SELECTION MODEL AND CONDITIONAL TREATMENT EFFECTS, INCLUDING ENDOGENOUS COVARIATES

Arthur Lewbel
Boston College

September 2001

Abstract

In a sample selection or treatment effects model, common unobservables may affect both the outcome and the probability of selection in unknown ways. This paper shows that the distribution function of potential outcomes and treatment effects may still be identified if an observed variable V affects the treatment or selection probability in certain ways and is conditionally independent of the unobservables. Estimators based on this identification are provided, which take the form of simple weighted averages.

A special case is a two stage least squares estimator of the coefficients in a linear selection model, which permits endogeneous or mismeasured regressors. An application to estimation of firm investment decisions is provided.

Portions of this paper were previously circulated under the title "Two Stage Least Squares Estimation of Endogeneous Sample Selection Models".

JEL Codes: C14, C25, C13. Keywords: Sample Selection, Treatment Effects, Censoring, Semiparametric, Endogeneous, Instrumental Variables, Switching Regressions, Heteroscedasticity, Latent Variable Models.

* This research was supported in part by the NSF, grant SES-9905010. I'd like to thank Yuriy Tchamourliyski for research assistance, and Edward Vytlacil, Jim Powell, Jim Heckman, Fabio Schiantarelli, and anonymous referees for helpful comments. Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu.

1 Introduction

Assume that for a sample of individuals we observe an indicator D that equals one if the individual is treated or selected, and zero otherwise. If the individual is treated or selected we observe an outcome or response Y, otherwise let $Y = 0$. Define Y^* to equal the observed outcome Y when $D = 1$, otherwise Y^* equals the outcome that would have been observed if D had equaled one (i.e, a counterfactual), so $Y = DY^*$. We may also observe a covariate vector X. Treatment or selection D may be unconditionally or conditionally correlated with Y^*, so Y^* and D may depend in unknown ways on common unobservables.

This paper provides estimators for conditional and unconditional moments of Y^*. Given an arbitrary function g, estimators for $E[g(Y^*, X) \mid X = x]$ and $E[g(Y^*, X)]$ are provided, along with their limiting normal distributions. In particular, letting $g(\psi, x) = I(\psi \leq y^*)$ yields estimators for $F(y^* \mid X = x)$ and $F(y^*)$, the conditional and unconditional distribution functions of Y^*. This paper shows how to construct a variable W such that $E[g(Y^*, X)] = E[Wg(Y, X)]/E(W)$.

This last result is of special interest because it implies that if the unobserved Y^* is given by $Y^* = X_1^T\beta + \varepsilon$ and if $E(Z\varepsilon) = 0$, then the coefficients β (including the intercept) can be estimated by an ordinary linear two stage least squares regression of WY on WX_1, using instruments Z. The instruments Z would only need to possess the usual properties of linear regression model instruments. This then permits estimation of β in a selection model in the presence of endogeneous or mismeasured regressors, and general forms of heteroscedasticity.

Consider the usual treatment model where Y° is an observed outcome and T is a binary treatment indicator. Then this paper's estimators can be applied with $Y = TY^\circ$ and $D = T$ to obtain features of the population (unconditional or conditioned on $X = x$) if everyone were treated, and the estimators can be reapplied with $Y = (1 - T)Y^\circ$ and $D = 1 - T$ to obtain the corresponding features of the population if no one were treated. This (along with readily estimated objects like the conditional distribution of Y° given $T = 1$) then permits recovery of average conditional treatment effects, effects of treatment on the treated, and general welfare calculations associated with treatment.

Many estimators exist for treatment, sample selection and censored regression models. Standard maximum likelihood estimation requires that the entire joint distribution of the unobservables, conditional on covariates or instruments, be finitely parameterized. In particular, the selection equation (and the endogeneous regressors as functions of instruments) would need to be completely specified. Parametric model estimators other than ML consist of specifying enough features of this conditional distribution to permit identification. See, e.g., Heckman (1974, 1976, 1979), Rubin (1974), Koul, Susarla, and van Ryzin (1981),

and Lee (1982).

Semiparametric estimators of sample selection models include Powell (1987), Newey (1988), (1999), Choi (1990), Cosslett (1991), Ichimura and Lee (1991), Lee (1992), Lee (1994), and Ahn and Powell (1993). Donald (1995), Wooldridge (1995), Kyriazidou (1997), Andrews and Schafgans (1998), Chen and Lee (1998), Das (1998), Vella and Verbeek (1999), and Das, Newey, and Vella (2000). Recent treatment related estimators include Imbens and Angrist (1994), Heckman, Ichimura, and Todd (1998), Hahn (1998), and Hirano, Imbens and Ridder. (2000) See also Heckman (1990), Manski (1994), and Chamberlain (1986). Surveys include Heckman and MaCurdy (1986), Wainer (1986), Powell (1994), and Vella (1998).

This paper's proposed estimators have many attractive features. Unlike other estimators, they do not require estimation of the propensity score (i.e., the selection equation, the conditional expectation of D given X). Very few assumptions about the unobservables are required. For example, unlike typical treatment model estimators, it is not required that D and Y^* be conditionally independent or have a finitely parameterized joint distribution, nor is the distribution of Y^* required to be either discrete or continuous. Correlated or common unobservables can affect both D and Y^* in unknown ways. Also, in the case where $Y^* = X_1^T\beta + \varepsilon$ and one or more elements of X_1 may be endogeneous, the estimator for β proposed here is numerically simple (equivalent to a linear two stage least squares regression), imposes few assumptions about X_1 as a function of instruments Z, and does not require estimation of $E(X_1 \mid Z)$.

The price paid for these advantages is that the estimators in this paper require the existence of an observed variable V that satisfies a strong set of conditions, summarized below. It also requires estimation of the conditional (on X) probability density function of V. Essentially, the estimators in this paper work by replacing the usual strong assumptions about all unobservables with strong assumptions about one observable covariate. Another disadvantage is that, for typical one sided censoring models, many of the estimators proposed here will possess an arbitrarily small but non-zero asymptotic bias.

The assumptions based on V differ markedly from the usual identifying assumptions regarding unobservables. The estimators proposed here may therefore prove useful as checks of model robustness, that is, one would have greater confidence in estimates produced by more conventional estimators if the estimators provided here, based on very different assumptions, yield similar results. Of course, in any particular application, these assumptions may be either more or less plausible than the usual ones.

The difficulty in estimating moments of Y^* is that D may covary with Y^*, even after conditioning on observed covariates. Equivalently, common unobservables may affect both Y^* and D (in other terminology, the unconfoundness assumption may not hold). This paper shows that conditional moments of Y^* can

be identified and estimated if a single well behaved observed covariate V exists such that $D = I(a_0 \leq M + V \leq a_1)$ where a_0 and a_1 are constants (one of which could be infinite), M is an unobserved latent variable, and I is the indicator function that equals one if its argument is true and zero otherwise. It will also be required that V have a large support, and be conditionally independent of M and Y^*, conditioning on X. Virtually no other restriction is placed on the joint distribution of M and Y^* (which corresponds to the joint distribution of D and Y^*). In particular, M and Y^* can depend in unknown ways on common unobservables. Identification is obtained by the presence of V.

As some motivation for this structure, it is shown that, given regularity, if the probability of $D = 1$ is monotonic in V, then the above expression for D holds. This structure also generalizes the usual parametric selection model, which is of this form with M linear in X and an additive error term.

Define *two sided censoring* to be the case where both a_0 and a_1 are finite, while *one sided censoring* is when either $a_0 = -\infty$ or $a_1 = \infty$. One sided censoring implies D is monotonic in V. Consistent estimators will be proposed for both one and two sided censoring, though consistency with one sided censoring will require some unbounded support and asymptotic trimming assumptions. A convenient result will be that the same bounded support estimators that are consistent with two sided censoring will only have an arbitrarily small asymptotic bias when applied (without change) to one sided censoring models. This implies that, in applications, the estimator could be used without even specifying whether the censoring is one or two sided.

The estimators proposed here can be applied, without change, when the unobservables in the outcome and selection equations are perfectly correlated. For example, the estimators may be applied to censored regression models. In that case we would have $M = Y^*$, and V would be the random censoring point.

This paper's model and associated estimators extend readily to the case of ordered selection or ordered treatment models. Let Y_j^* denote the outcome or response that would result from a treatment t_j, where $\{t_0, ..., t_J\}$ is the set of possible treatments. In the ordered selection or treatment model we observe covariates V, X, treatment T, and outcome Y°, where $Y^\circ = \sum_{j=1}^{J} Y_j^* I(T = t_j)$ and $T = t_j I(a_j \leq M + V < a_{j+1})$ for some constants $a_0, ..., a_{J+1}$ and unobserved latent M. The conditional or unconditional moments and distribution of each possible outcome Y_j^* could then be obtained by applying this paper's estimator with $D = I(T = t_j)$ and $Y = Y^\circ D$ for each possible treatment t_j. Similarly, if $Y_j^* = X^T \beta_j + \varepsilon_j$, then this corresponds to a switching regression model and the coefficient vectors β_j may be estimated even in the presence of endogeneous regressors.

The latent index $M + V$ may be interpreted as a monotonic transformation of a desired continuous level of treatment. Vytlacil (2000) demonstrates an equivalence between latent index treatment models and

alternative restrictions on the selection process. If V were a cost or a price, then differing values of M could correspond to the willingness to pay for different levels of treatment t_j. Using, e.g., results from Lewbel, Linton, and McFadden (2001) one may estimate the conditional distribution of this willingness to pay M given $X = x$.

Consider the classic wage equation as an example of a selection model (Gronau 1974, Heckman 1974, 1976), where Y^* is an individual's wage and $D = 1$ if the individual is employed, both of which depend on unobservables such as ability, as well as on observable covariates X such as measures of schooling or training. Then an appropriate V would be some form of nonwage income that, conditional on X, is independent of ability, e.g., government defined benefits.

Another class of examples are models involving a range of inaction. Here Y is some continuous decision variable such as quantity to consume or money to spend on investment. Action is taken, that is, $D = 1$, only if the amount to be spent or purchased (or some other decision variable) exceeds some threshold level. For example, investment only takes place if the return from the investment exceeds some fixed cost associated with investment. Then V would be a variable that affects the fixed cost, and hence the threshold, but doesn't affect the return on the investment. The selection variable D is then the indicator of whether the threshold is exceeded. Two sided censoring models can also arise in the context of range of action models.

This paper provides an empirical application in which plant level investment decisions of firms are modeled. The size of the plant V affects fixed costs of investment, and hence affects D, and a Tobin's Q type model determines Y when $D = 1$.

Other examples of models containing a suitable V are provided by Lewbel, Linton and McFadden (2001), Maurin (1999), and Alonso, Fernandez, and Rodriguez-Póo (1999). Each of these applications require a V to estimate binary choice models (equivalent in the present context to estimation of just the selection or treatment equation). Lewbel, Linton and McFadden (2001) consider applications like willingness to pay studies, where V is a bid determined by experimental design. Maurin (1999) applies Lewbel's (2000) estimator in a model of whether students repeat a grade in elementary school, using date of birth as V. Alonso, Fernandez, and Rodriguez-Póo (1999) use individual's age as V in a duration model application.

1.1 How the Estimator Works

To illustrate how identification is obtained here, consider estimation of $E(Y^*)$ in a simple case without covariates X. For this illustration, we observe draws of Y, D, V, where $Y = Y^*D$, $D = I(a_0 \leq M + V \leq a_1)$ with a_0 and a_1 finite (two sided censoring), and V is independent of M and Y^* with a large support.

The naive estimator of $E(Y^*)$ is $E(Y)/E(D)$, but this is biased in general because Y^* and D are correlated. By the law of iterated expectations, we have $E(Y) = E[Y^*E(D \mid Y^*)]$, so if we were lucky enough to have $E(D \mid Y^*)$ be constant (which here with D binary is equivalent to D and Y^* independent), then $E(Y)/E(D)$ would equal $E(Y^*)$. Given our identifying assumptions,

$$E(D \mid Y^*) = prob\{V \in [a_0 - M, a_1 - M] \mid Y^*\}$$

If V had an independent uniform distribution that included the interval $[a_0-M, a_1-M]$, then the probability that V lies in this interval would be proportional to $a_1 - a_0$, and hence constant. In that case $cov(Y^*, D)$ would equal zero and the naive estimator would work.

The key here is that the propensity score (the conditional probability of treatment or selection) equals the probability that V lies in an interval, and while this interval depends on the unknown latent M, the *length* of this interval is constant, so if V were uniform, then the conditional probability of treatment would be constant.

Now consider instead the estimator $E(WY)/E(W)$ where $W = D/f_v(V)$ and f_v is the pdf of V. This division by the density of v is equivalent to converting V to a uniformly distributed random variable, and so by the above logic this estimator will yield the desired $E(Y^*)$. Formally, we have

$$\begin{aligned}
E(WY) &= E[E(WY \mid M, Y^*)] \\
&= E\left[\int_{\text{supp}(V)} \frac{I(a_0 \leq M + v \leq a_1)Y^*}{f_v(v)} f_v(v \mid M, Y^*)dv\right] \\
&= E\left[\int_{\text{supp}(V)} I(a_0 \leq M + v \leq a_1)Y^* dv\right] \\
&= E\left[Y^* \int_{a_0-M}^{a_1-M} dv\right] = (a_1 - a_0)E(Y^*)
\end{aligned}$$

and similarly, $E(W) = (a_1 - a_0)$, so $E(WY)/E(W) = E(Y^*)$. The mean of the unobserved Y^* equals the weighted mean of the observed Y, with weights given by $W/E(W)$.

For estimands that depend on covariates X, the above results extend using weights of the form $W = D/f_v(V \mid X)$. Compare this to the more usual propensity score weight estimators (see, e.g., Koul, Susarla, and van Ryzin 1981), which look similar but employ weights of the form $W = D/E(D \mid X, Y)$, and for consistency require special structure on the joint distribution of Y^*, D, and X. In contrast, the weights proposed in this paper are based on the density of a covariate V that affects treatment, rather than directly on the probability of treatment.

The estimators in this paper extend the above idea to arbitrary moments of Y that include covariates, and either divide by the density of V or (equivalently) directly integrate over v. The main virtue of this procedure is that it avoids having to make assumptions about the joint distribution of Y^* and M, other than independence from V (given X).

2 Identification

ASSUMPTION 1. For a sample of individuals we observe a binary treatment or selection indicator D, a covariate vector X, a continuously distributed covariate scalar V, and an outcome Y where $Y = Y^*D$.

The form of the distribution of outcomes is not restricted, that is, Y^* could be continuous, discrete, or contain mass points. If Y^* is continuous then it is sufficient to only observe Y, V, and X, since in that case D could be defined by $D = I(Y \neq 0)$. The elements of X may also be continuous or discrete, or X could be empty.

ASSUMPTION 2. The indicator D is determined by

$$D = I(a_0 \leq M + V \leq a_1)$$

where a_0 and a_1 are (possibly infinite) constants and M is an unobserved latent variable.

One sided censoring is when either a_0 or a_1 is infinite. Assumption 2 is not as restrictive as it might appear. In particular, Proposition 1 below shows that, with some regularity, a sufficient (but not necessary) condition for Assumption 2 is that the probability of selection be monotonic in the covariate V.

