Simple Estimators For Hard Problems: Endogeneity and Dependence in Binary Choice Related Models

and

Endogenous Selection or Treatment Model Estimation

Arthur Lewbel
Boston College

This pdf file contains both of the above papers.

Simple Estimators For Hard Problems: Endogeneity in Discrete Choice Related Models

Arthur Lewbel Boston College

August 2004

Abstract

This paper describes numerically simple estimators that can be used to estimate binary choice and other related models (such as selection and ordered choice models) when some regressors are endogenous or mismeasured. Simple estimators are provided that allow for discrete or otherwise limited endogenous regressors, lagged dependent variables and other dynamic effects, heteroskedastic and autocorrelated latent errors, and latent fixed effects.

Keywords: Binary choice, Binomial Response, Endogeneity, Measurement Error, Dynamics, Autocorrelation, Fixed Effects, Panel models, Identification, Latent Variable Models.

I would like to thank Thierry Magnac, Kit Baum, Russell Davidson, Andrew Chesher, and Whitney Newey for helpful comments and suggestions. Any errors are my own.

Corresponding Author: Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, http://www2.bc.edu/~lewbel/

1 Introduction

This paper describes numerically simple estimators that can be used to estimate binary choice (binomial response) models when some regressors are endogenous or mismeasured, and when latent errors can be heteroskedastic and correlated with regressors. Two types of estimators are discussed: control function estimators and estimators that exploit a very exogenous regressor. The latter estimators allow endogenous regressors to be discrete, censored, truncated, and otherwise limited.

Independent, identically distributed observations are assumed for most of the paper, but extension sections show how the very exogenous regressor estimators can still remain simple while allowing for dynamic effects, fixed effects, and autocorrelated latent errors. Extensions are also provided to other limited dependent variable models such as ordered response and sample selection mdoels.

This paper focuses on descriptions of estimators for applied work, rather than limiting distribution theory. Most of the estimators take standard forms such as GMM, and are also simple enough to allow the use of ordinary boostrapping for generating test statistics and confidence intervals if desired. Readers that are primarily interested in applied work may want to skim the preliminary and theoretical sections of the paper to focus on the estimators themselves, which are described in simple 'recipe' forms.

Since much of this paper involves variants of existing estimators, in each section I will flag what material is new.

1.1 Why Simple Estimators?

Define a simple estimator as one that:

1. Requires few or no choices of smoothers such as kernels, bandwidths, or polynomial orders.
2. Closely resembles, (or consists of steps that each resemble) estimators that are already in common use.
3. Requires few or no numerical searches or numerical maximizations.

The simple estimators in this paper require more restrictive assumptions on the data generating process, than other, harder estimators. The goal here is to achieve substantial estimator simplification with only modest decreases in generality. Given that most econometric theory is devoted to the development of estimators that increase generality, why consider more restrictive, simple estimators?

1. Finite sample performance of hard estimators is often sensitive to the exact choice of smoother and of the choice of numerical search or optimization algorithm.

2. With simple estimators, it is numerically feasible to apply the bootstrap or other resampling techniques for generating confidence intervals and hypothesis tests. Ordinary asymptotic limiting variances may also be easier to estimate.

3. With simple estimators, there is less chance of coding mistakes, and greater chance of widespread use.

4. Simple estimators avoid numerical search failures. With hard estimators, particularly in problems with many parameters, grid searches are computationally infeasible, while hill climbing and other numerical search techniques often fail, due to features of complicated objective functions such as ridges, cliffs, inflection points, and multiple local maxima.

5. Simple estimators can provide good starting values for more general, difficult estimators.

2 Efficiency and Standard Errors for Two Step Estimators

Many simple estimators, including most of the estimators provided in this paper, take the form of two (or multi) step estimators. For example, consider the well known "Heckit" sample selection model estimator (Heckman 1976). The first step consists of estimating a probit model $D = I(X'\widehat{\gamma} + \varepsilon \geq 0)$ for selection by maximizing a likelihood function $L(D, X, \gamma)$, yielding estimated parameters $\widehat{\gamma}$. Then in a second step parameters $\widehat{\beta}, \widehat{\lambda}$ are estimated by applying ordinary least squares to the linear regression model $Y = X'\beta + m(X'\widehat{\gamma})\lambda + e$, where $m(X'\widehat{\gamma})$ is the estimated inverse mills ratio.

Two stage estimators like this one are generally inefficient, and the standard errors that result in the second stage are often incorrect, because they fail to account for estimation error in the first stage. This section describes a simple cure for both of these problems, from Newey (1984).

For data Z, Assume first stage estimates $\widehat{\gamma}$ are obtained from applying the method of moments (MM) or generalized method of moments (GMM) to moments of the form $E[g_1(Z, \gamma)] = 0$ for some m_1 vector of known functions g_1, and assume second stage estimates $\widehat{\beta}$ are obtained from applying MM or GMM to $E[g_2(Z, \beta, \gamma) \mid \gamma = \widehat{\gamma}] = 0$ for some m_2 vector of known functions g_2.

For example, in the Heckit procedure $\widehat{\gamma}$ is obtained by maximizing the probit likelihood function $L(Z, \gamma)$, which is equivalent to letting $\widehat{\gamma}$ be the value of γ that solves $n^{-1}\sum_{i=1}^{n} \partial L(Z_i, \gamma)/\partial\gamma = 0$, so the function $g_1(Z_i, \gamma) = \partial L(Z_i, \gamma)/\partial\gamma$ is the probit score vector. In the second step the ordinary least squares $\widehat{\beta}$ (and $\widehat{\lambda}$) are the solutions to $n^{-1}\sum_{i=1}^{n}\left[Y_i - X_i'\beta - \lambda m(X_i'\widehat{\gamma})\right](X_i', m(X_i'\widehat{\gamma}))' = 0$, which defines the function $g_2(Z_i, \beta, \gamma)$ such that $n^{-1}\sum_{i=1}^{n} g_2(Z_i, \beta, \widehat{\gamma}) = 0$ (redefining β to include both β and λ).

More generally, for two step estimation each step can take the form of maximum likelihood, linear or nonlinear regression, linear or nonlinear two stage least squares, GMM estimation, etc.,.

PROPOSITION 1: Define $g(Z, \beta, \gamma) = (g_1(Z, \gamma)', g_2(Z, \beta, \gamma)')'$, so $g(Z, \beta, \gamma)$ is the $m_1 + m_2$ vector consisting of all the functions $g_1(Z, \gamma)$ and $g_2(Z_i, \beta, \gamma)$. Applying efficient GMM to the moments $E[g(Z, \beta, \gamma)] = 0$ yields estimators for β and γ that have correct standard errors and are at least as efficient as two step estimation. Two step estimation can be used as the first stage of the standard GMM estimator, or as starting values for standard preprogrammed GMM packages.

For iid data, the joint GMM estimator takes the form

$$\widehat{\gamma}, \widehat{\beta} = \arg\min_{\gamma,\beta} \sum_{i=1}^{n} g(Z_i, \beta, \gamma)' \Omega_n \sum_{i=1}^{n} g(Z_i, \beta, \gamma). \tag{1}$$

Two step estimation can be written as a special case of this GMM estimator using a block triangular estimated weighting matrix Ω_n. Efficient estimation and consistent standard errors are obtained by using the standard formulas for the efficient choice of Ω_n. General analyses of two step estimators includes Newey and McFadden (1994, section 6) and Wooldridge (2002, section 12.4). These general analyses describe conditions under which the second step estimators are efficient or when they yield correct standard errors, and provide consistent standard error formulas.

Proposition 1 is not new (see Newey 1984), but it appears to be often overlooked. In particular, the fact that it can be applied when one of the steps is maximum likelihood (by taking g_1 or g_2 to be a score function) is rarely noted. Given currently existing computing power and readily available automated GMM estimation programs, this general procedure should be broadly applicable, especially since the two step estimators can themselves provide good, consistent starting values for estimation.

To illustrate, in the Heckit model $E[g(Z, \beta, \gamma)] = 0$ takes the form

$$E\left[\begin{array}{c} \partial L(Z,\gamma)/\partial\gamma \\ \left[Y - X'\beta - m(X'\gamma)\lambda\right] D\left(\begin{array}{c} X \\ m(X'\gamma) \end{array}\right) \end{array}\right] = 0$$

For moments involving derivatives, such as the score function $\partial L(Z, \gamma)/\partial\gamma$, either numerical or computer calculated analytic derivatives can be placed directly into GMM routines, to avoid manual derivative calculation.

Also, to minimize or avoid numerical searches if necessary, asymptotic efficiency can be obtained without iterating to convergence, e.g., Newey and McFadden (1994, section 3.4) show that asymptotic efficiency is obtained by just doing one iteration of the efficient GMM estimator. This result may be applied to all of the GMM estimators that will be proposed in this paper to avoid numerical searches, though iterating to convergence is likely to be preferable in practice.

3 Binomial Response Models: Some Preliminaries

3.1 Convenient representation

The representation described in this section is common in semiparametric work. The reasons and uses for this representation are summarized here.

Consider the linear latent variable binary choice or binomial response model

$$D = I(\widetilde{X}'\widetilde{\beta} + \widetilde{\varepsilon} \geq 0)$$

where D is an observed dummy variable that equals zero or one, $\widetilde{X}$ is a vector of observed regressors, $\widetilde{\beta}$ is a vector of coefficients to be estimated, $\widetilde{\varepsilon}$ is an unobserved error having variance equal one, and I is the indicator function that equals one if its argument is true and zero otherwise. The probit model has $\widetilde{\varepsilon} \sim N(0, 1)$, while for logit $\widetilde{\varepsilon}$ has a logistic distribution.

Let V be some conveniently chosen exogenous element of $\widetilde{X}$ that is known to have a positive coefficient, and let X be the vector of remaining elements of $\widetilde{X}$. Then the linear latent model can be equivalently written as

$$D = I(X'\beta + V + \varepsilon \geq 0)$$

where the variance of ε is some unknown constant σ_ε^2, and β is a vector of coefficients to be estimated. In the special case of the probit model, $\varepsilon \sim N(0, \sigma_\varepsilon^2)$.

These two representations of the model are completely equivalent, with parameters that are related by.$\widetilde{\beta} = (\beta, 1)/\sigma_\varepsilon$ (where V has been taken to be the last element of $\widetilde{X}$). One can easily rewrite probit or logit code to report estimates of β and σ_ε instead of $\widetilde{\beta}$. Specifically, the log likelihood function for the scaled probit is $\sum_{i=1}^{n} L(D_i, X_i, V_i, \beta, \sigma_\varepsilon)$ where

$$L(D, X, V, \beta, \sigma_\varepsilon) = D \ln \Phi\left(\frac{X'\beta + V}{\sigma_\varepsilon}\right) + (1 - D) \ln\left[1 - \Phi\left(\frac{X'\beta + V}{\sigma_\varepsilon}\right)\right]. \tag{2}$$

and Φ is the standard normal cumulative distribution function. Define the *scaled probit* model to be this representation of the probit model.

Choice probabilities are given by $\Pr(D = 1) = 1 - F_\varepsilon[-(X'\beta + V)]$ where F_ε is the CDF of ε, instead of the equivalent $\Pr(D = 1) = 1 - F_{\widetilde{\varepsilon}}[-(\widetilde{X}'\widetilde{\beta})]$. For normal errors the CDF's are $F_{\widetilde{\varepsilon}}(\cdot) = \Phi(\cdot)$ and $F_\varepsilon(\cdot) = \Phi(\cdot/\sigma_\varepsilon)$.

If one is not sure about the sign of the coefficient of V, one way it can be determined is as the sign of the estimated average derivative $E[\partial E(D \mid V, X)/\partial V]$. Signs of estimators can generally be estimated at faster than root n rates, so a first stage estimate of the sign need not affect the later distribution theory. Alternatively, in many applications the signs of some covariates are known a priori from economic theory.

In some applications the economics of the problem provide a natural scaling, for example, if D is the decision of a consumer to purchase a good and we take V to be the negative of the logged price of the good faced by the consumer, then $X'\beta$ is the log of the consumer's reservation price (that is, their willingness to pay) for the good.

Note also that the signs of the coefficients $(\beta, 1)$ are the same as the signs of the coefficients $\widetilde{\beta}$, and the relative magnitudes of the coefficients are also the same, that is $\widetilde{\beta}$ is proportional to $(\beta, 1)$, so the estimated marginal rates of substitution between any two elements of $\widetilde{X}$ is the same using either $(\beta, 1)$ or $\widetilde{\beta}$. In applications where the distribution of the latent error is unknown but independent of $\widetilde{X}$, this distribution, and resulting estimated choice probabilitites may be equivalently estimated by a nonparametric regression of D on either $X'\beta + V$ or on $\widetilde{X}'\widetilde{\beta}$.

In general $\widetilde{X}$, and therefore X, will be assumed to include a constant term, which is estimated as part of β. However, many semiparametric estimators, such as Klein and Spady (1993), cannot be used to estimate the constant. In applications where those estimators are used $\widetilde{\varepsilon}$ and ε can be assumed to have a nonzero mean or median.

A common strategy in semiparametric estimation is to construct an estimator

of β (either with or without a constant term), then separately estimate the error distribution by, e.g., nonparametrically regressing D on $X'\beta + V$. One reason for this separation is that β can often be estimated at a faster rate (even root n) than the distribution function. Examples of semiparmetric estimators of β include Manski (1975),(1985), Cosslett (1983), Ruud (1983), Powell, Stock and Stoker (1989), Horowitz (1992), (1993), Ichimura (1993), Klein and Spady (1993), Newey and Ruud (1994), Härdle and Horowitz (1996), Lewbel (2000), and Blundell and Powell (2003).

3.2 Binomial Response With Endogenous Regressors

Let Y be a vector of endogenous or mismeasured regressors, and let W be a vector of exogenous covariates. Let

$$\begin{aligned} X &= (Y', X_2')' \\ W &= (Z_1', X_2', V)' \end{aligned}$$

for vectors Z_1 and X_2, so the vector of instruments that do not appear in the structural model (excluded or outside instruments) is Z_1, the set of regressors in the model for D is X, V, and the set of all covariates is Y, W. The model is

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0) \\ &= I(Y'\beta_1 + X_2'\beta_2 + V + \varepsilon \geq 0) \end{aligned}$$

where the latent error ε is uncorrelated with the exogenous covariates Z, but may be correlated with the endogenous or mismeasured regressors Y. We observe a sample of n observations of D_i, Y_i, W_i. The latent error may also be heteroskedastic, having second and higher moments that may depend on W and Y.

One possible estimator with many well known attractive features is maximum likelihood. A disadvantage of maximum likelihood is that it requires a complete parametric specification of the joint distribution of ε and of all the endogenous regressors Y, conditional on all of the exogenous regressors W. Also, the resulting estimates of β can be very sensitive to nuisance parameters that are difficult to estimate, like the covariances between the latent error ε and the errors in the parameterized model of Y.

Commonly used coefficient estimators are to just do probit or logit, ignoring the endogeneity of Y and heteroskedasticity of ε, or to estimate a linear probability model, that is, run a linear two stage least squares regression of D on

X,V using instruments W. Both procedures are inconsistent, among other obvious drawbacks.

The goal here is to provide a range of alternative estimators, which are numerically very simple and consistent under reasonably general conditions.

3.3 Types of Endogeneity

Recalling that $X = (Y', X_2')$, the general class of of binomial response models with endogenous regressors to be considered is

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0) \\ Y &= g(W, U) \end{aligned}$$

for some function g and some vector of unobservable errors U. Endogeneity arises from correlations or other dependence between U and ε. This correlation could be due to measurement error in Y, or simultaneity in the determination of Y and D. Higher moments of ε could also depend on W and U.

Two classes of estimators for binomial response models with endogenous regressors will be discussed. These are control function estimators and very exogenous (or 'special') regressor estimators. Control function estimators assume models of the form

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0) \\ Y &= h(W) + U \end{aligned}$$

along with assumptions regarding the condition distribution ε given U, W, and entail explicitly estimating U.

Instrumental variables models do not impose restrictions on, nor explicitly estimate U, but instead assume only that we have instruments (elements of W) that are correlated with Y and uncorrelated with the modeling error ε. Instrumental variable estimators cannot be directly applied to binomial response models (other than the linear probability model), however, they can be applied to a certain transformation of D, if there exists one exogenous regressor in the model that has special properties, which can be taken to be V. These are instrumental variable estimators based on a very exogenous regressor.

3.4 Choice Probabilities With Endogenous Regressors

In this section, only the index choice probability concept is new material (and it has likely been used implicitly in the past).

One reason for estimating β in binary choice models is to then use the results to estimate choice probabilities, that is, the probability that $D = 1$.

With endogenous regressors or heteroskedastic errors, there are a few different possible choice probabilities one might construct. The probability that $D = 1$, conditional on covariates is

$$E(D \mid Y, W) = 1 - F_\varepsilon[-(X'\beta + V) \mid Y, W]$$

where $F_\varepsilon(\varepsilon \mid \cdot)$ denotes the conditional distribution function of ε, conditioning on the information set $(\cdot)$. Estimating choice probabilities using this definition requires knowing or modeling the entire conditional distribution of ε given Y, W. If little is known about this distribution, it may be simpler (and in some cases no less efficient) to ignore the β estimate and just nonparametrically regress D on Y, W. A common modeling assumption that simplifies estimation of the choice probability is to assume that $F_\varepsilon(\varepsilon \mid Y, W) = F_\varepsilon(\varepsilon \mid U)$ where U is the error term in a regression of Y on W. In this case the choice probability can be obtained by nonparametrically regressing $1 - D$ on $-(X'\widehat{\beta} + V)$ and on $\widehat{U}$, or by a parametric model if $F_\varepsilon(\varepsilon \mid U)$ has a known functional form.

An alternative choice probability definition is what Blundell and Powell (2000) call the average structural function. For the binomial response model the average structural function is

$$1 - F_\varepsilon[-(X'\beta + V)]$$

that is, the marginal distribution of ε evaluated at the regression function. The average structural function is roughly analogous to, in a linear model context, forecasting using the fitted values of two stage least squares regression; since in that situation we evaluate the error term at its marginal mean (zero), even though its conditional mean, conditioning on the regressors (including endogenous regressors), would be nonzero.

A third choice probability measure is to just condition on the estimated index $X'\beta + V$, that is

$$E(D \mid X'\beta + V) = 1 - F_\varepsilon[-(X'\beta + V) \mid X'\beta + V]$$

We might call this the index choice probability, since this will equal $E(D \mid Y, W)$ when the distribution of ε depends on Y, W only through the single linear index $X'\beta + V$, in which case β could be estimated using general single index model estimators such as Ichimura (1993). More generally, the index choice probability can be interpreted as a middle ground between the previous two measures (conditioning ε on all covariates versus on none). One advantage of the index choice

probability versus other measures is that it can be readily estimated even when the conditional or marginal distribution of ε is unknown, by a one dimensional nonparametric regression of D on $X'\widehat{\beta} + V$, given any consistent estimator $\widehat{\beta}$.

When ε is independent of Y, W all three choice probability measures are the same. When ε is independent of $X'\beta + V$ the second and third measures are the same, and when ε is depends on Y, W only through $X'\beta + V$, then the first and third measures are the same.

Implementing any of these choice probability measures assumes we have a consistent estimator of β. Regardless of which measure is used for evaluating the estimates, the step of estimating β should take into account any endogeneity in X and hence any dependence of ε on covariates. So, e.g., β estimation based either on control functions or a very exogenous regressor would be appropriate.

