Binary Models with Endogenous Explanatory Variables

Class Notes

Manuel Arellano

November 7, 2007

Revised: January 21, 2008

1 Introduction

In Part I we considered linear and non-linear models with additive errors and endogenous explanatory variables. A simple case was a linear relationship between Y and X and an error U, where U was potentially correlated with X but not with an instrument Z:

$$Y = \alpha + \beta X + U, \qquad E(U) = 0, E(ZU) = 0. \tag{1}$$

This setting was motivated in structural models.

Now we wish to make similar considerations for binary index models of the form

$$Y = \mathbf{1}(\alpha + \beta X + U \geq 0).$$

There are two important differences that we need to examine:

1. In the new models effects are heterogeneous, so that there is a difference between effects at the individual level and aggregate or average effects.

2. Instrumental variable techniques are no longer directly applicable because the model is not invertible and we lack an expression for U. So we have to consider alternative ways of addressing endogeneity concerns.

To discuss these issues it is useful first to go back to the linear setting and re-examine endogeneity using an explicit notation for potential outcomes.

Potential outcomes notation When we are interested in (1) as a structural equation, we regard it as a conjectural relationship that produces potential outcomes for every possible value $x \in \mathcal{S}$ of the right-hand-side variable:

$$Y(x) = \alpha + \beta x + U$$

So we imagine that each unit in the population has a value of U and hence a value of $Y(x)$ for each value of x. This is the way we think about structural models in economics, for example about a demand schedule that gives the conjectural demand $Y(x)$ for every possible price x.

For each unit we only observe the actual value X that occurs in the distribution of the data, so that $Y = Y(X)$.[1]

If the assignment of values of X to units in the population is such that X and U are uncorrelated, β coincides with the regression coefficient of Y on X. If the assignment of values of Z (a predictor of X) to units in the population is such that Z and U are uncorrelated, β coincides with the IV coefficient of Y on X using Z as instrument.

Consider two individuals in the population with errors U and $U^\dagger$. Their potential outcomes will differ:

$$\begin{aligned} Y(x) &= \alpha + \beta x + U \\ Y(x)^\dagger &= \alpha + \beta x + U^\dagger \end{aligned}$$

but the effect of a change from x to x' will be the same for all individuals:

$$Y\left(x'\right) - Y(x) = \beta\left(x' - x\right).$$

In this sense we say that in models with additive errors the effects are homogeneous across units. We now turn to consider the situation in binary models.

Heterogeneous individual effects and aggregate effects Potential outcomes in the binary model are given by

$$Y(x) = \mathbf{1}(\alpha + \beta x + U \geq 0).$$

The effect of a change from x to x' for an individual with error U is:

$$Y\left(x'\right) - Y(x) = \mathbf{1}\left(\alpha + \beta x' + U \geq 0\right) - \mathbf{1}(\alpha + \beta x + U \geq 0).$$

Suppose for the sake of the argument that $\beta > 0$ and $x' > x$. The possibilities are

value of U	$Y(x)$	$Y(x')$	$Y(x') - Y(x)$
$-U \leq \alpha + \beta x$	1	1	0
$-U > \alpha + \beta x'$	0	0	0
$\alpha + \beta x < -U \leq \alpha + \beta x'$	0	1	1

Depending on the value of U the effects can be zero or unity, therefore they are heterogeneous across units.

[1] Given the form of the model all the $Y(x)$ are observable when β is known:

$$Y(x) = \alpha + \beta x + (Y - \alpha - \beta X) = Y + \beta(x - X).$$

In these circumstances it is natural to consider an average effect:

$$\begin{aligned} E_U\left[Y\left(x'\right) - Y\left(x\right)\right] &= E_U\left[\mathbf{1}\left(\alpha + \beta x' + U \geq 0\right)\right] - E_U\left[\mathbf{1}\left(\alpha + \beta x + U \geq 0\right)\right] \\ &= \Pr\left(-U \leq \alpha + \beta x'\right) - \Pr\left(-U \leq \alpha + \beta x\right) = \Pr\left(\alpha + \beta x < -U \leq \alpha + \beta x'\right) \\ &= F\left(\alpha + \beta x'\right) - F\left(\alpha + \beta x\right) \end{aligned}$$

where F is the *cdf* of U. The average effect is simply the fraction of units in the population whose outcomes are affected by the change from x to x' (those with $\alpha + \beta x < -U \leq \alpha + \beta x'$).

Marginal effects If X is binary there is only one effect to consider. If X is continuous we can consider average marginal effects:

$$\frac{\partial E_U\left[Y\left(x\right)\right]}{\partial x} = \frac{\partial F\left(\alpha + \beta x\right)}{\partial x} = \beta f\left(\alpha + \beta x\right).$$

Marginal effects can be regarded as a random variable associated with X. In this sense it may be of interest to obtain summary measures of its distribution, like the mean or the median. For example,

$$E_X\left[\beta f\left(\alpha + \beta X\right)\right].$$

Identification and estimation In models with additive errors, moment conditions of the form $E\left(ZU\right) = 0$ are often sufficient for identification, and GMM estimates can be easily constructed from their sample counterparts. In non-invertible models, GMM estimators are not directly available. In fact, the availability of instruments (a variable Z that is independent of U) by itself does not guarantee point identification in general.

