Endogeneity Bias May be Contagious DA Freedman
Statistics 215 November 2007

1) Let x and w be fixed n-vectors with mean 0. Let (δ_i, ϵ_i) be IID pairs of normal random variables, with expectation 0, variance σ^2 and τ^2 respectively, and correlation $\rho \neq 0$. We consider a regression model where the design matrix X is $n \times 2$. The first column is fixed. It is x. The second column is random. It is $w + \delta$. The response variable in the model is

$$Y = X\begin{pmatrix} a \\ b \end{pmatrix} + \epsilon.$$

Suppose

$$\frac{1}{n}\sum_{i=1} x_i^2 \to 1, \quad \frac{1}{n}\sum_{i=1} w_i^2 \to 1, \quad \frac{1}{n}\sum_{i=1} x_i w_i \to r$$

with $-1 < r < 1$.

(a) Show that the first column of X is exogenous and the second is endogenous.

(b) Show that

$$\frac{1}{n}X'X \to \begin{pmatrix} 1 & r \\ r & 1+\sigma^2 \end{pmatrix}.$$

Here and below, convergence is almost sure, but you may elect just to demonstrate convergence in probability.

(c) Show that

$$n(X'X)^{-1} \to \frac{1}{1+\sigma^2-r^2}\begin{pmatrix} 1+\sigma^2 & -r \\ -r & 1 \end{pmatrix}.$$

(d) Show that

$$\frac{1}{n}X'Y \to \begin{pmatrix} a+br \\ ar+b(1+\sigma^2)+c\sigma^2 \end{pmatrix}$$

where $c = \rho\tau/\sigma$.

(e) Show that the OLS estimate $\hat{a}$ is asymptotically biased downward by

$$\frac{rc\sigma^2}{1+\sigma^2-r^2}.$$

(f) Show that endogeneity bias affects $\hat{a}$ unless $r = 0$.

(g) Can endogeneity bias be positive?

2) Consider the model

$$Y = X\beta + \epsilon$$

where β is a $p \times 1$ parameter vector. The design matrix X is $n \times p$, random, of full rank, but endogenous. The ϵ_i are IID for $i = 1, \ldots, n$ with $E(\epsilon_i) = 0$. Happily, $E(\epsilon|X)$ is a linear combination of the first p_0 columns of X, where $1 \leq p_0 < p$.

(a) Why isn't $E(\epsilon|X) = 0$?

(b) Show that endogeneity bias affects only the first p_0 components of β. Hint: Let $E(\epsilon|X) = X\gamma$. Then

$$Y = X(\beta + \gamma) + \big[\epsilon - E(\epsilon|X)\big].$$

What is $E\big\{(X'X)^{-1}X'Y\big|X\big\}$?

(c) Suppose $\text{var}(\epsilon|X) = \sigma^2 I_{n\times n}$. Can you get an unbiased estimate for σ^2?

(d) If n is large, can you get an approximate 95% confidence interval for β_p? You may assume that p is fixed and $X'X/n$ converges to a $p \times p$ matrix that is positive definite.

The big picture. If some regressors are endogenous, OLS estimates—even for the coefficients of exogenous regressors—are going to be biased. So the bias spreads from the endogenous regressors to the exogenous ones. Under supplementary conditions, the bias remains localized. Similar conclusions apply to IVLS. Generally, random errors like δ and ϵ would not be observable, and $E(\epsilon|X)$ would be unknown. Thus, conditions for localization of bias are not readily checkable. Also see

http://www.stat.berkeley.edu/users/census/socident.pdf

What about probits and logits? Let X_i, Z_i, W_i be independent $N(0, 1)$ variables for $i = 1, \ldots, n$, where n is large. Let $0 < \rho < 1$. Let $U_i = \rho X_i + \sqrt{1-\rho^2}W_i$. Let a, b, c be real numbers. Consider a probit model where U_i is the latent variable, and the response variable Y_i is defined as follows:

$$Y_i = 1 \text{ if } a + bX_i + cZ_i + U_i > 0,$$

else $Y_i = 0$.

(a) Show that U_i is $N(0, 1)$ and Z_i is independent of (X_i, U_i).

(b) Is X_i endogenous or exogenous? What about Z_i?