4 Control Function Estimators for Binomial Response Models With Endogenous Regressors

First consider the model

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0) \\ Y &= W'b + U \quad E(WU) = 0 \\ \varepsilon &= U'\gamma + \eta, \ \ \eta \perp U, W, \quad \eta \sim N(0, \sigma_\eta^2) \end{aligned}$$

Where b is an unknown constant vector (or matrix if Y is a vector), λ is an unknown constant scalar (or vector if Y is a vector) and U is a mean zero error uncorrelated with W. The error ε is assumed to be linear in U and in another error η which is a mean zero normal independent of both W and U. A leading special case in which these assumptions hold is when U and ε are jointly normal and independent of W, though in general U is not required to be normal or homoskedastic.

Substituting out ε in the D equation yields

$$D = I(X'\beta + V + U'\gamma + \eta \geq 0).$$

This is just a probit model with regressors X, V, and U. This suggests the following simple estimator.

ESTIMATOR A

1. For each observation i, construct data $\widehat{U}_i = Y_i - W_i'\widehat{b}$, which are the residuals of an ordinary least squares regression of Y on W (or seemingly unrelated regression if Y is a vector).

2. Let $\widehat{\beta}$ be the estimated coefficients of X in an ordinary scaled probit regression of D on X, V and $\widehat{U}$.

Recall that the scaled probit is a probit that normalizes the coefficient of V to be one instead of normalizing the variance of the latent error to be one. This conveniently keeps β unchanged, unlike the ordinary probit that, in this second step, would normalize the variance of η to be one, which is a different scale than normalizing the variance of ε to be one. Up to scaling, Estimator A is identical to Quong and Rivers (1988) and the control function estimator of Blundell and Smith (1986), which in turn is closely related to Nelson and Olsen (1978) and is the basic idea proposed by Heckman (1978).

The estimated coefficients are root n consistent and asymptotically normal, but the standard error estimates for $\widehat{\gamma}$ and $\widehat{\beta}$ generated by second stage probit fail to take into account the estimation error in the construction of $\widehat{U}$. Correct standard error formulas can be obtained by applying the general theory of two step estimators (see, e.g., Newey and McFadden, 1994, Theorem 6.2). Consistent standard errors can also be readily obtained by bootstrapping the data, which is practical given the numerical simplicity of the estimator. Bootstrapping the confidence intervals instead of standard errors has the added theoretical advantage in this context of providing a higher order approximation to the true limiting distribution.

Based on Proposition 1, more efficient estimates with correct standard errors can be obtained using a GMM estimator to combine the first and second steps in estimator A, as follows.

ESTIMATOR B.

Define

$$R(D, X, V, U, \beta, \gamma, \sigma_\eta) = D\frac{\phi\left[\left(X'\beta + V + U'\gamma\right)/\sigma_\eta\right]}{\Phi\left[\left(X'\beta + V + U'\gamma\right)/\sigma_\eta\right]} + (1-D)\frac{-\phi\left[\left(X'\beta + V + U'\gamma\right)/\sigma_\eta\right]}{1 - \Phi\left[\left(X'\beta + V + U'\gamma\right)/\sigma_\eta\right]}$$

Use ordinary GMM to estimate the parameters $b, \beta, \gamma, \sigma_\eta$ based on the mo-

ment conditions

$$\begin{aligned} E\left[W(Y - W'b)\right] &= 0 \\ E\left[R(D, X, V, Y - W'b, \beta, \gamma, \sigma_\eta)X\right] &= 0 \\ E\left[R(D, X, V, Y - W'b, \beta, \gamma, \sigma_\eta)\left(Y - W'b\right)\right] &= 0 \\ E\left[R(D, X, V, Y - W'b, \beta, \gamma, \sigma_\eta)V\right] &= 0 \end{aligned}$$

Estimator B consists of the moments that define b and the first order conditions for the maximum likelihood step that is, mean zero score functions. Estimator A can be applied first, both to provide consistent starting values for the GMM estimation, and to construct estimates of the efficient GMM weighting matrix. Efficient estimates can also be obtained by applying Amemiya's GLS as described by Newey (1987).

The control function method can be greatly generalized. Blundell and Powell (2003) provide a control function estimator for the model

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0) \\ Y &= h(W) + U, \quad E(U \mid W) = 0 \\ \varepsilon &\mid X, U \sim \varepsilon \mid U \end{aligned}$$

where the function h and the distribution of the errors is unknown. Estimation in this case is not simple, requiring a high dimensional nonparametric regression first step to estimate h, then a semiparametric multiple index estimator to obtain β.

In terms of modeling, the control function method is for the most part not applicable if the endogenous regressors Y are discrete, censored, truncated, or otherwise limited, because in such cases the distribution of U (and therefore its relationship to ε) will in general depend upon X. Also, unlike two stage least squares, if the assumptions hold for a set of instruments Z, then they will not hold in general using some smaller subset of instruments, so omitting variables that one is unsure about as instruments will result in inconsistent estimates, instead of just a loss of efficiency as in instrumental variables methods.

5 A Very Exogenous Regressor

Consider for the moment the linear probability model $D = X'\beta + \varepsilon$ where X includes a subvector of endogenous regressors Y. This model can be estimated

by linear two stage least squares. Two great advantages of linear two stage least squares estimation are its numerical simplicity, and the fact that it does not require explicit modeling of Y. Two stage least squares can be used without change regardless of whether endogenous regressors Y are continuous discrete, limited, truncated, etc.,. All that is needed are variables (instruments) Z that are correlated with Y and uncorrelated with ε. Unfortunately, two stage least squares is inconsistent for most limited dependent variable models such as logit, probit, ordered choice, tobit, etc.,.

This section describes estimators that preserve the above listed attractive features of two stage least squares without imposing a linear probability model. These estimators required the presence of one regressor, which without loss of generality is taken to be the V regressor, that is special in a sense that one might call "very exogenous." The definition of a very exogenous regressor is as follows.

Let S be a vector of covariates (including X and Z, so S can include endogenous regressors, ordinary exogenous regressors, and ordinary instrumental variables). Define V to be a very exogenous regressor if

1. $h(X) + V + \varepsilon$ is a latent variable in some model for some function h.
2. $V = g(\nu, S)$ for some function g that is differentiable and strictly monotonically increasing in its first element, with $\nu \perp S, \varepsilon$, and ν is continuously distributed.
3. The support of the conditional distribution of V given S contains the support of $-[h(X) + \varepsilon]$.

Condition 1 says that V is a regressor in the model, and the latent variable has been scaled to make its coefficient equal one. The latent variable is linear V and in an error ε.

The variable ν in Condition 2 can be interpreted as the error term in a model for V. Condition 2 is similar to the modeling assumptions in Matzkin (2003). It says that the error term ν in the model for V is independent of ε and of all the other covariates S, both exogenous and endogenous. This condition is a bit stronger than assuming that V and ε are conditionally independent, conditioning on S, which (when S is exogenous) is essentially the definition of ordinary exogeneity. Powell (1994), Section 2.5, discusses similar exclusion restrictions and their role in semiparametric identification. This condition is much weaker than the usual assumption that model errors ε are independent of all covariates, and arises naturally in some economic models. For example, in a labor supply model where ε represents unobserved ability, conditional independence is satisfied by

any variable V that affects labor supply decisions but not ability, such as government defined benefits. In demand models where ε represents unobserved preference variation, prices satisfy the conditional independence condition if they are determined by supply, such as under constant returns to scale production. Lewbel, Linton and McFadden (2001) consider applications like willingness to pay studies, where V is a bid determined by experimental design, and so satisfies the necessary restrictions by construction. An empirical application employing this assumption in a binary choice context (applying Lewbel 2000) is Cogneau and Maurin (2002), who analyze enrollment of children into school in Madagascar. For V they use the date of birth of the child within the relevant year, which strongly affects enrollment and is plausibly assumed to be conditionally independent of ε, which in this context consists of unobserved components of the child's abilities and unobserved socioeconomic factors. The continuity restriction in condition 2 can sometimes be relaxed. This is discussed in a later section on discrete V.

Condition 3 is a large support assumption. Requiring a regressor to have large or infinite support for identification is common in the literature on semiparametric limited dependent variable models. Examples include Manski (1975,1985) and Horowitz (1992) for heteroskedastic binary choice models, and Han (1987) and Cavanagh and Sherman (1998) for homoskedastic transformation models. The large support assumption can in some applications be replaced by assumptions regarding the distribution of the tail of ε. See, e.g., Magnac and Maurin (2003) and Chen (2002).

A simple model for V that will be used in this paper is

$$V = S'b + \nu, \quad \nu \perp S, \varepsilon, \quad \nu \sim N(0, \sigma^2)$$

This says that the model for V is linear with an independent, normal error. This can be easily generalized to allow for heteroskedastity in V as

$$V = S'b + \exp(S'c)\nu, \quad \nu \perp S, \varepsilon, \quad \nu \sim N(0, 1).$$

Both of these simple V models satisfy conditions 2 and 3.

5.1 Binomial Response Models With Endogenous Regressors and a Very Exogenous Regressor

The model and associated estimator described here is based on Lewbel (2000). It has somewhat different assumptions than the general model given in Lewbel

(2000), but as a result is very simple to implement. When all regressors are exogenous and g is linear this estimator simplifies to one of the estimators proposed in Lewbel, Linton, and McFadden (2003).

Let $Z = (Z_1', X_2')'$, so Z is the vector of all the available exogenous covariates except for V. Let $S = (Y', Z')'$, so S is all the available covariates except for V. Recall that $X = (Y', X_2')'$ is the set of all the regressors in the D equation except for V. The proposed estimators depend on the following Theorem, which is proved in Appendix.

THEOREM 1: Assume $D = I(X'\beta + V + \varepsilon \geq 0)$, $E(Z\varepsilon) = 0$, $supp(X'\beta + \varepsilon) \subseteq supp(-V \mid S)$, $V = g(\nu, S)$, g is differentiable and strictly monotonically increasing in its first element, $\nu \perp S, \varepsilon$, and ν is continuously distributed. Let $f(\nu)$ be the probability density function of ν. Define T and e by

$$\begin{aligned} T &= \frac{D - I(V \geq 0)}{f(\nu)} \frac{\partial g(\nu, S)}{\partial \nu} \\ e &= T - X'\beta \end{aligned}$$

Then $E(Ze) = 0$.

Theorem 1 says, instead of imposing strong restrictions on all of the endogenous regressors Y as in control function (or maximum likelihood) estimation, assume the standard latent variable form for binomial response models, assume we have ordinary instruments Z that are uncorrelated with the latent error term ε, and assume one of the regressors in the model is very exogenous. Then we can construct the variable T and estimate β by regressing T on X using linear two stage least squares with instruments Z.

For some intuition regarding Theorem 1, substitute out V in D to get $D = I[-(X'\beta + \varepsilon) \leq g(\nu, S)]$ so $E(D \mid S, \nu)$ equals the probability distribution function of $-(X'\beta + \varepsilon)$, conditioned on S, and evaluated at $V = g(\nu, S)$. Since V asymptotically takes on every value in the real line, this distribution function is identified everywhere, and the parameters β may be recovered from this distribution. This explains why the assumptions are sufficient for identification. To see why the resulting estimator has such a simple form, first note that the marginal density of ν corresponds to the conditional density of V, and that for expressions involving expectations, dividing by the conditional density of V is equivalent to converting V to a uniformly distributed random variable, that is, the conditional expectation of T (averaging over V), is equivalent to the conditional expectation of $D - I(V \geq 0)$ with a uniform V, which in turn is just

$\int [I(X'\beta + V + \varepsilon \geq 0) - I(V \geq 0)]dV$. When $X'\beta + \varepsilon > 0$ this integral is just $\int I(-X'\beta - \varepsilon \leq V \leq 0)dV = \int_{-X'\beta-\varepsilon}^{0} 1dV = X'\beta + \varepsilon$ and a similar result holds when $X'\beta + \varepsilon < 0$. It follows that the conditional expectation of T equals $X'\beta + \varepsilon$, so T equals $X'\beta + \varepsilon$ plus another error that has conditional mean zero. The error e is then just the sum of ε and this other error.

An implication of the large support assumption for V is that, for any value X and ε may take on, it is possible for V to be small enough to make $D = 0$, with probability one, or large enough to make $D = 1$ with probability one, This may not be plausible in some applications. Magnac and Maurin (2003) provides alternative restrictions that can be used to relax this large support assumption.

5.2 Simple Estimation of Binomial Response Models With Endogenous Regressors Based on a Very Exogenous Regressor

To make estimation based on Theorem 1 simple, a convenient parametric model is chosen here for g and f. Specifically, consider the model

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &= S'b + \nu, \quad \nu \perp S, \varepsilon, \quad \nu \sim N(0, \sigma^2) \end{aligned}$$

so V is linear in covariates plus a normal error. By Theorem 1, other regular parametric model for g and continuous distributions for ν could be assumed instead. This particular model is chosen for its simplicity.

Other than modeling restrictions involving the very exogenous regressor V, nothing more needs to be assumed for estimation except what would be required for a linear two stage least squares regression, namely, that $rank[E(ZX')] = rank[E(XX')]$ and $E(Z\varepsilon) = 0$. As a result, elements of Y and Z can be continuous, discrete, truncated, squared, interacted with each other, etc.,. Nothing else needs to be known or estimated regarding the data generating process of the endogenous regressors Y. For example, unlike control function models, this model and the resulting estimator can be used with a discrete endogenous regressor such as $Y = I(Z'\gamma + U \geq 0)$ with the joint distribution of U, ε unknown.

Based on this model and Theorem 1, we have the following simple estimator.

ESTIMATOR C

1. Make sure V_i takes on a range of both positive and negative values (V can be demeaned if not). Let $\widehat{b}$ be the estimated coefficients of an ordinary least

squares regression of V on S. For each observation i, construct data $\widehat{v}_i = V_i - S_i'\widehat{b}$, which are the residuals of this regression.

2. Let $\widehat{\sigma}^2$ be the sample mean of $\widehat{v}_i^2$, and for each observation i, let $f(\widehat{v}_i, \widehat{\sigma}^2)$ be the pdf of $\widehat{v}_i$, so

$$f(\widehat{v}_i) = \frac{1}{\sqrt{2\pi\widehat{\sigma}^2}} \exp\left(\frac{-\widehat{v}_i^2}{2\widehat{\sigma}^2}\right)$$

3. For each observation i construct data $\widehat{T}_i$ defined as

$$\widehat{T}_i = \frac{D_i - I(V_i \geq 0)}{f(\widehat{v}_i, \widehat{\sigma}^2)}$$

4. Let $\widehat{\beta}$ be the estimated coefficients of an ordinary linear two stage least squares regression of $\widehat{T}$ on X, using instruments Z.

This estimator differs from Lewbel (2000) mainly in that it assumes a parametric model for the very exogenous regressor V, while Lewbel (2000) used a nonparametric conditional density estimator for V. Lewbel (2000) is not strictly more general than the model assumed for Estimator C, since the above model allows V to depend on Y, while Lewbel (2000) assumed conditional independence.

Estimator C is a numerically trivial estimator, requiring no numerical searches, and involving nothing more complicated than linear regressions. It provides root n consistent, asymptotically normal estimates of the coefficients β. The correct standard errors are not equal to those that would be generated by ordinary two stage least squares in the last step, because they fail to take into account the estimation error in the construction of $\widehat{T}$. However, one may easily generate standard error and confidence interval estimates by bootstrapping, since the estimator itself is so simple. Alternatively, one may apply the GMM estimator described below.

Ordinary root n convergence in step 4 requires that ZT have a finite variance. Lewbel (2000) imposes this by assuming that $X'\beta + \varepsilon$ has bounded support, which makes the numerator of T identically zero for extreme values of v. Magnac and Maurin (2003) provides alternative sufficient conditions for root n convergence, which involve either a conditional moment restriction or a symmetry restriction in the tails of ε. These results, and Monte Carlo analyses in these papers, suggest that the estimator is sensitive to the choice of regressor V. In particular, the estimator depends heavily on variation in v, and so will tend to perform best when V has a large variance relative to the variance of $X'\beta + \varepsilon$.

This estimator may also be very sensitive to outliers, in particular, the density in the denominator defining T means that somewhat large values of ν_i will generate extremely large values of T_i. It may therefore be a good idea in practice to use robust moment estimators, for example, in the two stage least squares step, discarding all observations for which $f(\widehat{\nu}_i)$ is smaller than some tiny constant δ.

Based on Proposition 1, more efficient estimates with correct standard errors can be obtained by combining the above steps into a single GMM estimator as follows.

ESTIMATOR D

Make sure V_i takes on both positive and negative values, by demeaning it if necessary. Use GMM to estimate the parameters β, b, σ^2 based on the moment conditions

$$\begin{aligned} E[S(V - S'b)] &= 0 \\ E[\sigma^2 - (V - S'b)^2] &= 0 \end{aligned}$$

$$E\left[Z\left([D - I(V \geq 0)]\,(2\pi\sigma^2)^{1/2}\exp\left(\frac{(V - S'b)^2}{2\sigma^2}\right) - X'\beta\right)\right] = 0$$

The first two equations in Estimator D are equivalent to the score functions from maximum likelihood estimation of the V model, and the last equation is the moments corresponding to the two stage least squares regression of T on X with instruments Z.

The model for V here is parametric, and hence testable. Once could conduct a specification search from the particular model for V chosen here (linear with normal homoskedastic errors) to more general models such as those discussed below.

5.2.1 A More General V Model

The model of the previous section assumes the residuals $V - S'b$ of special regressor V model are homoskedastic, which may be unrealistic (the D model latent errors ε can be heteroskedastic and correlated with regressors). A more general model is

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &= S'b + \exp(S'c)\nu, \quad \nu \perp S, \varepsilon, \quad \nu \sim N(0,1) \end{aligned}$$

We may equivalently write the model for V as $V = S'b + \eta$, $\eta \sim N(0, \sigma_\eta^2)$, $\sigma_\eta = \exp(S'c)$, so V is linear in covariates plus a normal heteroskedastic error η, and the hetereoskedasticity has a simple multiplicative form. Again, by Theorem 1 other more general parametric models for V and continuous distributions for ν could be assumed, but this particular model is chosen for its combination of generality and simplicity.

Based on this model and Theorem 1, we now obtain the following multistep estimator.

ESTIMATOR C'

1. Make sure V_i takes on a range of both positive and negative values (V can be demeaned if not). Let $\widehat{b}$ be the estimated coefficients of an ordinary least squares regression of V on S. For each observation i, construct data $\widehat{\eta}_i = V_i - S_i'\widehat{b}$, which are the residuals of this regression.

2. Estimate $\widehat{c}$ as the coefficients of a nonlinear least squares regression of $\widehat{\eta}^2$ on $\exp(S'c)$ and for each observation i, and construct data $\widehat{\nu}_i = \widehat{\eta}_i \exp(-S_i'\widehat{c}/2)$

3. For each observation i, let $f(\widehat{\nu}_i)$ be the pdf of $\widehat{\nu}_i$, so

$$f(\widehat{\nu}_i) = \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\widehat{\nu}_i^2}{2}\right)$$

and construct data $\widehat{T}_i$ by

$$\widehat{T}_i = \frac{D_i - I(V_i \geq 0)}{f(\widehat{\nu}_i)} \frac{\widehat{\eta}_i}{\widehat{\nu}_i}$$

4. Let $\widehat{\beta}$ be the estimated coefficients of an ordinary linear two stage least squares regression of $\widehat{T}$ on X, using instruments Z.

The only nontrivial step now is step 2, and for that step good approximate starting values for $\widehat{c}$ may be obtained by linearly regressing $\ln(\widehat{\eta}^2)$ on S. We could have instead modeled σ_η^2 as $S'c$ instead of $\exp(S'c)$ to make step 2 a linear least squares regression, but that could lead to numerical problems because $S_i'\widehat{c}$ might be negative for some observations.