Next, we consider two specific models that fully specify the joint distribution of Y and X given Z. One is for a normally distributed X. It would have application to situations where X is a continuous variable. The other is for a binary X and leads to a multivariate probit model.

2 The normal endogenous explanatory variable probit model

The model is

$$\begin{aligned} Y &= \mathbf{1}\left(\alpha + \beta X + U \geq 0\right) \\ X &= \pi' Z + \sigma_v V \\ \begin{pmatrix} U \\ V \end{pmatrix} &\mid Z \sim \mathcal{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]. \end{aligned}$$

In this model X is an endogenous explanatory variable as long as $\rho \neq 0$. X is exogenous if $\rho = 0$.

Joint normality of U and V implies that the conditional distribution of U given V is also normal as follows:

$$U \mid V, Z \sim \mathcal{N}\left(\rho V, 1 - \rho^2\right)$$

or

$$\Pr\left(U \leq r \mid V, Z\right) = \Phi\left(\frac{r - \rho V}{\sqrt{1-\rho^2}}\right).$$

Therefore,

$$\Pr\left(Y = 1 \mid X, Z\right) = \Pr\left(\alpha + \beta X + U \geq 0 \mid V, Z\right) = \Phi\left(\frac{\alpha + \beta X + \rho V}{\sqrt{1-\rho^2}}\right).$$

Moreover, the density of $X \mid Z$ is just the normal linear regression density.

Thus, the joint probability distribution of Y and X given $Z = z$ is

$$f\left(y, x \mid z\right) = f\left(y \mid x, z\right) f\left(x \mid z\right)$$

or

$$\ln f\left(y, x \mid z\right) \propto y \ln \Phi\left(\frac{\alpha + \beta x + \rho v}{\sqrt{1-\rho^2}}\right) + (1-y) \ln\left[1 - \Phi\left(\frac{\alpha + \beta x + \rho v}{\sqrt{1-\rho^2}}\right)\right] - \frac{1}{2}\ln\sigma_v^2 - \frac{1}{2}v^2$$

where $v = \left(x - \pi' z\right)/\sigma_v$.

Therefore, the log likelihood of a random sample of N observations conditioned on the z variables is:

$$\begin{aligned} L\left(\alpha, \beta, \rho, \pi, \sigma_v^2\right) &= \sum_{i=1}^{N}\left\{y_i \ln \Phi\left(\frac{\alpha + \beta x_i + \rho v_i}{\sqrt{1-\rho^2}}\right) + (1-y_i)\ln\left[1 - \Phi\left(\frac{\alpha + \beta x_i + \rho v_i}{\sqrt{1-\rho^2}}\right)\right]\right\} \\ &\quad + \sum_{i=1}^{N}\left(-\frac{1}{2}\ln\sigma_v^2 - \frac{1}{2}v_i^2\right). \end{aligned}$$

Note that under exogeneity ($\rho = 0$) this log likelihood function boils down to the sum of the ordinary probit and normal OLS log-likelihood functions:

$$L\left(\alpha, \beta, 0, \pi, \sigma_v^2\right) = L_{probit}\left(\alpha, \beta\right) + L_{OLS}\left(\pi, \sigma_v^2\right).$$

3 The control function approach

3.1 Two-step estimation of the normal model

We can consider a two-step method:

- Step 1: Obtain OLS estimates $(\widehat{\pi}, \widehat{\sigma}_v)$ of the first stage equation and form the standardized residuals $\widehat{v}_i = \left(x_i - \widehat{\pi}' z_i\right) / \widehat{\sigma}_v$, $i = 1, ..., N$.
- Step 2: Do an ordinary probit of y on constant, x, and $\widehat{v}$ to obtain consistent estimates of $\left(\alpha^\dagger, \beta^\dagger, \rho^\dagger\right)$ where

$$\left(\alpha^\dagger, \beta^\dagger, \rho^\dagger\right) = \left(1 - \rho^2\right)^{-1/2} (\alpha, \beta, \rho).$$

Since there is a one-to-one mapping between the two, the original parameters can be recovered undoing the reparameterization. However, the fitted probabilities $\Phi\left(\widehat{\alpha}^\dagger + \widehat{\beta}^\dagger x_i + \widehat{\rho}^\dagger \widehat{v}_i\right)$ are in fact directly useful to get average derivative effects (see below).

In general, two-step estimators are asymptotically inefficient relative to maximum likelihood estimation, but they may be computationally convenient.

Ordinary probit standard errors calculated from the second step are inconsistent because estimated residuals are treated as if they were observations of the true first-stage errors. To get consistent standard errors, we need to take into account the additional uncertainty that results from using $(\widehat{\pi}, \widehat{\sigma}_v)$ as opposed to the truth (see the appendix).

Comparison with probit using fitted values Note that

$$Y = \mathbf{1}(\alpha + \beta X + U \geq 0) = \mathbf{1}\left(\alpha + \beta\left(\pi' Z\right) + \varepsilon \geq 0\right)$$

where $\varepsilon = U + \beta \sigma_v V$ is $\varepsilon \mid Z \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ with $\sigma_\varepsilon^2 = 1 + \beta^2 \sigma_v^2 + 2\beta \sigma_v \rho$.