(c) Let Φ be the standard normal distribution function. Show that

$$P(Y_i = 1|X_i, Z_i) = \Phi\left(\frac{a}{\sqrt{1-\rho^2}} + \frac{b+\rho}{\sqrt{1-\rho^2}}X_i + \frac{c}{\sqrt{1-\rho^2}}Z_i\right).$$

(d) An investigator fits a probit model to the data by the usual procedure, ignoring fine points like exogeneity of regressors. Show that the estimated intercept is nearly $a/\sqrt{1-\rho^2}$, the estimated coefficient of X_i is nearly $(b+\rho)/\sqrt{1-\rho^2}$, and the estimated coefficent of Z_i is nearly $c/\sqrt{1-\rho^2}$. This will take a fair amount of work; simulation might be easier.

Comments.

(i) If you use glmfit in the MATLAB toolbox, try small values for a, b, c, e.g., $\rho = .5$, $a = .1$, $b = .2$, $c = .3$. If you try $a = 1$, $b = 2$, $c = 3$ in release 7.0, you will see the dark side of numerical maximization; by release 7.4, the algorithm works much better.

(ii) Randomizing Z was just a convenient way to describe the data.

(iii) The probit is even more sensitive to endogeneity than OLS. In our example, conditioning on X changed the variance of U, which made the endogeneity bias spread from X to Z, even though Z is independent of X, U.

(iv) The endogeneity problem can easily be put into the response schedule framework. We make the construction more similar to the OLS example, as follows. Suppose the U_i are IID $N(0, 1)$ variables, while a, b, c are parameters. The response schedule for the 0–1 variable Y is

$$Y_{i,x,z} = 1 \text{ if } a + bx + cz + U_i > 0 \text{ else } Y_{i,x,z} = 0 \tag{$*$}$$

Let W_i be another sequence of IID random variables that are $N(0, 1)$ and independent of the U_i. Let s_i and z_i be sequences of fixed real numbers, with

$$\frac{1}{n}\sum_{i=1}^{n} s_i \to m_s, \quad \frac{1}{n}\sum_{i=1}^{n} s_i^2 \to m_{2,s}, \quad \frac{1}{n}\sum_{i=1}^{n} z_i \to m_z, \quad \frac{1}{n}\sum_{i=1}^{n} z_i^2 \to m_{2,z}, \quad \frac{1}{n}\sum_{i=1}^{n} s_i z_i \to m_{s,z}$$

We require all limits to be finite, and $m_{2,z} > m_z^2$. Let $-1 < \rho < 1$ be another parameter. To compute Y_i, Nature substitutes $X_i = s_i + \rho U_i + \sqrt{1-\rho^2}\,W_i$ for x and z_i for z in ($*$). The observables are

$$Y_i = Y_{i,X_i,z_i}, \quad X_i, \quad z_i$$

A probit regression of Y_i on X_i and z_i will produce biased estimates for a, b, c, because

$$P\{Y_i = 1 | X_i = x_i\} = \Phi\left(\frac{a - \rho s_i + (b+\rho)x_i + cz_i}{\sqrt{1-\rho^2}}\right)$$

Taking $s_i \equiv 0$ simplifies the calculations. Otherwise, there is another component of variance to deal with; if s_i is correlated with z_i, that has to be reckoned with as well.

(v) If we ignore small amounts of bias, $N(0, 1)$ latents are not de rigeur in the probit model. Conditional on the regressors, we really do need the latents to be nearly independent across subjects, with means that are nearly 0 and variances that are approximately constant. Near-symmetry seems to be called for, and tails that are not so different from the normal in length. By way of calibration, if the latent is rectangular rather than normal, but scaled to have mean 0 and variance 1, bias can be appreciable. The rectangular distribution is far from normal. If the mean of the latent changes across subjects, even in some way that is unrelated to the regressors, there are likely to be problems: see above. Haphazard changes in variance may make less of a difference, unless these are substantial: see below.

(vi) Suppose the 4-tuples $(X_i, Z_i, \sigma_i > 0, \zeta_i)$ are IID in i. Furthermore, (X_i, Z_i), $\sigma_i > 0$, ζ_i are independent for each i, with ζ_i being $N(0, 1)$. Let $Y_i = 1$ if $a + bX_i + cZ_i + \sigma_i\zeta_i > 0$, else $Y_i = 0$. If $\sigma_i \equiv 1$, this is the standard probit model, but we are allowing σ_i to be random. We fit a probit, ignoring this additional randomness. Perhaps with additional mild conditions, $\hat{a}$ is asymptotic to $a/E(\sigma)$, and so forth. The situation may be more complicated if σ_i is dependent on the regressors.