The corresponding efficient estimator is

ESTIMATOR D'

Make sure V_i takes on both positive and negative values, by demeaning it if necessary. Use GMM to estimate the parameters β, b, c based on the moment conditions

$$\begin{aligned} E[S(V - S'b)] &= 0 \\ E[S\left(\exp(2S'c) - \exp(S'c)(V - S'b)^2\right)] &= 0 \end{aligned}$$

$$E\left[Z\left([D - I(V \geq 0)]\,(2\pi\,\exp(S'c))^{1/2}\exp\left(\frac{(V - S'b)^2}{2\exp(S'c)}\right) - X'\beta\right)\right] = 0$$

Again the first two equations in Estimator D are equivalent to the score functions from maximum likelihood estimation of the V model, and the last equation is the moments corresponding to the two stage least squares regression of T on X with instruments Z.

For simplicity, most of the remaining choice estimators in this paper will assume homoskedastic V model errors, and so will be variants of estimators C and D instead of C' and D', but all may easily generalized as above to allow for conditionally hetereokedastic V as in this section's estimators C' and D'.

5.2.2 More Than One Very Exogenous Regressor

If we are lucky enough to have more than one very exogenous regressor in the model, then we can efficiently use the information in both by simply taking all the moments of estimator D based on each such regressor, and doing a large GMM on the entire set. Note that only one of very exogenous regressors can have a coefficient that is normalized to equal one.

To illustrate, assume we have two such regressors. The model is then

$$\begin{aligned} D &= I(X'\beta + V_1 + V_2\alpha + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V_1 &= S'b_1 + V_2c_1 + \nu_1, \quad \nu_1 \perp \nu_2, S, \varepsilon, \quad \nu_1 \sim N(0, \sigma^2_{\nu_1}) \\ V_2 &= S'b_2 + V_1c_2 + \nu_2, \quad \nu_2 \perp \nu_1, S, \varepsilon, \quad \nu_2 \sim N(0, \sigma^2_{\nu 2}) \end{aligned}$$

where some element of b_1, b_2, c_1, c_2 is set to zero for identification of the V equations. It follows from Theorem 1 that the model parameters satisfy the moments

$$\begin{aligned} E[S(V_1 - S'b_1 - V_2c_1)] &= 0 \\ E[\sigma^2_{\nu_1} - (V_1 - S'b_1 - V_2c_1)^2] &= 0 \end{aligned}$$

$$E\left[\begin{pmatrix} Z \\ V_2 \end{pmatrix}\left([D - I(V_1 \geq 0)]\exp\left(\frac{(V_1 - S'b_1 - V_2c_1)^2}{2\sigma^2_{\nu_1}}\right) - X'\beta - \alpha V_2\right)\right] = 0$$

$$\begin{aligned} E[S(V_2 - S'b_2 - V_1c_2)] &= 0 \\ E[\sigma^2_{\nu_2} - (V_2 - S'b_2 - V_1c_2)^2] &= 0 \end{aligned}$$

$$E\left[\begin{pmatrix} Z \\ V1 \end{pmatrix}\left([D - I(V_2 \geq 0)]\exp\left(\frac{(V_2 - S'b_2 - V_1c_2)^2}{2\sigma^2_{\nu_2}}\right) - \frac{X'\beta + V_1}{\alpha}\right)\right] = 0$$

The assumption that a given regressor has large support relative to the rest of latent variable cannot hold for more than one regressor, so root n convergence of GMM estimation using these moments will require tail assumptions as in Magnac and Maurin (2003).

Alternatively, we may estimate the model by defining a single special regressor to be a weighted average of the candidate very exogenous regressors, with weights that could either be selected arbitrarily for convenience, or be determined by minimum chi squared estimation, that is, choose weights to minimize the estimated variance of $\widehat{\beta}$.

5.2.3 Generalized Very Exogenous Regressor Estimation

The binomial response model with a very exogenous regressors generalizes to

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &= S'b + \nu, \quad \nu \perp S, \varepsilon, \quad supp(S'b + X'\beta + \varepsilon) \subseteq supp(-\nu) \end{aligned}$$

Here the distribution of ν is unknown, so the large support assumption must be made explicitly.

The estimator remains the same, except that a more general estimator of the density function of ν is required. This is just a one dimension density, and so is relatively simple to estimate. A kernel density estimator could be used, but an even simpler estimator that does not entail choosing a kernel or bandwidth is the following. Given n observations of ν_i, sort these observations from lowest to highest. For each observation ν_i, let ν_i^+ be the value of ν that, in the sorted data, comes immediately after ν_i and similarly let ν_i^- be the value that comes immediately before ν_i. Then

$$f(\nu_i) = \frac{\partial F(\nu_i)}{\partial \nu} \approx \frac{F(\nu_i^+) - F(\nu_i^-)}{\nu_i^+ - \nu_i^-} \approx \frac{2/n}{\nu_i^+ - \nu_i^-}$$

where the last step replaces the true distribution function F with the empirical distribution function. This suggests the estimator $\widehat{f}(v_i) = (v_i^+ - v_i^-)n/2$. Lewbel and Schennach (2003) show that, although this is not a consistent estimator of the density function $f(v_i)$, with sufficient regularity sample averages that divide by this estimator are root n consistent.

The result is the following estimator.

ESTIMATOR E

1. Make sure V_i takes on both positive and negative values (V can be demeaned if not). For each observation i, construct data $\widehat{v}_i = V_i - S_i'\widehat{b}$ as the residuals of an ordinary least squares regression of V on S.

2. For each observation i, define $\widehat{v}_i^-$ and $\widehat{v}_i^+$ as the values adjacent to $\widehat{v}_i$ when the $\widehat{v}$ estimates are sorted in increasing order, and construct data $\widehat{T}_i$ by

$$\widehat{T}_i = \frac{[D_i - I(V_i \geq 0)](\widehat{v}_i^+ - \widehat{v}_i^-)n}{2}$$

3. Let $\widehat{\beta}$ be the estimated coefficients of an ordinary linear two stage least squares regression of $\widehat{T}$ on X, using instruments Z.

Once again, no numerical searches are involved, and no calculations more difficult than sorting data or linear regression are required, so bootstrapping is numerically practical. Estimator E is the same estimator proposed in Lewbel and Schennach (2003), except that paper assumed independence for V itself rather than for $\widehat{v}$, so the first stage regression step was not needed. Based on Lewbel (2000) and Lewbel and Schennach (2003), root n convergence of this estimator requires strong assumptions regarding the support and tail thickness of the distribution of the very exogenous regressor.

As with control functions, the very exogenous regressor methodology can be further generalized, for example, we may let

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &= g(S) + v, \quad v \perp S, \varepsilon, \quad supp(g(S) + X'\beta + \varepsilon) \subseteq supp(-v) \end{aligned}$$

for unknown function g by using estimator E, except that in step 1 $\widehat{v}_i$ is now the residuals from a nonparametric regression of V on S.

A further generalization is the model

$$\begin{aligned} D &= I(X'\beta + V + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &\mid S, \varepsilon \sim V \mid S, \quad supp(X'\beta + \varepsilon) \subseteq supp(-V \mid S) \end{aligned}$$

now the estimator is to let $\widehat{f}_V(V \mid S)$ be a kernel or other nonparametric estimator of the conditional density of V given S, construct $\widehat{T}_i$ by

$$\widehat{T}_i = \frac{D_i - I(V_i \geq 0)}{\widehat{f}_V(V_i \mid S_i)}$$

and let $\widehat{\beta}$ be the estimated coefficients of an ordinary linear two stage least squares regression of $\widehat{T}$ on X, using instruments Z. This is a hard estimator in terms of requiring a high dimensional nonparametric component, but it is simple in that no numerical optimization or searches are required This last estimator is more general than Lewbel (2000), in that it permits V to correlate with the endogenous regressors.

5.2.4 A Discrete Very Exogenous Regressor

The very exogenous regressor estimators depend on V only through $I(V \geq 0)$ and the conditional density function of V. These estimators can therefore be used with a discrete V if that V is itself a function of an unobserved underlying variable $\widetilde{V}$ with a uniform (or uniform within cells) distribution. For example, consider the model

$$\begin{aligned} D &= I(X'\beta + \widetilde{V} + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V &= I(\widetilde{V} \geq 0), \quad \widetilde{V} \perp S, \varepsilon, \quad \widetilde{V} \sim U[-K, 1-K] \end{aligned}$$

where $\widetilde{V}$ is unobserved and $U[-K, 1-K]$ denotes a uniform distribution on the interval $[-K, 1-K]$ for some postive constant K. The probability density function of $\widetilde{V}$ is $f_{\widetilde{V}}(\widetilde{V}) = 1$ so by Theorem 1, $E(Ze) = 0$ where $e = T - X'\beta$ and

$$T = \frac{D - I(\widetilde{V} \geq 0)}{f_{\widetilde{V}}(\widetilde{V})} = (D - V)$$

Here D depends on an unobserved continuous very exogenous regressor $\widetilde{V}$, while the observed very exogenous regressor V is discrete and only takes the values zero and one. The corresponding consistent estimator of β is then just a linear two stage least squares regression of $D_i - V_i$ on X_i using instruments Z_i.

To illustrate, consider the following example, which is similar to a very exogenous regressor model in Cogneau and Maurin (2003). Suppose we have a data set of students in the same grade cohort at school. Students must be five years old on

September 1 to start school. We observe the calendar year in which each student in the class is born, but not their exact birthdate. Let $V_i = 1$ for the older students in the class and zero for the younger students, based on their observed birthyear. Define $\widetilde{V}_i$ to be exact age minus six of student i on December 31 of the year they start school. Then $\widetilde{V}_i$ is a value from $-2/3$ to $1/3$, and $V_i = I(\widetilde{V}_i \geq 0)$.

Let D be a schooling outcome measure such as an indicator for not repeating a grade or for graduating, and let X be a vector of determinants of schooling outcome, other than age, such as parent's income, data on siblings etc.,.We then obtain the above model, assuming that births are uniformly distributed throughout the year, that birthdays are independent of other observable and unobservable determinents of schooling outcome, and that the latent variable determining the outcome is linear in the age of the student. In this example β is consistently estimated just by linearly regressing $D_i - V_i$ on X_i. If some elements of X_i are endogeneous, then this linear regression would be done by two stage least squares with ordinary appropriate instruments, e.g., Cogneau and Maurin (2003) use grandparents socioeconomic status as an instrument for the endogenous parents income regressor.

The uniform distribution assumption can be relaxed given more information regarding $\widetilde{V}_i$. For example, if we know the month of birth of each student i, then we need only assume that births are uniformly distributed within each month. Letting F_i be the fraction of students that have birthdays in the same month as student i, the density of $\widetilde{V}_i$ is then $f_{\widetilde{V}}(\widetilde{V}_i) = 12F_i$ and so we would then estimate β by regressing $(D_i - V_i)/(12F_i)$ on X_i Similar estimators could be constructed using other discretized data, for example, $\widetilde{V}_i$ could be the true log distance to school when we only observe a few different distance intervals, or log income when we only observe income brackets.

Other results regarding a discrete very exogenous regressor are provided by Magnac and Maurin (2004), including set identification results for β when the uniform distribution assumption within cells is dropped.

6 Ordered Choice

Both the control function and very exogenous regressor estimators can be applied to other limited dependent variable models with endogenous regressors. For ex-

ample, the ordered choice model with possible choices $k^* = 0, ..., K$ is

$$k^* = \sum_{k=1}^{K} kI\left[\alpha_{k-1} \leq -\left(X'\beta + V + \varepsilon\right) \leq \alpha_k\right]$$

where $\alpha_0 = -\infty$ and α_1,...,α_K are threshold constants. Here again the free normalization is used in which a regressor V has a coefficient of one instead of normalizing the variance of the error ε to be one. This ordered choice model can be rewritten as a collection of binary choices as

$$D_k = I(\alpha_k + X'\beta + V + \varepsilon \geq 0)$$

where D_k for $k = 1, ..., K$ are dummy variables such that $D_k = 1$ if the individual chooses $k^* \leq k$ and zero otherwise, so the individual's choice is $k^* = \sum_{k=1}^{K} D_k$. Without loss of generality, the element of β corresponding to the constant term in X is set to zero, so the constant term in each D_k equation then equals the choice k threshold α_k.

Estimator B or D provides a collection of moment conditions for estimating any one of these D_k models with endogenous regressors, based either on control functions or on a very exogenous V. Using either estimator, we may apply GMM to the collection of all the moments corresponding to every D_k model to estimate the parameters β and $\alpha_1, ..., \alpha_K$. Lewbel (2000) describes an example of this estimator for the very exogenous regressor model with a general estimator for the conditional density of V.

To illustrate, suppose $K = 2$, that is, ordered choice with three possible choices $k = 0, 1, 2$. With a very exogenous V, based on estimator B the moments for GMM estimation are

$$\begin{aligned} E[S(V - S'b)] &= 0 \\ E[\sigma^2 - (V - S'b)^2] &= 0 \end{aligned}$$

$$\begin{aligned} E\left[Z\left([D_1 - I(V \geq 0)]\,(2\pi\sigma^2)^{1/2}\exp\left(\frac{\left(V - S'b\right)^2}{2\sigma^2}\right) - \alpha_1 - X'\beta\right)\right] &= 0 \\ E\left[Z\left([D_2 - I(V \geq 0)]\,(2\pi\sigma^2)^{1/2}\exp\left(\frac{\left(V - S'b\right)^2}{2\sigma^2}\right) - \alpha_2 - X'\beta\right)\right] &= 0 \end{aligned}$$

If instead we use control functions, based on estimator B the moments for GMM estimation are

$$\begin{aligned}
E\left[W(Y - W'b)\right] &= 0 \\
E\left[S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)\right] &= 0 \\
E\left[S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)X\right] &= 0 \\
E\left[S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)\left(Y - W'b\right)\right] &= 0 \\
E\left[S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)V\right] &= 0
\end{aligned}$$

$$\begin{aligned}
E\left[S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)\right] &= 0 \\
E\left[S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)X\right] &= 0 \\
E\left[S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)\left(Y - W'b\right)\right] &= 0 \\
E\left[S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)V\right] &= 0
\end{aligned}$$

where the function S is defined as

$$S(D, X, V, U, \alpha, \beta, \gamma, \sigma_\eta) = D\frac{\phi\left[\left(\alpha + X'\beta + V + U'\gamma\right)/\sigma_\eta\right]}{\Phi\left[\left(\alpha + X'\beta + V + U'\gamma\right)/\sigma_\eta\right]} + (1-D)\frac{-\phi\left[\left(\alpha + X'\beta + V + U'\gamma\right)/\right.}{1 - \Phi\left[\left(\alpha + X'\beta + V + U'\gamma\right)\right.}$$

7 Selection Models With Endogenous Regressors

The Heckman sample selection model estimator described earlier provides a simple estimator for selection models with exogenous regressors, and Proposition 1 showed how that estimator could be rewritten in a GMM form to increase efficiency and provide correct standard errors. Here a very exogenous regressor estimator is provided for linear two stage least squares estimation of sample selection models with endogenous regressors.

Continue to let $Z = (Z_1', X_2')'$ and $S = (Y', Z')'$, so Z is the vector of all the available exogenous covariates except for a very exogenous V, and S is all the available covariates except for V. Now also define P to be an outcome, that is, an endogenous variable, that is only observed, or selected, when $D = 1$. The vector $X = (Y', X_2')'$ will now be the set of all the regressors in the P equation (instead of the D equation) except for V. Consider the selection model

$$\begin{aligned}
P &= (X'\beta + V\gamma + \varepsilon)D, \quad E(Z\varepsilon) = 0 \\
D &= I(a_0 \leq M(S, e) \pm V \leq a_1) \\
V &= S'b + \nu, \quad \nu \perp S, \varepsilon, e, \quad \nu \sim N(0, \sigma_\nu^2)
\end{aligned}$$

where M is an unknown function, a_0 and a_1 are unknown, possibly infinite constants, and e and ε are errors with a joint unknown distribution. The conditional distribution of ε, e conditional on S is unknown. This model and associated simple estimator are a special case of Lewbel (2003), which uses a nonparametric specification of the conditional distribution of V instead of a linear model with a normal error ν. Details regarding limiting distribution theory and the relationship of this estimator to others in the literature may also be found there.

An example is a standard Heckman (1976) wage model with P equalling observed wage and D the indicator of employment. In this case $a_0 = 0$, $a_1 = \infty$, M is linear (typically), the covariates would include variables like training, education, demographics, and nonwage income, and the errors ε and e are correlated with each other because they depend on common components such as unobservable abilities and motivation. An endogenous covariate Y might be a spouse's or parent's income, and a very exogenous regressor V could be age, or a Card (1995) type access (cost, distance) to schooling measure, though note that the model allows the outcome P (but not ε) to depend on V.

Lewbel (2003) provides another empirical application in which P is factory investment rate, D indicates nonzero investment which occurs when returns to investment are positive, V is plant size, and the endogenous regressor Y is the factory profit rate, which proxies for Tobin's Q.

Ordered selection or ordered treatment models have both a_0 and a_1 finite. These are models where selection or treatment is determined by an ordered choice or response. For example, $M(S, e) \pm V$ could be a latent variable representing desired schooling, D could indicate graduating high shool but not college (a latent variable value less than a_0 would index not graduating high school, and greater than a_1 would correspond to attaining a college degree). P could then be an earnings equation for high school graduates without college degrees.

THEOREM 2: Assume $P = (X'\beta + V\gamma + \varepsilon)D$, $E(Z\varepsilon) = 0$, $I(a_0 \leq M(S, e) \pm V \leq a_1)$, $V = S'b + \nu$, $\nu \perp S, \varepsilon, e$, and ν is continuously distributed with support equal to the real line. Assume a_0 and a_1 are finite. Let $f(\nu)$ be the probability density function of ν. Then

$$E\left[Z\frac{D}{f(v)}(P - X'\beta + V\gamma)\right] = 0$$

Theorem 2 as stated applies only when a_0 and a_1 are finite. When one of them is infinite, then the result can be preserved by adding an asymptotic trim-

ming parameter, or one may assume a large but finite support distribution for ν (e.g., a trimmed normal), and in the latter case the resulting estimator has a small bias term that is proportional to the inverse of the largest value $|V|$ can take on. Since this largest value can be arbitrarily large, the resulting bias can be arbitrarily small. Lewbel (2003) provided details, and finds that ignoring this boundedness or asymptotic trimming technicality makes very little difference in practice.

The following simple estimator E, and associated efficient estimator F, are based directly on Theorem 2.

ESTIMATOR E

1. Let $\widehat{b}$ be the estimated coefficients of an ordinary least squares regression of V on S. For each observation i, construct data $\widehat{\upsilon}_i = V_i - S_i'\widehat{b}$, which are the residuals of this regression.

2. Let $\widehat{\sigma}^2$ be the sample mean of $\widehat{\upsilon}_i^2$, and for each observation i, let $f(\widehat{\upsilon}_i, \widehat{\sigma}^2)$ be the pdf of $\widehat{\upsilon}_i$, so

$$f(\widehat{\upsilon}_i) = \frac{1}{\sqrt{2\pi\,\widehat{\sigma}^2}} \exp\left(\frac{-\widehat{\upsilon}_i^2}{2\widehat{\sigma}^2}\right).$$

3. For each observation i, construct instruments $\widehat{Z}_i = Z_i D_i / f(\widehat{\upsilon}_i, \widehat{\sigma}_\upsilon^2)$.