PROPOSITION 1. Given regularity, if the probability that $D = 1$ is monotonic in a continuously distributed V having a large support, then Assumption 2 holds with one sided censoring.

A formal statement and proof of Proposition 1 is given in the appendix as Lemma 1. Intuitively, the result follows because, conditioning on everything other than V that determines D, we can define M to equal the negative of whatever value of V is just large (or just small) enough to cause D to change from zero to one.

While monotonicity in V is sufficient for Assumption 2, it is not necessary. In particular, monotonicity does not hold with two sided censoring.

Assumption 2 implies that all of the observables and unobservables that determine D, other than V, can be subsumed into a scalar M. If M were linear in X and in an additive independent error, then Assumption 2 would be equivalent to a standard parametric choice model for selection.

ASSUMPTION 3. Conditioning on $X = x$, the covariate V is continuously distributed and conditionally independent of M and Y^*.

In a linear simultaneous system of two equations, a standard means of obtaining identification is by exclusion restrictions, where coefficients in one equation are identified by having an exogeneous variable appear only in the other equation. Given Assumptions 1 and 2, Assumption 3 is as an exclusion restriction, in which moments of Y^* are identified by having a variable V that only affects D.

ASSUMPTION 4. Let δ_0 and δ_1 be constants that satisfy $\delta_0 \leq a_0 - \sup[\text{supp}(M)]$ and $\delta_1 \geq a_1 - \inf[\text{supp}(M)]$, and either

A. The support of V contains the interval (δ_0, δ_1) and δ_0 and δ_1 are both finite, or

B. The support of V contains the interval (δ_0, δ_1), or

C. The support of V is a bounded interval, and contains the interval $[-\tau, \tau]$ for some large scalar τ.

Assumption 4, which has three variants, defines the sense in which V is required to have a large support. Assumption 4A implies that a_0 and a_1 are finite, and so applies only to two sided censoring, while Assumptions 4B and 4C will be used for both one and two sided censoring results. Assumptions 4A and 4B imply that V can take on any value in the interval $(a_0 - M, a_1 - M)$,.and therefore the conditional probability of $D = 1$ can take on any value from zero to one, which is generally a requirement for full identification of treatment effects. Assumption 4C will be used for estimators that are not consistent, but rather have asymptotic bias of order $O(\tau^{-1})$, and hence require τ large to make this bias negligible.

The above assumptions, and hence identification results based on them, do not require independent or identically distributed observations, though the estimators provided later will assume i.i.d. observations.

2.1 Unconditional moments

First consider identification and estimation of unconditional expectations of the form $E[g(Y^*, X)]$ for a given function g. Let $f_v(v \mid x)$ denote the conditional pdf of V evaluated at $V = v$ and conditioning on $X = x$. Define W by

$$W = \frac{D}{f_v(V \mid X)}$$

ASSUMPTION 5. *The expectations* $E[g(Y^*, X)]$ *and* $E[g(0, X)]$ *exist.*

THEOREM 1. *Given a function* $g(\psi, x)$, *define* ω by

$$\omega = \frac{E[Wg(Y, X)]}{E(W)} \tag{1}$$

If Assumptions 1, 2, 3, 4A, and 5 hold then $E[g(Y^*, X)] = \omega$.

By Theorem 1, the mean of the unobservable $g(Y^*, X)$ can be consistently estimated by a weighted average of the observable $g(Y, X)$, using weights W, where these weights are functions of observables D, V, X and the pdf of V. An interesting feature of Theorem 1 is that it does not require estimation of the propensity score, that is, we do not need to construct or estimate $E(D \mid V, X)$.

In Theorem 1, defining $g(\psi, x) = I(\psi \leq y^*)$ for any constant y^* makes ω equal the unconditional distribution function of Y^*, evaluated at y^*. Theorem 1 thereby provides a direct estimator of the distribution function of the latent Y^*. Theorem 1 will later be applied with $g(\psi, x) = x\psi$ to estimate $E(XY^*)$ in a regression specification for Y^*.

If a_0 or a_1 is infinite, then the definitions of δ_0 and δ_1 will require that either δ_0 or δ_1 be infinite. The difficulty with this case is then $E[Wg(Y, X)]$ and $E(W)$ may not exist. The following corollaries deal with this complication.

COROLLARY 1. *If Assumptions 1, 2, 3, 4C, and 5 hold then* $E[g(Y^*, X)] = \omega + O(\tau^{-1})$.

Corollary 1 implies that if V has bounded support that contains the interval $(-\tau, \tau)$, for some large τ, then the same estimator ω can be used for either one or two sided censoring, and the resulting bias, if any, will be of order τ^{-1}. This asymptotic bias, which (for sufficiently large τ) is only present with one sided censoring, can be made arbitrarily small by having the support of V be arbitrarily large. This result implies that, in applications, the estimator could be used without even specifying whether the censoring is one or two sided.

It is possible to estimate $E[g(Y^*, X)]$ without this one sided censoring induced bias term, but a somewhat more complicated estimator is required. Define

$$\omega(\tau) = \frac{E[I(|V| \leq \tau)Wg(Y, X)]}{E[I(|V| \leq \tau)W]}$$

COROLLARY 2. *If Assumptions 1, 2, 3, 4B and 5 hold then* $E[g(Y^*, X)] = \omega(\tau) + O(\tau^{-1})$ *and* $E[g(Y^*, X)] = \lim_{\tau \to \infty} \omega(\tau)$.

The exact expression for the bias term $E[g(Y^*, X)] - \omega(\tau)$ is given in the proof of Corollary 2. For example, if $g(0, X) = 0$ and $a_1 = \infty$ then the bias term is given by

$$E[g(Y^*, X)] - \omega = \frac{cov[M, g(Y^*, X)]}{\tau - a_0 + E(M)}$$

This expression is also an upper bound for the bias term in Corollary 1. Similar expressions are obtained if $g(0, X) \neq 0$ or $a_0 = -\infty$.

Corollary 2 implies that, given either one sided or two sided censoring, $E[g(Y^*, X)]$ can be consistently estimated by a sample weighted average of $g(Y, X)$, with weights given by $I(|V| \leq \tau)W$ divided by the sample average of $I(|V| \leq \tau)W$, and letting $\tau \to \infty$ as the sample size grows to infinity. For one sided censoring, consistency of this estimator requires V to have infinite support, which will later be shown to imply a slower than root n rate of convergence.

An interesting question for future research would be determination of an optimal trimming rule, i.e. a data dependent procedure for choosing τ that minimizes some root mean squared error criterion, balancing the contribution to variance of observations having very large values of $I(|V| \leq \tau)W$ against the $O(\tau^{-1})$ bias term.

For the present, although Corollary 2 provides consistency in the case of unbounded supports, to avoid technical problems associated with vanishing densities and to obtain root n limiting distributions, most of the estimators provided later will assume that the support of V is bounded, corresponding to either Theorem 1 for consistent estimation with two sided censoring, or Corollary 1 for estimation with an arbitrarily small but nonzero bias when censoring is one sided. Consideration of limiting distributions with infinite support is deferred to the extensions section of the paper.

2.2 Linear Outcome and Possibly Endogeneous Regressors

Now consider a linear selection model, so the unobserved Y^* is given by $Y^* = X_1^T \beta + \varepsilon$. Given instruments Z, Corollary 3 below shows that β can be estimated by an ordinary two stage least squares linear regression of WY on WX_1, using instruments Z.

ASSUMPTION 6. *Assume $Y^* = X_1^T\beta + \varepsilon$, where ε is an unobserved error. The vector X contains the regressors X_1 and instruments Z, where $E(\varepsilon Z) = 0$, $E(ZZ^T)$ exists and is nonsingular, and the rank of $E(X_1Z^T)$ is J, the dimension of X_1.*

Define Σ_{wxz}, Σ_{zz}, and Δ by $\Sigma_{wxz} = E(WX_1Z^T)$, $\Sigma_{zz} = E(ZZ^T)$, and

$$\Delta = (\Sigma_{wxz}\Sigma_{zz}^{-1}\Sigma_{wxz}^T)^{-1}\Sigma_{wxz}\Sigma_{zz}^{-1}$$

Similarly, let $\Sigma_{wxz}(\tau) = E[I(|V| \le \tau)WX_1Z^T]$ and define $\beta(\tau) = [\Sigma_{wxz}(\tau)\Sigma_{zz}^{-1}\Sigma_{wxz}^T(\tau)]^{-1}\Sigma_{wxz}(\tau)\Sigma_{zz}^{-1}E[I(|V| \le \tau)WZY]$.

COROLLARY 3. *If Assumptions 1, 2, 3, 4A and 6 hold then*

$$\beta = \Delta E(WZY)$$

If Assumptions 1, 2, 3, 4B and 6 hold then $\beta(\tau) = \beta + O(\tau^{-1})$ so $\beta = \lim_{\tau\to\infty}\beta(\tau)$, and if Assumptions 1, 2, 3, 4C and 6 hold then $\beta = \Delta E(WZY) + O(\tau^{-1})$.

In short, Corollary 3 says that β can be estimated by an ordinary two stage least squares linear regression of WY on WX_1, using instruments Z. In the special case where $Z = X_1$, this reduces to an ordinary (weighted) least squares regression.

For two sided censoring, this two stage least squares estimator is consistent. With one sided censoring, this estimator will have an arbitrarily small asymptotic bias if V has a large but not infinite support, or alternatively, consistency may be obtained in the one sided censoring case if V has infinite support by replacing W with $I(|V| \le \tau)W$, and letting τ grow with the sample size.

Corollary 3 permits estimation under weak assumptions regarding the errors ε. No restriction is placed on the relationship between ε and M, other than both being conditionally independent of V given X, so the same unobservables are permitted to effect outcomes and selection or treatment, and to do so in unknown ways.

If ε and X_1 are uncorrelated, then we may take $Z = X_1$, and Corollary 3 permits general forms of heteroscedasticity of ε, so higher moments of ε may depend on X_1 in arbitrary ways. For example, Assumption 6 is satisified with $Z = X_1$ given classical random coefficients in Y^*, since if $Y^* = X_1^T(\beta + \varepsilon^*)$ with mean zero ε^* independent of X_1, then then $\varepsilon = X_1^T\varepsilon^*$ and $E(X_1\varepsilon) = 0$.

More interesting is the case where ε and X_1 may be correlated, as would occur if the regressors X_1 are endogeneous or mismeasured. Assumption 6 is identical to the minimal assumptions that would be made

about covariates X_1 and instruments Z if Y^* were observed and β was to be estimated by ordinary linear two stage least squares. The errors ε do not need to be continously distributed, and can have moments that depend in arbitrary ways on X_1 and Z, as long as $E(\varepsilon Z) = 0$.

We do not need to construct $E(X_1 \mid Z)$, nor do we require any assumptions regarding the 'instrument equation' errors $X_1 - E(X_1 \mid Z)$, other than the conditional independence from V implied by Assumption 3. For example, if X_1 is an arbitrary function of Y^*, Z, and a vector of unobservables ε_1 (as would be the case for classical measurement errors or for a simultaneous system of equations for X_1 and Y^*), and M is an arbitrary function of X and unobservables e, then the required conditional independence from V will hold if the set of unobservables $\varepsilon, \varepsilon_1, e$ is conditionally independent of V, conditioning on Z. Alternatively, if X_1 is an arbitrary function of V, Z, and ε_1, then having ε, e be conditionally independent of ε_1, V conditioning on Z would suffice.

It is notable that the estimator does not require specification or estimation of either the instrument fits $E(X_1 \mid Z)$ or the propensity score $E(D \mid V, X)$, nor does it require any consideration or specification of the joint distribution of errors or unobservables in the selection, outcome and instrument equations, other than the conditional independence and support assumptions regarding V.

The vectors of regressors X_1 and instruments Z may each include a constant term (so location is estimated along with other coefficients), and they may contain discretely distributed variables such as dummy variables. Squares and interaction terms are also permitted, e.g., the third element of X_1 could equal the square of the second element, or equal the product of the first two elements. In addition, X_1 and Z can be correlated with V, though Assumption 3 rules out having elements of X_1 or Z be deterministic functions of V.

2.3 Conditional moments

Now consider conditional expectations of the form $E[g(Y^*, X) \mid X = x]$ for a given function g. The natural extension of Theorem 1 would be based on $E[Wg(Y, X) \mid X = x]/E(W)$. A consistent estimator of this form can be constructed, but its limiting distribution will be needlessly complicated because the density $f_v(V \mid X)$ must first be estimated to construct W, followed by a nonparametric regression of $Wg(Y, X)$ on X. Theorem 2 below provides an alternative that has a simpler limiting distribution.

THEOREM 2. *Let Assumptions 1, 2, 3 and 4B hold. Given a function* $g(\psi, x)$, *define* $\mu(x)$ *by*

$$\mu(x) = g(0,x) + \frac{\int_{\delta_0}^{\delta_1} E\left([g(Y,X) - g(0,X)]D \mid V = v, X = x\right) dv}{E\left[\int_{\delta_0}^{\delta_1} E\left(D \mid V = v, X\right) dv\right]}. \tag{2}$$

Then $\mu(x) = E[g(Y^*, X) \mid X = x]$ *if the expectations and integrals in equation (2) exist.*

Let $F(y^* \mid X = x)$ denote the conditional distribution function of Y^*. An immediate implication of Theorem 2 is that F is identified, since defining $g(\psi, x) = I(\psi \leq y^*)$ for any constant y^* makes $\mu(x) = F(y^* \mid X = x)$.

A consistent estimator based on Theorem 2 can be constructed by replacing the conditional expectations in equation (2) with nonparametric regressions, and replacing the unconditional outer expectation in the denominator of equation (2) with a sample average.

In addition to providing an estimator for the conditional distribution of Y^* given X, Theorem 2 more generally provides an estimator for conditional means, which is useful because many objects of interest can thereby be directly estimated. For example, in the treatment model discussed in the introduction, take $g(\psi, x) = \psi$, define $\mu_1(x)$ to equal $\mu(x)$ in Theorem 2 with $D = T$ and $Y = Y^\circ T$, and define $\mu_0(x)$ to equal $\mu(x)$ in Theorem 2 with $D = (1 - T)$ and $Y = Y^\circ(1 - T)$. Then $\mu_1(x) - \mu_0(x)$ will equal the conditional average treatment effect, conditioning on $X = x$.

Care must be taken when applying Theorem 2 with one sided censoring (as in the above treatment example), since Assumption 4B may then require δ_0 or δ_1 to be infinite, and so the integrals in equation (2) may not exist. In that case, define the left side of equation (2) to be $\mu(x, \delta_0, \delta_1)$. Then, as long as $\mu(x)$ exists, $\mu(x)$ will equal $\lim_{\tau\to\infty} \mu(x, -\tau, \delta_1)$ or $\lim_{\tau\to\infty} \mu(x, \delta_0, \tau)$, exactly analogous to Corollary 2. Although this extension provides consistency in the case of unbounded support, to avoid technical problems associated with vanishing densities, the limiting distribution for estimation based on Theorem 2 provided later will assume that the support of V is bounded. Lemma 2 in the appendix shows that, regardless of whether δ_0 and δ_1 are finite or infinite (i.e., for one or two sided censoring), with τ finite we have $\mu(x, -\tau, \tau) = E[g(Y^*, X) \mid X = x] + O(\tau^{-1})$ provided that $[-\tau, \tau]$ is in the support of V, so any asymptotic bias induced by one sided censoring can be made arbitrarily small by having the support of V be arbitrarily large.

