

7. Endogeneity

Extra Reading: M. Murray, (2006), "Avoiding Invalid Instruments and Coping with Weak Instruments", *Journal of Economic Perspectives*, Volume 20, Number 4, Fall, Pages 111–132

Is the term given to the situation when one or more of the regressors in the model are correlated with the error term such that

$$E(X'u) \neq 0$$

The 3 main causes of endogeneity are:

- i) Measurement error in the right hand side variables
- ii) Simultaneity (2 way causality) between dependent and right hand side variables
- iii) Omitted variables

In practice the solutions to these problems and the properties of the resulting estimation techniques can only be established for large samples (asymptotically)

In the absence of endogeneity OLS will produce a consistent estimator of the true parameter values β

$$\text{Given } \hat{\beta} = (X'X)^{-1} X'y = \beta + (X'X)^{-1} X'u$$

$$= \beta + \left(\frac{X'X}{N} \right)^{-1} \left(\frac{X'u}{N} \right)$$

The behaviour of this estimator as the sample size gets larger can be determined by taking the probability limits

$$p \lim(\hat{\beta}) = p \lim \left[\beta + \left(\frac{X'X}{N} \right)^{-1} \left(\frac{X'u}{N} \right) \right] \quad (1)$$

$$\text{Since } \left(\frac{X'X}{N} \right) = \frac{1}{N} \sum_{i=1}^N x_i x_i'$$

where x_i' is the i^{th} row of X

then this sample average will not go to zero as the sample size gets larger (it converges to a finite value, say Q_x)

$$\text{so } p \lim \left(\frac{X'X}{N} \right) = p \lim \left(\frac{1}{N} \sum_{i=1}^N x_i x_i' \right) = Q_x$$

However

$$\left(\frac{X'u}{N} \right) = \frac{1}{N} \sum_{i=1}^N x_i u_i$$

should give an average value of zero as the sample size gets larger (tends to infinity) **if** the Gauss Markov assumption about the X variables and the residual being uncorrelated is true.

So

$$p \lim \left(\frac{X'u}{N} \right) = p \lim \left(\frac{1}{N} \sum_{i=1}^N x_i u_i \right) = 0$$

Hence in (1)

$$p \lim(\hat{\beta}) = \beta$$

and OLS said to be a consistent estimator of β

If this assumption is violated then OLS will be inconsistent

(effectively the residual term becomes a function of X so that

$$y = bx + u(x)$$

hence

$$dy/dx = b + du/dx$$

so it is difficult to distinguish between the contribution of the two channels of influence of x)

Errors in Variables

One common situation where the assumption of no endogeneity may be violated is when one or more of the X variables is measured with error

Simplest to illustrate this using the 2 variable model before generalising

i) Dependent Variable Measured with Error

$$\text{True:} \quad y^{\text{true}} = bx + u \quad (1)$$

$$\text{Observe:} \quad y = y^{\text{true}} + e \quad (2)$$

ie dependent variable measured with error e

e is a random residual term just like u, so $e \sim N(0, \sigma_e^2)$

and errors in measurement are assumed to be uncorrelated with other residuals

Sub. (2) into (1)

$$\begin{aligned} y - e &= bx + u \\ y &= bx + u + e \\ y &= bx + v \quad \text{where } v = u + e \end{aligned} \quad (3)$$

Ok to estimate (3) by OLS, since

$$\begin{aligned} E(u) &= E(e) = 0 \\ \text{Cov}(x, u) &= \text{Cov}(x, e) = 0 \end{aligned}$$

(nothing to suggest values of the X variable correlated with meas. error in dependent variable)

So OLS estimates are unbiased in this case

But

standard errors are larger than would be in absence of meas. error

$$\text{Var}(v) = \text{Var}(u + e) = \text{Var}(u) + \text{Var}(e) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$$

the residual variance in presence of measurement error in dependent variable now also contains an additional contribution from error in y variable, σ_e^2

and so t, F values smaller than should be leading to Type II error (accept false null)

Measurement Error in Explanatory Variable

$$\text{True:} \quad y^{\text{true}} = bx^{\text{true}} + u \quad (1)$$

$$\text{Observe:} \quad x = x^{\text{true}} + w \quad (2)$$

ie rhs var. measured with error (w)

sub. (2) into (1)

$$\begin{aligned} y^{\text{true}} &= b(x-w) + u \\ y^{\text{true}} &= bx - bw + u \\ y^{\text{true}} &= bx + v \end{aligned} \quad (3)$$

where now $v = -bw + u$

(so residual term again consists of 2 components)

OLS on (3) gives

$$\begin{aligned} \hat{\beta} &= \frac{\sum xy^{\text{true}}}{\sum x^2} \\ &= \frac{\sum x(bx^{\text{true}} + u)}{\sum x^2} = b \frac{\sum xx^{\text{true}}}{\sum x^2} + \frac{\sum xu}{\sum x^2} \end{aligned} \quad (4)$$

Interested in the asymptotic properties of this estimator so take probability limit.