If we run a probit of y on constant and $\widehat{x} = \widehat{\pi}' z$ we get consistent estimates of $\overline{\alpha} = \alpha / \sigma_\varepsilon$ and $\overline{\beta} = \beta / \sigma_\varepsilon$. Note that from estimates of $\overline{\alpha}$, $\overline{\beta}$, and σ_v we cannot back up estimates of α and β due to not knowing ρ. We cannot get average derivative effects either. So estimation of parameters of interest from this method (other than relative effects) is problematic.

3.2 The linear case: 2SLS as a control function estimator

The 2SLS estimator for the linear IV model is $\widehat{\theta} = \left(\widehat{X}' \widehat{X}\right)^{-1} \widehat{X}' y$ where $\widehat{X} = Z \widehat{\Pi}'$ and $\widehat{\Pi} = X' Z \left(Z' Z\right)^{-1}$. The matrix of first-stage residuals is $\widehat{V} = X - Z \widehat{\Pi}'$. Typically, X and Z will have some columns in common. For those variables, the columns of $\widehat{V}$ will be identically zero. Let us call $\widehat{V}_1$ the subset of non-zero columns of $\widehat{V}$ (those corresponding to endogenous explanatory variables).

It can be shown that 2SLS coincides with the estimated θ in the OLS regression $y = X\theta + \widehat{V}_1\gamma + \xi$ (see appendix). Therefore, linear 2SLS can be regarded as a control function method.

In the binary situation we obtained a similar estimator from a probit regression of y on X and first-stage residuals. An important difference between the two settings is that 2SLS is robust to misspecification of the first stage model whereas two-step probit is not. Examples of misspecifications occur if $E(X \mid Z)$ is nonlinear, if $Var(X \mid Z)$ is non-constant (heteroskedastic), or if $X \mid Z$ is non-normal.

Another difference is that in the linear case the control-function approach and the fitted-value approach lead to the same estimator (2SLS) whereas this is not true for probit.

3.3 A semiparametric generalization

Consider the model

$$\begin{aligned} Y &= \mathbf{1}(\alpha + \beta X + U \geq 0) \\ X &= \pi' Z + \sigma_v V \end{aligned}$$

and assume that

$$U \mid X, V \sim U \mid V.$$

In the previous parametric model we additionally assumed that $U \mid V$ was $\mathcal{N}\left(\rho V, 1-\rho^2\right)$ and V was $\mathcal{N}(0,1)$. The semiparametric generalization consists in leaving the distributions of $U \mid V$ and V unspecified.

In this way

$$\Pr(Y = 1 \mid X, V) = \Pr(-U \leq \alpha + \beta X \mid X, V) = \Pr(-U \leq \alpha + \beta X \mid V)$$

Thus

$$E(Y \mid X, V) = F(\alpha + \beta X, V)$$

where $F(., V)$ is the conditional *cdf* of $-U$ given V. The function $F(., V)$ can be estimated non-parametrically using estimated first-stage residuals. This is a bivariate-index generalization of the semiparametric approaches to estimating single-index models with exogenous variables (cf. Blundell and Powell, 2003).

3.3.1 Constructing policy parameters

To construct a policy parameter we need $p(x) = \Pr(-U \leq \alpha + \beta x)$. Note that

$$\Pr(-U \leq \alpha + \beta x) = \int \Pr(-U \leq \alpha + \beta x \mid v)\, dF_v \equiv E_V\left[F(\alpha + \beta x, V)\right].$$

In the normal model

$$\Pr(-U \leq \alpha + \beta x) = \Phi(\alpha + \beta x)$$

But this means that

$$\Phi(\alpha + \beta x) = E_V\left[\Phi\left(\frac{\alpha + \beta x + \rho V}{\sqrt{1-\rho^2}}\right)\right] \equiv E_V\left[\Phi\left(\alpha^\dagger + \beta^\dagger x + \rho^\dagger V\right)\right]. \tag{2}$$

A simple consistent estimate of $\Phi(\alpha + \beta x)$ is $\Phi\left(\widehat{\alpha} + \widehat{\beta} x\right)$, where $\widehat{\alpha}$ and $\widehat{\beta}$ are consistent estimates. For example, using that $1 - \rho^2 = 1/\left(1 + \rho^{\dagger 2}\right)$, we may use

$$\left(\widehat{\alpha}, \widehat{\beta}\right) = \left(1 + \widehat{\rho}^{\dagger 2}\right)^{-1/2}\left(\widehat{\alpha}^\dagger, \widehat{\beta}^\dagger\right)$$

where $\left(\widehat{\alpha}^\dagger, \widehat{\beta}^\dagger, \widehat{\rho}^\dagger\right)$ are two-step control function estimates.