3 Limiting Distributions

Assume that a random sample D_i, Y_i, V_i, X_i for $i = 1, \ldots, n$ is observed, where D_i is a realization of D, and similarly for Y_i, V_i, and X_i. Let Assumptions 1,2, and 3 hold.

3.1 Unconditional Moment Estimation

Assume a function g has been chosen, and consider estimation of $E[g(Y^*, X)]$, based on Theorem 1 or Corollary 1. The simplest case is when ω defined by equation (1) is estimated assuming f_v is known and V has bounded support. Define this estimator, $\widehat{\omega}_1$, by

$$W_i = \frac{D_i}{f_v(V_i \mid X_i)}$$

$$\widehat{\omega}_1 = \frac{\sum_{i=1}^n W_i g(Y_i, X_i)}{\sum_{i=1}^n W_i}$$

THEOREM 3. *Assume that $f_v(v \mid x)$ is bounded away from zero and $E[W^2 g(Y, X)^2]$ exists. Then*

$$\frac{\sqrt{n}(\widehat{\omega}_1 - \omega)}{\sqrt{E[W^2(g(Y, X) - \omega)^2]}/E(W)} \xrightarrow{d} N(0, 1).$$

This estimator converges to ω at rate root n. By Theorem 1 and Corollary 1, ω will either exactly equal $E[g(Y^*, X)]$, or under one sided censoring will differ from it by an arbitrarily small bias term. Later in an extensions section, an estimator based on Corollary 2 will be provided that consistently estimates $E[g(Y^*, X)]$ for either one or two sided censoring, but converges at slower than rate root n.

3.1.1 Estimation With Unknown Density

Now consider estimation of ω when f_v is not known and must be estimated.

Given an arbitrary S_i and a sufficiently regular nonparametric estimator $\hat{f}_v(v \mid x)^{-1}$, Lewbel (2000a) and Honoré and Lewbel (2001) provide the following root n limiting distribution.

$$\frac{\sqrt{n}\left[\left(n^{-1}\sum_{i=1}^n S_i \hat{f}_v(V_i \mid X_i)^{-1}\right) - E[f_v(V \mid X)^{-1} S]\right]}{var\left[f_v(V \mid X)^{-1} S + E[f_v(V \mid X)^{-1} S \mid X] - E[f_v(V \mid X)^{-1} S \mid V, X]\right]} \xrightarrow{d} N(0, 1) \qquad (3)$$

This is a two step estimator with a nonparametric first step. Examples of root n convergence of similar estimators involving a kernel estimated first step include Robinson (1988), Powell, Stock, and Stoker (1989), Hardle and Stoker (1989), Newey and McFadden (1994), Newey (1994), Newey and Ruud (1994), Sherman (1994), Lewbel (1995), Andrews (1995), and Hardle and Horowitz (1996). The root n limiting distribution theory for such estimators is well known. See, e.g., Theorems 8.2 and 8.12 of Newey and McFadden (1994) for a set of high level assumptions, and a corresponding set of kernel estimator assumptions, yielding root n normality for this type of two step estimator.

The difficulty in applying generic results like these to estimands of the form $E[f_v(V \mid X)^{-1}S]$ is that remainder terms in the expansions generally cannot be bounded sufficiently unless $f_v(V \mid X)$ itself is bounded away from zero, but bounding $f_v(V \mid X)$ away from zero introduces boundary effects in the density estimation, which also interferes with sufficiently fast shrinkage of remainder terms, unless S equals zero in the neighborhood of the boundary (i.e., fixed trimming).

Lewbel (2000a) deals with this difficulty by bounding $f_v(V \mid X)$ away from zero and introducing an asymptotic trimming function that sets to zero all terms in the average having data within a distance t of the boundary. The estimator sends t to zero more slowly than the bandwidth to eliminate boundary effects from kernel estimation, but also has t shrink to zero faster than $n^{-1/2}$, which makes the volume of the trimmed space vanish quickly enough to send the trimming induced bias to zero. A closely related alternative is Hong and White (2000), who, based on Rice (1986), use jackknife boundary kernels in place of asymptotic trimming.

The resulting kernel estimator for $\hat{f}_v(V \mid X)^{-1}$ has the form

$$\hat{f}_v(v \mid x)^{-1} = \frac{I_t(v,x)b\sum_{i=1}^{n} K\left(\frac{x-X_i}{b}\right)}{\sum_{i=1}^{n} k\left(\frac{v-V_i}{b}\right)K\left(\frac{x-X_i}{b}\right)} \tag{4}$$

where k is a kernel function, $K(t) = \prod_{j=1}^{d} k(t_j)$, b is a bandwidth, and $I_t(v,x)$ is a trimming function defined to equal zero if (v,x) is within a distance t of the boundary of the support of V, X, and one otherwise. Theorem 1 in Lewbel (2000a) then provides sufficient regularity conditions to obtain equation (3), assuming i.i.d. draws of V, X, S. These conditions consist of existence of moments, densities bounded away from zero, local Lipschitz conditions, kernels of order p, and rates $nb^J \to \infty$, $nb^{2p} \to 0$, $b/t \to 0$, and $nt^2 \to 0$.

To keep the estimation simple, in the later empirical application no trimming is employed, so $I_t(v,x)$ is set equal to one. Hardle and Stoker (1989) also report insensitivity to trimming in applications. Similarly,

root n convergence calls for higher order kernels, but ordinary kernels typically perform better in practice. The empirical results are also relatively insensitive to bandwidth choice.

Define

$$\widehat{W}_i = D_i \hat{f}_v(V_i \mid X_i)^{-1} \tag{5}$$

$$\widehat{\omega} = \frac{\sum_{i=1}^n \widehat{W}_i g(Y_i, X_i)}{\sum_{i=1}^n \widehat{W}_i}$$

$$Q_\omega = (g(Y,X) - \omega)W + E[(g(Y,X) - \omega)W \mid X] - E[(g(Y,X) - \omega)W \mid V, X] \tag{6}$$

THEOREM 4. *Assume $f_v(v \mid x)$ is bounded away from zero, $E(Q_\omega^2)$ exists, and equation (3) holds for $S = D$* and for *$S = g(Y,X)$. Then*

$$\frac{\sqrt{n}(\widehat{\omega} - \omega)}{\sqrt{E(Q_\omega^2)}/E(W)} \xrightarrow{d} N(0,1). \tag{7}$$

The variance in equation (7) can be estimated by replacing ω and W in equation (6) with $\widehat{\omega}$ and $\widehat{W}$, then replacing the expectations in that equation with nonparametric regressions evaluated at V_i and X_i to define $\widehat{Q}_{\omega i}$, and finally replacing the expectations in $E(Q_\omega^2)^{1/2}/E(W)$ with sample averages of $\widehat{Q}_{\omega i}^2$ and $\widehat{W}_i$.

In place of a kernel estimator, consistent (though perhaps not root n) estimates could be obtained using a series expansion based density estimator of f_v, as in Gallant and Nychka (1987).

3.1.2 Very Simple Estimators

This section describes a computationally trivial "ordered data" estimator for the density f_v which does not require kernels or bandwidths.

ASSUMPTION 7. *Assume there exists a vector $\widetilde{\beta}$ such that $V = X^T\widetilde{\beta} + e_v$, where e_v is continuously distributed, has bounded support, and is independent of X.*

A special case of Assumption 7 would be if V were independent of X, which would then make e_v equal V. Let f_{ev} denote the unconditional density function of e_v. If Assumption 7 holds then $f_{ev}(e_v) = f_v(V \mid X)$. Define $\widehat{e}_{vi}$ as the residuals from linearly regressing V on X, so

$$\widehat{e}_{vi} = V_i - X_i(\Sigma_{i=1}^n X_i X_i^T)^{-1} \Sigma_{i=1}^n X_i V_i$$

Let $\widehat{e}_{vi}^{+}$ denote the smallest element of $\{\widehat{e}_{v1}, ...\widehat{e}_{vn}\}$ that is greater than $\widehat{e}_{vi}$, and let $\widehat{e}_{vi}^{-}$ denote the largest element of element of $\{\widehat{e}_{v1}, ...\widehat{e}_{vn}\}$ that is less than $\widehat{e}_{vi}$. In other words, if the data $\widehat{e}_{v1}, ...\widehat{e}_{vn}$ are sorted in ascending order, the number immediately preceeding $\widehat{e}_{vi}$ would be $\widehat{e}_{vi}^{-}$, and the number immediately following $\widehat{e}_{vi}$ would be $\widehat{e}_{vi}^{+}$. Endpoints may be dealt with by letting $\widehat{e}_{vi}^{-}$ equal $\widehat{e}_{vi}$ if there is no element of $\{\widehat{e}_{v1}, ...\widehat{e}_{vn}\}$ that is smaller than $\widehat{e}_{vi}$, and similarly for the largest element.

Define the estimator

$$\widetilde{f}_v(V_i \mid X_i)^{-1} = (\widehat{e}_{vi}^{+} - \widehat{e}_{vi}^{-})n/2 \tag{8}$$

Now i/n is an estimate of the distribution of e_v evaluated at $\widehat{e}_{vi}$, so $\widetilde{f}_v(V_i \mid X_i)^{-1} \approx f_{ev}(e_{vi})^{-1} = f_v(V_i \mid X_i)^{-1}$. Although $\widetilde{f}_v(v \mid x)^{-1}$ is not a consistent estimator of $f_v(v \mid x)^{-1}$, it is the case given Assumption 7 that for arbitrary S, with iid data, $n^{-1}\sum_{i=1}^{n} \widetilde{f}_v(V_i \mid X_i)^{-1} S_i$ is a consistent estimator of $E[f_v(V \mid X)^{-1} S]$ (see section 4.1 of Lewbel 2000).

Therefore, using equation (8) instead of $\widehat{f}_v$ in the definition of $\widehat{W}$ yields a numerically very simple estimator. In particular, $\widehat{\omega}$ then simplifies to

$$\widehat{\omega} = \frac{\sum_{i=1}^{n} g(Y_i, X_i)(\widehat{e}_{vi}^{+} - \widehat{e}_{vi}^{-})D_i}{\sum_{i=1}^{n} (\widehat{e}_{vi}^{+} - \widehat{e}_{vi}^{-})D_i}$$

which will be a consistent estimator of ω.

This estimator is convenient for its numerical simplicity, but it requires the extra Assumption 7 for consistency. This assumption limits the permitted dependence of V on X. An application in which this additional assumption may be satisfied by construction is Lewbel, Linton, and McFadden (2001), where a special regressor is determined by experimental design. Another application is Maurin's (1999) example where V is a child's exact date of birth and X is a vector of socioeconomic attributes of the child's family.

3.1.3 Estimation with a Parametric Density

Suppose $f_v(V \mid X)$ is not known but is finitely parameterized. For example, the income distribution is known to be well approximated by a lognormal distribution with a Pareto tail, so this specification might be used when v is income. Let $f_v(V \mid X, \theta)$ be a parameterization of f_v in terms of a parameter vector θ, with estimator $\widehat{\theta}$ where

$$\sqrt{n}(\widehat{\theta} - \theta) \Rightarrow N[0, var(Q_\theta)] \tag{9}$$

for some influence function Q_θ. For example, θ might consist of means or other moments of V, X and $\widehat{\theta}$ would be the corresponding sample moments. The standard limiting distribution theory for parametric two

step estimation can now be applied (see, e.g., section 6 of Newey and McFadden 1994). The result is again equation (7) but this time with Q_ω defined by

$$Q_\omega = [g(Y, X) - \omega]W \left[1 - Q_\theta^T \frac{\partial f_\upsilon(V \mid X, \theta)/\partial\theta}{f_\upsilon(V \mid X, \theta)}\right].$$

In the case where f_υ is known this further simplifies to $Q_\omega = [g(Y, X) - \omega]$, equivalent to Theorem 3.

Alternatively, instead of first estimating θ, one could simply stack the moment conditions defining $\widehat{\theta}$ with the moment condition $E[(g(Y, X) - \omega)D/f_\upsilon(V \mid X, \theta)] = 0$, and apply an ordinary GMM estimator to the stack.

3.2 Estimation With Possibly Endogeneous Regressors

Now consider root n estimation of β, which for this section will be defined as $\beta = \Delta E(WZY)$. By Corollary 3, this definition of β equals the coefficients in the linear outcome model $Y^* = X_1^T\beta + \varepsilon$ under two sided censoring, and under one sided censoring differs from these coefficients by an arbitrarily small amount.

Let $\hat{f}_\upsilon(V \mid X)^{-1}$ be nonparametrically estimated as in Theorem 4, define $\widehat{W}$ by equation (5), and define

$$\hat{\Delta} = \left[\left(\frac{\Sigma_{i=1}^n \widehat{W}_i X_{1i} Z_i^T}{n}\right)\left(\frac{\Sigma_{i=1}^n Z_i Z_i^T}{n}\right)^{-1}\left(\frac{\Sigma_{i=1}^n \widehat{W}_i Z_i X_{1i}^T}{n}\right)\right]^{-1}\left(\frac{\Sigma_{i=1}^n \widehat{W}_i X_{1i} Z_i^T}{n}\right)\left(\frac{\Sigma_{i=1}^n Z_i Z_i^T}{n}\right)^{-1}$$

$$\widehat{\beta} = \hat{\Delta}\left(\frac{\Sigma_{i=1}^n \widehat{W}_i Z_i Y_i}{n}\right) \tag{10}$$

$$Q_\beta = WZY + E(WZY \mid X) - E(WZY \mid V, X)$$

THEOREM 5. *Define $\beta = \Delta E(WZY)$. Assume $f_\upsilon(\upsilon \mid x)$ is bounded away from zero and equation (3) holds. Then*

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N[0, \Delta var\left(Q_\beta - WZX_1^T\beta\right)\Delta^T]. \tag{11}$$

This $\widehat{\beta}$ is numerically identical to a linear two stage least squares regression of $\widehat{W}Y$ on $\widehat{W}X_1$ using instruments Z. If Q_β equaled WZY, then the variance in equation (11) would also be the same as the

variance of two stage least squares (with heteroscedastic errors). The additional terms in Q_β are due to the estimation error from using $\widehat{W}$ instead of W.

If $X_1 = Z$ (which by Corollary 3 permits arbitrary heteroscedasticity in ε but not endogeneity of X_1), then $\hat{\Delta}$ simplifies to $\left(\Sigma_{i=1}^n \widehat{W}_i X_{1i} X_{1i}^T / n\right)^{-1}$ and $\widehat{\beta}$ becomes numerically identical to a linear weighted least squares regression of Y on X_1 using weights $\widehat{W}$.

The variance in equation (11) can be estimated as follows. In the definition of Q_β replace W with $\widehat{W}$ and replace expectations with nonparametric regressions to obtain $\widehat{Q}_\beta$. Then the variance of $\sqrt{n}(\widehat{\beta} - \beta)$ may be estimated as $\hat{\Delta} \widehat{var}\left(\widehat{Q}_\beta - \widehat{W} Z X_1^T \widehat{\beta}\right) \hat{\Delta}^T$, where $\widehat{var}$ denotes the sample variance.

