Statistical Methods III: Spring 2013

Jonathan Wand

Stanford University

Choice models with endogeneity

Outline

Bivariate normal: derivation

- Take $X, Z \sim N(0,1)$, independent
- Set $Y = \rho X \; + \; \sqrt{1-\rho^2}\, Z$
- Note: $E(X) = E(Y) = 0$,
  $Var(X) = Var(Y) = 1, \; Corr(X,Y) = \rho$

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \cdot \begin{bmatrix} X \\ Z \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$$

- $\Rightarrow$ (X, Y) has the "standard bivariate normal" distribution

Conditional expectations

Things you need to know to derive mills ratio,

$$E(z \mid z > c) = \int_c^\infty \frac{z\phi(z)}{[1-\Phi(c)]} dz$$

$$E(z \mid z > c) = \frac{1}{(1-\Phi(c))} \int_c^\infty \frac{z}{\sqrt{2\pi}}.\exp(\frac{-z^2}{2}) dz$$

$$E(z \mid z > c) = \frac{1}{(1-\Phi(c))} \int^\infty -(\frac{d\phi(z)}{dz}) dz$$

$$\frac{d\phi(z)}{dz} = \frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) \cdot -z$$

$$\int_c^\infty -(\frac{d\phi(z)}{dz}) dz = \int_c^\infty -\frac{1}{\sqrt{2\pi}} \exp(-\frac{z^2}{2}) = 0 + \frac{1}{\sqrt{2\pi}} \exp(-\frac{c^2}{2}) = \phi(c)$$

Endogenous assignment

$$C_i = 1 \text{ if } a + bX_i + U_i > 0, \text{ else } C_i = 0. \tag{1}$$

In application, $C_i = 1$ means that subject i self-selects into treatment. The second equation defines the subject's response to treatment:

$$Y_i = 1 \text{ if } c + dZ_i + eC_i + V_i > 0, \text{ else } Y_i = 0. \tag{2}$$

Endogenous assignment

Step 1. Estimate the probit model (1) by likelihood techniques.

Step 2. To estimate (2), fit the expanded probit model

$$P(Y_i = 1|X_i, Z_i, C_i) = \Phi(c + dZ_i + eC_i + fM_i) \tag{3}$$

to the data, where

$$M_i = C_i \frac{\phi(a + bX_i)}{\Phi(a + bX_i)} - (1 - C_i) \frac{\phi(a + bX_i)}{1 - \Phi(a + bX_i)}. \tag{4}$$

Endogenous assignment

Consider (1–2). We can represent V_i as $\rho U_i + \sqrt{1-\rho^2}W_i$, where W_i is an $N(0, 1)$ random variable, independent of U_i. Then

$$\begin{aligned} E\left\{V_i \middle| X_i = x, C_i = 1\right\} &= E\left\{\rho U_i + \sqrt{1-\rho^2}W_i \middle| U_i > -a - bx_i\right\} \\ &= \rho E\{U_i | U_i > -a - bx_i\} \\ &= \rho \frac{1}{\Phi(a + bx_i)} \int_{-a-bx_i}^{\infty} x\phi(x)\mathrm{d}x \\ &= \rho \frac{\phi(a + bx_i)}{\Phi(a + bx_i)} \end{aligned} \tag{9}$$

because $P\{U_i > -a - bx_i\} = P\{U_i < a + bx_i\} = \Phi(a + bx_i)$. Likewise,

$$E\left\{V_i \middle| X_i = x, C_i = 0\right\} = -\rho \frac{\phi(a + bx_i)}{1 - \Phi(a + bx_i)}. \tag{10}$$

Endogenous assignment

Step 1. Estimate the probit model (1) by likelihood techniques.

Step 2. To estimate (2), fit the expanded probit model

$$P(Y_i = 1|X_i, Z_i, C_i) = \Phi(c + dZ_i + eC_i + fM_i) \tag{3}$$

to the data, where

$$M_i = C_i \frac{\phi(a + bX_i)}{\Phi(a + bX_i)} - (1 - C_i) \frac{\phi(a + bX_i)}{1 - \Phi(a + bX_i)}. \tag{4}$$

Endogenous assignment

Table 1 Simulation results

	c	d	e	ρ
True values				
	−1.0000	0.7500	0.5000	0.6000
Raw estimates				
Mean	−1.5901	0.7234	1.3285	
SD	0.1184	0.0587	0.1276	
Two-step				
Mean	−1.1118	0.8265	0.5432	
SD	0.1581	0.0622	0.2081	
MLE				
Mean	−0.9964	0.7542	0.4964	0.6025
SD	0.161	0.0546	0.1899	0.0900

Notes. Correcting endogeneity bias when the response is binary probit. There are 500 repetitions. The sample size is 1000. The correlation between latents is $\rho = 0.60$. The parameters in the selection equation (1) are set at $a = 0.50$ and $b = 1$. The parameters in the response equation (2) are set at $c = -1$, $d = 0.75$, and $e = 0.50$. The response equation includes the endogenous dummy C_i defined by (1). The correlation between the exogenous regressors is 0.40. MLE computed by VGAM 0.7-6.

Endogenous assignment

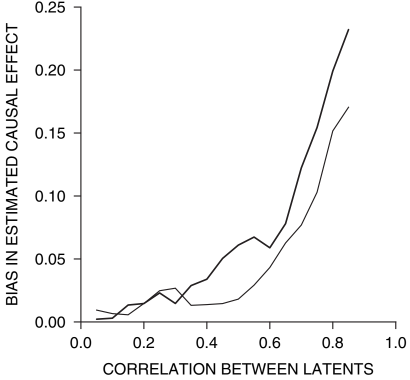

Fig. 1 The two-step correction. Graph of bias in $\hat{e}$ against ρ, the correlation between the latents. The light lower line sets the correlation between regressors to 0.40, and the heavy upper line sets the correlation to 0.60. Other parameters as for Table 1. Below 0.35, the lines crisscross.

Endogenous selection

$$P(Y_i = 1|X_i, Z_i) = \Phi(c + dZ_i + fM_i) \tag{7}$$

to the data on subjects i with $C_i = 1$. This time,

$$M_i = \frac{\phi(a + bX_i)}{\Phi(a + bX_i)}. \tag{8}$$

Endogenous selection

Table 2 Simulation results

	c	d	ρ
True values			
	−1.0000	0.7500	0.6000
Raw estimates			
Mean	−0.7936	0.7299	
SD	0.0620	0.0681	
Two-step			
Mean	−1.0751	0.8160	
SD	0.1151	0.0766	
MLE			
Mean	−0.9997	0.7518	0.5946
SD	0.0757	0.0658	0.1590

Notes. Correcting endogeneity bias in sample selection when the response is binary probit. There are 500 repetitions. The sample size is 1000. The correlation between latents is $\rho = 0.60$. The parameters in the selection equation (5) are set at $a = 0.50$ and $b = 1$. The parameters in the response equation (6) are set at $c = -1$, and $d = 0.75$. Response data are observed only when $C_i = 1$, as determined by the selection equation. This will occur for about 64% of the subjects. The correlation between the exogenous regressors is 0.40. MLE computed using *Stata* 9.2.