Alternatively, using expression (2) a consistent estimate of $\Phi(\alpha + \beta x)$ in the normal model can be obtained as

$$\widehat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \Phi\left(\widehat{\alpha}^\dagger + \widehat{\beta}^\dagger x + \widehat{\rho}^\dagger \widehat{v}_i\right).$$

In the semiparametric model this result generalizes to

$$\widetilde{p}(x) = \frac{1}{N}\sum_{i=1}^{N} \widehat{F}\left(\widetilde{\alpha} + \widetilde{\beta} x, \widehat{v}_i\right)$$

where $\left(\widetilde{\alpha}, \widetilde{\beta}\right)$ are semiparametric control function estimates and $\widehat{F}(.,.)$ is a non-parametric estimate of the conditional *cdf* of $-U$ given V.

4 The endogenous dummy explanatory variable probit model

The model is

$$\begin{aligned} Y &= \mathbf{1}(\alpha + \beta X + U \geq 0) \\ X &= \mathbf{1}\left(\pi' Z + V \geq 0\right) \\ \begin{pmatrix} U \\ V \end{pmatrix} &\mid Z \sim \mathcal{N}\left[0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right]. \end{aligned}$$

In this model X is an endogenous explanatory variable as long as $\rho \neq 0$. X is exogenous if $\rho = 0$.

Let us introduce a notation for standard normal bivariate probabilities: $\Phi_2(r, s; \rho) = \Pr(U \leq r, V \leq s)$.

The joint probability distribution of Y and X given Z consists of four terms:

$$p_{00} = \Pr(Y = 0, X = 0) = \Pr(\alpha + \beta X + U < 0, X = 0) = \Pr\left(\alpha + U < 0, \pi' Z + V < 0\right) = \Phi_2\left(-\alpha, -\pi' Z; \rho\right)$$

$$\begin{aligned} p_{01} &= \Pr(Y = 0, X = 1) = \Pr(\alpha + \beta + U < 0, X = 1) = \Pr(X = 1 \mid \alpha + \beta + U < 0) \Pr(\alpha + \beta + U < 0) \\ &= [1 - \Pr(X = 0 \mid \alpha + \beta + U < 0)] \Pr(\alpha + \beta + U < 0) \\ &= \Pr(\alpha + \beta + U < 0) - \Pr(\alpha + \beta + U < 0, X = 0) = \Phi(-\alpha - \beta) - \Phi_2\left(-\alpha - \beta, -\pi' Z; \rho\right) \end{aligned}$$

$$\begin{aligned} p_{10} &= \Pr(Y = 1, X = 0) = \Pr(\alpha + U \geq 0, X = 0) = \Pr(\alpha + U \geq 0 \mid X = 0) \Pr(X = 0) \\ &= [1 - \Pr(\alpha + U < 0 \mid X = 0)] \Pr(X = 0) = \Pr(X = 0) - \Pr(\alpha + U < 0, X = 0) = \Phi\left(-\pi' Z\right) - p_{00} \end{aligned}$$

$$p_{11} = 1 - p_{00} - p_{01} - p_{10}.$$

Therefore, the log-likelihood is given by

$$L = \sum_{i=1}^{N} \{(1 - y_i)(1 - x_i) \ln p_{00i} + (1 - y_i) x_i \ln p_{01i} + y_i (1 - x_i) \ln p_{10i} + y_i x_i \ln p_{11i}\}.$$

In this model there are only two potential outcomes:

$$\begin{aligned} Y(1) &= \mathbf{1}(\alpha + \beta + U \geq 0) \\ Y(0) &= \mathbf{1}(\alpha + U \geq 0) \end{aligned}$$

The average probability effect of interest is given by

$$\theta = E[Y(1) - Y(0)] = \Phi(\alpha + \beta) - \Phi(\alpha).$$

In less parametric specifications $E[Y(1) - Y(0)]$ may not be point identified, but we may still be able to estimate average effects for certain sub-populations of interest (cf. Imbens and Angrist, 1994; Vytlacil, 2002).

A straightforward extension of this model is an ordered probit with dummy endogenous explanatory variable (see Appendix C, which also discusses pseudo ML estimation of ordered probit models with or without endogeneity using binary probit methods).

Local average treatment effects (LATE) Consider the case where

$$X = \mathbf{1}(\pi_0 + \pi_1 Z + V \geq 0)$$

and Z is a scalar 0–1 instrument, so that there are only two potential values of X:

$$\begin{aligned} X(1) &= \mathbf{1}(\pi_0 + \pi_1 + V \geq 0) \\ X(0) &= \mathbf{1}(\pi_0 + V \geq 0). \end{aligned}$$

Suppose without lack of generality that $\pi_1 \geq 0$. Then we can distinguish three subpopulations depending on an individual's value of V:

- Never-takers: Units with $V < -\pi_0 - \pi_1$. They have $X(1) = 0$ and $X(0) = 0$. Their mass is $\Phi(-\pi_0 - \pi_1) = 1 - \Phi(\pi_0 + \pi_1)$.
- Compliers: Units with $V \geq -\pi_0 - \pi_1$ but $V < -\pi_0$. They have $X(1) = 1$ and $X(0) = 0$. Their mass is $\Phi(-\pi_0) - \Phi(-\pi_0 - \pi_1) = \Phi(\pi_0 + \pi_1) - \Phi(\pi_0)$.
- Always-takers: Units with $V \geq -\pi_{00}$. They have $X(1) = 1$ and $X(0) = 1$. Their mass is $1 - \Phi(-\pi_0) = \Phi(\pi_0)$.