The limiting distribution in Theorem 5 assumes X is continuously distributed. Discrete elements of X (having a finite number of mass points) can be readily handled in the estimation of f_v using cell means, or as in, e.g., Racine and Li (2000).

As before, if instead of being nonparametrically estimated we have f_v parameterized as $f_v(V \mid X, \theta)$ with equation (9), then equation (11) will still hold, but now with

$$Q_\beta = ZYW\left[1 - Q_\theta^T \frac{\partial f_v(V \mid X, \theta)/\partial\theta}{f_v(V \mid X, \theta)}\right].$$

and, in particular, if f_v and hence W is known, then this simplifies to $Q_\beta = ZYW$.

3.2.1 A Very Simple Estimator of Beta

It follows from the earlier section on very simple estimators that, if Assumption 7 holds then we can consistently estimate $\beta = \Delta E(WZY)$ using the density estimator in equation (8) in the definition of $\widehat{W}$. This results in the following extremely simple consistent estimator for β: 1. define $\widehat{e_v}$ as the residuals from regressing V on X using ordinary least squares, 2. sort the $\widehat{e_v}$ data from smallest to largest to obtain $\widehat{e_{vi}^+}$ and $\widehat{e_{vi}^-}$ for each observation i, and 3. Let $\hat{\beta}$ be the estimated coefficients from linearly regressing $(\widehat{e_{vi}^+} - \widehat{e_{vi}^-})Y_i$ on $(\widehat{e_{vi}^+} - \widehat{e_{vi}^-})D_i X_{1i}$ using two stage least squares with instruments Z_i.

3.3 Nonparametric Conditional Moment Estimation

Assume a function g has been chosen, and consider estimation of $\mu(x)$ defined by equation (2). By Theorem 2 and Lemma 2, either $\mu(x) = E[g(Y^*, X) \mid X = x]$ or $\mu(x)$ differs from this expectation by an arbitrarily small bias term.

Define $\widetilde{Y}$ by

$$\widetilde{Y} = [g(Y, X) - g(0, X)]D$$

Define the functions $m(v,x)$ and $m^*(v,x)$ by

$$\begin{aligned} m(v,x) &= E[\widetilde{Y} \mid V = v, X = x] \\ m^*(v,x) &= E[D \mid V = v, X = x] \end{aligned}$$

Let $\widehat{m}(v,x)$ be a consistent estimator of m, that is, a nonparametric regression of $\widetilde{Y}$ on X, V, evaluated at x, v. Similarly let $\widehat{m}^*(v,x)$ be a consistent estimator of m^*. Define $\widehat{v}_0$ and $\widehat{v}_1$ by

$$\begin{aligned} \widehat{v}_0 &= \min\{V_1, ..., V_n\} \\ \widehat{v}_1 &= \max\{V_1, ..., V_n\} \end{aligned}$$

Define κ, $\widehat{\kappa}$ and $\widehat{\mu}(x)$ by

$$\kappa = E\left[\int_{\delta_0}^{\delta_1} m^*(v,X)dv\right]$$

$$\widehat{\kappa} = \int_{\widehat{v}_0}^{\widehat{v}_1} n^{-1}\sum_{i=1}^{n} \widehat{m}^*(v,X_i)dv$$

$$\widehat{\mu}(x) = g(0,x) + \widehat{\kappa}^{-1}\int_{\widehat{v}_0}^{\widehat{v}_1} \widehat{m}(v,x)dv$$

The integrals involved in the definition of $\widehat{\mu}(x)$ are one dimensional, and so can be readily evaluated numerically.

Given an iid sample of $(\widetilde{Y}_i, V_i, X_i)$, the limiting distribution theory for estimators of the form $\int_{\delta_0}^{\delta_1} \widehat{m}(v,x)dv$ is known. This linear functional of a nonparametric regression is in the class of marginal integration/partial mean estimators sometimes used for estimating additive nonparametric regression models. See, e.g., Linton and Nielsen (1995), Newey (1994), and Tjøstheim and Auestad (1994). Based on this work, and using results in Masry (1996a), (1996b), and Gozalo and Linton (2000), Lewbel, Linton, and McFadden (2001) provide the limiting normal distribution for $\int_{\delta_0}^{\delta_1} \widehat{m}(v,x)dv - \int_{\delta_0}^{\delta_1} m(v,x)dv$, both under high level assumptions regarding the nonparametric regression estimator $\widehat{m}(v,x)$, and for the particular case of a kernel regression. The latter results are applied here in Theorem 6 below.

Define the kernel regression estimator

$$\widehat{m}(v,x) = \frac{\sum_{i=1}^{n} \widetilde{Y}_i k\left(\frac{v-V_i}{b}\right) K\left(\frac{x-X_i}{b}\right)}{\sum_{i=1}^{n} k\left(\frac{v-V_i}{b}\right) K\left(\frac{x-X_i}{b}\right)} \tag{12}$$

where k is a kernel function and $K(t) = \prod_{j=1}^{d} k(t_j)$, and define $\widehat{m}^*(v,x)$ analogously.

ASSUMPTION K. k is a symmetric probability density with bounded support, and is Lipschitz continuous on its support, i.e.,

$$|k(t) - k(s)| \leq \widetilde{c}|t - s|$$

for some constant $\widetilde{c}$. The variables (V, X) are continuously distributed with Lebesgue density $f_{V,X}(v, x)$ that satisfies $\inf_{\delta_0 \leq v \leq \delta_1} f_{V,X}(v, x) > 0$. The functions m, m^*, and $f_{V,X}$ are twice continuously differentiable for all v with $\delta_0 \leq v \leq \delta_1$. The set $[\delta_0 \leq v \leq \delta_1] \times \{x\}$ is strictly contained in the support of (V, X).

Let ∇, ∇^2 denote the first and second derivative operators, and define

$$\overline{\beta}(x) = \frac{\int t^2 k(t) dt}{2\kappa} \int_{\delta_0}^{\delta_1} \operatorname{tr}\Big[\nabla^2 m(v, x) + \nabla m(v, x) \nabla \log f_{V,X}(v, x)\Big] dv$$

$$\overline{\omega}(x) = \|K\|^2 \int_{\delta_0}^{\delta_1} \frac{var[\widetilde{Y} - m(V, X) \mid V = v, X = x]}{\kappa^2 f_{V,X}(v, x)} dv$$

THEOREM 6. *Let Assumptions 1, 2, 3, 4A, and K hold. Assume that the bandwidth sequence $b = b(n)$ satisfies $b \to 0$ and $nb^{d+2}/\log n \to \infty$. Then,*

$$\frac{\widehat{\mu}(x) - \mu(x) - b^2 \overline{\beta}(x))}{\sqrt{n^{-1} b^{-d} \overline{\omega}(x)}} \xrightarrow{d} N(0, 1). \tag{13}$$

Estimation error in $\widehat{v}_0$ and $\widehat{v}_1$ does not contibute to the limiting distribution, because they converge to values outside the range $[\delta_0, \delta_1]$, and $m(v, x)$ equals zero outside that range. Similarly, boundary effects of kernel estimators are not relevant here.

Theorem 6 requires Assumption 4A, and so only applies to the case of two sided censoring, making $\mu(x) = E[g(Y^*, X) \mid X = x]$. However, it can be readily verified that Theorem 6 will also hold if, instead of Assumption 4A, it is assumed that the support of V contains the interval $[-\tau, \tau]$ and if δ_0 and δ_1 (and $\widehat{v}_0$ and $\widehat{v}_1$) are replaced by $-\tau$ and τ in assumption K and in the definitions of κ, $\widehat{\kappa}$ $\mu(x)$, and $\widehat{\mu}(x)$. In that case, by Lemma 2 $\mu(x)$ will equal $E[g(Y^*, X) \mid X = x] + O(\tau^{-1})$ even if the censoring is one sided.

4 Extensions

A few brief extensions will be listed here first, followed by some sections describing lengthier results. The intent is primarily to indicate possible further potential of the estimators and to suggest areas for future research.

Theorems 1 and 2 can be used to recover information about the distribution of the outcome error ε. Given an estimate of β based on Corollary 3, $E[h(\varepsilon, X)]$ or $E[h(\varepsilon, X) \mid X = x]$ can be estimated for a given function h by letting $g(\psi, x) = h(\psi - x_1^T\beta, x)$ in Theorem 1 or Theorem 2.

The parameters β in a nonlinear outcome equation $Y^* = g^*(X, \beta)+\varepsilon$ could be estimated if $E(\varepsilon \mid x) = 0$ by first estimating $\mu(x)$ using Theorem 2, and then minimizing a quadratic form in $\mu(x) - g^*(X, \beta)$. Alternatively, analogous to Corollary 3, if $E(Z\varepsilon) = 0$ then Theorem 1 provides moment conditions $E[WZ(Y - g^*(X, \beta))] = 0$, and GMM could be applied to sample versions of these moments to estimate β. Equation (6) provides the appropriate influence functions to account for estimation error in W when constructing corresponding sample moments.

Many of the estimators provided here could be generalized to permit nonindependent and nonidentically distributed observations, essentially by adding i subscripts to supports, densities, and the expectation operator. For the estimators based on W, the conditional density function f_v should be assumed constant (or its variation finitely parameterized) across observations. Many results exist providing limiting distribution theory for semiparametric estimators when observations are not independently or identically distributed. See, e.g., Andrews (1995).

4.1 Bias Elimination With Unbounded Support

Most of the estimators provided here assume bounded support for V, resulting in an arbitrarily small but nonzero asymptotic bias under one sided censoring. To assess the cost of eliminating this bias, this section provides an estimator for $E[g(Y^*, X)]$ based on the limit as $\tau \to \infty$ of $\omega(\tau)$, and so by Corollary 2 is consistent for two or one sided censoring, provided in the one sided case that V has infinite support. It will be shown that this estimator has the same root n limiting distribution as $\widehat{\omega}_1$ if the support of V is bounded, but otherwise, the rate of convergence is slower than root n. The fastest possible rate depends on existence of moments of $g(Y^*, X)$ and the thickness of the tails of f_v, with thicker tails permitting faster convergence.

Define $\widehat{\omega}_\tau$ by

$$\widehat{\omega}_\tau = \frac{\sum_{i=1}^n W_i g(Y_i, X_i) I(|V_i| \leq \tau)}{\sum_{i=1}^n W_i I(|V_i| \leq \tau)}$$

If supp(V) is bounded then for large enough τ, $\widehat{\omega}_\tau$ will equal $\widehat{\omega}_1$, and so have the same limiting distribution. To simplify the expression of the limiting distribution of $\widehat{\omega}_\tau$ in the infinite support case, assume that $D = I(0 \leq M + V)$, and sup[supp(V)] $= \infty$ and that for some constant c, $f_v(v \mid x) = f_v(v)$ for all $v \geq c$. A variant of Theorem 7 below will still hold without this simplifying assumption that the tail of the density of V not depend on X, but in that case $\Gamma_{\tau\nu}$ below will need to be replaced by the more complicated equation (19) in the appendix, resulting in a rate function γ_τ that depends upon the function g and the distribution of Y^* and X.

Define $\Gamma_{\tau\nu}$ and γ_τ by

$$\begin{aligned} \Gamma_{\tau\nu} &= \int_c^\tau f_v(v)^{1-\nu} dv \\ \gamma_\tau &= \tau^{-2}\Gamma_{\tau 2} \end{aligned}$$

The rate of convergence of $\widehat{\omega}_\tau$ will be $(n/\gamma_\tau)^{1/2}$. Note that with f_v known, the functions $\Gamma_{\tau\nu}$ and γ_τ are known. If D is decreasing in V instead of increasing, then with unbounded support $\Gamma_{\tau\nu}$ and γ_τ will need to be defined analogously for the lower tail of v.

THEOREM 7. *Assume that $f_v(v \mid x)$ is bounded away from zero except at $v = \infty$, and that for some constant c, $f_v(v \mid x) = f_v(v)$ for all $v \geq c$. Assume that $D = I(0 \leq M + V)$, and for some $\nu > 2$, $E[|g(Y^*, X)|^\nu]$ exists. If $\tau \to \infty$ and $\Gamma_{\tau 2}^{-\nu/2}\Gamma_{\tau\nu} n^{1-\nu/2} \to 0$ then*

$$\frac{\sqrt{n/\gamma_\tau}[\widehat{\omega}_\tau - E[g(Y^*, X)]]}{\sqrt{var[g(Y^*, X)^2]}} \xrightarrow{d} N(0, 1).$$

Replacing $g(Y_i, X_i)$ with $g(Y_i, X_i)^2$ in the definition of $\widehat{\omega}_\tau$ provides a consistent estimator of $E[g(Y^*, X)^2]$, so the variance in Theorem 7 is readily estimated.

To illustrate the rates of convergence implied by Theorem 7, suppose $f_v(v \mid x)$ has a polynomial tail, so $f_v(v \mid x) = \kappa v^{-(1+\lambda)}$ for all $v \geq c$, for some positive constants κ and λ. Then $\Gamma_{\tau\nu} = \kappa_{v0} + \kappa_{v1}\tau^{v-\lambda+v\lambda}$ for some constants κ_{v0} and κ_{v1}, which implies $\gamma_\tau = O(\tau^\lambda)$ and $\Gamma_{\tau 2}^{-\nu/2}\Gamma_{\tau\nu} = O\left(\tau^{\frac{v}{2}+(\frac{v}{2}-1)\lambda}\right)$. The required rate condition on τ is therefore $\tau^{\lambda+\frac{v}{v-2}} n^{-1} \to 0$ for some $\nu > 2$ such that $E[|g(Y^*, X)|^\nu]$ exists, and the resulting rate of convergence of $\widehat{\omega}_\tau$ is $(n\tau^{-\lambda})^{1/2}$. The smaller λ is, and hence the thicker the tail of f_v, the faster is this rate of convergence. A necessary condition for root n convergence is $\lambda \leq 0$ (or more generally a tail that is thicker than κv^{-1}) but existence of the distribution function f_v requires $\lambda > 0$, so the root n

rate of convergence cannot be attained. However, having $E[|g(Y^*, X)|^\nu]$ exist for arbitrarily large ν and having λ arbitrarily close to zero means that a rate arbitrarily close to root n is possible. These results are very closely related to the Andrews and Schafgans (1998) analyses of rates of convergence of location estimators in censored models.

4.2 Bias Elimination and Selection Equation Estimation

An advantage of the estimators proposed in this paper is that they do not require estimation of the features of the selection or treatment equation, such as the.propensity score. However, features of the selection equation are often of interest, so this section provides estimators for the selection equation. This section also shows that if the selection equation is parameterized and estimated, then those estimates can provide another method to reduce or eliminate the asymptotic bias in $\widehat{\beta}$ from one sided censoring.

Suppose that, with one sided censoring, the selection equation is parameterized as

$$D = I(0 \leq V + X_1^T \gamma + e)$$

for some vector γ. Note that the constant term is included in X_1. Theorem 8 provides an estimator for γ, and hence for a general semiparametric binary choice model, which can be employed even when some or all of the regressors X_1 are endogeneous.

