

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 Cheshire, 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 models.

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 bootstrapping 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 $D = I(X'\hat{\gamma} + \varepsilon \geq 0)$ for selection by maximizing a likelihood function $L(D, X, \gamma)$, yielding estimated parameters $\hat{\gamma}$. Then in a second step parameters $\hat{\beta}, \hat{\lambda}$ are estimated by applying ordinary least squares to the linear regression model $Y = X'\beta + m(X'\hat{\gamma})\lambda + e$, where $m(X'\hat{\gamma})$ 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 $\hat{\gamma}$ 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 $\hat{\beta}$ are obtained from applying MM or GMM to $E[g_2(Z, \beta, \gamma) | \gamma = \hat{\gamma}] = 0$ for some m_2 vector of known functions g_2 .

For example, in the Heckit procedure $\hat{\gamma}$ is obtained by maximizing the probit likelihood function $L(Z, \gamma)$, which is equivalent to letting $\hat{\gamma}$ 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 $\hat{\beta}$ (and $\hat{\lambda}$) are the solutions to $n^{-1} \sum_{i=1}^n [Y_i - X_i' \beta - \lambda m(X_i' \hat{\gamma})] (X_i', m(X_i' \hat{\gamma}))' = 0$, which defines the function $g_2(Z_i, \beta, \gamma)$ such that $n^{-1} \sum_{i=1}^n g_2(Z_i, \beta, \hat{\gamma}) = 0$ (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 $g_2(Z, \beta, \gamma)$. 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

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

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

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

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

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

where D is an observed dummy variable that equals zero or one, \tilde{X} is a vector of observed regressors, $\tilde{\beta}$ is a vector of coefficients to be estimated, $\tilde{\varepsilon}$ 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 $\tilde{\varepsilon} \sim N(0, 1)$, while for logit $\tilde{\varepsilon}$ has a logistic distribution.

Let V be some conveniently chosen exogenous element of \tilde{X} that is known to have a positive coefficient, and let X be the vector of remaining elements of \tilde{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 $\tilde{\beta} = (\beta, 1)/\sigma_\varepsilon$ (where V has been taken to be the last element of \tilde{X}). One can easily rewrite probit or logit code to report estimates of β and σ_ε instead of $\tilde{\beta}$. 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

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

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 $\Pr(D = 1) = 1 - F_{\tilde{\varepsilon}}[-(\tilde{X}'\tilde{\beta})]$. For normal errors the CDF's are $F_{\tilde{\varepsilon}}(\cdot) = \Phi(\cdot)$ 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 $E[\partial E(D | V, X)/\partial V]$. 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 $\tilde{\beta}$, and the relative magnitudes of the coefficients are also the same, that is $\tilde{\beta}$ is proportional to $(\beta, 1)$, so the estimated marginal rates of substitution between any two elements of \tilde{X} is the same using either $(\beta, 1)$ or $\tilde{\beta}$. In applications where the distribution of the latent error is unknown but independent of \tilde{X} , this distribution, and resulting estimated choice probabilities may be equivalently estimated by a nonparametric regression of D on either $X'\beta + V$ or on $\tilde{X}'\tilde{\beta}$.

In general \tilde{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 $\tilde{\varepsilon}$ 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 semiparametric 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 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

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

where $F_\varepsilon(\varepsilon | \cdot)$ 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 $F_\varepsilon(\varepsilon | Y, W) = F_\varepsilon(\varepsilon | U)$ 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 $-(X'\hat{\beta} + V)$ and on \hat{U} , or by a parametric model if $F_\varepsilon(\varepsilon | U)$ 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

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

We might call this the index choice probability, since this will equal $E(D | Y, W)$ 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 $X'\hat{\beta} + V$, given any consistent estimator $\hat{\beta}$.

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

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

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

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

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

ment conditions

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

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

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

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. $V = g(v, S)$ for some function g that is differentiable and strictly monotonically increasing in its first element, with $v \perp S, \varepsilon$, and v 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 v 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 v 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

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

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

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

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

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

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 $D = I[-(X'\beta + \varepsilon) \leq g(v, S)]$ so $E(D \mid S, v)$ equals the probability distribution function of $-(X'\beta + \varepsilon)$, conditioned on S , and evaluated at $V = g(v, S)$. 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 v 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

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

so V is linear in covariates plus a normal error. By Theorem 1, other regular parametric model for g and continuous distributions for v 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 \hat{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 v_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 $f(\widehat{v}_i)$ 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

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

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. One 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

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

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 heteroskedasticity has a simple multiplicative form. Again, by Theorem 1 other more general parametric models for V and continuous distributions for v 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 \hat{b} be the estimated coefficients of an ordinary least squares regression of V on S . For each observation i , construct data $\hat{\eta}_i = V_i - S_i'\hat{b}$, which are the residuals of this regression.

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

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

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

and construct data \hat{T}_i by

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

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

The only nontrivial step now is step 2, and for that step good approximate starting values for \hat{c} may be obtained by linearly regressing $\ln(\hat{\eta}^2)$ 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 $S_i'\hat{c}$ 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

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

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 be generalized as above to allow for conditionally heteroskedastic 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

$$\begin{aligned} D &= I(X'\beta + V_1 + V_2\alpha + \varepsilon \geq 0), \quad E(Z\varepsilon) = 0 \\ V_1 &= S'b_1 + V_2c_1 + v_1, \quad v_1 \perp v_2, S, \varepsilon, \quad v_1 \sim N(0, \sigma_{v_1}^2) \\ V_2 &= S'b_2 + V_1c_2 + v_2, \quad v_2 \perp v_1, S, \varepsilon, \quad v_2 \sim N(0, \sigma_{v_2}^2) \end{aligned}$$