4. Let $\widehat{\beta}$ and $\widehat{\gamma}$ the estimated coefficients of an ordinary linear two stage least squares regression of P on X and V, using instruments $\widehat{Z}$

ESTIMATOR F

Use GMM to estimate the parameters β, γ, b, σ_υ^2 based on the moment conditions

$$\begin{aligned} E[S(V - S'b)] &= 0 \\ E[\sigma_\upsilon^2 - (V - S'b)^2] &= 0 \end{aligned}$$

$$E\left[ZD\left(P - X'\beta + V\gamma\right)\exp\left(\frac{(V - S'b)^2}{2\sigma_\upsilon^2}\right)\right] = 0$$

In addition to simplicity, these estimators are also convenient in that they do not require specifying or estimating the selection equation, in addition to not specifying or estimating models of the endogenous regressors Y. Similarly, the joint distribution of errors in the models of outcome, selection, and endogenous regressors is not specified or estimated. Only the distribution of the single very exogenous regressor needs to be modeled.

The economic interpretation of the estimated coefficients β and γ is straightforward. We may define $P^* = X'\beta + V\gamma + \varepsilon$ to be a latent outcome, which is only observed for individuals having $D = 1$. If P^* were observable for all individuals then β and γ could be estimated by an ordinary linear two stage least squares regression of P^* on X and V using instruments Z. The limiting values of $\widehat{\beta}$ and $\widehat{\gamma}$ in Estimators E and F are the same as the limiting values from this hypothetical regression. In short, the weighting of instruments by $D/f(\nu)$ corrects for selection, and the two stage least squares corrects in the usual way for regressor endogeneity.

8 Dynamic Binary Choice Panel Models With Fixed Effects and Endogenous Regressors

The estimator described in this section is the present paper's parameterized very exogenous regressor and GMM framework applied to the panel binary choice estimator in Honore and Lewbel (2002).

To handle panel data, individual i and time t subscripts are now used, where $i = 1, ..., n$ and $t = 1, ..., T$. The asymptotics assume T fixed and $n \to \infty$. The model is

$$\begin{aligned} D_{it} &= I(X_{it}'\beta + V_{it} + \alpha_i + \varepsilon_{it} \geq 0), \quad E[Z_{it}(\varepsilon_{it} - \varepsilon_{it-1})] = 0 \\ V_{it} &= S_{it}'b_t + \nu_{it}, \quad \nu_{it} \perp S_{it}, \alpha_i + \varepsilon_{it}, \quad \nu_{it} \sim N(0, \sigma_{tv}^2) \end{aligned}$$

Here D_{it} is the binary variable being modeled and the vector of regressors is $X_{it} = (Y_{it}', X_{2it}')'$, where Y_{it} is a vector of endogenous or mismeasured regressors and X_{2it} is a vector of exogenous regressors. The vector Y_{it} can include lags of the dependent variable such as D_{it-1}, so the panel model can be dynamic. We also have a very exogenous regressor V_{it}. The vector of instruments Z_{it} are exactly the same variables that would be used as instruments in a linear panel model after differencing out fixed effects, for example, Z_{it} could include lagged values of X_{2it}.

If the panel is dynamic, so Y_{it} includes P_{it-1}, then Z_{it} could consist of X_{2it-k} for $k > 1$. The panel can also be dynamic in that the errors ε_{it} can be autocorrelated. The dependence of ε_{it} on lagged values of ε_{it} is arbitrary; it does not need to be specified or estimated.

The vector of variables S_{it} includes X_{it}, Z_{it}, Z_{it-1}, and possibly additional lagged values of these variables (note that many or all elements of b_t could be

zero). The unobserved parameters α_i are treated as fixed effects, in that they will be differenced out upon estimation, and their distribution is not modeled. However, it is assumed that the data generating process makes ν_{it} conditionally independent of $\alpha_i + \varepsilon_{it}$, conditioning on S_{it}.

Honore and Lewbel (2002) make a similar assumption; see that paper for further discussion and economic examples. Here α_i represent permanent effects for each individual i, so possible V_{it} variables could be transitory effects, e.g., if D_{it} are purchase decisions, V_{it} might be transitory income. If D_{it} are production or investment decisions, V_{it} could be a temporary cost shock.

COROLLARY 1: Assume $D_{it} = I(X_{it}'\beta + V_{it} + \alpha_i + \varepsilon_{it} \geq 0)$, $E[Z_{it}(\varepsilon_{it} - \varepsilon_{it-1}) = 0$, $V_{it} = S_{it}'b_t + \nu_{it}$, $\nu_{it} \perp S_{it}, \alpha_i + \varepsilon_{it}$, and for each t, ν_{it} is continuously distributed with support equal to the real line. Let $f_t(\nu_{it})$ be the probability density function of ν_{it}. Then

$$E\left[Z_{it}\left(\frac{D_{it} - I(V_{it} \geq 0)}{f_t(\nu_{it})} - \frac{D_{it-1} - I(V_{it-1} \geq 0)}{f_{t-1}(\nu_{it-1})} - (X_{it} - X_{it-1})'\beta\right)\right] = 0$$

Corollary 1 is a direct extension of Theorem 1 with a linear g model (the extension to nonlinear g is straightforward and is omitted only for simplicity). This result does not require V_{it} to vary over time. If V_{it}, and therefore ν_{it}, is fixed over time then we get the simplification

$$\frac{D_{it} - I(V_{it} \geq 0)}{f_t(\nu_{it})} - \frac{D_{it-1} - I(V_{it-1} \geq 0)}{f_{t-1}(\nu_{it-1})} = \frac{D_{it} - D_{it-1}}{f(\nu_i)}$$

Estimators corresponding to Corollary 1 are

ESTIMATOR G

1. In each time period t, let $\widehat{b_t}$ be the estimated coefficients of an ordinary least squares regression of V_{it} on S_{it}. For each observation it, construct data $\widehat{\nu}_{it} = V_{it} - S_{it}'\widehat{b}_t$, which are the residuals of this regression.

2. In each time period t, let $\widehat{\sigma}_t^2$ be the sample mean of $\widehat{\nu}_{it}^2$, and for each observation it, let $f(\widehat{\nu}_{it}, \widehat{\sigma}_t^2)$ be the pdf of $\widehat{\nu}_{it}$, so

$$f(\widehat{\nu}_{it}) = \frac{1}{\sqrt{2\pi\widehat{\sigma}_t^2}} \exp\left(\frac{-\widehat{\nu}_{it}^2}{2\widehat{\sigma}_t^2}\right)$$

3. For each observation it construct data $\widehat{T}_{it}$ by

$$\widehat{T}_{it} = \frac{D_{it} - I(V_{it} \geq 0)}{f(\widehat{v}_{it}, \widehat{\sigma}_t^2)}$$

4. For each time period t, let $\widehat{\beta}_t$ be the estimated coefficients of an ordinary linear two stage least squares regression of $\widehat{T}_{it} - \widehat{T}_{it-1}$ on $X_{it} - X_{it-1}$, using instruments Z_{it}. Let $\widehat{\beta}$ be a weighted average of the estimates $\widehat{\beta}_t$.

Each $\widehat{\beta}_t$ in step 4 is a consistent estimator of β, so these estimates can just be averaged together. Weights can be chosen to minimize standard errors (that is, minimum chi squared estimation) if desired, or efficiency can be obtained using the following GMM estimator.

ESTIMATOR H

Use GMM to estimate the parameters $\beta, b_1, ..., b_T, \sigma_1^2, ..., \sigma_T^2$ based on the moment conditions

$$\begin{aligned} E[S_t(V_t - S_t' b_t)] &= 0 \\ E[\sigma_t^2 - (V_t - S_t' b_t)^2] &= 0 \end{aligned}$$

$$E\left[Z_t \left(\begin{array}{l} [D_t - I(V_t \geq 0)] (2\pi\sigma_t^2)^{1/2} \exp\left(\frac{(V_t - S_t' b_t)^2}{2\sigma_t^2} \right) \\ - \left[D_{t-1} - I(V_{t-1} \geq 0) \right] (2\pi\sigma_{t-1}^2)^{1/2} \exp\left(\frac{(V_{t-1} - S_{t-1}' b_{t-1})^2}{2\sigma_{t-1}^2} \right) - (X_t - X_{t-1})\beta \end{array} \right) \right] = 0$$

For $t = 1, ..., T$, where V_t denotes a draw of the random variable V in time period t, and similarly for S_t, X_t, D_t, and Z_t.

Note that Z_t can have different numbers of elements in different time periods. We can similarly construct estimators of panels of sample selection models with endogenous regressors and fixed effects, by combining the results of this and the previous section.

9 Large T Dynamic Binary Choice Models With Endogenous Regressors

Now consider dynamic binomial response models. The data have $t = 1, ..., T$, and the asymptotics assume stationarity and $T \to \infty$. The model is

$$\begin{aligned} D_t &= I(X_t'\beta + V_t + \varepsilon_t \geq 0), \quad E(Z_t\varepsilon_t) = 0 \\ V_t &= S_t'b + \exp(S_t'c)\nu_t, \quad \nu_t \perp S_t, \varepsilon_t, \ \nu_t \sim N(0,1) \end{aligned}$$

Here D_t is the binary variable being modeled and the vector of regressors is $X_t = (Y_t', X_{2t}')'$ where Y_t is a vector of endogenous or mismeasured regressors and X_{2t} is a vector of exogenous regressors. The model is dynamic, in that the vector Y_t can include lags of the dependent variable such as D_{t-1}, and interactions of lags with other regressors such as $D_{t-1}X_{2t}$. The vector of instruments Z_t are exactly the same variables that would be used as instruments in a linear dynamic model, for example, Z_{it} could include lagged values of X_{2t}. The vector of variables S_t includes X_t, Z_t (and hence lags of D_t) and possibly additional lagged values of these variables. S_t could also include lagged values of V_t.

We essentially have an ARCH model for the regressor V_t. This regressor is very exogenous in that the errors ν_t in the model for V_t (after removing multiplicative heteroskedasticity) are independent of regressors and of the D_t model error. This imposes strong restrictions on any possible dependence of ν_t over time, since ν_t is conditionally independent of S_t, which can include D_{t-1}, which itself is a function of V_{t-1} and hence of ν_{t-1}. To satisfy this restriction it may be assumed that ν_t are independently distributed over time, noting that the model for V_t can include lags such as V_{t-1}.

Theorem 1 now directly applies to this model, and so Estimators C' and D' can be directly applied. The errors in the moments of Estimator D' will no longer be iid, so for efficiency and correct standard errors time series versions of GMM will need to be used.

10 Appendix: Proofs

PROOF OF PROPOSITION 1: Proposition 1 can be stated as: Assume the conditions of Newey and McFadden 1994, Theorem 6.1 hold. Then the conditions of their Theorem 3.4 also hold, and Proposition 1 follows by applying their Theorem 3.4.

PROOF OF THEOREM 1: Define $D^* = X'\beta + \varepsilon$ so $D = I(D^* + V \geq 0)$. Then, by the definition of conditional expectation

$$\begin{aligned} E(T \mid S, \varepsilon) &= \int_{supp(\nu)} \frac{I(D^* + g(\nu, S) \geq 0) - I(g(\nu, S) \geq 0)}{f(\nu)} \frac{\partial g(\nu, S)}{\partial \nu} f(\nu \mid S, \varepsilon) d\nu \\ &= \int_{supp(\nu)} \left[I(D^* + g(\nu, S) \geq 0) - I(g(\nu, S) \geq 0) \right] \frac{\partial g(\nu, S)}{\partial \nu} d\nu \\ &= \int_{-\infty}^{\infty} \left[I(D^* + V \geq 0) - I(V \geq 0) \right] dV \end{aligned}$$

where the second equality follows from $\nu \perp S, \varepsilon$, and the third from the change of variables from ν to V. If $D^* \geq 0$ then

$$E(T \mid S, \varepsilon) = \int_{-\infty}^{\infty} I(-D^* \leq V \leq 0) dV = \int_{-D^*}^{0} 1 dV = D^*$$

and if $D^* \leq 0$ then

$$E(T \mid S, \varepsilon) = \int_{-\infty}^{\infty} -I(0 \leq V \leq -D^*) dV = -\int_{0}^{-D^*} 1 dV = D^*$$

This proves that $E(T \mid S, \varepsilon) = X'\beta + \varepsilon$. Now $e = T - X'\beta$ so

$$\begin{aligned} E(Ze) &= E[Z(T - X'\beta)] \\ &= E[E(Z(T - X'\beta) \mid S, \varepsilon)] \\ &= E[Z(E(T \mid S, \varepsilon) - X'\beta)] \\ &= E(Z\varepsilon) = 0. \end{aligned}$$

PROOF OF THEOREM 2: By the definition of conditional expectation,

$$\begin{aligned} E\left[Z(P - X'\beta + V\gamma) \frac{D}{f(v)} \mid S, \varepsilon, e \right] &= \int_{-\infty}^{\infty} \left(\frac{Z\varepsilon I(a_0 \leq M(S, e) \pm V \leq a_1)}{f(V - S'b)} \right) f_V(V \mid S, \varepsilon, e) dV \\ &= \int_{-\infty}^{\infty} Z\varepsilon I(a_0 \leq M(S, e) \pm V \leq a_1) dV \end{aligned}$$

where the second equality follows from the same logic as in Theorem 1. If the indicator function is increasing in V then

$$\begin{aligned} E\left[Z(P - X'\beta + V\gamma) \frac{D}{f(v)} \mid S, \varepsilon, e \right] &= \int_{a_0 - M(S,e)}^{a_1 - M(S,e)} Z\varepsilon dV \\ &= (a_1 - a_0) Z\varepsilon \end{aligned}$$

and if the indicator function is decreasing in V we obtain $(a_0 - a_1)Z\varepsilon$. Either way, the Theorem then follows from the law of iterated expectations and $E(Z\varepsilon) = 0$.

PROOF OF COROLLARY 1: Define

$$T_{it} = \frac{D_{it} - (V_{it} \geq 0)}{f_t(v_{it})}$$

by Theorem 1, $E(T_{it} \mid S_{it}, \alpha_i + \varepsilon_{it}) = X'_{it}\beta + \alpha_i + \varepsilon_{it}$, so by the law of iterated expectations $E(Z_{it}T_{it-k}) = E[Z_{it}\left(X'_{it-k}\beta + \alpha_i + \varepsilon_{it-k}\right)]$ for $k = 0, 1$, and therefore $E[Z_{it}(T_{it} - T_{it-1})] = E[Z_{it}\left(\left(X'_{it}\beta + \alpha_i + \varepsilon_{it}\right) - \left(X'_{it-1}\beta + \alpha_i + \varepsilon_{it-1}\right)\right)]$.

ENDOGENOUS SELECTION OR TREATMENT MODEL ESTIMATION

Arthur Lewbel
Boston College

Revised April 2005

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, conditional on covariates, can be identified given an observed variable V that affects the treatment or selection probability in certain ways and is conditionally independent of the error terms in a model of potential outcomes. Selection model estimators based on this identification are provided, which take the form of either simple weighted averages or GMM or two stage least squares. These estimators permit endogenous and mismeasured regressors. Empirical applications are provided to estimation of a firm investment model and a returns to schooling wage model.

Portions of this paper were previously circulated under other titles including, "Two Stage Least Squares Estimation of Endogenous Sample Selection Models."
JEL Codes: C14, C25, C13. Keywords: Sample Selection, Treatment Effects, Censoring, Semiparametric, Endogeneity, Instrumental Variables, Switching Regressions, Heteroscedasticity, Latent Variable Models, Ordered Choice, Investment, Returns to Schooling.
* This research was supported in part by the NSF, grant SES-9905010. I'd like to thank Yuriy Tchamourliyski for research assistance, and Hidehiko Ichimura, Edward Vytlacil, Jim Powell, Jim Heckman, Fabio Schiantarelli, Stacey Chen, Jinyong Hahn, Alberto Abadie, Shakeeb Khan, and anonymous referees for data and helpful comments. Arthur Lewbel, Department of Economics, Boston College, 140 Commonwealth Ave., Chestnut Hill, MA, 02467, USA. (617)-552-3678, lewbel@bc.edu, , http://www2.bc.edu/~lewbel/

1 Introduction

Assume that for a sample of individuals $i = 1, ..., n$ we observe an indicator D that equals one if an individual is treated, selected, or completely observed, and zero otherwise. If $D = 1$ we observe some scalar or vector P, otherwise let $P = 0$. Define P^* to equal the observed P when $D = 1$, otherwise P^* equals the value of P that would have been observed if D had equaled one, that is, either a counterfactual or an unobserved response. Then $P = P^*D$. We also observe covariates, though selection on observables is not assumed. Treatment or selection D may be unconditionally or conditionally correlated with P^*, so P^* and D may depend on common unobservables. Rubin (1974) type restrictions like unconfoundedness or ignorability of selection are not assumed.

To illustrate, in a classic wage model (Gronau 1974, Heckman 1974, 1976), $D = 1$ if the individual is employed, P^* is the wage an individual would get if employed, and P is the observed wage, which is zero for the unemployed. Both P^* and D depend on common unobservables such as ability, as well as on observable covariates such as measures of schooling or training.

Another example is models based on data sets where some regressors are missing, not at random. For example, models of individual's consumption or purchasing decisions depend on income P^*, and in surveys many individuals do not report their income. Failure to report income ($D = 0$) is likely to be correlated with income, even after conditioning on other observed covariates.

For simplicity, refer to D as selection, though more generally it is just an indicator of not observing P^* for whatever reason. Assume that selection D is given by

$$D = I(0 \leq M^* + V \leq A^*) \tag{1}$$

where the unobserved A^* can be a constant, a random variable, or infinity, V is an observed, continuously distributed covariate (or known function of covariates) with large support, M^* is an unobserved latent variable, and I is the indicator function that equals one if its argument is true and zero otherwise. Typical parametric or semiparametric models of selection are special cases of equation (1) where M^* is assumed to be linear in other covariates X and a well behaved error term e, but that structure is not imposed here.

Setting the lower bound to zero in equation (1) is a free normalization, since no location assumptions are imposed on M^* and A^*. Similarly, setting the coefficient of V to one is (apart from sign) a free scale normalization that is imposed without loss of generality.

In the wage model example, the typical assumption is that one chooses to work if the gains in utility from working, indexed by the latent $M^* + V$, are sufficiently large, so in that case A^* is infinite. Examples in which A^* is finite arise in ordered treatment or ordered selection models. For example, if an ordered choice model with latent variable $M^* + V$ determines an individual's years of schooling and D indexes having exactly 12 years of schooling then individuals with $M^* + V < 0$ choose 11 or fewer years while those with $M^* + V > A^*$ choose 13 or more years. We might then be interested in modelling the returns P from having just 12 years of schooling. Examples of models like this with A^* random include Cameron and Heckman (1998) and Carneiro, Hansen, and Heckman (2003).

A convenient feature of the proposed estimators is that they will not require specifying, modeling or estimating the D (propensity score) model, apart from assuming equation (1) holds. For example, any dependence of M^* on X can be unknown, and the estimator is the same regardless of whether A^* is constant, random, or infinite. Empirical applications with both finite A^* and infinite A^* are provided.

Regarding outcomes, define U^* and U by

$$\psi(P^*, X, V, \theta_0) = U^* \tag{2}$$

$$\psi(P, X, V, \theta_0)D = U$$

for some known vector valued function ψ. The initial goal will be estimation of $E(U^*)$ which in turn is used to estimate the parameter vector θ_0. For example, if ψ is defined by $P^* = U^*$, then an estimate of $E(U^*) = E(P^*)$ is an estimate of what the mean outcome in the population would be if everyone in the population were treated, selected, or observed. More generally, suppose that θ_0 uniquely satisfies $E[\psi(P^*, X, V, \theta_0)] = 0$. If there were no selection problem, so if P^* instead of $P = P^*D$ was observable, then the generalized method of moments (GMM) could be applied, minimizing a quadratic form in the sample mean of $\psi(P^*, X, V, \theta_0)$, to consistently estimate θ_0. In the presence of selection this GMM is infeasible, but it becomes feasible given an estimator of $E(U^*)$.