Let us obtain the average treatment effect for the subpopulation of compliers:

$$\theta_{LATE} = E[Y(1) - Y(0) \mid X(1) - X(0) = 1] \equiv E[Y(1) - Y(0) \mid -\pi_0 - \pi_1 \leq V < -\pi_0].$$

We have

$$\begin{aligned} E[Y(1) \mid -\pi_0 - \pi_1 \leq V < -\pi_0] &= \Pr(\alpha + \beta + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0) \\ &= 1 - \Pr(U \leq -\alpha - \beta \mid -\pi_0 - \pi_1 \leq V < -\pi_0) \\ &= 1 - \frac{\Pr(-\pi_0 - \pi_1 \leq V < -\pi_0 \mid U \leq -\alpha - \beta)\Pr(U \leq -\alpha - \beta)}{\Pr(-\pi_0 - \pi_1 \leq V < -\pi_0)} \\ &= 1 - \frac{\Pr(U \leq -\alpha - \beta, V \leq -\pi_0) - \Pr(U \leq -\alpha - \beta, V \leq -\pi_0 - \pi_1)}{\Pr(V \leq -\pi_0) - \Pr(V \leq -\pi_0 - \pi_1)} \end{aligned}$$

and similarly

$$\begin{aligned} E[Y(0) \mid -\pi_0 - \pi_1 \leq V < -\pi_0] &= \Pr(\alpha + U \geq 0 \mid -\pi_0 - \pi_1 \leq V < -\pi_0) \\ &= 1 - \frac{\Pr(U \leq -\alpha, V \leq -\pi_0) - \Pr(U \leq -\alpha, V \leq -\pi_0 - \pi_1)}{\Pr(V \leq -\pi_0) - \Pr(V \leq -\pi_0 - \pi_1)}, \end{aligned}$$

so that

$$\theta_{LATE} = \frac{\Phi_2(-\alpha, -\pi_0; \rho) - \Phi_2(-\alpha, -\pi_0 - \pi_1; \rho) - \Phi_2(-\alpha - \beta, -\pi_0; \rho) + \Phi_2(-\alpha - \beta, -\pi_0 - \pi_1; \rho)}{\Phi(-\pi_0) - \Phi(-\pi_0 - \pi_1)}.$$

This provides a formal connection with the IV estimand since we know that

$$\theta_{LATE} = \frac{E(Y \mid Z = 1) - E(Y \mid Z = 0)}{E(D \mid Z = 1) - E(D \mid Z = 0)}.$$

The nice thing about θ_{LATE} is that it is identified from the Wald formula in the absence of joint normality. In fact, it does not even require the index model assumption for $Y(1)$ and $Y(0)$. So we do not need monotonicity in the relationship between Y and X. The relevance of θ_{LATE} partly depends on how large the probability of compliers is, and partly on its policy relevance.

We have

$$\begin{aligned}\theta_{ATE} &= \theta_{LATE} \Pr(\text{compliers}) + E[Y(1) - Y(0) \mid \text{never-takers}] \Pr(\text{never-takers}) \\ &\quad + E[Y(1) - Y(0) \mid \text{always-takers}] \Pr(\text{always-takers}).\end{aligned}$$

There is a connection with fixed-effects identification in binary-choice panel data models.

Two related references are J. Angrist (2001, *JBES*, 19, 2–16) on LDV models, and Ed Vytlacil on the identification content of enforcing the index model assumption on $Y(1)$ and $Y(0)$.

References

[1] Blundell, R. and J. L. Powell (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models". In: M. Dewatripont, L. P. Hansen, and S. J. Turnovsky (eds.), *Advances in Economics and Econometrics, Eighth World Congress*, vol. II, Cambridge University Press.

[2] Imbens, G. W. and J. Angrist (1994): "Identification and Estimation of Local Average Treatment Effects", *Econometrica*, 62, 467-475.

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

A 2SLS as a control function estimator

The 2SLS estimator for the single equation model

$$y_i = x_i'\theta + u_i \qquad E(z_i u_i) = 0$$

is given by

$$\widehat{\theta} = (X'MX)^{-1} X'My = \left(\widehat{X}'\widehat{X}\right)^{-1} \widehat{X}'y$$

where X is $N \times k$, Z is $N \times r$, y is $N \times 1$, and $M = Z(Z'Z)^{-1} Z'$. Moreover, the fitted values are $\widehat{X} = Z\widehat{\Pi}'$ where $\widehat{\Pi} = X'Z(Z'Z)^{-1}$, and the corresponding $N \times k$ matrix of first-stage residuals:

$$\widehat{V} = X - Z\widehat{\Pi}' = (I_N - M)X \equiv QX.$$

Typically, X and Z will have some columns in common (e.g. a constant term). For those variables, the corresponding columns of $\widehat{V}$ will be identically zero. That is, letting $X = (X_1, X_2)$ and $Z = (Z_1, X_2)$,

$$\widehat{V} = Q(X_1, X_2) = \left(\widehat{V}_1, 0\right)$$

where $\widehat{V}_1 = QX_1$ is the subset of non-zero columns of $\widehat{V}$ (those corresponding to endogenous explanatory variables).