Assuming u , x^{true} and w are independent (so that level of X gives no indication about the size/sign of the error in measurement) then

$$p \lim \left(\frac{1}{N} \right) \sum xu = p \lim \left(\frac{1}{N} \right) \sum (x^{\text{true}} + w)u = 0$$

so 2nd term in (4) vanishes

but

$$p \lim \left(\frac{1}{N} \right) \sum x^2 = p \lim \left(\frac{1}{N} \right) \sum (x^{\text{true}} + w)^2 = p \lim \left(\frac{1}{N} \right) \sum x^{\text{true}^2} + p \lim \left(\frac{1}{N} \right) \sum w^2 = \sigma_{x^{\text{true}}}^2 + \sigma_w^2 \neq 0$$

$$p \lim \left(\frac{1}{N} \right) \sum x x' = p \lim \left(\frac{1}{N} \right) \sum (x' + w) x' = p \lim \left(\frac{1}{N} \right) \sum x'^2 = \sigma_{x'}^2$$

so

$$p \lim(\hat{\beta}) = b \left[\frac{\sigma_{x'}^2}{\sigma_{x'}^2 + \sigma_w^2} \right] \neq b$$

or

$$p \lim(\hat{\beta}) = b \left[1 - \frac{\sigma_w^2}{\sigma_{x'}^2 + \sigma_w^2} \right] = \frac{b}{\left(1 + \frac{\sigma_w^2}{\sigma_{x'}^2} \right)}$$

If $b > 0$ then $\hat{b} < b$

If $b < 0$ then $\hat{b} > b$

so that measurement error in the right hand side variable means that the OLS estimates will be inconsistent and suffer from **attenuation bias** (closer to zero in absolute values)

The ratio

$$\left[1 - \frac{\sigma_w^2}{\sigma_{x'}^2 + \sigma_w^2} \right] = \left[\frac{\sigma_{x'}^2}{\sigma_{x'}^2 + \sigma_w^2} \right]$$

is called the reliability ratio (the proportion of variation in the unobserved true variable that is accounted for by the variation in the observed x variable)

and the ratio $\frac{\sigma_w^2}{\sigma_{x'}^2}$ is called the noise to signal ratio

An increase in reliability (fall in noise-signal) means a fall in measurement error and the OLS estimates are closer to their true values.

Generalising to the k variable model

$$\text{True:} \quad y^{\text{true}} = X^{\text{true}}\beta + u \quad (1)$$

$$\text{Observe:} \quad X = X^{\text{true}} + w \quad (2)$$

Where now w is an $N \times k$ matrix of measurement errors and the i^{th} row of w corresponds to the measurement errors on all the X variables associated with the i^{th} observation.

Consistency of the OLS estimator depends on the behaviour of

$$p \lim(\hat{\beta}) = p \lim \left[\left(\frac{X'X}{N} \right)^{-1} \left(\frac{X'y}{N} \right) \right]$$

$$p \lim(\hat{\beta}) = p \lim \left[\left(\frac{(X^t + w)'(X^t + w)}{N} \right)^{-1} \left(\frac{(X^t + w)'(X^t\beta + u)}{N} \right) \right]$$

Since x^t , w and u are independent can ignore cross-products, so

$$p \lim(\hat{\beta}) = p \lim \left[\left(\frac{(X^{t'} X^t + w'w)}{N} \right)^{-1} \left(\frac{(X^{t'} X^t \beta)}{N} \right) \right]$$

$$= [Q_x^t + \Lambda_w]^{-1} Q_x^t \beta$$

$$= \beta - [Q_x^t + \Lambda_w]^{-1} \Lambda_w \beta$$

which is a mixture of all the parameters in the model

NB 1. Can show that this general result also holds when the mis-measured X is a dummy variable

NB 2. In the simple variable model can obtain a bound for the true value of β

$$p \lim(\hat{\beta}) < \beta < p \lim(1/\hat{\delta})$$

where δ is the coefficient on y in a reverse regression of x on y