For example, let X be the union of elements in vectors Y and Z, where Y is a vector of regressors and Z is a vector of instruments. Suppose potential wages are determined by the model $P^* = Y'\beta_Y + V\beta_V + \varepsilon$, where some of the regressors in Y may be mismeasured, endogenous, or otherwise correlated with the error ε, and the error ε could also have unspecified heteroskedasticity. Given ordinary instruments Z that are uncorrelated with ε and correlated with V, Y (Z may include exogenous elements of Y), define ψ by $Z(P^* - Y'\beta_Y - V\beta_V) = U^*$. If

P^* were observable for everyone then $\beta = \beta_V, \beta_Y$ could be estimated by GMM, an example of which is an ordinary linear two stage least squares regression of P^* on V, Y using instruments Z. The difficulty is that this estimator is infeasible because of the selection problem, that is, we only observe P instead of P^*, and unobservables that determine the selection such as M^* are correlated with P^* and U^*. Feasible estimation of β requires an estimator for $E(U^*)$.

Define the weighting scalar W by

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

where f is the conditional probability density function of V given X. This paper shows that under general conditions

$$E(U^*) = plim_{n\to\infty} \frac{\sum_{i=1}^{n} U_i W_i}{\sum_{i=1}^{n} W_i} \tag{3}$$

so $E(U^*)$ is consistently estimated as the weighted average of the observed U_i (including $U_i = 0$ for all unobserved outcomes) using weights W. Based on equation (3), the infeasible GMM for estimating θ_0 is converted into a feasible GMM using the observable $W\psi(P, X, V, \theta_0)$ in place of the unobservable function $\psi(P^*, X, V, \theta_0)$. In the above two stage least squares example, this means β would be consistently estimated by an ordinary linear two stage least squares regression of WP on WV, WY using instruments Z.

The main assumptions required for equation (3) to hold are that the support of $V|X$ contains the supports of both $-M^*|X$ and $A^* - M^*|X$ (these could all equal the real line, for example), and that

$$V \mid X, U^*, M^*, A^* \sim V \mid X \tag{4}$$

that is, V is conditionally independent of the unobserved latent variables of the model, conditioning on the set of covariates X.

To give some intuition for equation (3), and why the above restrictions on V are required, suppose for the moment that $A^* = a$ is constant and that V has a uniform distribution with constant density f, independent of M^*, U^*. In that case we would have $E(U) = E[E(I(0 \le M^* + V \le a)U^* \mid M^*, U^*)] = E\left(\int_{-M^*}^{a-M^*} U^* f dv\right) = aE(U^*)$ and similarly $E(D) = a$, so in that case there would be no selection problem (or more precisely, unconditional propensity score weighting would then fix the selection problem), because we would then have

$E(U^*) = E(DU)/E(D)$. The key is that M^*, the source of unobserved correlation between D and U^*, drops out when V is independently, uniformly distributed. In the general problem, V is not an independent uniform, but weighting by W, and hence scaling by the conditional density of f, is algebraically equivalent to converting V to a uniform. Given equation (4), it suffices to condition the density of V on the observable X, and the support assumption on V ensures the bounds of the integral are not cut off.

Equation (3) resembles propensity score weighting estimators (see, e.g., Horvitz and Thompson (1952), Koul, Susarla, and van Ryzin (1981), Hahn (1998), and Hirano, Imbens and Ridder (2003)), but equation (3) holds even though the independence and conditioning assumptions required for consistency of propensity score weighted estimators are not imposed here. Specifically, we cannot use $E(U)/E(D)$ or $E[E(U \mid V, X)/E(D \mid V, X)]$ to estimate $E(U^*)$, because D and U^* (or equivalently, M^* and U^*) can covary, even after conditioning on observables like X, V. However, averaging after V density weighting is equivalent to integrating over V, that is equation (3) implies $E(U^*) = E[\int_{-\infty}^{\infty} E(U \mid V, X)dV]/E[\int_{-\infty}^{\infty} E(D \mid V, X)dV]$ when these expectations exist, so the proposed estimator is equivalent to weighting an integral of the conditional expectation of U by an integral of the conditional propensity score.

One interpretation of equation (4) is simply that V is an exogenous covariate, in the strong sense of being conditionally (conditioning on other covariates X) independent of the unobservables in the model, and hence conditionally independent of the errors if the model were parameterized. More generally, equation (4) is an example of an exclusion restriction, of the sort that is commonly used to identify models in simultaneous systems. Section 2.5 of Powell's (1994) survey discusses the use of exclusion assumptions in semiparametric estimators. Magnac and Maurin (2003) call this a partial independence assumption. In models where the errors are independent of regressors, every regressor is exogenous and so satisfies equation (4). Blundell and Powell (2004) and Heckman and Vytlacil (2004) use exclusion assumptions for identification in binary choice and treatment models.

Requiring a regressor to have support containing a large or infinite interval, such as containing the supports of other variables, is also common in the semiparametric limited dependent variable model literature. Examples include Manski (1975,1985), Han (1987), Horowitz (1992), and Cavanagh and Sherman (1998).

The estimator here weights observations by the density of a regressor V that satisfies exclusion and large support assumptions. Lewbel (1998, 2000) and Khan

and Lewbel (2005) use a similar idea to estimate linear index, binary choice and truncated regression models, respectively. Magnac and Maurin (2003) prove semiparametric efficiency of Lewbel (2000), and Jacho-Chavez (2005) shows semiparametric efficiency of a general class of density weighted estimators. Magnac and Maurin (2003) also show in the binary choice context that large support can be replaced by a tail symmetry restriction, and that identification based on either is observationally equivalent. Lewbel, Linton, and McFadden (2002) extend and apply the binary choice estimator in Lewbel (2000) to recover general features of a distribution from binary outcomes, with application to contingent valuation and willingness to pay studies. Other empirical applications of the methodology in discrete choice applications include Anton, Fernández, and Rodriguez-Póo (2001) and Cogneau and Maurin (2001). The latter analyze enrollment of children into school in Madagascar, using the date of birth of the child within the relevant year as the regressor V that satisfies exclusion and large support.

No restriction is placed on the joint distribution of M^* and U^* other than equation (4). In the wage example, this means that no restriction is placed on the joint distribution of unobservables such as ability that determine employment status and wages, other than that these unobservables are (conditionally on X) independent of the one covariate V. This is a markedly weaker restriction on the joint distribution of outcomes and selection than is required by other selection or treatment estimators. In particular, the estimators proposed here do *not* assume unconfoundedness, selection on observables, independence of errors and covariates, or any parameterization of the joint distribution of selection and outcome errors. The assumptions permit general forms of endogeneity, mismeasurement, and heteroskedasticity in both selection and outcomes.

Estimation based on equation (3) is consistent whether A^* in equation (1) is finite or infinite. The estimator is numerically the same in either case. When A^* is finite the estimator has an ordinary root n limiting distribution. However, when A^* and the support of M^* are both infinite, then the identification is only at infinity, as in Heckman (1990) and Andrews and Schafgans (1998). If A^* is infinite and the support of $V|X$ is bounded, then the estimator can still be used and will again have an ordinary root n limiting distribution, though in this case the estimator will be asymptotically biased, with bias of order $1/\tau$ where τ is the largest value V can take on. Since τ can be arbitrarily large this bias, when it is present, can be arbitrarily small. In particular, with a finite sample size one could never tell if V really had infinite support, or if it had bounded support with τ large enough to make the resulting bias smaller than any computer roundoff error, or any number of significant digits one chose for reporting estimates.

Many estimators exist for treatment, sample selection, missing data, and other related 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 endogenous regressors as functions of instruments) would need to be completely specified. Alternative assumptions like unconfoundedness or selection on observables likewise impose strong restrictions on the joint behavior of unobservables affecting selection and outcomes.

Related parametric estimators of sample selection models include Horvitz and Thompson (1952), Heckman (1974, 1976, 1979, 1990), Rubin (1974), Koul, Susarla, and van Ryzin (1981), Lee (1982), and Rosenbaum and Rubin (1985). Related semiparametric estimators of sample selection, and other probability weighted models, include Newey (1988), (1999), Cosslett (1991), Ichimura and Lee (1991), Lee (1992, 1994), and Ahn and Powell (1993), Angrist and Imbens (1995), Donald (1995), Wooldridge (1995), Kyriazidou (1997), Andrews and Schafgans (1998), Hahn (1998), Chen and Lee (1998), Abadie (2001), Hirano, Imbens and Ridder (2003), and Das, Newey, and Vella (2003). Surveys include Wainer (1986), Manski (1994), and Vella (1998).

2 Mean Estimation

Selection or treatment is determined by the model $D = I(0 \leq M^* + V \leq A^*)$. The following lemma gives an alternate characterization for the case where A^* is infinite. Proofs of Lemmas, Theorems, and Corollaries are in the Appendix.

LEMMA 1. If $pr(D = 1 \mid V, X)$ is strictly monotonically increasing in V, then there exists an M^* such that $E(D \mid V, X) = E[I(0 \leq M^* + V) \mid V, X]$ and $V \perp X, M^*$.

Lemma 1 shows that the assumption that selection is determined by a model of the form $D = I(0 \leq M^* + V)$ with an independent M^* is observationally equivalent to just assuming that, conditional on V, X, the probability of selection is monotonically increasing in V. Closely related equivalence results are Vytlacil (2002) and Magnac and Maurin (2003). In this case where A^* is infinite, D equals one with probability approaching one as V goes to infinity, which provides identification at infinity for the outcome model.

When A^* is finite, the probability of selection is not monotonic in V. Instead,

this probability goes to zero for both small and large V, so in this case the outcome model is not identified at infinity.

ASSUMPTION A1. D is a binary treatment or selection indicator, X is a covariate vector, and V is a covariate scalar. U^* is an unobserved random vector with finite mean. $U = U^*D$. The indicator D is determined by $D = I(0 \leq M^* + V \leq A^*)$, where A^* and M^* are unobserved latent variables and $E(U^* \mid A^*) = E(U^*)$. The random scalar V is continuously distributed conditional on X. $V \mid X, U^*, M^*, A^* \sim V \mid X$. $0 < E(D) < 1$. Either $E(A^*)$ is finite or $E(M^*U^*)$ and $E(M^*)$ are finite.

Let $f(V \mid X)$ denote the probability density function of V given X and define $W = D/f(V \mid X)$.

ASSUMPTION A2. The support of $V|X$ is an interval on the real line and contains the supports of $-M^*|X$ and $A^* - M^*|X$.

Define the estimator

$$\widehat{\mu} = \frac{\sum_{i=1}^{n} U_i W_i}{\sum_{i=1}^{n} W_i} \tag{5}$$

THEOREM 1. Let Assumptions A1 and A2 hold. Given n independent, identically distributed draws of U_i, W_i,

$$E(U^*) = plim\ \widehat{\mu} \tag{6}$$

The proof given for Theorem 1 in the appendix separately considers the three cases where $E(A^*)$ is finite (which of course includes the case where A^* is a constant), where A^* is infinite, and other cases such as A^* infinite with probability between zero and one, as would arise if some individuals have $D = I(0 \leq M^* + V \leq A^*)$ for finite A^* while others have $D = I(0 \leq M^* + V)$. A more concise proof combining these cases is possible, but intermediate results in the proof as provided are used later.

Theorem 1 provides a consistent estimator of the mean of P^* when P^* is (conditional on X) independent of V, by letting $P^* = U^*$. This conditional independence can be rather restrictive, so later extensions relax this, instead imposing only that the errors in a model that includes P^* be conditionally independent of V.

Theorem 1 immediately implies identification of the entire distribution function of U^* (and hence of P^* under conditional independence of V) because, for any constant c, Theorem 1 can be applied replacing U_i^* with $I(U_i^* \leq c)$ and replacing U_i with $I(U_i \leq c)D_i$, from which it then follows that

$$plim \frac{\sum_{i=1}^n I(U_i \leq c)W_i}{\sum_{i=1}^n W_i} = E\left[I(U_i^* \leq c)\right] = prob(U_i^* \leq c).$$

Similarly, Theorem 1 can also be used to directly estimate any moments of U^* by replacing U_i^* and U_i with $(U_i^*)^c$ and $(U_i)^c$ respectively.

ASSUMPTION A3. The support of $V|X$ contains the interval $(-\tau, \tau)$ for some positive constant τ.

COROLLARY 1. Let Assumptions A1 and A3 hold. Then, given n independent, identically distributed draws of U_i, W_i, $plim\ \widehat{\mu} = E(U^*) + O(\tau^{-1})$.

Assumption A2 cannot be directly tested since M^* is unobserved, but Corollary 1 shows that as long as the observable V has a large support as defined by Assumption A3 for large τ, the estimator will have at most a small asymptotic bias even if the support of V is not large enough to satisfy Assumption A2.

ASSUMPTION A4. $\sup(supp(W))$ and $E\left(U^{*2}\right)$ are finite. Define $\mu = plim\ \widehat{\mu}$.

Assumption A4 implies that the means and variances of W and UW are finite.

COROLLARY 2. Let Assumptions A1 and A4 hold. Then, given n independent, identically distributed draws of U_i, W_i,

$$\sqrt{n}(\widehat{\mu} - \mu) \xrightarrow{d} N\left(0, \frac{var[(U-\mu)W]}{E(W)^2}\right).$$

Corollary 2 shows that with finite fixed or random A^*, the estimator $\widehat{\mu}$ is root n consistent and asymptotically normal. Assumption A4 conflicts with Assumption A2 when A^* is infinite. However, by Corollary 1, if we limit V to a very large but not infinite support, the resulting asymptotic bias $\mu - E(U^*)$ will be tiny. If V has large enough bounded support, this bias can be made smaller than any computer

roundoff error, while still preserving Assumption A4 and hence a root n normal limiting distribution.

The difficulty with allowing A^* to be infinite while satisfying Assumption A2 (which would then require the upper bound on the support of V to be infinite) is not only that the rate of convergence of $\widehat{\mu}$ becomes slower than root n, but also that the Lindeberg condition for asymptotic normality at any rate will generally be violated. This problem could be overcome using asymptotic trimming, replacing the weights W_i in the estimator with $W_{\tau i}$ as defined in the proof of Theorem 1, and letting $\tau \to \infty$ at an appropriate rate (slower than the rate in the proof of Theorem 1, and hence using an estimator that is not asymptotically equivalent to $\widehat{\mu}$). The resulting estimator would then be essentially equivalent to the Andrews and Schafgans (1998) identification at infinity estimator (using f in place of their weighting function), and so is not pursued further here. So, although the estimator can be consistent even when A^* is infinite by Theorem 1, for the sake of obtaining simple limiting distributions, avoiding asymptotic trimming, exploiting Corollary 1 in place of the untestable Assumption A2, and avoiding duplication of existing identification at infinity estimators, Assumption A4 will be maintained for the remainder of this paper, with the understanding that the estimand will therefore suffer from an arbitrarily small asymptotic bias in applications where A^* is infinite.

3 GMM Estimation

ASSUMPTION A5. Let P be an observed outcome satisfying $P = P^*D$ for some latent, unobserved P^*. Let $U^* = \psi(P^*, X, V, \theta^*)$ and $U = \psi(P, X, V, \theta^*)D$, where the vector valued function $\psi(P, X, V, \theta)$ is known and continuously differentiable in a parameter vector θ. Define Θ to be the set of possible values of θ and Ω to be a positive definite matrix. Among all $\theta \in \Theta$, $E[\psi(P^*, X, V, \theta)] = 0$ only if $\theta = \theta^*$. For all $\theta \in \Theta$, $\nu(\theta_0)'\Omega\nu(\theta)$ is nonsingular, where $\nu(\theta)$ and θ_0 are given by

$$\nu(\theta) = E\left(W\frac{\partial\psi(P, X, V, \theta)}{\partial\theta}\right)/E(W) \tag{7}$$

$$\theta_0 = \arg\min_{\theta\in\Theta} E[W\psi(P, X, V, \theta)]'\Omega E[W\psi(P, X, V, \theta)] \tag{8}$$

THEOREM 2. Let Assumptions A1, A3, A4, and A5 hold. If $\mu = E(U^*)$ then $\theta_0 = \theta^*$, and if $\mu = E(U^*) + O(\tau^{-1})$ then $\theta_0 = \theta^* + O(\tau^{-1})$.

COROLLARY 3. Let Assumptions A1, A3, A4, and A5 hold. Assume n independent, identically distributed draws of W_i, P_i, X_i, V_i. Θ is compact and θ_0, which is uniquely defined by equation (8), is in the interior of Θ. The second moment of $W\psi(P, X, V, \theta)$ is finite. $W\partial\psi(P, X, V, \theta)/\partial\theta$ is bounded in absolute value by a function $b(W, P, X, V)$ that has finite mean. Ω_n is a sequence of positive definite matrices with $p\lim \Omega_n = \Omega$. Define $\widehat{\theta}$ by

$$\widehat{\theta} = \arg\min_{\theta\in\Theta} \left[\sum_{i=1}^{n} W_i \psi(P_i, X_i, V_i, \theta)\right]' \Omega_n \left[\sum_{i=1}^{n} W_i \psi(P_i, X_i, V_i, \theta)\right]$$

Then

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \left(S_0'\Omega S_0\right)^{-1} S_0'\Omega\Sigma_0\Omega S_0 \left(S_0'\Omega S_0\right)^{-1}\right).$$

where $S_0 = E[W\partial\psi(P, X, V, \theta_0)/\partial\theta]$ and $\Sigma_0 = E[W^2\psi(P, X, V, \theta_0)\psi(P, X, V, \theta_0)']$.

Efficiency is obtained in the usual way by two step GMM, constructing Ω_n so that $\Omega = \Sigma_0^{-1}$.

As discussed in the introduction, if there were no selection problem so P^* could be observed, θ would be estimated by applying GMM to the moments $E[\psi(P^*, X, V, \theta^*)] = 0$. This estimator is infeasible, but Theorem 2 and Corollary 3 describe the corresponding feasible GMM that replaces P^* with the observable P and corrects for selection by multiplying the moments by W, i.e., using the feasible moments $E[W\psi(P, X, V, \theta_0)] = 0$. The resulting estimand θ_0 equals the desired θ^* when A^* has finite mean and the support of V is sufficiently large, otherwise the difference between θ_0 and θ^* (the asymptotic bias) is of order $O(\tau^{-1})$ where τ is the largest value V can take on. As discussed earlier, this bias can be assumed to be smaller than any printed coefficient roundoff error, by a support assumption that could never be falsified with a finite data set. Of course the real question, addressed later in the empirical applications and Monte Carlo, will be whether the above root n limiting distribution theory provides a good approximation to the distribution of $\widehat{\theta}$.

4 Example Models

The example models in this section will all be consistent with the assumptions of Theorem 2. These models use the following definitions and assumptions. Let Z be

a vector of covariates, which are exogenous in the sense that they are uncorrelated with the errors in a model of P^*. Assume Z does not include V (more generally, V is not a deterministic function of Z). Let Y be a vector of covariates, some of which may be endogenous, in that they may be correlated with the errors in a model of P^*. The vector Y can include elements of Z. Define X to be the union of all the elements of Z and Y. The data consist of a sample of observations of X, V, P, D, which implies that Z and Y are also observed.