Now consider the coefficients of X in the OLS regression of y on X and $\widehat{V}_1$. Using the formulae for partitioned regression, these are given by

$$\widetilde{\theta} = \left(X'\left[I_N - \widehat{V}_1\left(\widehat{V}_1'\widehat{V}_1\right)^{-1}\widehat{V}_1'\right]X\right)^{-1} X'\left[I_N - \widehat{V}_1\left(\widehat{V}_1'\widehat{V}_1\right)^{-1}\widehat{V}_1'\right]y.$$

Clearly, $\widetilde{\theta} = \widehat{\theta}$ since

$$\begin{aligned}\widetilde{\theta} &= \left(X'\left[I_N - QX_1\left(X_1'QX_1\right)^{-1}X_1'Q\right]X\right)^{-1} X'\left[I_N - QX_1\left(X_1'QX_1\right)^{-1}X_1'Q\right]y \\ &= \left(X'X - X'QX\right)^{-1}\left(X'y - X'Qy\right) = \left(X'MX\right)^{-1}X'My.\end{aligned}$$

Note that, due to $QX_2 = 0$, we have

$$X'QX_1\left(X_1'QX_1\right)^{-1}X_1'QX = \begin{pmatrix} X_1'QX_1 & 0 \\ 0 & 0 \end{pmatrix} = X'QX.$$

The conclusion is that 2SLS estimates can be obtained from the OLS regression of y on X and $\widehat{V}_1$. Therefore, linear 2SLS can be regarded as a control function method.

B Consistent standard errors for two-step estimators

Consider an estimator $\widehat{\theta}$ that maximizes an objective function that depends on estimated parameters:

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{N} \ell_i\left(\theta, \widehat{\gamma}\right).$$

The estimated parameters themselves are obtained by solving

$$\widehat{\gamma} = \arg\max_{\gamma} \sum_{i=1}^{N} \zeta_i\left(\gamma\right).$$

Denote true values as (θ_0, γ_0), and consider the following notation for score and Hessian terms:

$$\ell_i^{\theta}\left(\theta,\gamma\right) = \frac{\partial \ell_i\left(\theta,\gamma\right)}{\partial \theta}, \qquad \zeta_i^{\gamma}\left(\gamma\right) = \frac{\partial \zeta_i\left(\gamma\right)}{\partial \gamma}$$

$$H_{\theta\theta} = -\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \ell_i\left(\theta_0,\gamma_0\right)}{\partial\theta\partial\theta'}, \quad H_{\theta\gamma} = -\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \ell_i\left(\theta_0,\gamma_0\right)}{\partial\theta\partial\gamma'}, \quad H_{\gamma\gamma} = -\operatorname*{plim}_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \frac{\partial^2 \zeta_i\left(\gamma_0\right)}{\partial\gamma\partial\gamma'}.$$

Moreover, assume that the sample scores are asymptotically normal:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \begin{pmatrix} \ell_i^{\theta}\left(\theta_0,\gamma_0\right) \\ \zeta_i^{\gamma}\left(\gamma_0\right) \end{pmatrix} \xrightarrow{d} \mathcal{N}\left[0, \begin{pmatrix} \Upsilon_{\theta\theta} & \Upsilon_{\theta\gamma} \\ \Upsilon_{\gamma\theta} & \Upsilon_{\gamma\gamma} \end{pmatrix}\right].$$

Under standard regularity conditions, an expansion of the first-order conditions for $\widehat{\theta}$ gives

$$0 = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \ell_i^{\theta}\left(\widehat{\theta},\widehat{\gamma}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \ell_i^{\theta}\left(\theta_0,\gamma_0\right) - H_{\theta\theta}\sqrt{N}\left(\widehat{\theta}-\theta_0\right) - H_{\theta\gamma}\sqrt{N}\left(\widehat{\gamma}-\gamma_0\right) + o_p\left(1\right). \quad (3)$$

A similar expansion of the first-order conditions for $\widehat{\gamma}$ gives

$$0 = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i^{\gamma}\left(\widehat{\gamma}\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i^{\gamma}\left(\gamma_0\right) - H_{\gamma\gamma}\sqrt{N}\left(\widehat{\gamma}-\gamma_0\right) + o_p\left(1\right)$$

or equivalently

$$\sqrt{N}\left(\widehat{\gamma}-\gamma_0\right) = H_{\gamma\gamma}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i^{\gamma}\left(\gamma_0\right) + o_p\left(1\right). \quad (4)$$

Substituting (4) into (3) we get

$$H_{\theta\theta}\sqrt{N}\left(\widehat{\theta}-\theta_0\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N} \ell_i^{\theta}\left(\theta_0,\gamma_0\right) - H_{\theta\gamma}H_{\gamma\gamma}^{-1}\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \zeta_i^{\gamma}\left(\gamma_0\right) + o_p\left(1\right)$$

and

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right) = H_{\theta\theta}^{-1}\left(I, -H_{\theta\gamma}H_{\gamma\gamma}^{-1}\right)\frac{1}{\sqrt{N}}\sum_{i=1}^{N} \begin{pmatrix} \ell_i^{\theta}\left(\theta_0,\gamma_0\right) \\ \zeta_i^{\gamma}\left(\gamma_0\right) \end{pmatrix} + o_p\left(1\right).$$