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

$$\begin{aligned} E[S(V_1 - S'b_1 - V_2c_1)] &= 0 \\ E[\sigma_{v_1}^2 - (V_1 - S'b_1 - V_2c_1)^2] &= 0 \end{aligned}$$

$$E \left[\begin{pmatrix} Z \\ V_2 \end{pmatrix} \left([D - I(V_1 \geq 0)] \exp \left(\frac{(V_1 - S'b_1 - V_2 c_1)^2}{2\sigma_{v_1}^2} \right) - X'\beta - \alpha V_2 \right) \right] = 0$$

$$\begin{aligned} E[S(V_2 - S'b_2 - V_1 c_2)] &= 0 \\ E[\sigma_{v_2}^2 - (V_2 - S'b_2 - V_1 c_2)^2] &= 0 \end{aligned}$$

$$E \left[\begin{pmatrix} Z \\ V_1 \end{pmatrix} \left([D - I(V_2 \geq 0)] \exp \left(\frac{(V_2 - S'b_2 - V_1 c_2)^2}{2\sigma_{v_2}^2} \right) - \frac{X'\beta + V_1}{\alpha} \right) \right] = 0$$

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 $\hat{\beta}$.

5.2.3 Generalized Very Exogenous Regressor Estimation

The binomial response model with a very exogenous regressors generalizes to

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

Here the distribution of v 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 v 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 v_i , sort these observations from lowest to highest. For each observation v_i , let v_i^+ be the value of v that, in the sorted data, comes immediately after v_i and similarly let v_i^- be the value that comes immediately before v_i . Then

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

where the last step replaces the true distribution function F with the empirical distribution function. This suggests the estimator $\hat{f}(v_i) = (v_i^+ - v_i^-)n/2$. 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 $\hat{v}_i = V_i - S_i'\hat{b}$ as the residuals of an ordinary least squares regression of V on S .

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

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

3. Let $\hat{\beta}$ be the estimated coefficients of an ordinary linear two stage least squares regression of \hat{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 \hat{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

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

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

A further generalization is the model

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

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

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

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 positive 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 \tilde{V}_i to be exact age minus six of student i on December 31 of the year they start school. Then \tilde{V}_i is a value from $-2/3$ to $1/3$, and $V_i = I(\tilde{V}_i \geq 0)$.

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 determinants 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 endogenous, 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 \tilde{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 \tilde{V}_i is then $f_{\tilde{V}}(\tilde{V}_i) = 12F_i$ 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, \tilde{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 $k^* = 0, \dots, K$ is

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

where $\alpha_0 = -\infty$ and $\alpha_1, \dots, \alpha_K$ 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 $k = 1, \dots, K$ 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 $\alpha_1, \dots, \alpha_K$. 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

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

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

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

$$\begin{aligned}
E [W(Y - W'b)] &= 0 \\
E [S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)] &= 0 \\
E [S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)X] &= 0 \\
E [S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)(Y - W'b)] &= 0 \\
E [S(D_1, X, V, Y - W'b, \alpha_1, \beta, \gamma, \sigma_\eta)V] &= 0 \\
\\
E [S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)] &= 0 \\
E [S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)X] &= 0 \\
E [S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)(Y - W'b)] &= 0 \\
E [S(D_2, X, V, Y - W'b, \alpha_2, \beta, \gamma, \sigma_\eta)V] &= 0
\end{aligned}$$

where the function S is defined as

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

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

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

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 v . 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 school 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 $P = (X'\beta + V\gamma + \varepsilon)D$, $E(Z\varepsilon) = 0$, $I(a_0 \leq M(S, e) \pm V \leq a_1)$, $V = S'b + v$, $v \perp S$, ε , e , and v is continuously distributed with support equal to the real line. Assume a_0 and a_1 are finite. Let $f(v)$ be the probability density function of v . 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 \hat{b} be the estimated coefficients of an ordinary least squares regression of V on S . For each observation i , construct data $\hat{v}_i = V_i - S_i' \hat{b}$, which are the residuals of this regression.

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

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

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

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

ESTIMATOR F

Use GMM to estimate the parameters $\beta, \gamma, b, \sigma_v^2$ based on the moment conditions

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

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

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 $D/f(v)$ 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 $i = 1, \dots, n$ and $t = 1, \dots, T$. The asymptotics assume T fixed and $n \rightarrow \infty$. The model is

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

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

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

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 v_{it} , is fixed over time then we get the simplification

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

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 $\widehat{v}_{it} = V_{it} - S'_{it}\widehat{b}_t$, which are the residuals of this regression.

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

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

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 $\beta, b_1, \dots, b_T, \sigma_1^2, \dots, \sigma_T^2$ based on the moment conditions

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

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

For $t = 1, \dots, T$, 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 $t = 1, \dots, T$, and the asymptotics assume stationarity and $T \rightarrow \infty$. The model is

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

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 v_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 v_t over time, since v_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 v_{t-1} . To satisfy this restriction it may be assumed that v_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

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

where the second equality follows from $v \perp S, \varepsilon$, and the third from the change of variables from v 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, $E(T_{it} | S_{it}, \alpha_i + \varepsilon_{it}) = X'_{it}\beta + \alpha_i + \varepsilon_{it}$, so by the law of iterated expectations $E(Z_{it}T_{it-k}) = E[Z_{it}(X'_{it-k}\beta + \alpha_i + \varepsilon_{it-k})]$ for $k = 0, 1$, and therefore $E[Z_{it}(T_{it} - T_{it-1})] = E[Z_{it}((X'_{it}\beta + \alpha_i + \varepsilon_{it}) - (X'_{it-1}\beta + \alpha_i + \varepsilon_{it-1}))]$.