THEOREM 8. *Let Assumptions 1, 2, 3, 4B, and 6 hold (except that all mentions of Y and Y^* can be omitted) with $D = I(0 \leq V + X_1^T \gamma + e)$, and assume $\delta_0 < 0$. Then*

$$\gamma = \left[E(X_1 Z^T)E(ZZ^T)^{-1}E(ZX_1^T)\right]^{-1} E(X_1 Z^T)E(ZZ^T)^{-1}E\left[Z\frac{D - I(V > 0)}{f_v(V \mid X)}\right]$$

Theorem 8 is proved as Theorem 1' in Lewbel (2000). Theorem 8 shows that the parameters γ can be estimated by a linear two stage least squares regression of $f_v(V \mid X)^{-1}[D - I(V > 0)]$ on X_1 using instruments Z. Lewbel 2000 also provides the limiting distribution for this estimator. If the regressors are endogeneous, then given γ, propensity scores could be estimated using, e.g., the Blundell and Powell (1999) control function methodology.

More generally if M is not parameterized, then Lewbel, Linton and McFadden (2001) may be applied to estimate moments and features of the distribution of M in the model $D = I(0 \leq V + M)$, in results

roughly analogous to Theorems 1 and 2, just as Theorem 8 is a selection equation analog to the estimation of β in Corollary 3.

Now consider using estimates of the selection equation to mitigate the bias from one sided censoring.

COROLLARY 4. *Let Assumptions 1, 2, 3, 4B, and 6 hold with the first element of X_1 identically equal to one, and $D = I(0 \leq V + X_1^T\gamma + e)$. Assume $cov(Z, e\varepsilon) = 0$. Let τ be a constant satisfying* $\sup(M) < \tau < \delta_1$. *Define b to be the vector of all zeros except that the first element of b is $E(e\varepsilon)/\tau$. Then*

$$\left[E(X_1Z^T)E(ZZ^T)^{-1}E(ZX_1^T)\right]^{-1}E(X_1Z^T)E(ZZ^T)^{-1}\frac{E[I(V \leq \tau - X_1^T\gamma)WZY]}{E[I(V \leq \tau - X_1^T\gamma)W]} = \beta + b$$

The assumption in Corollary 4 that $cov(Z, e\varepsilon) = 0$ limits the degree of heteroscedasticity that is permitted in e and ε. For example, if $e = a\varepsilon + \varepsilon^*$ where ε and ε^* are (conditional on Z) uncorrelated with each other, then the assumption would require either that $a = 0$ or that ε^2 be uncorrelated with Z.

Corollary 4 implies that if we first estimate γ, we may then construct

$$\widehat{\Psi}_i = \frac{I(V_i \leq \tau - X_{1i}^T\widehat{\gamma})\widehat{W}_iY_i}{n^{-1}\sum_{i=1}^n I(V_i \leq \tau - X_{1i}^T\widehat{\gamma})\widehat{W}_i}$$

and estimate β by a linear two stage least squares regression of $\widehat{\Psi}$ on X_1 using instruments Z. The result will consistently estimate all of the elements of β except the constant term, which will be biased by the small quantity $E(e\varepsilon)/\tau$.

Any bias reducing procedure like this may have the unwanted side effect of increasing variance, due to the extra estimation errors that are involved. As an alternative to eliminating bias, it may be preferable to choose a value for the trimming parameter τ that minimizes some mean squared error criterion. For example, it should be possible to estimate the asymptotic bias from trimming based on equation (18) (which is itself a function of M), and choose a τ to minimize a function of the estimated bias and estimated variance of β.

4.3 Panel Models with Fixed Effects

Consider the panel sample selection or treatment model

$$\begin{aligned} Y_{it} &= (X_{1t}^T\beta + \alpha_i + \widetilde{\varepsilon}_{it})D_{it} \\ D_{it} &= I(a_{0t} \leq M_{it} + V_{it} \leq a_{1t}) \end{aligned}$$

It is assumed that the number of individuals N is large relative to the number of time periods T, so the asymptotic theory assumes T is fixed and N goes to infinity. Related models include Heckman and Honore (1990), Kyriazidou (1997), and Hansen (1999).

The structural model has an explicit individual specific effect α_i, while the selection equation may have individual specific effects implicitly incorporated into M_{it}. These individual specific effects will be treated as fixed effects, in that their distribution will not be specified or parameterized.

The covariate V_{it} is assumed to be strongly exogeneous. This V_{it} need not vary by time. The regressors X_{1it} may be endogeneous or weakly exogeneous.

Given two time periods r and s, let Z_i be instruments that are uncorrelated with both $\widetilde{\varepsilon}_{ir}$ and $\widetilde{\varepsilon}_{is}$. In particular, if some of the regressors X_{1it} are weakly exogeneous, then Z_i could include those weakly exogeneous regressors from time periods t that preceed times r and s.

For $t = r$ and $t = s$, define weights W_{it} by

$$W_{it} = \frac{D_{it}}{f_{vt}(V_{it} \mid X_i)}$$

where X_i contains all of the distinct elements in X_{1is}, X_{1ir}, and Z_i and f_{vt} denotes the conditional density of V_{it}.

Let $\varepsilon_{it} = \alpha_i + \widetilde{\varepsilon}_{it}$. Applying Theorem 1 for $t = r$ or $t = s$ yields

$$E[ZW_t(Y_t - X_{1t}^T\beta)]/E(W_t) = E(Z\alpha)$$

and it follows that, similar to Corollary 3,

$$E\left[Z\left(\frac{W_rY_r}{E(W_r)} - \frac{W_sY_s}{E(W_s)}\right)\right] = E\left[Z\left(\frac{W_rX_{1r}}{E(W_r)} - \frac{W_sX_{1s}}{E(W_s)}\right)\right]^T\beta$$

so β can be estimated by a linear two stage least squares regression of $W_{ir}Y_{ir}/\overline{W}_r - W_{is}Y_{is}/\overline{W}_s$ on $W_{ir}X_{1ir}/\overline{W}_r - W_{is}X_{1is}/\overline{W}_s$, using instruments Z_i, where $\overline{W}_t = \sum_{i=1}^n W_{it}/n$. A similar method is used by Honoré and Lewbel (2001) to estimate a binary choice panel model based on Theorem 8. They provide some economic examples of possible choices for V_{it}.

This estimator of β is consistent with two sided censoring, and under one sided censoring (applying Corollary 1) has a bias that is $O(\tau^{-1})$, which can be made arbitrarily small by having the support of V_{it} be arbitrarily large. In addition, with one sided censoring if the bias defined by $E[ZW_t(Y_t - X_{1t}^T\beta)]/E(W_t) - E(Z\alpha)$ is constant over time (as would be the case, e.g., if a trimmed W is used as in Corollary 4 and $E(e_t\varepsilon_t)$ is constant over time) then the estimator of β remains consistent under one sided censoring, because the differencing that eliminates $E(Z\alpha)$ will also eliminate this bias.

5 An Investment Model

This section describes an empirical application of Abel and Eberly's (1994) investment model, using the present paper's weighted two least squares estimator to control for possible endogeneity and for sample selection of unknown functional form. The application entails one sided censoring, and hence in theory provides a more challenging testbed for the estimator than would a two sided censoring application.

5.1 Investment Theory

Let Y_i be the rate of investment in manufacturing plant i, defined as the level of investment in a year divided by the beginning of the year value of the plant's capital, and let Q_i be Tobin's Q for the plant. Classical models of firm behavior (e.g., Eisner and Strotz 1963) imply Y_i proportional to Q_i, where the constant of proportionality is inversely related to the magnitude of adjustment costs. However, simple estimates of this relationship at varying levels of aggregation typically find a very low constant of proportionality (see, e.g., Summers 1981 or Hayashi 1982), implying implausibly large adjustment costs.

Another empirical finding inconsistent with proportionality is that plant or firm level data on investment show many periods of zero or near zero investment, alternating with periods of high investment. See, e.g., Doms and Dunne (1998) and Nilsen and Schiantarelli (2000). These empirical findings are generally attributed to discontinuous costs of adjustment, due to factors such as irreversibility or indivisibility of investments. See Blundell, Bond, and Meghir (1996) for a survey.

One difficulty in applying Q models to disaggregate data is that accurate measures of an appropriate firm or plant level marginal Q are difficult to construct. Typical proxies for Q are sales or profit rates. Let P_i be the profit rate of plant i, defined as profits derived from the plant in a year divided by the beginning of the year capital. A problem with the use of a proxy like P_i is that it may be endogeneous, since profits depend on the level of investment.

Let C_i be the cost of investment in plant i in a year, divided by capital at the beginning of the year. Based on the model of Abel and Eberly (1994), assume plant i has investment costs of the form

$$C_i = a_{1i} I(Y_i \neq 0) + a_{2i} Y_i + a_3 Y_i^2$$

The term a_{1i} is plant i's fixed (per unit of capital) cost associated with any nonzero investment, a_{2i} is the price of investment, which can vary across plants, and a_3 is a quadratic adjustment cost parameter. Following the logic of Abel and Eberly (1994), given the above investment cost function the firm chooses

investment Y_i to maximize the present value of current and expected future profits, resulting in a model of the form

$$\begin{aligned} Y_i &= [g^*(a_{2i}) + \beta_1^* Q_i] D_i \\ D_i &= I[Q_i > g(a_{1i}, a_{2i})] \end{aligned}$$

Where the functions g^* and g and the parameter β_1^* depend on features of the firm's intertemporal profit function. Abel and Eberly's model also implies disinvestment ($Y_i < 0$) if Q_i is below some lower bound. Very few firms in the data set have negative investment, so that outcome will not be explicitly modeled. The above equations for Y and D hold as written for all firms if Y_i is set to zero for any firm having negative investment.

Note in this model that profit maximization results in the features that the outcome Y is linear in Q when $D = 1$, and that the fixed cost parameter a_{1i} appears only in the expression for D.

This theoretical model implies one sided censoring, but one could imagine more elaborate versions that would give rise to two sided censoring, e.g., if the benefits from investment were sufficiently large then one might choose to build an entirely new plant rather than invest more in the old one.

Marginal plant level Tobin's Q is not observed, and so will be proxied by the profit rate P_i. Specifically, Q_i is assumed to be linear in P_i, X_{2i}, and an additive error, where X_{2i} is a vector of observable attributes of the firm or plant. The function $g^*(a_{2i})$ is also assumed to be linear in X_{2i} and an additive error. This yields the outcome model

$$Y_i = (P_i \beta_1 + X_{2i}^T \beta_2 + \varepsilon_i) D_i \tag{14}$$

The error term ε_i will be independent of profits, or nearly so, if a collection of restrictive assumptions hold (including constant returns to scale, competitive product markets, and a first order autoregressive model for P_i. See Abel and Eberly for details). Because these assumptions are unlikely to hold in practice, the estimator here will not require ε_i to be independent of P_i, i.e., the estimator will allow for possible endogeneity of profits.

Let Z_i be a vector of instruments, comprised of Z_{1i} defined as the lagged profit rate, and plant characteristics $Z_{2i} = X_{2i}$. Define the function H by $H(z) = E(P \mid Z = z)$, and define ε_{pi} by

$$P_i = H(Z_i) + \varepsilon_{pi} \tag{15}$$

The function H is unknown. Because of endogeneity of profits, the error term ε_{pi} may be correlated with ε_i, and is not assumed to be independent of Z_i.

Let V_i be a measure of the size of plant i. In standard Q models, the relationship of the investment rate Y to Q does not depend on the size of the firm or plant, except to the extent that both Y and Q are expressed in "per unit of capital" terms. However, in empirical applications it is generally found that size does matter. The Abel and Eberly model provides an explanation, by allowing V_i to affect the fixed cost of investments a_{1i}. In particular, a_{1i} is the fixed cost per unit of capital, so if true fixed costs (in absolute terms) are present, then a_{1i} will be a decreasing function of V_i. Nilsen and Schiantarelli (2000) find strong statistical evidence of this relationship, including much greater incidences of zero investments in small versus large plants. They attribute this relevance of plant size both to the presence of absolute as well as relative fixed costs and to potential indivisibilities in investment. Many other studies confirm the relevance of size on the decision to invest, but most cannot separate plant level effects from other factors, because they use more aggregated firm or industry level data.

Based on the above, it is assumed that a_{1i} depends on V_i, and may also depend on X_{2i} and on unobserved characteristics of the plant, firm, or industry. Consistent with the presence of absolute fixed costs, Nilsen and Schiantarelli (2000) find strong evidence that D is monotonically increasing in V, so (recalling Lemma 1) we may write the resulting selection equation as

$$D_i = I[0 \leq V_i + M(P_i, X_{2i}, e_i)] \tag{16}$$

for some function M, where e_i denotes a vector of unobserved variables or errors that affect the decision to invest. The unobservables e_i will in general be correlated with the other unobservables in the system, ε_i and ε_{pi}.Also, in the Abel and Eberly model the function g is nonlinear in a_{1i} (it's related to a root of a quadratic equation) and a_{1i} itself is an unknown, possibly nonlinear function of V_i. Therefore M, which is based on g, a_{1i}, and a_{2i}, is an unknown function that is likely to be nonlinear.

The goal is estimation of the parameters β of the outcome equation (14), given the selection equation (16) and the instrument equation (15). The coefficient of P_i, which is β_1, is of particular interest as the proxy for the relationship between investment and Q.

5.2 Data and Estimation

The outcome equation is estimated using data from Norwegian manufacturing plants in 1986, ISIC codes (industry numbers) 300-390. The available sample consists of $n = 974$ plants. See Nilsen and Schiantarelli (2000) for a full data description. The main advantage over more conventional investment data sets is that the data here are available at the level of individual manufacturing plants, rather than firm level data

that is aggregated across plants. This is important because the theory involving fixed costs applies at the plant level, and averaging this nonlinear model across plants or firms may introduce aggregation biases, particularly in the role of variables affecting D_i, such as V_i.

Y_i is investment just in equipment in plant i in 1986, divided by the beginning of the year's capital stock in the plant. The investment rate Y_i equals zero in about tweny per cent of the plants. Around two percent of plants have negative investment. Consistent with the model, negative investment plants have Y_i set to zero. The selection function is then $D_i = I(Y_i > 0)$.

The variable P_i is profits attributable to plant i in 1986, divided by the beginning of the year's capital stock. Plant characteristics X_{2i} consist of a constant term, dummy variables for two digit ISIC code, and dummies indicating whether the firm is a single plant or multiplant firm, and if multiplant, whether plant i is the primary manufacturing facility or a secondary plant. The instruments Z_i are comprised of $Z_{2i} = X_{2i}$, and Z_{1i} defined as lagged P_i, so Z_{1i} is the profit rate for the plant in 1985. The size variable V_i is taken to be the log of employment at plant i in 1978 (or later for some plants for which 1978 data were unavailable).

To apply this paper's estimator for β, we need the assumptions of Corollary 3 to hold. The structural model is equations (14), (15), and (16). Plant size V appears only in the selection equation (16) of this model, as required. This is consistent with previous studies (using more aggregated data) that, without controlling for selection, find Y correlated with size.