Thus,

$$\sqrt{N}\left(\widehat{\theta}-\theta_0\right)\xrightarrow{d}\mathcal{N}\left(0,V\right)$$

where

$$\begin{aligned}V &= H_{\theta\theta}^{-1}\left(I,-H_{\theta\gamma}H_{\gamma\gamma}^{-1}\right)\left(\begin{array}{cc}\Upsilon_{\theta\theta} & \Upsilon_{\theta\gamma}\\ \Upsilon_{\gamma\theta} & \Upsilon_{\gamma\gamma}\end{array}\right)\left(\begin{array}{c}I\\ -H_{\gamma\gamma}^{-1}H_{\gamma\theta}\end{array}\right)H_{\theta\theta}^{-1}\\ &= H_{\theta\theta}^{-1}\left[\Upsilon_{\theta\theta}+H_{\theta\gamma}\left(H_{\gamma\gamma}^{-1}\Upsilon_{\gamma\gamma}H_{\gamma\gamma}^{-1}\right)H_{\gamma\theta}-H_{\theta\gamma}H_{\gamma\gamma}^{-1}\Upsilon_{\gamma\theta}-\Upsilon_{\theta\gamma}H_{\gamma\gamma}^{-1}H_{\gamma\theta}\right]H_{\theta\theta}^{-1}.\end{aligned}$$

A consistent estimator of V is:

$$\widehat{V}=\widehat{H}_{\theta\theta}^{-1}\left(I,-\widehat{H}_{\theta\gamma}\widehat{H}_{\gamma\gamma}^{-1}\right)\left(\begin{array}{cc}\widehat{\Upsilon}_{\theta\theta} & \widehat{\Upsilon}_{\theta\gamma}\\ \widehat{\Upsilon}_{\gamma\theta} & \widehat{\Upsilon}_{\gamma\gamma}\end{array}\right)\left(\begin{array}{c}I\\ -\widehat{H}_{\gamma\gamma}^{-1}\widehat{H}_{\gamma\theta}\end{array}\right)\widehat{H}_{\theta\theta}^{-1}$$

where

$$\widehat{H}_{\theta\theta}=-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2\ell_i\left(\widehat{\theta},\widehat{\gamma}\right)}{\partial\theta\partial\theta'},\qquad \widehat{H}_{\theta\gamma}=-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2\ell_i\left(\widehat{\theta},\widehat{\gamma}\right)}{\partial\theta\partial\gamma'},\qquad H_{\gamma\gamma}=-\frac{1}{N}\sum_{i=1}^{N}\frac{\partial^2\zeta_i\left(\widehat{\gamma}\right)}{\partial\gamma\partial\gamma'}.$$

$$\widehat{\Upsilon}_{\theta\theta}=\frac{1}{N}\sum_{i=1}^{N}\ell_i^{\theta}\left(\widehat{\theta},\widehat{\gamma}\right)\ell_i^{\theta}\left(\widehat{\theta},\widehat{\gamma}\right)',\qquad \widehat{\Upsilon}_{\theta\gamma}=\frac{1}{N}\sum_{i=1}^{N}\ell_i^{\theta}\left(\widehat{\theta},\widehat{\gamma}\right)\zeta_i^{\gamma}\left(\widehat{\gamma}\right)',\qquad \widehat{\Upsilon}_{\gamma\gamma}=\frac{1}{N}\sum_{i=1}^{N}\zeta_i^{\gamma}\left(\widehat{\gamma}\right)\zeta_i^{\gamma}\left(\widehat{\gamma}\right)'.$$

Note that $H_{\theta\theta}^{-1}\Upsilon_{\theta\theta}H_{\theta\theta}^{-1}$ is the asymptotic variance of the infeasible estimator that maximizes $\sum_{i=1}^{N}\ell_i\left(\theta,\gamma_0\right)$, and that $H_{\gamma\gamma}^{-1}\Upsilon_{\gamma\gamma}H_{\gamma\gamma}^{-1}$ is the asymptotic variance of $\sqrt{N}\left(\widehat{\gamma}-\gamma_0\right)$.

If the information identities hold ($H_{\theta\theta}^{-1}=\Upsilon_{\theta\theta}$ and $H_{\gamma\gamma}^{-1}=\Upsilon_{\gamma\gamma}$), given consistent estimates of $H_{\theta\theta}^{-1}$ and $H_{\gamma\gamma}^{-1}$, all we need to construct a consistent estimate of V are consistent estimates of the cross-terms $H_{\theta\gamma}$ and $\Upsilon_{\theta\gamma}$.