4.1 Examples of Selection Models

This paper's estimators assume equation (1) holds, but they do not require specifying, modeling or estimating the resulting probability of selection (propensity score). However, as an illustration consider $D = I(a_0 + e_0 \leq M(X, e) + V \leq a_1 + e_1)$ with errors e_1, e_0, and e and unknown function M. Then equation (1) holds with $A^* = a_1 - a_0 + e_1 - e_0$ and $M^* = -a_0 + M(X, e) + e_0$. This is a random thresholds ordered selection model. For example, $M(X, e)$ could equal the benefits of college and $-V$ could be a measure of the cost of college. If benefits minus cost for an individual are low, below $a_0 + e_0$, then the individual does not get a college degree, while if the benefits are very high, above $a_1 + e_1$, then the individual goes on to graduate school. D would then the indicator of just getting a college degree, and the associated P^* could be some outcome like earnings associated with getting just a college degree. The possible randomness in the thresholds, e_0 and e_1, could embody unobserved heterogeneity in the utility of education.

More common models like fixed threshold ordered choice $D = I(a_0 \leq X'\beta + V + e \leq a_1)$ or standard binary choice $D = I(0 \leq X'\beta + V + e)$, are special cases that are included in this general framework. The estimator allows covariates other than V to be endogenous, and errors to be heteroskedastic. For example, in all of the above models X could be endogenous or mismeasured, with the joint distribution of e_1, e_0, e, X unknown. More generally, the estimator does not require modeling or estimating the dependence of M^* on X, and the estimator is the same regardless of whether A^* is constant, random, or infinite. Empirical applications with both finite A^* and infinite A^* are provided.

4.2 Examples of Outcome Models

Suppose for a known function h that

$$h(P^*, Y, V, \beta) = \varepsilon, \quad E(\varepsilon Z) = 0, \quad \varepsilon \mid V, X \sim \varepsilon \mid X \tag{9}$$

Some or all of the elements of P^*,Y may be endogenous and hence correlated with ε, so estimation is based on $E(\varepsilon Z) = 0$, that is, if P^* were observed the parameters β would be estimated by applying GMM to the moments $E[Zh(P^*, Y, V, \beta)] = 0$. The unobservables that affect selection D (that is, M^* and A^*) can be correlated in unknown ways with ε. This model fits the assumptions of Theorem 2 by defining the function ψ as

$$\psi(P^*, X, V, \theta) = Zh(P^*, Y, V, \beta).$$

For example, $h(P^*, Y, V, \beta) = P^* - H(Y, V, \beta)$ could be a model of wages P^* where Y includes some endogenous regressors, e.g., spouse's income or transfers from parents, and V (which helps determine labor force participation D) can be regressor in the h model.

Another example is $h(P^*, Y, V, \beta) = Y - H(P^*, Z, V, \beta)$, which could be a model of consumption of a vector of goods Y where P^* is income that is not reported by a significant number of individuals in the sample, and where income nonresponse is correlated with ε, even after conditioning on observables.

Another class of models that can estimated using Corollary 3 are models that could have been estimated by maximum likelihood if P^* were observed. For example, suppose that

$$P^* = H(X, \beta, \varepsilon), \quad \varepsilon \mid V, X \sim \varepsilon \mid X, \quad \varepsilon \sim F_\varepsilon(\varepsilon \mid X, \delta) \tag{10}$$

So H is a known parametric model for P^* having latent errors ε, The errors ε are conditionally independent of V given X (so V is exogenous), and the conditional distribution function of ε given X, denoted by F_ε, is known up to a parameter vector δ. The unobservables that affect D can be correlated in unknown ways with ε. Let $\theta = (\beta, \delta)$. Assuming each $\varepsilon \mid X$ is independently and identically distributed, we can construct a corresponding log likelihood function $\sum_{i=1}^{n} L(P_i^*, X_i, \theta)$ that could be used to estimate θ by maximum likelihood if P^* were observable. Given ordinary maximum likelihood regularity, define the function ψ by $\psi(P^*, X, V, \theta) = \partial L(P^*, X, \theta)/\partial\theta$, (the score function) and θ would be identified from $E[\psi(P^*, X, V, \theta_0)] = 0$. Corollary 3 can then be used to estimate θ given P instead of P^*

For an example of model (10), suppose that

$$P^* = I(\beta' X + \varepsilon \geq 0), \quad \varepsilon \perp V, X, \quad \varepsilon \sim N(0, 1)$$

that is, the unobserved outcome P^* is determined by a probit model. Equivalently, we may interpret this model as one where an individual makes two binary decisions or choices, D and P^*, and we can only observe the outcome of the second

decision, P^*, when the first decision is $D = 1$. The unobservables that affect both decisions are related in unknown ways, so we do not know the joint distribution of ε and errors in the D model, nor do we know how those errors jointly vary with X. It is only the marginal distribution of ε that is specified here. In this example $\psi(P^*, X, V, \beta)$ would just be the ordinary probit score function for each observation of this P^*. As this example shows, we do not require continuity of P^*, so the methodology can be used without change when the unobserved potential outcome P^* is discrete, censored, or otherwise limited.

The main impact of the exclusion restriction (4) for the P^* model is the implication that $U^* \mid V, X \sim U^* \mid X$ for $U^* = \psi(P^*, X, V, \theta)$. In the above examples, the assumption that either $\varepsilon \mid V, X \sim \varepsilon \mid X$ or $\varepsilon \perp V, X$ ensures that this requirement is satisfied.

4.3 Examples of Density Models

The GMM estimator in Corollary 3 assumes W and therefore the density $f(V \mid X)$ is known. Estimation remains straightforward if $f(V \mid X)$ is finitely parameterized. In this case, denote the conditional density of V as $f(V \mid X, \lambda)$, and let the vector θ include both the set of unknown parameters in the P^* model ψ and the parameters λ of the distribution of f. Let $R(V, X, \theta)$ be any vector valued function having the property that λ is identified from the moments $E[R(V, X, \theta)] = 0$. In particular, we could let

$$R(V, X, \theta) = \frac{\partial \ln f(V \mid X, \lambda)}{\partial \lambda}$$

in which case $R(V, X, \theta)$ is the score function associated with the maximum likelihood estimator of the parameters of f. Estimation of the model then proceeds by applying GMM to the set of moments

$$E\begin{pmatrix} \psi(P, X, V, \theta) D / f(V \mid X, \theta) \\ R(V, X, \theta) \end{pmatrix} = 0 \tag{11}$$

and standard GMM limiting distribution theory applies. See, e.g., Newey (1984) or Wooldridge (2001), p. 425.

For example, suppose that we can model V in terms of other covariates as

$$V = g(X, \alpha) + \sigma(X, \gamma)\eta, \quad \eta \perp X, \varepsilon \tag{12}$$

for some known functions g and σ with $\sigma(X, \gamma) > 0$, unknown parameter vectors α and γ, and unobserved error term η having some known density function f_η with

mean zero and variance one. We may think of η as an exogenous covariate in the P^* model. Then

$$f(V \mid X) = \frac{1}{\sigma(X,\gamma)} f_\eta\left(\frac{V - g(X,\alpha)}{\sigma(X,\gamma)}\right) \tag{13}$$

and equation (11) then becomes

$$E\begin{pmatrix} \frac{\sigma(X,\gamma)\psi(P,X,V,\theta)D}{f_\eta\left(\left[V-g(X,\alpha)\right]/\sigma(X,\gamma)\right)} \\ V - g(X,\alpha) \\ \left[V - g(X,\alpha)\right]^2 - \sigma(X,\gamma)^2 \end{pmatrix} = 0 \tag{14}$$

where θ (which includes α and γ along with whatever parameters appear in the model ψ) is estimated by applying standard GMM to this set of moments. A leading case would be taking f_η to be standard normal, though other (in particular more flexible classes of densities), could also be used.

If $f(V \mid X)$ is not finitely parameterized, the GMM estimator can still be applied by replacing $W = D/f(V \mid X)$ with $\widehat{W} = D/\widehat{f}(V \mid X)$ in Corollary 3, where $\widehat{f}(V \mid X)$ is a nonparametric estimator of $f(V \mid X)$, such as a kernel density estimator. The general limiting distribution theory for these types of semiparametric two step estimators (where the first step is a nonparametric plug in like this) is given by Newey and McFadden (1984). The relevant result is that, with sufficient regularity (see, e.g., Khan and Lewbel 2005, Theorem A.1, who provide one example of detailed regularity conditions that suffice, including asymptotic trimming to deal with boundary effects in the kernel estimation of f in the denominator),

$$\begin{aligned} &\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{T_i}{\widehat{f}(V_i \mid X_i)} - E\left(\frac{T}{f(V \mid X)}\right)\right] \\ &\xrightarrow{d} N\left(0, var\left[\frac{T}{f(V \mid X)} + E\left(\frac{T}{f(V \mid X)} \mid X\right) - E\left(\frac{T}{f(V \mid X)} \mid V, X\right)\right]\right) \end{aligned}$$

where T is a random vector. To apply this result to the GMM estimator, let $T = D\psi(P, X, V, \theta)$ to obtain the limiting distribution of the sample average of $\widehat{W}\psi(P, X, V, \theta)$. The result is if $\widehat{\theta}$ is given by

$$\widehat{\theta} = \arg\min_{\theta\in\Theta}\left[\sum_{i=1}^{n}\frac{D_i\psi(P_i, X_i, V_i, \theta)}{\widehat{f}(V_i \mid X_i)}\right]' \Omega_n \left[\sum_{i=1}^{n}\frac{D_i\psi(P_i, X_i, V_i, \theta)}{\widehat{f}(V_i \mid X_i)}\right] \tag{15}$$

with independent, identically distributed observations, then

$$\sqrt{n}(\widehat{\theta} - \theta_0) \xrightarrow{d} N\left(0, \left(S_0'\Omega S_0\right)^{-1} S_0'\Omega\widetilde{\Sigma}_0\Omega S_0 \left(S_0'\Omega S_0\right)^{-1}\right)$$

where Ω and S_0 are as before and

$$\widetilde{\Sigma}_0 = \upsilon ar\left[\frac{D\psi(P, X, V, \theta)}{f(V \mid X)} + E\left(\frac{D\psi(P, X, V, \theta)}{f(V \mid X)} \mid X\right) - E\left(\frac{D\psi(P, X, V, \theta)}{f(V \mid X)} \mid V, X\right)\right].$$

Efficiency would be obtained in the usual GMM way by having $plim\ \Omega_n = \widetilde{\Sigma}_0^{-1}$. In very closely related contexts, Magnac and Maurin (2003) and Jacho-Chavez (2005) show that the estimator that plugs in a nonparametric estimator of f is semiparametrically efficient, being more efficient than estimation using the true f (similar to Hirano, Imbens, and Ridder 2003).

Nonparametric estimation of $f(V \mid X)$ may be problematic in applications where X has moderate or high dimension. In those cases f could be semiparametrically estimated. For example, suppose equation (12) holds but the distribution of η is unknown. One could then estimate the parameters of equation (12) by GMM, apply a one dimensional nonparametric density estimator to the estimated residuals $\widehat{\eta}$ from that equation to obtain $\widehat{f_\eta}$, then estimate $\widehat{\theta}$ by equations (13) and (15). Lewbel and Schennach (2005) provide root n limiting distribution theory for a numerically simple "sorted data" estimator of this form (one that does not require selection of kernels or bandwidths), using

$$\frac{1}{\widehat{f_\eta}(\widehat{\eta}_i)} = \frac{\widehat{\eta}_{[i+1]} - \widehat{\eta}_{[i-1]}}{2n} \tag{16}$$

where $\widehat{\eta}_{[i+1]}$ is the smallest value of $\widehat{\eta}_1$,...,$\widehat{\eta}_n$ that is greater than $\widehat{\eta}_i$ and $\widehat{\eta}_{[i-1]}$ is the largest value of $\widehat{\eta}_1$,...,$\widehat{\eta}_n$ that is smaller than $\widehat{\eta}_i$.

4.4 Two Stage Least Squares

Suppose that

$$P^* = Y'\beta_Y + V\beta_V + \varepsilon, \quad E(\varepsilon Z) = 0, \quad \varepsilon \mid V, X \sim \varepsilon \mid X. \tag{17}$$

This is just the special case of model (9) where $h(P^*, Y, V, \beta)$ is $P^* - Y'\beta_Y - V\beta_V$. As before, some of the elements of Y may be endogenous and hence correlated with ε, so estimation is based on $E(\varepsilon Z) = 0$, which for this linear model

means that if P^* were observed, then the parameters $\beta = \beta_Y, \beta_V$ could be estimated by regressing P^* on Y, V using linear two stage least squares with instruments Z.

With this model, the GMM estimator of Theorem 2 and Corollary 3 is based on the moments

$$E\left[ZW(P - Y'\beta_Y - V\beta_V)\right] = 0$$

and so simplifies to estimating β by linearly regressing WP on WY, WV using two stage least squares with instruments Z. Define

$$\begin{aligned}\Delta &= \left[E\left(W\begin{pmatrix}Y\\V\end{pmatrix}Z'\right)E(ZZ')^{-1}E\left(WZ\begin{pmatrix}Y\\V\end{pmatrix}'\right)\right]^{-1}E\left(W\begin{pmatrix}Y\\V\end{pmatrix}Z'\right)E(ZZ')^{-1}\\ \beta &= \Delta E(WZP)\end{aligned}$$

The estimator is just these equations, replacing expectations with sample averages.

Suppose $f(V \mid X)$ is parameterized as $f(V \mid X, \lambda)$ and we have some estimator for the vector λ satisfying

$$\sqrt{n}(\widehat{\lambda} - \lambda) \xrightarrow{d} N[0, var(Q_\lambda)] \tag{18}$$

for some influence function Q_λ. For example, λ might consist of means or other moments of V, X and $\widehat{\lambda}$ could be the corresponding sample moments, or λ could be estimated by a separate GMM, or by maximum likelihood as before. Then the estimator $\widehat{\beta}$ is a linear two stage least squares regression of $PD/f(V \mid X, \widehat{\lambda})$ on $(Y, W)D/f(V \mid X, \widehat{\lambda})$ using instruments Z. With independent, identically distributed observations, standard limiting distribution theory for parametric two step estimation (see, e.g., section 6 of Newey and McFadden 1994) then gives

$$\sqrt{n}(\widehat{\beta} - \beta) \xrightarrow{d} N[0, \Delta var\left(Q_\beta - WZY'\beta\right)\Delta'] \tag{19}$$

where

$$Q_\beta = WZP\left(1 - Q_\lambda' \frac{\partial \ln f(V \mid X, \lambda)}{\partial \lambda}\right)$$

If instead of a parametric density, we use a kernel or other sufficiently regular nonparametric density estimator $\widehat{f}(V \mid X)$, then let $\widehat{W} = D/\widehat{f}(V \mid X)$ and estimate β by regressing $\widehat{W}P$ on $\widehat{W}Y$, $\widehat{W}V$ using linear two stage least squares with instruments Z. This is then just a special case of GMM with a nonparametric density estimator as described in the previous section, which yields the same limiting

distribution (19) as the parametric density case, except that now

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

These two stage least squares estimators do not require numerical searches, so it would also be computationally feasible to estimate alternative limiting distributions by bootstrapping.

4.5 A Numerically Trivial Estimator

Consider the weighted least squares model with a semiparametric specification of f, specifically,

$$\begin{aligned} P^* &= Y'\beta_Y + V\beta_V + \varepsilon, \quad E(\varepsilon Z) = 0, \\ V &= X'\alpha + \eta, \quad \eta \perp \varepsilon, X \end{aligned}$$

and D given by equation (1). Assume the distributions of errors ε, η and unobservables A^*, M^* are unknown. This model is a special case of equations (17) and (12), and provides a compromise between parametric vs nonparametric V density estimation. A numerically trivial estimator (which combines estimators described in the previous sections) consists of the following steps.

First estimate a by linearly regressing V on X using ordinary least squares, and then let $\widehat{\eta}_i = V_i - X_i'\widehat{\alpha}$ at each data point i. By equation (13) in this model we have $W = D/f_\eta(\eta)$, so by equation (16) at each data point i we may construct $\widehat{W}_i$ using

$$\widehat{W}_i = \frac{\widehat{\eta}_{[i+1]} - \widehat{\eta}_{[i-1]}}{2n} D_i \tag{20}$$

where $\widehat{\eta}_{[i+1]}$ is the smallest value of $\widehat{\eta}_1,...,\widehat{\eta}_n$ that is greater than $\widehat{\eta}_i$ and $\widehat{\eta}_{[i-1]}$ is the largest value of $\widehat{\eta}_1,...,\widehat{\eta}_n$ that is smaller than $\widehat{\eta}_i$. To handle the endpoints, for $\widehat{\eta}_j = \min(\widehat{\eta}_1,...,\widehat{\eta}_n)$ let $\widehat{\eta}_{[j-1]} = \widehat{\eta}_j$ and for $\widehat{\eta}_j = \max(\widehat{\eta}_1,...,\widehat{\eta}_n)$ let $\widehat{\eta}_{[j+1]} = \widehat{\eta}_j$. Alternatively, instead of equation (20), $f_\eta\left(\eta_i\right)$ could be estimated using a one dimensional nonparametric kernel density estimator. Finally, estimate β by linearly regressing $\widehat{W}P$ on $\widehat{W}Y$, $\widehat{W}V$ using two stage least squares with instruments Z. Lewbel and Schennach (2005) provides relevant asymptotic distribution theory, however, since the estimator consists only of sorting data and some linear regressions, the alternative of bootstrapping it is computationally trivial.

4.6 Estimation Notes

The estimators entail division by a density $f(V \mid X)$. This raises potential numerical issues when the density is close to zero. Formally, asymptotic trimming is required for high dimensional nonparametric density plug ins, and while not required in theory for root n convergence with a parametric $\widehat{f}$, may still be advisable in terms of mean squared error. Essentially, the issue is that W_i will be extremely small, and hence U_i may be extremely large, for observations i where V_i is moderately large, and such observations may dominate the associated sample averages. Basically, such observations i will be outliers. This suggests that one should look for and possibly discard outliers in the moments U_i or two stage least squares errors. One could more formally replace the weighted means in the GMM or two stage least squares with robust estimators of such means, e.g., discarding some small percentage of the observations having the largest residuals.

Alternatively, one could simply use the median of the bootstrap distribution of parameters as an estimator, that is, randomly draw an integer j from 1 to n. Do this n times, and then use the resulting n observations of data D_j, P_j, X_j, V_j to estimate the parameter vector θ. Repeat this procedure a large number of times, and for each element θ_k of θ use the median of the resulting set of estimates of θ_k as the estimator $\widehat{\theta}_k$, and use the $2.5th$ and $97.5th$ quantile of the set of estimates of θ_k as a ninety five percent confidence interval for $\widehat{\theta}_k$.

5 A Factory Investment Model

Let P_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 P_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 (2003). These empirical findings are generally attributed to discontinuous costs of adjustment, due to factors such as irreversibility or indi-

visibility 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 R_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 R_i is that it may be endogenous, 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

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

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 derivations in Abel and Eberly (1994), given the above investment cost function the firm chooses investment P_i to maximize the present value of current and expected future profits, resulting in a model of the form

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

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 ($P_i < 0$) 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 P and D hold as written for all firms if P_i is set to zero for any firm having negative investment.

The above equations show that profit maximization yields a model that has a sample selection structure. It also has the features that the outcome P is linear in Q when $D = 1$, and that the fixed cost parameter a_{1i} appears only in the expression for D.

Comparing this model for D with equation (1) shows that A^* will be infinite, but one could construct more elaborate versions that would give rise to a finite upper bound, e.g., if it were the case that a firm would build a new plant rather than invest in the old one when the benefits from investment were sufficiently large.