Assumption 3 requires that the unobservables in the model, e, ε, and ε_p, be conditionally independent of V, conditioning on Z. Some unobservables, such as those determining a_2, are independent of V by construction of the underlying theoretical model. It is likely that the error terms ε and ε_p are also at least close to conditionally independent of V, because they are additive errors in rate equations, while V is a level or size variable. Also, profits and lagged profits are dated 1986 and 1985, respectively, while V is measured in 1978. Still it certainly possible in this application that V does not completely satisfy the required conditional independence assumptions.

The underlying supports of the variables in this model are unknown, so the required support conditions cannot be directly verified. However, in this data set V takes on a large range of values relative to the other covariates, and hence the asymptotic bias from one sided trimming is likely to be small. For example, the standard deviation of V is 1.16, while the profit rate P has a standard deviation of .17. In the applications where, for comparison, the selection equation is parameterized, the systematic component of $M(X_2, e)$, modeled as $X_2^T\gamma$, has a standard deviation comparable to that of V, ranging from .80 to 1.40 depending on the model and the estimator. In a Monte Carlo analysis of the related estimator given in Theorem 8, Lewbel

(2000) found that the estimator generally performed well when the standard deviation of V was comparable in magnitude to the standard deviation of M.

Strong alternative assumptions are required to estimate β by other means, such maximum likelihood. The model can be rewritten in a partly reduced form as

$$\begin{aligned} P_i &= H(Z_i) + \varepsilon_{pi} \\ Y_i &= [H(Z_i)\beta_1 + X_{2i}^T\beta_2 + (\varepsilon_{pi}\beta_1 + \varepsilon_i)]D_i \\ D_i &= I[0 \leq V_i + M(H(Z_i) + \varepsilon_{pi}, X_{2i}, e_i)] \end{aligned}$$

The parametric model that will be estimated for comparison is

$$\begin{aligned} P_i &= Z_i^T b + \widetilde{\varepsilon}_{pi} \\ Y_i &= [(Z_i^T b)\beta_1 + X_{2i}^T\beta_2 + \widetilde{\varepsilon}_i)D_i \\ D_i &= I[0 \leq V_i + (Z_i^T b)\gamma_1 + X_{2i}^T\gamma_2 + \widetilde{e}_i] \end{aligned}$$

where the errors $(\widetilde{e}_{pi}, \widetilde{\varepsilon}_i, \widetilde{e}_i)$ are assumed to be trivariate normal and independent of Z_i and V_i. Unlike the general semiparametric specification, this parametric model assumes that the functions H and M are linear, that the errors and unobservables ε_{pi} and e_i can be subsumed into a single additive error $\widetilde{e}_i$, and that the errors are jointly normal and independent of Z. Assumptions like these are required for estimation of the model by standard methods such as maximum likelihood, although they are not well motivated in terms of the economics of the problem. For example, linearity of the function M with a scalar error is inconsistent with the theoretical derivation of the model. This illustrates the value of the proposed semiparametric estimator, which does not require such assumptions.

5.3 Empirical Results

Let X_{1i} denote the vector consisting of P_i and the elements of X_{2i}, and correspondingly β is the vector of β_1 and β_2.

Table 1 reports results for six different estimators. The first and second estimators ignore the sample selection problem, and just estimate the equation $Y_i = X_{1i}^T\beta + \widetilde{\varepsilon}_i$ by ordinary least squares and two stage least squares, respectively (the latter using instruments Z_i).

The third estimator controls for sample selection parametrically, but does not control for possible endogeneity. This is the two equation parametric model $Y_i = (X_{1i}^T\beta + \widetilde{\varepsilon}_i)D_i$ and $D_i = I[0 \leq V_i + X_{1i}^T\gamma + \widetilde{e}_i]$,

assuming $\widetilde{\varepsilon}_i$ and $\widetilde{e}_i$ are jointly normal and independent of V_i and X_{1i}. This third estimator is the standard Heckman model, estimated using maximum likelihood.

The fourth estimator is maximum likelihood estimation of the entire parametric model, which entails simultaneously estimating the parametric selection, outcome, and instrument equations, assuming $\widetilde{e}_{pi}$, $\widetilde{\varepsilon}_i$, and $\widetilde{e}_i$ are jointly normal and independent of Z_i and V_i.

The fifth estimator is the semiparametrically weighted ordinary least squares estimator of β, that is, $\widehat{\beta} = \left(\Sigma_{i=1}^{n} \widehat{W}_i X_{1i} {X_{1i}}^T\right)^{-1} \Sigma_{i=1}^{n} \widehat{W}_i X_{1i} Y_i$, using weights $\widehat{W}_i = \hat{f}(V_i \mid X_{1i})^{-1} D_i$. This semiparametrically controls for selection but not endogeneity, and so corresponds to estimating β when the true model is defined by the system of two equations (14) and (16), assuming ε_i is uncorrelated with P_i and X_{2i}.

The final estimator is the semiparametrically weighted two stage least squares estimator given by equation (10). Here the weights are $\widehat{W}_i = \hat{f}(V_i \mid X_i)^{-1} D_i$, where $X_i = X_{1i}, Z_{1i}$. This estimator semiparametrically controls for both selection and endogeneity, and so corresponds to estimating β when the true model is defined by the full general structure of equations (14), (15), and (16).

The density estimator $\hat{f}(V_i \mid X_i)^{-1}$ is given by equation (4) with no trimming, so $I_t(V_i, X_i) = 1$ for all observations, and a quartic kernel, with bandwidth chosen by ordinary cross validation. Estimates were also generated with bandwidth's constructed using the procedure described in Lewbel (2000), and by halving the cross validated bandwidths to undersmooth as required for root n convergence. Those are not reported, since the resulting coefficient estimates were not very sensitive to bandwidth choice.

The semiparametric estimators were computationally quick and straightforward, since they only entail kernel density estimation and linear two stage least squares. In contrast, the maximum likelihood estimates were quite difficult to obtain, with frequent numerical problems and failures to converge. The difficulty with maximum likelihood is that some parameters are intrinsically difficult to identify, in particular correlations between the latent selection error $\widetilde{e}_i$ and the other model errors, and many structural parameters were sensitive to the estimates of these correlations. The semiparametric estimator does not require estimation of these difficult to obtain nuisance parameters.

In both the parametric and semiparametric models, controlling for selection and for endogeneity each raise the estimate of β_1 (recall the empirical finding in this literature is that naive estimates of this coefficient are implausibly low). The semiparametric estimates are comparable to, though generally higher than, the corresponding parametric model estimates.

This one sided censoring model is in theory less favorable for the estimator than two sided censoring would be, and one could easily question whether V satisfies all of the required conditional independence

assumptions in this application. Of course the maximum likelihood estimators also require some rather suspect, though very different, strong assumptions. Still, the empirical results are sensible, suggesting at a minimum that the semiparametric estimator produced plausible results here. Moreover, the similarity in estimates obtained by the parametric and semiparametric estimators should increase confidence in at least rough validity of the underlying model.

6 Conclusions

If a binary selection or treatment indicator D is monotonic in a continuous covariate V, then under mild regularity conditions either $D = I(0 \leq M + V)$ or $D = I(M + V \leq 0)$ for some latent M. Let $Y = DY^*$ for some unobserved Y^*. This paper assumes the general structure $D = I(a_0 \leq M + V \leq a_1)$ for either finite or infinite a_0 and a_1, and shows that identification of the entire conditional distribution of Y^*, conditioned on covariates X, can be obtained by conditional independence and support assumptions regarding the single covariate V. In short, strong assumptions about one observed covariate can replace the usual strong assumptions about the joint distribution of D and Y^*.

In particular, the mean of Y^* or of XY^* can be estimated as a weighted average of Y or XY, with weights $W/E(W)$, where W equals D divided by the conditional density of V given X. As a result, linear estimators that ordinarily could only be applied to Y^* if Y^* were observed, such as least squares, two stage least squares, kernel regressions, or differencing out fixed effects in panel models, can instead be applied to $WY/E(W)$. Essentially, this weighting converts expectations of the censored data into expectations of uncensored data. As a result, any estimator that is based on expectations can then be applied to the weighted, censored data. Rather than weighing by a propensity score estimate, it is sufficient to weight by the density of a covariate V that affects the propensity score.

The usefulness of these results in any application of course depends on whether an appropriate covariate V exists. This paper provided one empirical application, and cited other studies that possess a plausible candidate V. It seems likely that, in at least some applications, one would be more comfortable making assumptions about an observed covariate than the alternative, which requires assumptions about the joint distribution of all the unobservables that affect both selection and outcomes. If nothing else, one would have more confidence in the results produced by more conventional estimators if the very different identifying assumptions employed here yield comparable estimates.

7 Appendix

This Appendix provides proofs of theorems, lemmas, and corollaries, and provides the statements of some additional required lemmas.

LEMMA 1. Assume D is a random variable that takes on the value zero or one, and V is a continuously distributed random scalar. Assume there exists a random vector U and a function φ such that $D = \varphi(V, U)$ where $\varphi(V, U)$ is monotonic in V. Assume there exists (possibly infinite) constants v_0 and v_1 such that $prob[\varphi(v_j, U) = j] = 1$ for $j = 0, 1$, and that the support of V contains the interval $[v_0, v_1]$ (or the interval $[v_1, v_0]$ if $v_1 < v_0$). Then there exists a function $M(U)$ such that either $D = I[0 \leq M(U) + V]$ or $D = I[M(U) + V \leq 0]$.

PROOF OF LEMMA 1. Consider first the case where $\varphi(v, u)$ is increasing in v. For all u in the support of U, define the function M by $M(u) = -\inf\{v \mid v \in [v_0, v_1], \varphi(v, u) = 1\}$. Then $D = I[0 \leq M(U) + V]$. If $\varphi(v, u)$ is decreasing in v then let $M(u) = -\sup\{v \mid v \in [v_0, v_1], \varphi(v, u) = 0\}$ to obtain $D = I[M(U) + V \leq 0]$. ■

PROOF OF THEOREM 1. Assume for a given function $h(Y^*, X)$ that $E[h(Y^*, X)W]$ exists. Then, given Assumptions 1, 2, 3, and 4A,

$$\begin{aligned}
E[h(Y^*, X)W] &= E\left(\frac{h(Y^*, X)D}{f_v(V \mid X)}\right) \\
&= E\left[E\left(\frac{h(Y^*, X)I(a_0 \leq M + V \leq a_1)}{f_v(V \mid X)} \mid X, Y^*, M\right)\right] \\
&= E\left(\int_{\text{supp}(v)} \frac{h(Y^*, X)I(a_0 \leq M + v \leq a_1)}{f_v(v \mid X)} f_v(v \mid X, Y^*, M)dv\right) \\
&= E\left(\int_{\text{supp}(v)} h(Y^*, X)I(a_0 \leq M + v \leq a_1)dv\right) \\
&= E\left(h(Y^*, X)\int_{\text{supp}(v)} I(a_0 - M \leq v \leq a_1 - M)dv\right) \\
&= (a_1 - a_0)E[h(Y^*, X)]
\end{aligned}$$

Taking $h(Y^*, X)$ to equal one above yields $E(W) = (a_1 - a_0)$, and therefore for any function $h(Y^*, X)$ we have

$$E[h(Y^*, X)W] = E(W)E[h(Y^*, X)] \tag{17}$$

Now consider a function $g(Y, X) = g(Y^*D, X)$. Recalling that $W = 0$ whenever $D = 0$ we have

$$\begin{aligned} E[Wg(Y,X)] &= E[W(g(Y,X) - g(0,X)] + E[Wg(0,X)] \\ &= E[W(g(Y^*,X) - g(0,X)] + E[Wg(0,X)] \\ &= E(W)E[g(Y^*,X) - g(0,X)] + E(W)g(0,X)] \\ &= E(W)E[g(Y^*,X)] \end{aligned}$$

where the third equality above follows from applying equation (17) twice, once with $h(Y^*, X) = g(Y^*, X) - g(0, X)$ and once with $h(Y^*, X) = g(0, X)$. ■

PROOF OF COROLLARY 1. Let $[\delta_0^*, \delta_1^*]$ denote the support of V. Follow the same steps as in the proof of Corollary 2 below, replacing $-\tau$ and τ with δ_0^* and δ_1^*, respectively. In particular $WI(|V| \leq \tau)$ is then replaced with $WI(\delta_0^* \leq V \leq \delta_1^*) = W$, and the resulting bias term is then $O(\delta_0^{*-1} + \delta_1^{*-1}) = O(\tau^{-1})$. ■

PROOF OF COROLLARY 2. Define $\zeta(M, \tau) = \min(a_1 - M, \tau) - \max(a_0 - M, -\tau)$. Following the same steps as in the proof of Theorem 1, for any function $h(Y^*, X)$, we have

$$\begin{aligned} E[h(Y^*,X)WI(|V| \leq \tau)] &= E\left(\int_{\text{supp}(v)} h(Y^*,X)I(a_0 \leq M + v \leq a_1)I(|v| \leq \tau)dv\right) \\ &= E\left(h(Y^*,X)\int_{\text{supp}(v)} I[\max(a_0 - M, -\tau) \leq v \leq \min(a_1 - M, \tau)]dv\right) \\ &= E[\zeta(M,\tau)h(Y^*,X)] \end{aligned}$$

and therefore,

$$\frac{E[h(Y^*,X)WI(|V| \leq \tau)]}{E[WI(|V| \leq \tau)]} = E[h(Y^*,X)] + \frac{cov[\zeta(M,\tau), h(Y^*,X)]}{E[\zeta(M,\tau)]}$$

If a_1 and a_0 are both finite then for sufficiently large τ we will have $a_1 - M \leq \tau$ and $a_0 - M \geq -\tau$ for all M, which makes $\zeta(M, \tau) = (a_1 - a_0)$ which is constant and nonzero, and hence $cov[\zeta(M, \tau), h(Y^*, X)]/E[\zeta(M, \tau)] = 0$. If $a_1 = \infty$ and a_0 is finite then for sufficiently large τ we will have $\zeta(M, \tau) = \tau - a_0 + M$, in which case

$$\frac{cov[\zeta(M,\tau), h(Y^*,X)]}{E[\zeta(M,\tau)]} = \frac{cov[M, h(Y^*,X)]}{\tau - a_0 + E(M)} = O(\tau^{-1}) \tag{18}$$

The remaining case of $a_0 = -\infty$ proceeds in the same way. The proof is finished by following the same steps as in the proof of Theorem 1 to go from $h(Y^*, X)$ to $g(Y, X)$. ■

PROOF OF THEOREM 2. Let f_v denote the probability density function of v and let F_{*m} denote the joint distribution function of Y^* and M. For any function h we have

$$\int_{\delta_0}^{\delta_1} E[h(Y^*, x)D \mid X = x, V = v]dv$$

$$= \int_{\delta_0}^{\delta_1} \int_{\text{supp}(Y^*, M)} I(a_0 \leq m + v \leq a_1) h(y^*, x) dF_{*m}(y^*, m \mid X = x, V = v) dv$$

$$= \int_{\text{supp}(Y^*, M)} \left[\int_{\delta_0}^{\delta_1} I(a_0 - m \leq v \leq a_1 - m) dv \right] h(y^*, x) dF_{*m}(y^*, m \mid X = x)$$

$$= \int_{\text{supp}(Y^*, M)} [(a_1 - a_0)] h(y^*, x) dF_{*m}(y^*, m \mid X = x)$$

$$= (a_1 - a_0) E[h(Y^*, X) \mid X = x]$$

Next, observe that $[g(Y, X) - g(0, X)]D = [g(Y^*, X) - g(0, X)]D$, so we may apply the above result with $h(\psi, x) = g(\psi, x) - g(0, x)$. We may also apply it with $h(\psi, x) = 1$, which yields

$$\mu(x) = g(0, x) + \frac{(a_1 - a_0) E[g(Y^*, X) - g(0, X) \mid X = x]}{E[(a_1 - a_0) E(1 \mid X)]} = E[g(Y^*, X) \mid X = x]$$

■

LEMMA 2. Let Assumptions 1,2, and 3 hold. Assume $[-\tau, \tau]$ is in the support of V, and that $cov[M, g(Y^*, X) - g(0, X) \mid X = x]$ and $E(M \mid X = x)$ exist. Then $\mu(x, -\tau, \tau) = E[g(Y^*, X) \mid X = x] + O(\tau^{-1})$.