C Estimating an ordered probit model as binary probit

M. Arellano, November 17, 2007

Ordered Probit with Exogeneity. Consider three alternatives:

$$\begin{aligned}
\Pr(y_1 = 1) &= \Pr\left(x'\beta + u \leq c_1\right) = \Phi\left(c_1 - x'\beta\right) \\
\Pr(y_2 = 1) &= \Pr\left(c_1 < x'\beta + u \leq c_2\right) = \Phi\left(c_2 - x'\beta\right) - \Phi\left(c_1 - x'\beta\right) \\
\Pr(y_3 = 1) &= \Pr\left(x'\beta + u > c_2\right) = 1 - \Phi\left(c_2 - x'\beta\right)
\end{aligned}$$

Note that binary probit of $(y_2 + y_3)$ on x and a constant provides consistent estimates of β and $-c_1$. Also, binary probit of y_3 on x and a constant provides consistent estimates of β and $-c_2$. The trouble with this is that we are not enforcing the restriction that the β coefficients in the two cases are the same. To do so we form a duplicated dataset as follows:

	Unit	w	b_1	b_2	X
Full-time	1	1	1	0	X_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	N_F	1	1	0	X_{N_F}
Part-time	$N_F + 1$	1	1	0	X_{N_F+1}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$N_F + N_P$	1	1	0	$X_{N_F+N_P}$
No-work	$N_F + N_P + 1$	0	1	0	$X_{N_F+N_P+1}$
	$N_F + N_P + 2$	0	1	0	$X_{N_F+N_P+2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$N_F + N_P + N_O$	0	1	0	X_N
Full-time	1	1	0	1	X_1
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	N_F	1	0	1	X_{N_F}
Part-time	$N_F + 1$	0	0	1	X_{N_F+1}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$N_F + N_P$	0	0	1	$X_{N_F+N_P}$
No-work	$N_F + N_P + 1$	0	0	1	$X_{N_F+N_P+1}$
	$N_F + N_P + 2$	0	0	1	$X_{N_F+N_P+2}$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$N_F + N_P + N_O$	0	0	1	X_N

N_F is the number of units working full-time, N_P the number of units working part-time, and N_O the number of non-working units, so that $N = N_F + N_P + N_O$ is the total number of observations.

The proposed method is to form the artificial sample of size $2N$ shown in the Table and run a binary probit of w on b_1,b_2, and X to obtain estimates of $-c_1, -c_2$, and β. These estimates are consistent and asymptotically normal but not as efficient as ordered probit ML, because they are maximizing a pseudo-likelihood as opposed to the full-likelihood. The advantage is that they can be obtained from a binary probit routine, while enforcing the constraint on β across groups.

Note that since one of the equations is redundant the model can be written as

$$\begin{aligned} E(y_2 \mid x) &= \Phi\left(c_2 - x'\beta\right) - \left(c_1 - x'\beta\right) \\ E(y_3 \mid x) &= 1 - \Phi\left(c_2 - x'\beta\right) \end{aligned}$$

or equivalently

$$\begin{aligned} E(y_2 + y_3 \mid x) &= \Phi\left(-c_1 + x'\beta\right) \\ E(y_3 \mid x) &= \Phi\left(-c_2 + x'\beta\right), \end{aligned}$$

which is a system of two probit equations. Optimal GMM in this system is asymptotically equivalent to ordered probit ML. GMM without taking into account the dependence between the two equation moments is asymptotically equivalent to the method suggested above.

Ordered Probit with Endogeneity

Another advantage is that it is possible to use the same trick to estimate the ordered probit model with dummy endogenous explanatory variable using the bivariate probit routine.[2] The model is

$$y_2 = \mathbf{1}\left(c_1 < x\alpha + z_1'\gamma + u \leq c_2\right) \tag{5}$$

$$y_3 = \mathbf{1}\left(x\alpha + z_1'\gamma + u > c_2\right) \tag{6}$$

$$x = \mathbf{1}\left(z'\pi + v > 0\right) \tag{7}$$

$$\begin{pmatrix} u \\ v \end{pmatrix} \mid z \sim N\left(0, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right) \tag{8}$$

with $z = (z_1', z_2')'$ where z_1 and are exogenous controls and z_2 are excluded instruments.

The first equation can be replaced by

$$y_2 + y_3 = \mathbf{1}\left(x\alpha + z_1'\gamma + u > c_1\right). \tag{9}$$

Equations (9),(7), and (8) describe a standard bivariate probit with log-likelihood $L_{23}(\alpha, \gamma, c_1, \rho, \pi)$. Equations (6),(7), and (8) describe another standard bivariate probit with log-likelihood $L_3(\alpha, \gamma, c_2, \rho, \pi)$. The following estimator of $(\alpha, \gamma, c_1, c_2, \rho)$ is consistent and can be obtained as ordinary bivariate probit in the duplicated sample, having fixed the first-stage coefficients at its probit estimates $\widehat{\pi}$:

$$(\widehat{\alpha}, \widehat{\gamma}, \widehat{c_1}, \widehat{c_2}, \widehat{\rho}) = \arg\max\left\{L_{23}(\alpha, \gamma, c_1, \rho, \widehat{\pi}) + L_3(\alpha, \gamma, c_2, \rho, \widehat{\pi})\right\}.$$

[2] The same is true for estimating an ordered logit model with fixed effects from panel data.