Marginal plant level Tobin's Q is not observed, and is proxied by the profit rate R_i. Specifically, Q_i is assumed to be linear in R_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 $P_i = (R_i\beta_1 + X'_{2i}\beta_2 + \varepsilon_i)D_i$. 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 R_i. See Abel and Eberly 1994 for details). Because these assumptions are unlikely to hold in practice, the estimator here will not require ε_i to be independent of the profit rate R_i, and so 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 $H(z) = E(R \mid Z = z)$, and define ε_{Ri} by $R_i = H(Z_i) + \varepsilon_{Ri}$. The function H is unknown. Because of endogeneity of profits, the error term ε_{Ri} 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 P to Q does not depend on the size of the firm or plant, except to the extent that both P 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 one possible 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 (2003) 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. 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. Since V_i is a plant characteristic, it may also appear everywhere in that model that X_{2i} appears.

Based on the above, it is assumed that a_{1i} depends on V_i and may also depend on other characteristics of the plant, firm, or industry, both observed X_{2i} and unobserved e_i. Consistent with the presence of absolute fixed costs, Nilsen and Schiantarelli (2003) find that D is monotonically increasing in V, so (recalling Lemma 1) we may write the resulting selection equation as $D_i = I[0 \leq V_i + M(R_i, X_{2i}, e_i)]$ 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_{Ri}. 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 above derivations yield the following system of equations

$$P_i = (R_i\beta_1 + X_{2i}'\beta_2 + V_i\beta_3 + \varepsilon_i)D_i \tag{21}$$

$$D_i = I[0 \leq V_i + M(R_i, X_{2i}, e_i)] \tag{22}$$

$$R_i = H(Z_i) + \varepsilon_{Ri} \tag{23}$$

The profit rate R_i is endogenous, correlated with ε_i, and the selection unobservables vector e_i is correlated with both the investment rate error ε_i and the profit rate error ε_{Ri}. The joint distribution of these errors, and the functions M and H, are unknown. The goal is estimation of the parameters β. The coefficient of R_i, which is β_1, is of particular interest as the proxy for the relationship between investment and Q.

Equation (21) takes the form $P = P^*D$ with P^* linear, so this paper's two stage least squares estimators will be used. In the notation of the previous sections of this paper, P, D, V, and Z are the same, Y is R, X_2, and X is the union of the elements of R, X_2, and Z.

5.1 Data and Estimation

The model 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 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.

P_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 P_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 P_i set to zero. The selection function is then $D_i = I(P_i > 0)$.

The variable R_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 R_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 an earlier year. Experiments with other measures of size, such as lagged capital stock, yielded similar results.

To apply this paper's estimator for β, we need the assumptions of Corollary 3 to hold. The structural model is equations (21), (22), and (23). This requires that the unobservables in the model, e, ε, and ε_R, be conditionally independent of V, conditioning on Z. The most likely source of violation of this assumption is from the replacement of Q with R (in part because Q should subsume all dynamic effects, but R need not). This was also the motivation for inclusion of the term $\beta_3 V_i$ in equation (21).

The required support conditions for V imply that, at any given time, some plants could be so small that they will not invest regardless of their values for X_2 and e, while other plants could be so large that they definitely invest. Over time plant sizes change, both through the model via investment, and by depreciation, closings, technology change, etc., so the model does not require the existence of plants that are permanently static or permanently growing.

Empirically, the supports of the continuous variables in this model are unknown, so the required support conditions cannot be directly verified. However, indirect evidence is favorable. In this data set V takes on a large range of values relative to the other covariates. For example, as a measure of data spread, the standard deviation of V is 1.16, while the profit rate R has a standard deviation of .17. In the applications where, for comparison, the selection equation is parameterized, the systematic component of M^*, modeled as $X_2'\gamma$, 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, Lewbel (2000) found that the similar binary choice estimator generally performed well when the standard deviation of V was comparable in magnitude to the standard deviation of M^*.

Very strong alternative assumptions are required to estimate β by other means,

such as maximum likelihood. The model can be rewritten as

$$\begin{aligned} R_i &= H(Z_i) + \varepsilon_{Ri} \\ P_i &= [H(Z_i)\beta_1 + X_{2i}'\beta_2 + V_i\beta_3 + (\varepsilon_{Ri}\beta_1 + \varepsilon_i)]D_i \\ D_i &= I[0 \leq V_i + M(H(Z_i) + \varepsilon_{Ri}, X_{2i}, e_i)] \end{aligned} \tag{24}$$

The parametric model that will be estimated for comparison is

$$\begin{aligned} R_i &= Z_i'b + \widetilde{\varepsilon}_{Ri} \\ P_i &= [(Z_i'b)\beta_1 + X_{2i}'\beta_2 + V_i\beta_3 + \widetilde{\varepsilon}_i)D_i \\ D_i &= I[0 \leq V_i + (Z_i'b)\gamma_1 + X_{2i}'\gamma_2 + \widetilde{e}_i] \end{aligned} \tag{25}$$

where the errors $(\widetilde{e}_{Ri}, \widetilde{\varepsilon}_i, \widetilde{e}_i)$ are assumed to be trivariate normal and independent of Z_i and V_i. Unlike the general semiparametric specification, this parametric model requires the simultaneous estimation of the equations for R_i and D_i along with the P_i equation. The parametric model also assumes that the functions H and M are linear, that the errors and unobservables ε_{Ri} 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 any standard method 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 semiparametric estimation, which does not require such assumptions.

5.2 Empirical Results

Table 1 summarizes results for six different estimators. For brevity, Table 1 only reports estimates of the coefficient of interest, β_1. A complete list of all parameter estimates in all equations, along with the Gauss code used to generate them, are available from the author on request.

Let $X_{1i} = R_i$, X_{2i} and let β denote the corresponding coeficients in equation (21). The first and second estimators in Table 1, labeled OLS and TSLS, ignore the sample selection problem, and just estimate the equation $P_i = X_{1i}'\beta + \widetilde{\varepsilon}_i$ by ordinary least squares and two stage least squares, respectively (the latter using instruments Z_i).

The third estimator, labeled Heckman, controls for sample selection parametrically, but does not control for possible endogeneity. This is the two equation

parametric model $P_i = (X'_{1i}\beta + \widetilde{\varepsilon}_i)D_i$ and $D_i = I[0 \leq V_i + X'_{1i}\gamma + \widetilde{e}_i]$, 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, labeled Endogenous ML, is maximum likelihood estimation of the entire three equation parametric model (25), which entails simultaneously estimating the parametric selection, outcome, and instrument equations, assuming $\widetilde{e}_{Ri}$, $\widetilde{\varepsilon}_i$, and $\widetilde{e}_i$ are jointly normal and independent of Z_i and V_i.

The remaining estimators are this paper's estimators from section 4.4. The fifth estimator, Weighted OLS, is a linear least squares regression of $\widehat{W}_i P_i$ on $\widehat{W}_i X_{1i}$, where the weights are $\widehat{W}_i = D_i / \hat{f}(V_i \mid X_{1i})$. This semiparametrically controls for selection but not for endogeneity, and so corresponds to estimating β when the true model is defined by the system of two equations (21) and (22), assuming $\beta_3 = 0$ (see below for more on this point) and ε_i is uncorrelated with X_{1i}.

The final estimator, Weighted TSLS, is a linear two stage least squares regression of $\widehat{W}_i P_i$ on $\widehat{W}_i X_{1i}$ using instruments Z_{1i}, where the weights are $\widehat{W}_i = D_i / \hat{f}(V_i \mid X_i)$ with $X_i = X_{1i}, Z_{1i}$. This estimator semiparametrically controls for both selection and endogeneity, and so corresponds to estimating β when the true model is defined by the general structure of equations (21), (23), and (22).

A kernel density estimator is used to construct $\hat{f}(V_i \mid X_i)$. A quartic kernel is used for continuous regressors, calculated for each cell of the discrete regressors and averaged across cells. The kernel bandwidth is chosen by ordinary cross validation. No density trimming was used. 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 are 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. Attempts to replicate the analysis for a different year of data failed because no converged values for the maximum likelihood estimator could be obtained. The difficulty with maximum likelihood is that some parameters are intrinsically difficult to identify in the sense that the likelihood function is relatively flat in directions that involve changing these parameters. These parameters include 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.

Table 1 reports estimates imposing $\beta_3 = 0$ in equation (21). This is a necessary assumption for the weighted OLS estimator (because U, which equals ε times instruments, must be conditionally independent of V, and in weighted OLS the instruments are the regressors, which include V when β_3 is nonzero). The other estimators do not require this assumption in theory, however, the Endogenous ML estimates failed to converge when β_3 was allowed to be nonzero, and when the other estimators were redone allowing β_3 to be nonzero, the resulting estimates of β_3 were tiny and completely insignificant statistically. Also, imposing $\beta_3 = 0$ only slightly changed the resulting estimates of β_1. Note that having $\beta_3 = 0$ is consistent with a model where V_i only affects fixed costs of investment.

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.

One could easily question whether V satisfies all of the required 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 produces 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.

5.3 Monte Carlo Simulation

To assess the performance of the proposed estimator, a Monte Carlo simulation based on the investment model application is provided. For the simulation, the true model is taken to be the three equation parametric model (25), without plant or industry dummies, and $\beta_3 = 0$. Parameter values are taken to equal the estimated coefficients from applying maximum likelihood to the investment data (the Endogenous ML model in Table 1) with the full set of plant and industry dummies included in X_{2i}. The intercept term in the outcome equation is then taken to equal the mean in the real data of $X'_{2i}\beta_2$, and the intercepts for the other two equations are defined analogously. The exogenous variables V and Z, corresponding to the size and lagged profit rate variables, are drawn as independent normals with means and variances matching those in the data. The covariance matrix of the normal model errors $(\widetilde{e}_{Ri}, \widetilde{\varepsilon}_i, \widetilde{e}_i)$ is then constructed so that the means, variances,

and covariances of the endogenous variables R_i, P_i, and D_i generated by the parametric model match those in the real data. The sample size is the same as the real data, 974 observations.

Simulated data were drawn in this way five thousand times, and each of the six estimators described in Table 1 were applied to each replication. With each replication, the same code that was used on the real data was applied to the simulated data to provide estimates of both the coefficients and the standard errors.

Table 2 reports summary statistics on the distribution of the estimated profit coefficient in the outcome equation from these simulations. Corresponding summary statistics on all of the estimated parameters, along with the Gauss code used to generate them, is available from the author on request. Reported summary statistics include moments, quantiles, root mean squared errors, and mean and median absolute errors. Also reported is the mean across replications of the estimated standard errors, and the fraction of simulations in which the true coefficient was within two estimated standard errors of the estimated coefficient.

By the Monte Carlo design, the endogenous ML estimator is consistent and efficient, and so provides an asymptotically best case benchmark. The results show that this ML estimator is mean and median unbiased, with a smaller root mean square error than the other estimators, as expected. One way in which ML behaved poorly was that its estimated standard errors were much too large, providing 100% coverage of what is supposed to be a 95% confidence interval. This illustrates the problem noted in the real data analysis that the ML estimates are sensitive to the estimated covariance matrix of the model errors, which in turn is estimated imprecisely because one of the errors is latent. Equivalently, in this application ML is a highly nonlinear function, making the linearization required for standard error estimation a poor approximation.

The OLS, TSLS, Heckman, and Weighted OLS estimators are inconsistent for this design. The simulated estimates of each of these estimators show considerable bias, with means and medians that are very similar to estimates reported with real data (compare the next to last column of Table 1 with the mean and median columns of Table 2). The estimated standard errors of these estimators also closely match the real data estimated standard errors. The Weighted OLS estimator delivers estimates close to those of the Heckman model, as it should.

The Weighted TSLS estimator has about a ten percent mean and median bias, and a standard deviation about double that of ML. This is the price paid for the generality of the semiparametric estimator. Unlike ML, the standard errors of the weighted TSLS are quite accurate, resulting in 96 percent.of coverage for what is supposed to be a 95% confidence interval. This paper's proposed weighted

TSLS estimator does not require estimation of the latent errors (indeed, it does not involve any estimation at all of the selection equation), which may explain its better behavior regarding standard error estimation.

In this Monte Carlo design all of the variables and errors, including V, have unbounded support and no asymptotic trimming was applied. These estimates can be consistent based on Theorem 1, but formally the unbounded support violates our root n limiting distribution theory (since in these designs A^* is infinite). Nevertheless, the Monte Carlo results suggest that the root n limiting distribution theory (with a small limiting bias) provides a good approximation to the observed sampling distribution. Also, these results show that the finite sample bias from the proposed estimator is much smaller than that of other simple biased estimators which ignore either selection or endogeneity. Some experiments with asymptotic trimming were performed, but they are not reported because they did not produce any improvements in the simulations.

6 Wages and Schooling

This section describes an empirical application in which A^* is finite. Let $-V_i$ be the log cost of a year of school, and let M_i^* denote an individual i's unobserved utility from education (comparably normalized), so the larger $M_i^* + V_i$ is, the more education individual i will choose to obtain. Let D_i equal one if i has an undergraduate degree and no post graduate education, and zero otherwise. Then $D_i = I(0 \leq M_i^* + V_i \leq A_i^*)$, where i does not get an undergraduate degree if $M_i^* + V_i < 0$ and gets some graduate education if $M_i^* + V_i \geq A_i^*$. This simple model of the selection equation ignores dynamic optimization issues in schooling choice, but does allow thresholds to vary either randomly or systematically across individuals (see, e.g., Cameron and Heckman 1998 or Carneiro, Hansen, and Heckman 2003), and leaves unspecified the many observables and unobservables that affect utility and thresholds, that is, M_i^* and A_i^*.

Let the potential outcome $P_i^* = Y_i'\beta + \varepsilon_i$ be the log wages individual i would get if he or she chose to obtain an undergraduate degree but no graduate education, where Y_i is a vector of observed covariates and $E(Y_i\varepsilon_i) = 0$, so we do not have endogenous regressors in this example. The goal is estimation of β and hence the returns from obtaining an undergraduate degree. The selection problem is that we can only observe P_i^* for individuals having $D_i = 1$, and we can expect M_i^* and possibly also A_i^* to correlate with P_i^* in unknown ways. We may therefore directly apply the estimators described in section 4.4, to obtain $\widehat{\beta}$ by a linear ordinary least

squares regression of $\widehat{W}_i P_i$ on $\widehat{W}_i Y_i$, where P_i is individual i's observed log wage and $\widehat{W}_i$ is D_i divided by an estimate of the conditional density of V_i given X_i. This density is estimated three ways. The first uses the same nonparametric estimator as in the investment application. The second assumes $V_i = X_i'\alpha + \eta_i$ where η_i is an independent normal error, with the V and WP equations estimated jointly by GMM. The third is the numerically trivial estimator of section 4.5, which again assumes $V_i = X_i'\alpha + \eta_i$, but now with the independent error η_i having an unknown density that is nonparametrically estimated using Lewbel and Schennach (2005). This last estimator is sequential, where first V is linearly regressed on X, then the errors in that regression are sorted and differenced to construct $\widehat{W}$ using equation (20), and last β is estimated by a linear least squares regression of $\widehat{W}_i P_i$ on $\widehat{W}_i Y_i$.

For comparison, estimates are also obtained by just regressing P_i on Y_i for those individuals having $D_i = 1$. This regression suffers from selection bias, unless the $D_i = 0$ observations are missing at random. Also reported are maximum likelihood estimates of a two equation system where A_i^* is a constant and M_i^* is modeled as $X_i'\gamma + e_i$ and assuming e_i, ε_i are bivariate normal, independent of Y_i and X_i. The results are all in Table 3.

The data set used here, and the choice of regressors Y_i, X_i, and V_i, is from Chen (2003), constructed primarily from the National Longitudinal Survey for Youth (NLSY). V_i is minus the log of the total expense of attending a local in state public college, deflated by the local average hourly wage of unskilled workers that prevailed when i was 17 years old. Alternative choices for V_i such as distance to schools as in Card (1995) could be used, but did not vary as much as this cost measure. X_i consists of a constant term, a scholastic ability index (constructed as a composite of test scores), dummy variable indicators for a parent that went to college, whether i is black, whether i is male, and whether i's cohort is from the 1980's or the 1990's. Y_i equals X_i plus additional dummies indicating one to five years of work experience and over five years of work experience. The total sample size is 7013 individuals, with 3775 of them having $D_i = 1$. See Chen (2003) and Chen and Khan (2003) for more details on the construction and use of this data set, and Kane and Rouse (1995) for related results on NLSY data.

The estimates from all the estimators are roughly comparable, which shows that the density weighted estimators are at least not generating wild estimates. A possible exception is the normal weighted OLS, which has a few implausible coefficient estimates, such as negative effect on over five years of work experience. Normality may not be a reasonable assumption for η_i.

One substantial difference across the estimates is that OLS gives a significant 5 percent increase in wages resulting from a parent having a college education,

while MLE gives an implausible negative 5 percent effect. The semiparametric kernel and sorted density estimates are near zero and completely insignificant statistically (unlike every other coefficient, the sign of this coefficient in the kernel estimator changes when a different bandwidth is used). Selection bias may cause the OLS estimate to be too high, because parent's education is a strong determinant of whether the child goes to college.

Another notable (though not statistically significant) difference is that all the semiparametric estimators say the increase from the 1980's to the 1990's in real wages from having a degree, after controlling for other covariates, is around 12 percent or more, while the MLE and OLS give gains of only 9 and 10 percent. The semiparametric estimates also have higher scholastic test score effects on wages than MLE (though not as high as OLS).

Most of the OLS estimates are not very different from the others, which suggests that in this application the effects of selection bias may not be very large. It may be the case that, with two sided censoring, the selection bias due to censoring from above partially offset the selection bias due to censoring from below.

Similar models could be estimated for other amounts of schooling. One caveat on interpreting these results is that only employed individuals are included in the data set, so the results are conditional on finding employment. The estimators in this paper could also be used to estimate the differences in probabilities of employment resulting from schooling, by defining P_i to be an indicator of employment and estimating a nonlinear or nonparametric discrete choice model, again controlling for selection by V density weighting.

7 Conclusions

Instead of weighting by a propensity score, this paper shows that selection can be addressed through weighting by the conditional density of one covariate V. Strong support and independence assumptions about V replace the usual strong assumptions about the joint distribution of unobservables affecting selection or treatment and outcomes. Essentially, this density weighting converts expectations of data censored by D into expectations of uncensored data. As a result, selection problems can be handled in conjunction with any estimator that is based on expectations. This paper focused on GMM type estimators, including least squares, instrumental variables, and maximum likelihood, but the method could also be used with other estimators based on expectations. For example, Theorem 1 and its corollaries can be extended to identify and estimate $E(U^* \mid X)$ (assuming

$A^* \perp X$), essentially by replacing the numerator and denominator of equation (5) with nonparametric regressions of UW on X and of W on X, respectively. Another example is estimation of panel models with fixed effects and selection. If $P_{it}^* = Y_{it}'\beta + \alpha_i + \varepsilon_{it}$ and $P_{it} = P_{it}^* D_i$ then β can be estimated by regressing $P_{it} - P_{it-1}$ on $Y_{it-1} - Y_{it}$ with weights W_i, thereby differencing out the fixed effects despite the selection problem.

The usefulness of these results in any application of course depends on whether an appropriate covariate V exists. This paper provided two empirical applications, one with A^* finite, and the other with the more common case of infinite A^*. It seems likely that, in at least some applications, one would be more comfortable making strong assumptions about a single observed covariate than the alternative, which requires strong assumptions regarding 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.