PROOF OF LEMMA 2. If $-\tau \leq \delta_0$ and $\tau \geq \delta_1$ then $\mu(x, -\tau, \tau) = \mu(x) = E[g(Y^*, X) \mid X = x]$ by Theorem 2. Consider next the case where $-\tau \leq \delta_0$ and $\tau \leq \delta_1$. Following the same steps as in the proof of Theorem 2 gives

$$\int_{-\tau}^{\tau} E[h(Y^*, x)D \mid X = x, V = v]dv$$
$$= E[(\tau - a_0 + M) h(Y^*, X) \mid X = x]$$

so

$$\begin{aligned} \mu(x, -\tau, \tau) &= g(0, x) + \frac{E[\tau - a_0 + M)(g(Y^*, X) - g(0, X)) \mid X = x]}{E[\tau - a_0 + M) \mid X = x]} \\ &= E[g(Y^*, X) \mid X = x] + \frac{cov[M, g(Y^*, X) - g(0, X) \mid X = x]}{\tau - a_0 + E[M \mid X = x]} \end{aligned}$$

which proves the result. The case of $-\tau > \delta_0$ works in the same way. ■

PROOF OF COROLLARY 3. Applying Theorem 1 with $g(Y, X) = Z(Y - X_1^T\beta)$ yields $E[WZ(Y - X_1^T\beta)]/E(W) = E[Z(Y^* - X_1^T\beta)] = E(Z\varepsilon) = 0$, and $\beta = \Delta E(WZY)$ follows immediately. Applying Corollary 1 with this g yields the result $\beta = \Delta E(WZY) + O(\tau^{-1})$, and the $\beta(\tau)$ result follows from similarly applying Corollary 2 to $E[I(|v| \leq \tau)WZY]/E[I(|v| \leq \tau)W]$ and to $E[I(|v| \leq \tau)WZX_1^TY]/E[I(|v| \leq \tau)W]$. ■

PROOF OF THEOREM 3. Let $\overline{W}$ and $\overline{Wg}$ denote the sample means of W_i and $W_i g(Y_i, X_i)$, respectively. By algebra,

$$(\widehat{\omega}_1 - \omega) = \frac{\overline{Wg} - \overline{W}\omega}{E(W)} + R_n$$

where the remainder term R_n is given by

$$R_n = \left(\frac{\overline{Wg} - \overline{W}\omega}{E(W)}\right)[\overline{W} - E(W)]\left(\frac{-1}{\overline{W}}\right)$$

Both $\overline{Wg} - \overline{W}\omega$ and $\overline{W} - E(W)$ are sample means of mean zero, iid random variables having finite variances (the latter is ensured by conditions (i) and (ii)), so by the Lindeberg-Levy central limit theorem each is $O_p(n^{-1/2})$. Now for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \Pr[n^{-1/2}\overline{W}^{-1} < \varepsilon] = \lim_{n\to\infty} \Pr[n^{1/2}(\overline{W} - E(W)) > \varepsilon^{-1} - n^{1/2}E(W)] = 1$$

where the second equality follows because $n^{1/2}(\overline{W} - E(W))$ goes to a normal, and so the limit goes to probability that a normal is greater than $-\infty$. We therefore have that $\overline{W}^{-1} = o_p(n^{1/2})$, so $R_n = o_p(n^{-1/2})$, and the central limit theorem may now be applied to the above expression for $(\widehat{\omega}_1 - \omega)$. ■

PROOF OF THEOREM 4. Apply equation (3) with $S = D$ to show that $n^{-1}\sum_{i=1}^n \widehat{W}_i - E(W)$ has a mean zero, root n limiting distribution. Follow the same steps as in the proof of Theorem 3 to show that $\widehat{\omega}$ has the same limiting distribution as $n^{-1}\sum_{i=1}^n \widehat{W}_i[g(Y_i, X_i) - \omega]$. Apply equation (3), now with $S = [g(Y, X) - \omega]D$ to obtain equation (7). ■

PROOF OF THEOREM 5. Apply equation (3) with $S = ZYD$ and $S = DZX_1^T$ to obtain the limiting distributions for $\hat{\eta} = n^{-1}\sum_{i=1}^n Z_i Y_i \widehat{W}_i$ and for $n^{-1}\sum_{i=1}^n \widehat{W}_i Z_i X_{1i}^T$. These yield $\sqrt{n}(\hat{\Delta} - \Delta) = O_p(1)$ (analogous to the treatment of $(n^{-1}\sum_{i=1}^n \widehat{W}_i)^{-1}$ in Theorem 3) and $\sqrt{n}\hat{\eta} = n^{-1/2}\Sigma_{i=1}^n Q_{\beta i} + o_p(1)$, so

$$\begin{aligned}\sqrt{n}\hat{\beta} &= \sqrt{n}\hat{\Delta}\hat{\eta} = n^{-1/2}\hat{\Delta}\Sigma_{i=1}^n[W_i Z_i X_{1i}^T\beta + (Q_{\beta i} - W_i Z_i X_{1i}^T\beta)] + o_p(1) \\ &= \sqrt{n}\beta + n^{-1/2}\hat{\Delta}\Sigma_{i=1}^n(Q_{\beta i} - W_i Z_i X_{1i}^T\beta)) + o_p(1)\end{aligned}$$

which yields the distribution for $\hat{\beta}$. ■

PROOF OF THEOREM 6. Consider first the estimator $\widetilde{\mu}(x) = g(0,x) + \kappa^{-1}\int_{\delta_0}^{\delta_1} \widehat{m}(v,x)dv$. It follows from Theorem 4 of Lewbel, Linton, and McFadden (2001) that, with an arbitrary $\widetilde{Y}$,

$$\frac{\widetilde{\mu}(x) - \mu(x) - b^2\overline{\beta}(x))}{\sqrt{n^{-1}b^{-d}\overline{\omega}(x)}} \xrightarrow{d} N(0,1).$$

so all that remains is to show that $\widetilde{\mu}(x)$ and $\widehat{\mu}(x)$ have the same limiting distribution. By assumption, there exists some constant c such that $\delta_0 \geq \inf[\text{supp}(V)] + c$ and $\delta_1 \leq \sup[\text{supp}(V)] - c$. Therefore the probability that $\widehat{v}_0 < \delta_0$ and that $\widehat{v}_1 > \delta_1$ goes to one at a fast rate. Also, $m(v,x) = m^*(v,x) = 0$ for all $v < \delta_0$ and all $v > \delta_1$, and as a result, use of $\widehat{v}_0$ in place of δ_0 and $\widehat{v}_1$ in place of δ_1 will have no effect on the limiting distributions of $\widehat{\kappa}$ and $\widehat{\mu}(x)$. Next, we have that estimation of $\widehat{\kappa}$ entails averaging over X, which therefore converges at a faster rate than $\int_{\delta_0}^{\delta_1} \widehat{m}(v,x)dv$, so estimation error in $\widehat{\kappa}$ is also asymptotically irrelevant. The result is that $\widetilde{\mu}(x)$ and $\widehat{\mu}(x)$ have the same limiting distribution. ■

The following Lemma will be used in the proof of Theorem 7.

LEMMA 3: Assume that $f_v(v \mid x)$ is bounded away from zero except at $v = \infty$, and that for some constant c, $f_v(v \mid x) = f_v(v)$ for all $v \geq c$. Assume that $D = I(0 \leq M + V)$. Define $Z_{\tau i}$ by

$$Z_{\tau i} = \frac{\phi_{\tau i} D_i I(|V_i| \leq \tau)}{\tau f_v(V_i \mid X_i)}$$

where $\phi_{ti} = \phi(Y_i^*, X_i) + O(\tau^{-1})$ for some function $\phi(Y_i^*, X_i)$. Assume that for some $\nu > 2$, $E[|\phi_\tau|^\nu]$ exists for all large τ and in the limit as $\tau \to \infty$. Let $\overline{Z}_\tau$ denote the sample mean of $Z_{\tau i}$. Define

$$\Gamma_\nu(\phi, t) = E[\phi_\tau^\nu \int_c^t f_v(v \mid X)^{1-\nu} dv] \tag{19}$$

Define $\Gamma_{\tau\nu} = \Gamma_\nu(1,t)$ and $\gamma_\tau = \tau^{-2}\Gamma_{\tau 2}$. If $\tau \to \infty$ and $\Gamma_{\tau 2}^{-\nu/2}\Gamma_{\tau\nu} n^{1-\nu/2} \to 0$ then

$$\frac{\sqrt{n/\gamma_\tau}[\overline{Z}_\tau - E[\phi(Y^*, X)] + O(1/\tau)]}{\sqrt{E[\phi(Y^*, X)^2] + O(\Gamma_{\tau 2}^{-1})}} \xrightarrow{d} N(0,1).$$

PROOF OF LEMMA 3. The assumption that $f_v(v \mid x) = f_v(v)$ for $v \geq c$ makes $\Gamma_\nu(\phi_\tau, \tau) = E(\phi_\tau^\nu)\Gamma_{\tau\nu}$. Positive densities must have a finite integral, so $f_v(\tau) = o(\tau^{-1})$, and therefore $\tau^{-\nu}\Gamma_{\tau\nu} \to \infty$ and

$\Gamma_{\tau\xi}^{-1}/\Gamma_{\tau\nu} \to \infty$ for any $\nu > \xi \geq 1$. For sufficiently large τ, $DI(|V| \leq \tau) = I(-M \leq V \leq \tau)$, so for any ν we have $E(Z_\tau^\nu) = [\Gamma_\nu(\phi_\tau, \tau) - \Gamma_\nu(\phi_\tau, -M)]\tau^{-\nu}$, which for $\nu = 1$ simplifies to $E(Z_\tau) = E(\phi) + O(\tau^{-1})$ and for $\nu > 1$ gives $\tau^{-\nu}\Gamma_{\tau\nu}E(Z_\tau^\nu) = E(\phi^\nu) + O(\Gamma_{\tau\nu}^{-1})$ and $\gamma_\tau var(Z_\tau) = E(\phi^2) + O(\Gamma_{\tau 2}^{-1})$.

By the central limit theorem for double arrays, $\sqrt{n}[\overline{Z_\tau} - E(Z_\tau)]/\sqrt{var(\overline{Z_\tau})} \xrightarrow{d} N(0,1)$ if $E[|[Z_\tau - E(Z_\tau)]/\sqrt{var(\overline{Z_\tau})}|^\nu]n^{1-\nu/2} \to 0$ is satisfied for some $\nu > 2$. Define $Z_{\tau i}^*$ the same as $Z_{\tau i}$, except with $|\phi_{\tau i}|$ in place of $\phi_{\tau i}$. Now $|Z_{\tau i} - E(Z_\tau)| \leq Z_{\tau i}^* + E(Z_\tau^*)$ and so by a Taylor expansion $E[|Z_\tau - E(Z_\tau)|^\nu] \leq E(Z_{\tau i}^{*\nu}) + O[E(Z_{\tau i}^{*\nu-1})] = O(\tau^{-\nu}\Gamma_{\tau\nu})$, so the required moment condition holds if $\Gamma_{\tau 2}^{-\nu/2}\Gamma_{\tau\nu}n^{1-\nu/2} \to 0$, as assumed. ■

PROOF OF THEOREM 7. Let $W_\tau = WI(|V| \leq \tau)/\tau$ and let $\overline{W}_\tau$ denote the sample mean of $W_{\tau i}$. Applying Lemma 3 with $\phi_{ti} = 1$ shows that $E(W_\tau) = 1 + O(\tau^{-1})$, $\overline{W}_\tau - E(W_\tau) = O_p[(n/\gamma_\tau)^{-1/2}]$, and that, similar to the proof of Theorem 3, $\overline{W}_\tau^{-1} = o_p[(n/\gamma_\tau)^{1/2}]$.

Now let $\phi_i = [g(Y_i^*, X_i) - \omega]$, $\phi_{ti} = E(W_\tau)^{-2}\phi_i$ and $Z_{ti} = W_\tau\phi_{ti}$. Let $\overline{W}_\tau$ and $\overline{Z_\tau}$ denote the sample means of $W_{\tau i}$ and $Z_{\tau i}$, respectively. We have $E(\phi) = 0$ and so by Lemma 3, $\overline{Z_\tau} = O_p[(n/\gamma_\tau)^{-1/2}]$. Now

$$(\widehat{\omega}_\tau - \omega) = \overline{Z_\tau} + R_n$$

where the remainder term R_n is given by

$$R_n = -\overline{W_\tau}^{-1}\overline{Z_\tau}[\overline{W_\tau} - E(W_\tau)] = o_p[(n/\gamma_\tau)^{1/2}]$$

so the rate $(n/\gamma_\tau)^{1/2}$ limiting distribution of $(\widehat{\omega}_\tau - \omega)$ equals the limiting distribution of $\overline{Z_\tau}$, which is given by Lemma 3. ■

PROOF OF COROLLARY 4. It can be readily verified that the proof of Corollary 2, and hence Corollary 3, holds replacing τ with $\tau - X_1^T\gamma$. It then follows from Corollary 3 and equation (18) that

$$\begin{aligned}
\frac{E[I(V \leq \tau - X_1^T\gamma)WZY]}{E[I(V \leq \tau - X_1^T\gamma)W]} &= E(ZX_1^T)\beta + \frac{E[I(V \leq \tau - X_1^T\gamma)WZ(Y - X_1^T\beta)]}{E[I(V \leq \tau - X_1^T\gamma)W]} \\
&= E(ZX_1^T)\beta + \frac{cov(\tau - X_1^T\gamma + M, Z\varepsilon)}{E(\tau - X_1^T\gamma + M)} \\
&= E(ZX_1^T)\beta + E(Z)\frac{E(e\varepsilon)}{\tau} = E(ZX_1^T)(\beta + b)
\end{aligned}$$

■

References

[1] ABEL, A. B. AND J. C. EBERLY, (1994) "A Unified Model of Investment Under Uncertainty," *American Economic Review*, 84, 1369-1384.

[2] AHN, AND J. L. POWELL, (1993), "Semiparametric Estimation of Censored Models with a Nonparametric Selection Mechanism," *Journal of Enometrics*, 58, 3-29.

