0.499	-0.497	-0.493	-0.499	-0.507	-0.509	-0.508
sd	0.148	0.090	0.098	0.097	0.112	0.120	0.204	0.073	0.073
se	0.154	0.089	0.092	0.098	0.107	0.122	0.243		
90%	0.917	0.896	0.890	0.906	0.891	0.915	0.948		
AIO-SP	-0.501	-0.503	-0.503	-0.499	-0.497	-0.500	-0.513	-0.501	-0.501
sd	0.125	0.080	0.082	0.063	0.092	0.096	0.181	0.026	0.027
se	0.125	0.079	0.080	0.064	0.092	0.094	0.183		
90%	0.899	0.895	0.893	0.898	0.900	0.892	0.900		
Tobit	-0.498	-0.504	-0.502	-0.498	-0.495	-0.499	-0.514	-0.501	-0.501
sd	0.120	0.077	0.074	0.056	0.082	0.081	0.159	0.024	0.026
Unadjusted	-0.279	-0.244	-0.212	-0.183	-0.155	-0.136	-0.118	-0.191	-0.191
sd	0.121	0.070	0.073	0.072	0.076	0.082	0.141	0.047	0.047