If more than one plausible candidate for V is present, they could in general be combined. For example, if $D = I(0 \leq M^* + b_1 V_1 + b_2 V_2)$, then we could let $V = \widehat{b}_1 V_1 + \widehat{b}_2 V_2$ using some consistent (up to scale) estimators for $\widehat{b}_1$ and $\widehat{b}_2$ such as Powell, Stock, and Stoker's (1989) weighted average derivatives. Alternatively, with GMM estimation we could write one set of moments for estimating θ using V_1 as V, and a second set of moments for estimating θ using V_2 as V, and then estimate a single GMM with both sets of moments simultaneously to efficiently combine the information in both sets (though in this case the relative supports of V_1 and V_2 are an issue).

Magnac and Maurin (2003) showed that, for the related binary choice estimator in Lewbel (2000), the large support assumption for V could be relaxed by adding an error tail symmetry assumption, and that the two assumptions (large support vs tail symmetry) are observationally equivalent. As discussed earlier, many semiparametric estimators require a regressor to have a large or infinite support, but it would still be desirable to search for alternatives that could relax the large support requirement in the present sample selection context.

8 Appendix: Proofs

PROOF OF LEMMA 1: Define $\pi(V, X) = pr(D = 1 \mid V, X)$. Let e have uniform distribution on [0, 1], independent of V, X. Define $M^* = \pi^{-1}(e, X)$ and $\widetilde{D} =$

$I(0 \leq M^* + V)$. Then

$$\begin{aligned} pr(\widetilde{D} &= 1 \mid V, X) = pr[\pi^{-1}(e, X)] \leq V \mid V, X] \\ &= pr[e \leq \pi(V, X) \mid V, X] = \pi(V, X) \end{aligned}$$

■

PROOF OF THEOREM 1. First consider the case where $E(A^*)$ is finite. Then

$$\begin{aligned} E(UW) &= E\left(\frac{DU^*}{f(V \mid X)}\right) \\ &= E\left[E\left(\frac{I(0 \leq M^* + V \leq A^*)U^*}{f(V \mid X)} \mid X, U^*, M^*, A^*\right)\right] \\ &= E\left[\int_{supp(V|X,U^*,M^*,A^*)} \frac{I(0 \leq M^* + v \leq A^*)U^*}{f(V \mid X)} f(V \mid X, U^*, M^*, A^*)dv\right] \\ &= E\left[\int_{supp(V|X)} I(-M^* \leq v \leq A^* - M^*)U^* dv\right] \\ &= E\left[\int_{-M^*}^{A^*-M^*} 1 dv U^*\right] = E[\left(A^* - M^* + M^*\right)U^*] \\ &= E\left(A^*U^*\right) = E\left(A^*\right)E(U^*) \end{aligned}$$

and by the same logic

$$E(W) = E\left(\frac{D}{f(V \mid X)}\right) = E(A^*)$$

so

$$p\lim \frac{\sum_{i=1}^n U_i W_i}{\sum_{i=1}^n W_i} = \frac{E(UW)}{E(W)} = \frac{E(A^*)E\left(U^*\right)}{E(A^*)} = E\left(U^*\right)$$

Now consider the case where A^* is infinity. In that case $E(UW)$ and $E(W)$ are both infinite. To deal with this complication, define

$$W_{\tau i} = \frac{I(V_i \leq \tau)W_i}{\tau} = \frac{I(V_i \leq \tau)D_i}{\tau f(V_i \mid X_i)}$$

$$\widehat{\mu}_{\tau 1} = \frac{1}{n}\sum_{i=1}^n U_i W_{\tau i}, \quad \widehat{\mu}_{\tau 2} = \frac{1}{n}\sum_{i=1}^n W_{\tau i}$$

where $\tau = \tau(n)$ is an asymptotic trimming parameter. Let $\tau \to \infty$ at a rate that makes $[\inf_{xsupp(X)} F(\tau \mid x)]^n \to 1$ where F is the cumulative distribution function of $V|X$. By the definition of $\widehat{\mu}_{\tau 1}$,

$$\begin{aligned}
\Pr\left(\widehat{\mu}_{\tau 1} = \frac{1}{n\tau}\sum_{i=1}^{n} U_i W_i\right) &= \Pr\left(\frac{1}{n\tau}\sum_{i=1}^{n} U_i W_i I(V_i > \tau_i) = 0\right) \\
&\leq \prod_{i=1}^{n} \Pr(V_i \leq \tau) \\
&\leq \prod_{i=1}^{n} \inf_{X,U^*,M^*,A^* \in supp(X,U^*,M^*,A^*)} F(\tau \mid X, U^*, M^*, A^*) \\
&\leq [\inf_{xsupp(X)} F(\tau \mid x)]^n \to 1
\end{aligned}$$

so $\widehat{\mu}_{\tau 1} - (n\tau)^{-1}\sum_{i=1}^{n} U_i W_i \to 0$ with probability one. The same logic replacing U_i with one shows that $\widehat{\mu}_{\tau 2} - (n\tau)^{-1}\sum_{i=1}^{n} W_i \to 0$ with probability one, and therefore $(\widehat{\mu}_{\tau 1}/\widehat{\mu}_{\tau 2}) - \widehat{\mu} \to 0$ with probability one, given the fast rate that $\tau \to \infty$. It follows that

$$p\lim \widehat{\mu} = p\lim \frac{\widehat{\mu}_{\tau 1}}{\widehat{\mu}_{\tau 2}} = \lim_{\tau \to \infty} \frac{E(\widehat{\mu}_{\tau 1})}{E(\widehat{\mu}_{\tau 2})}$$

assuming $E(\widehat{\mu}_{\tau 1})$ and $E(\widehat{\mu}_{\tau 2})$ are finite for any given sufficiently large τ . Now

$$\begin{aligned}
E(\widehat{\mu}_{\tau 1}) &= E\left(\frac{I(V \leq \tau) D U^*}{\tau f(V \mid X)}\right) \\
&= E\left[E\left(\frac{I(0 \leq M^* + V) I(V \leq \tau) U^*}{\tau f(V \mid X)} \mid X, U^*, M^*\right)\right] \\
&= E\left[\int_{supp(V|X,U^*,M^*)} \frac{I(-M^* \leq V \leq \tau) U^*}{\tau f(V \mid X)} f(V \mid X, U^*, M^*) dv\right] \\
&= E\left[\int_{supp(V|X)} \frac{I(-M^* \leq V \leq \tau) U^*}{\tau} dv\right] \\
&= E\left[\int_{-M^*}^{\tau} \frac{1}{\tau} dv U^*\right] \\
&= E\left(U^*\right) + \frac{E\left(M^* U^*\right)}{\tau}
\end{aligned}$$

and by the same logic

$$E(\widehat{\mu}_{\tau 2}) = 1 + \frac{E\left(M^*\right)}{\tau}$$

so $\lim_{\tau\to\infty} E(\widehat{\mu}_{\tau 1})/E(\widehat{\mu}_{\tau 2}) = E\,(U^*)$.

The remaining case to consider is where A^* is random and has infinite mean. This can include the case where A^* has a positive but less than one probability of being infinite, corresponding to the case where some fraction of the population has selection that is not bounded from above. For this case, consider the more complicated weighting function $W^*_{\tau i}$ defined by

$$W^*_{\tau i} = \frac{I(V_i \leq \tau)W_i}{(1 - I^*_i)\tau + I^*_i A^*_i}$$

where I^*_i equals one if A^*_i is finite and zero otherwise, and assume $\tau \to \infty$ sufficiently fast so that $prob(\tau < A^*_i - M^*_i)$ is zero when $I^*_i = 1$. Following the same logic as before

$$E\left(\frac{I(V \leq \tau)W}{(1 - I^*)\tau + I^*A^*}\right) = E\left(\frac{I(V \leq \tau)DU^*}{[(1 - I^*)\tau + I^*A^*]\,f(V \mid X)}\right)$$

$$\begin{aligned}
&= E\left[\int_{supp(V|X,U^*,M^*,A^*)} \frac{I(-M^* \leq V \leq \min(\tau,\, A^* - M^*))U^*}{[(1 - I^*)\tau + I^*A^*]\,f(V \mid X)} f(V \mid X, U^*, M^*, A^*)dv\right] \\
&= E\left[\int_{-M^*}^{\min(\tau, A^* - M^*)} \frac{1}{(1 - I^*)\tau + I^*A^*} dv U^*\right] \\
&= E\left[I^*\int_{-M^*}^{A^* - M^*} \frac{1}{A^*} dv U^* + (1 - I^*)\int_{-M^*}^{\tau} \frac{1}{\tau} dv U^*\right] \\
&= E\left[I^*U^* + (1 - I^*)U^* + \frac{(1 - I^*)M^*U^*}{\tau}\right] \\
&= E\left(U^*\right) + \frac{E\left[(1 - I^*)M^*U^*\right]}{\tau}
\end{aligned}$$

and the remainder of the proof also follows as before. ■

The proof of Corollary 1 is omitted to save space, since it follows the same logic as the proof of Theorem 1 in the case where A^* is infinite, with a fixed instead of asymptotic τ.

PROOF OF COROLLARY 2. Let $\overline{W}$ and $\overline{WU}$ denote the sample means of W_i and W_iU_i, respectively, and let $c = \sup(supp(W))$. Then $E((WU)^2) =$

$E((WU^*)^2) \leq c^2 E\,(U^*)^2$, and similarly $E((W)^2) \leq c^2$ so W_i and $W_i U_i$ have finite second moments. Assumption A1 also implies $E(W) > 0$. Corollary 2 then follows from applying the Lindeberg-Levy central limit theorem to $(\overline{WU}, \overline{W})$, and the delta method. ■

PROOF OF THEOREM 2. θ_0 is equivalently given by

$$\theta_0 = \arg\min_{\theta\in\Theta} \left(\frac{E[W\psi(P, X, V, \theta)]}{E(W)}\right)' \Omega \left(\frac{E[W\psi(P, X, V, \theta)]}{E(W)}\right)$$

The first order condition for θ_0 and the mean value theorem give

$$\begin{aligned} 0 &= \nu(\theta_0)'\Omega\frac{E[W\psi(P, X, V, \theta_0)]}{E(W)} \\ &= \nu(\theta_0)'\Omega\left(\frac{E[W\psi(P, X, V, \theta^*)]}{E(W)} + \nu(\widetilde{\theta})(\theta_0 - \theta^*)\right) \\ &= \nu(\theta_0)'\Omega\left[\mu + \nu(\widetilde{\theta})(\theta_0 - \theta^*)\right] \end{aligned}$$

where $\widetilde{\theta}$ lies between θ^* and θ_0. Solving for $\theta_0 - \theta^*$ gives

$$\theta_0 - \theta^* = -\left[\nu(\theta_0)'\Omega\nu(\widetilde{\theta})\right]^{-1} \nu(\theta_0)'\Omega\mu$$

Now $E(U^*) = 0$, so if $\mu = E(U^*)$ then $\theta_0 - \theta^* = 0$, while if $\mu = E(U^*) + O(\tau^{-1})$ then

$$\theta_0 - \theta^* = -\left[\nu(\theta_0)'\Omega\nu(\widetilde{\theta})\right]^{-1} \nu(\theta_0)'\Omega O(\tau^{-1}) = O(\tau^{-1})$$

■

PROOF OF COROLLARY 3. This is standard GMM limiting distribution theory with iid data. See, e.g., Newey and McFadden (1984) or Wooldridge (2001) Theorems 14.1 and 14.2.

References

[1] ABADIE, A., (2001), "Semiparametric Instrumental Variable Estimation of Treatment Response Models," Unpublished Manuscript, Harvard University.

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

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

[4] ANTON, A. A., S. A. FERNÁNDEZ, AND J. RODRIGUEZ-PÓO (2001), "Semiparametric Estimation of a Duration Model," *Oxford Bulletin of Economics & Statistics,* 63, 517-533.

[5] 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.

[6] ANGRIST, J. AND G. IMBENS, (1995), "Two-Stage Least Squares Estimation of Average Causal Effects in Models With Variable Treatment Intensity," *Journal of the American Statistical Association* 90, 430–442.

[7] 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.

[8] BLUNDELL, R. AND J. L. POWELL (2004), "Endogeneity in Semiparametric Binary Response Models," *Review of Economic Studies*, 71, 655-679.

[9] CAMERON, S. V., AND J. J. HECKMAN. (1998), "Life Cycle Schooling and Dynamic Selection Bias: Models and Evidence for Five Cohorts of American Males," *Journal of Political economy* 106, 262-333.

[10] CARNEIRO, P., K. T. HANSEN, AND J. J. HECKMAN (2003), "Estimating Distributions of Treatment Effects With an Application to the Returns to Schooling," *International Economic Review,* 44, 361-422.

[11] CAVANAGH, C. AND R. P. SHERMAN (1998), "Rank Estimators for Monotonic Index Models", *Journal of Econometrics*, 84, 351-381

[12] CHEN, S. H., (2003), "Estimating Wage Volatilities for College versus High School Careers," SUNY Albany unpublished manuscript.

[13] CHEN, S. H., AND S. KHAN, (2003), "Nonparametric Estimation of Volatility Differentials in Selection Models with an Application to Returns to Schooling," SUNY Albany unpublished manuscript.

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

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

[16] COGNEAU, D. AND E. MAURIN, (2001), Parental Income and School Attendence in a Low-Income Country: a Semi-parametric Analysis," Unpublished Manuscript.

[17] 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.

[18] DAS, M., W. K. NEWEY, AND F. VELLA (2003), "Nonparametric estimation of sample selection models," *Review of Economic Studies,* 70, 33-58.

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

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

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

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

[23] HAN, A. K. (1987) "Non-parametric Analysis of a Generalized Regression Model," *Journal of Econometrics*, 35, 303–31.

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

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

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

[27] 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

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

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

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

[31] HECKMAN, J. AND E. VYTLACIL (2004), "Structural Equations, Treatment Effects, and Econometric Policy Evaluation," unpublished manuscript.

[32] HIRANO, K., G. W. IMBENS AND G. RIDDER, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica* 71, 1161-1189.

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

[34] HORVITZ, D.G. AND D. J. THOMPSON, (1952), "A generalization of sampling without replacement from a finite universe" *Journal of the American Statistical Association*, 47, 663-685.

[35] 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.

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

[37] JACHO-CHAVEZ, D. T. (2005), "Efficiency Bounds in Semiparametric Models defined by Moment Restrictions using an Estimated Conditional Probability Density, " unpublished manuscript.

[38] KANE, T. J. AND C. E. ROUSE (1995), "Labor-Market Returns to Two and Four-Year College," *American Economic Review* 85, 600-614.

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

[40] KHAN, S. AND A. LEWBEL (2005), "Weighted and Two Stage Least Squares Estimation of Semiparametric Truncated Regression Models," unpublished manuscript.

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

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

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

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

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

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

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

[48] LEWBEL, A., AND S. SCHENNACH (2005), "A Simple Ordered Data Estimator for Inverse Density Weighted Expectations," *Journal of Econometrics*, forthcoming

[49] MAGNAC, T. AND E. MAURIN. (2003), "Identification and Information in Monotone Binary Models," unpublished manuscript.

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

[51] MANSKI, C. F. (1985), "Semiparametric Analysis of Discrete Response: Asymptotic Properties of Maximum Score Estimation", *Journal of Econometrics*, 27, 313-334.

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

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

[54] 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.

[55] NEWEY, W. K. (1984), "A Method of Moments Interpretation of Sequential Estimators," *Economics Letters*, 14, 201-206.

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

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

[58] 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.

[59] NILSEN, O. A. AND F. SCHIANTARELLI (2003), ”Zeros And Lumps In Investment: Empirical Evidence On Irreversibilities And Nonconvexities," *The Review of Economics and Statistics*, 85, 1021-1037.

[60] 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.

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

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

[63] ROSENBAUM, P. AND D. RUBIN, (1985), “Reducing Bias in Observational Studies Using Subclassification on the Propensity Score,” *Journal of the American Statistical Association*, 79, 516-524.

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

[65] TCHAMOURLIYSKI, Y. (2002), ”Returns to Experience and Seniority” Evidence From a Panel Selection Model With Endogenous Regressors and No Exclusion Restriction,” Boston College Unpublished Manuscript.

[66] VELLA, F. (1998), ”Estimating Models With Sample Selection Bias: A Survey,” *Journal of Human Resources*, 33, 127-169.

[67] VYTLACIL, E. (2002), "Independence, Monotonicity, and Latent Index Models: An Equivalence Result,” *Econometrica*, 70, 331-341.

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

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

[70] WOOLDRIDGE, J. M. (2001), Econometric Analysis of Cross Section and Panel Data, Cambridge: MIT Press.

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
Endogenous ML	.468	.061	.403	.062	.413	.057
Weighted OLS	.323	.062	.317	.059	.316	.051
Weighted TSLS	.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.

Table 2. Simulations of the Outcome Equation Profit Coefficient

	MEAN	SD	LQ	MED	UQ	RMSE	MAE	MDAE	MESE	%2SE
OLS	.231	.017	.221	.232	.245	.183	.182	.182	.036	.000
TSLS	.362	.023	.346	.362	.377	.056	.052	.051	.051	.987
Heckman ML	.268	.020	.254	.268	.282	.146	.145	.145	.102	.997
Endogenous ML	.414	.028	.394	.413	.433	.028	.023	.020	.145	1.00
Weighted OLS	.243	.043	.213	.243	.273	.177	.171	.170	.049	.060
Weighted TSLS	.399	.061	.358	.397	.440	.063	.050	.043	.066	.961

Notes: In these simulations the 'true' value of the coefficient is .4132. The reported statistics are as follows. MEAN and SD are the mean and standard deviation of the estimates across the simulations. LQ, MED, and UQ are the 25% (lower) 50% (median) and 75% (upper) quartiles. RMSE, MAE, and MDAE are the root mean squared error, mean absolute error and median absolute error of the estimates. MESE is the mean estimated standard error, and %2SE is the percentage of simulations in which the true coefficient was within two estimated standard errors of the estimated coefficient.

Table 3. Estimates of Wage Outcome Equations

	OLS		MLE		Kernel Weighted OLS	
Constant	1.37	.053	1.69	.111	1.40	.371
Test Score	.120	.009	.059	.020	.109	.061
Parent College	.048	.016	−.048	.034	−.011	.122
Black	−.017	.027	−.087	.035	−.047	.140
Male	.249	.015	.266	.017	.225	.113
Urban	.161	.020	.132	.023	.167	.104
90's cohort	.097	.018	.086	.019	.122	.117
1 to 5 years work.	.001	.038	−.000	.038	.027	.273
Over 5 years	.342	.051	.337	.051	.350	.306

	Normal Weighted GMM		Sorted Weighting OLS	
Constant	1.64	.166	1.48	.204
Test Score	.164	.028	.133	.031
Parent College	.115	.059	.012	.103
Black	−.152	.077	−.135	.152
Male	−.071	.055	.151	.108
Urban	.023	.064	.047	.088
90's cohort	.215	.079	.115	.197
1 to 5 years work.	−.064	.119	−.004	.149
Over 5 years	−.141	.181	.294	.329

Notes: In each block, the first number is the coefficient, and the second number is the estimated standard error. OLS is the wage equation only using data on college graduates with no graduate education, and so does not control for any selection bias. ML is a parametric two equation system of selection and wages with normal errors. Weighted OLS is the density weighted semiparametric estimator of the wage equation, using a kernel estimator of the conditional density of V given X. Normal Weighted GMM models V as linear in X with an independent normal error, estimating the V and density weighted wage equations simultaneously by GMM. Sorted Weighting OLS models V as linear in X with an independent error of unknown density that is estimated using the numerically trivial sorted data estimator.