[3] AI, C., AND MCFADDEN, D. (1997), "Estimation of Some Partially Specified Nonlinear Models," *Journal of Econometrics* 76, 1-37.

[4] AIT-SAHALIA, Y., P. J. BICKEL, AND T. M. STOKER (1997), "Goodness of Fit Tests For Regression Using Kernel Methods," Unpublished Manuscript.

[5] ALONSO, A. A., S. A. FERNÁNDEZ, AND J. RODRIGUEZ-PÓO (1999), "Semiparametric Estimation of a Duration Model, Universidad del País Vasco and Universidad de Cantabria unpublished manuscript

[6] AMEMIYA, T. (1973), "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997–1016.

[7] AMEMIYA, T. (1985) *Advanced Econometrics*. Harvard University Press.

[8] ANDREWS, D.W.K. (1994) Asymptotics for Semiparametric Econometric Models by Stochastic Equicontinuity. Econometrica 62, 43-72.

[9] ANDREWS, D. W. K., (1995), "Nonparametric Kernel Estimation for Semiparametric Models," *Econometric Theory*, 11, 560–596.

[10] ANDREWS, D. W. K. AND M. M. A. SCHAFGANS (1998), "Semiparametric Estimation of the Intercept of a Sample Selection Model," *Review of Economic Studies*, 65, 497-517.

[11] BICKEL, P.J., C.A.J. KLAASSEN, J. RITOV, AND J. WELLNER (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer: Berlin.

[12] BLUNDELL, R., S. BOND, AND C. MEGHIR (1996), "Econometric Models of Company Investment," in The econometrics of panel data: A handbook of the theory with applications. Matyas, Laszlo Sevestre, Patrick, eds., Second edition, London: Kluwer Academic. 685-710.

[13] BLUNDELL, R. AND J. L. POWELL (1999), Endogeneity in Single Index Models, unpublished manuscript.

[14] CHAMBERLAIN, G. (1986), "Asymptotic Efficiency in Semiparametric Models With Censoring," *Journal of Econometrics*, 32, 189-218.

[15] CHAMBERLAIN, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics* 34, 305-334.

[16] CHAUDHURI, P. (1991). "Nonparametric estimates of regression quantiles and their local Bahadur representation," *Annals of Statistics* 19, 760-777.

[17] CHEN, S. AND L. F. LEE. (1998), "Efficient Semiparametric Scoring of Sample Selection Models," *Econometric Theory*, 14, 423-462.

[18] CHOI, K. (1990) The Semiparametric Estimation of the Sample Selection Model Using Series Expansion and the Propensity Score," University of Chicago manuscript.

[19] COSSLETT, S. R. (1991), "Semiparametric Estimation of a Regression Model with Sample Selectivity," in W. A. Barnett, J. L. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.

[20] DAS, M., (1998), "Nonparametric Estimation Methods for Sample Selection and Instrumental Variables," Massachusetts Institute of Technology PH.D. thesis.

[21] DAS, M., W. K. NEWEY, AND F. VELLA (2000), "Nonparametric Estimation of Sample Selection Models," unpublished manuscript.

[22] DOMS, M. AND T. DUNNE, (1998), "Capital Adjustment Patterns in Manufacturing Plants," *Review of Economic Dynamics*, 1, 409-429.

[23] DONALD, S. (1995), "Two Step Estimation of Heteroskedastic Sample Selection Models," *Journal of Econometrics*, 65, 347-380.

[24] EISNER, R. AND R. STROTZ, (1963) "Determinent of Investment Behavior," in *Impact of Monetary Policy*. Englewood Cliffs,NJ: Prentice-Hall.

[25] FAN, J., AND I. GIJBELS (1996), *Local Polynomial Modelling and Its Applications* Chapman and Hall.

[26] GALLANT, A. R. AND D. W. NYCHKA (1987), "Semi-nonparametric Maximum Likelihood Estimation," *Econometrica* 55, 363-390.

[27] GOZALO, P., AND O.B. LINTON (2000). Local Nonlinear Least Squares: Using Parametric Information in Nonparametric Regression. *Journal of Econometrics, 99, 63-106.*

[28] GRONAU, R. (1974), "Wage Comparisons - A Selectivity Bias," *Journal of Political Economy,* 82, 1119-1144.

[29] HAHN, J. (1998), On the Role of The Propensity Score in Efficient Semiparametric Estimation of Average Treatement Effects, *Econometrica*, 66, 315-331.

[30] HANSEN, B. (1999), "Threshold Effects in Non-Dynamic Panels: Estimation, Testing, and Inference," *Journal of Econometrics*, 93, 345-368.

[31] HARDLE, W., J. HART, J. S. MARRON, AND A. B. TSYBAKOV, (1992) "Bandwidth Choice for Average Derivative Estimation," *Journal of the American Statistical Association*, 87, 218-226.

[32] HARDLE, W. AND J. L. HOROWITZ (1996), "Direct Semiparametric Estimation of Single-Index Models With Discrete Covariates," *Journal of the American Statistical Association*, 91, 1632-1640.

[33] HECKMAN, J. (1974), "Shadow Prices, Market Wages, and Labor Supply," *Econometrica*, 42, 679-693

[34] HECKMAN, 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, 475-495

[35] HECKMAN, J. (1976), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153-161.

[36] HECKMAN, J. (1990), "Varieties of Selection Bias," *American Economic Review: Papers and Proceedings*, 313-318.

[37] HECKMAN, J. AND B. E. HONORÉ, (1990), "The Empirical Content of the Roy Model," *Econometrica*, 58, 1121-1149.

[38] HECKMAN, J., H. ICHIMURA, AND P. TODD (1998), Matching as an Econometric Evaluation Estimator, *Review of Economic Studies*, 65, 261-294.

[39] HECKMAN, J. AND T. E. MACURDY (1986), "Labor Econometrics," Handbook of Econometrics, vol. 3, ed. by Z. Griliches and M. D. Intriligator, pp. 1917-1977, Amsterdam: Elsevior.

[40] HIRANO, K., G. W. IMBENS AND G. RIDDER (2000), Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score, unpublished manuscript.

[41] HONORÉ, B. E. AND A. LEWBEL (2000), "Semiparametric Binary Choice Panel Data Models Without Strictly Exogeneous Regressors," unpublished manuscript.

[42] HOROWITZ, J. L., (1998), "Nonparametric estimation of a generalized additive model with an unknown link function," Iowa City Manuscript.

[43] HAYASHI, F., (1982) "Tobin's Marginal q and Average q: A Neoclassical Interpretation." *Econometrica*, 50, 213-224.

[44] ICHIMURA, H. AND L. F. LEE (1991), "Semiparametric Least Squares Estimation of Multiple Index Models: Single Equation Estimation," in W. A. Barnett, J. L. Powell, and G. Tauchen, eds., *Nonparametric and Semiparametric Methods in Econometrics and Statistics*. Cambridge: Cambridge University Press.

[45] IMBENS, G. W., AND J. ANGRIST (1994), Identification and Estimation of Local Average Treatment Effects, *Econometrica*, 62, 476-476.

[46] KOUL, H. L., V. SUSARLA AND J. VAN RYZIN (1981), "Regression Analysis With Randomly Right Censored Data," *Annals of Statistics* 9, 1276-1288.

[47] KHAN, S. AND A. LEWBEL (1999), "Weighted and Two Stage Least Squares Estimation of Semiparametric Censored and Truncated Regressions," unpublished manuscript.

[48] KIM, W., O. LINTON AND N. HENGARTNER (1999). A computationally efficient oracle estimator for additive nonparametric regression with bootstrap confidence intervals. *The Journal of Computational and Graphical Statistics* 8, 278-297.

[49] KYRIAZIDOU, E. (1997), "Estimation of a Panel Data Sample Selection Model," *Econometrica* 65, 1334-1364

[50] LEE, L. F. (1982), "Some Approaches to the Correction of Selectivity Bias," *Review of Economic Studies*, 49, 355-372.

[51] LEE, L. F. (1992), "Semiparametric Two Stage Estimation of Sample Selection Models Subject to Tobit-type Selection Rules," *Journal of Econometrics*, 61, 305–344.

[52] LEE, L. F. (1994), "Semiparametric Instrumental Variables Estimation of Simultaneous Equation Sample Selection Models," *Journal of Econometrics*, 63, 341–388..

[53] LEWBEL, A. (1995), "Consistent Nonparametric Tests With An Application to Slutsky Symmetry," *Journal of Econometrics*, 67, 379–401.

[54] LEWBEL, A. (1997), "Semiparametric Estimation of Location and Other Discrete Choice Moments," *Econometric Theory, 13, 32-51.*

[55] LEWBEL, A. (1998), "Semiparametric Latent Variable Model Estimation With Endogenous or Mismeasured Regressors," *Econometrica*, 66, 105–121.

[56] LEWBEL, A. (2000), "Semiparametric Qualitative Response Model Estimation With Unknown Heteroscedasticity or Instrumental Variables," *Journal of Econometrics*, 97, 145-177.

[57] LEWBEL, A. (2000a), "Asymptotic Trimming for Bounded Density Plug-in Estimators," Unpublished manuscript.

[58] LEWBEL, A., O. LINTON, AND D. MCFADDEN (2001), "Estimating Features of a Distribution From Binomial Data," Unpublished manuscript.

[59] LINTON, O. (1997) Efficient estimation of additive nonparametric regression models. *Biometrika*, 84, 469-474.

[60] LINTON, O. (2000). Efficient estimation of generalized additive nonparametric regression models. *Econometric Theory* 16, 502-523.

[61] LINTON, O. AND J.P. NIELSEN (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82, 93-100.

[62] MADDALA, G. S. (1983), *Limited Dependent and Qualitative Variables in Econometrics*, Econometric Society Monograph No. 3, Cambridge: Cambridge University Press.

[63] MANSKI, C. (1994), "The Selection Problem," In Sims, C. Ed., *Advances in Econometrics,* Cambridge: Cambridge University Press.

[64] MASRY, E. (1996a), "Multivariate local polynomial regression for time series: Uniform strong consistency and rates," *J. Time Ser. Anal.* 17, 571-599.

[65] MASRY, E., (1996b), "Multivariate regression estimation: Local polynomial fitting for time series. *Stochastic Processes and their Applications* 65, 81-101.

[66] MAURIN, E. (1999), "The Impact of Parental Income on Early Schooling Transitions: A Reexamination Using Data Over Three Generations," CREST-INSEE unpublished manuscript.

[67] MCFADDEN, D. L. (1984), "Econometric Analysis of Qualitative Response Models," Handbook of Econometrics, vol. 2, ed. by Z. Griliches and M. D. Intriligator, pp. 1395-1457, Amsterdam: Elsevior.

[68] MCFADDEN, D. L. (1993), "Estimation of Social Value From Willingness-To-Pay Data," Unpublished Manuscript.

[69] NEWEY, W. K. (1985), "Semiparametric Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables," Annales de l'INSEE, n59-60, 219-237.

[70] NEWEY, W. K. (1988), "Two Step Estimation of Sample Selection Models," Princeton University manuscript.

[71] NEWEY, W. K. (1994), "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

[72] NEWEY, W. K. (1999), "Consistency of Two-Step Sample Selection Estimators Despite Misspecification of Distribution," *Economics Letters*, 63, 129-132.

[73] NEWEY, W. K. AND D. MCFADDEN (1994), "Large Sample Estimation and Hypothesis Testing," in Handbook of Econometrics, vol. iv, ed. by R. F. Engle and D. L. McFadden, pp. 2111-2245, Amsterdam: Elsevier.

[74] NEWEY, W. K. AND P. A. RUUD (1994), "Density Weighted Linear Least Squares," University of California at Berkeley working paper.

[75] NILSEN, O. A. AND F. SCHIANTARELLI (2000), "Zeroes and Lump Sums in Investment: Empirical Evidence on Irreversibilities and Non-Convexities," Unpublished Manuscript, Boston College.

[76] POWELL, J. L., (1994), "Estimation of Semiparametric Models," Handbook of Econometrics, vol. 4, ed. by R. F. Engle and D. L. McFadden, pp. 2443-2521, Amsterdam: Elsevior..

[77] POWELL, J. L., (1987), "Semiparametric Estimation of Bivariate Latent Variable Models," University of Wisconsin Working Paper no. 8704.

[78] POWELL, J. L., J. H. STOCK, AND T. M. STOKER (1989), "Semiparametric Estimation of Index Coefficients," *Econometrica* 57, 1403–1430.

[79] POWELL, J. L., AND T. M. STOKER (1996), "Optimal Bandwidth Choice For Density-Weighted Averages," *Journal of Econometrics*, 75, 291-316.

[80] RACINE, J. AND Q. LI (2000), "Nonparametric Estimation of Conditional Distributions With Mixed Categorical and Continuous Data," Unpublished manuscript.

[81] ROBINSON, PETER M. (1988), "Root-N-Consistent Semiparametric Regression," *Econometrica*, 56, 931–954.

[82] RUBIN, D. (1974), Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies, *Journal of Educational Psychology*, 66, 688-701.

[83] SERFLING, R. J. (1980), Approximation Theorems of Mathematical Statistics, New York: John Wiley and Sons.

[84] SILVERMAN, B. (1986), Density estimation for statistics and data analysis. London, Chapman and Hall.

[85] STOKER, T. M. (1991), "Equivalence of Direct, Indirect and Slope Estimators of Average Derivatives," in *Nonparametric and Semiparametric Methods in Econometrics and Statistics*, W. A. Barnett, J. Powell, and G. Tauchen, Eds., Cambridge University Press.

[86] SUMMERS, L. H. (1981), "Taxation and Corporate Investment: A q-Theory Approach," *Brookings Papers on Economic Activity,* 1, 67-127.

[87] VELLA, F. (1999), "Estimating Models With Sample Selection Bias: A Survey," *Journal of Human Resources*, 33, 127-169.

[88] VELLA, F. AND M. VERBEEK (1999), "Two-Step Estimation of Panel Data Models With Censored Endogeneous Variables and Selection Bias," *Journal of Econometrics*, 90, 239-263.

[89] VYTLACIL, E. (2001), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result," *Econometrica,* forthcoming.

[90] WAINER, H. (1986), Drawing inferences from self-selected samples New York Springer-Verlag.

[91] WOOLDRIDGE, J. M. (1995), "Selection Corrections for Panel Data Models Under Conditional Mean Independence Assumptions," *Journal of Econometrics*, 68, 115-132.

Table 1. Estimates of the Outcome Equation Profit Coefficient

	no dummies		plant type dummies		types & ISIC dummies	
OLS	.231	.036	.219	.035	.221	.035
2SLS	.383	.051	.353	.050	.355	.050
Heckman	.298	.087	.287	.092	.298	.094
Endogeneous ML	.468	.061	.403	.062	.413	.057
Weighted OLS	.323	.062	.317	.059	.316	.051
Weighted 2SLS	.470	.070	.431	.073	.411	.080

Notes: In each block, the first number is β_1, the coefficient of the profit rate in the outcome equation, and the second number is the estimated standard error. In the first pair of columns, X_2 and Z_2 consist only of the constant term. In the second pair of columns, X_2 and Z_2 also include plant type dummies, and in the third pair of columns, X_2 and Z_2 contain dummies both for plant type and for two digit industry (ISIC) code.