13 Errors in variables. Endogeneity. IV

13.0.1 Classical measurement error

Suppose we observe noisy versions of the variables we would like to observe. We obtain data y_i and x_i for $i = 1, ..., n$ while the true values are y_i^* and x_i^*. Also assume:

$$x_i = x_i^* + \nu_i$$

$$y_i = y_i^* + \eta_i$$

where:

$$E(\nu_i) = E(\eta_i) = 0$$
$$E(x_i^* \nu_i) = 0, E(y_i^* \eta_i) = 0$$
$$E(x_i^* \eta_i) = 0, E(y_i^* \nu_i) = 0$$
$$E(\nu_i \eta_i) = 0$$

Given that $E(y_i^*|x_i^*) = x_i^* \beta$, if we proceeded as if there were no measurement error we would estimate the following by OLS:

$$\hat{\beta}_{OLS} = (X'X)^{-1} X'Y$$
$$= (\frac{1}{n} \sum_{i=1}^{n} x_i x_i')^{-1} \frac{1}{n} \sum_{i=1}^{n} x_i y_i'$$
$$= (\frac{1}{n} \sum_{i=1}^{n} (x_i^* + \nu_i)(x_i^* + \nu_i)')^{-1} \frac{1}{n} \sum_{i=1}^{n} (x_i^* + \nu_i)(y_i^* + \eta_i)'$$
$$= \left(\frac{1}{n} \sum_{i=1}^{n} \underset{\downarrow p}{x_i^* x_i'^*} + \underset{\downarrow p}{x_i^* \nu_i'} + \underset{\downarrow p}{\nu_i x_i'^*} + \underset{\downarrow p}{\nu_i \nu_i'} \right)^{-1} \frac{1}{n} \sum_{i=1}^{n} \underset{\downarrow p}{x_i^* y_i'^*} + \underset{\downarrow p}{x_i^* \eta_i'} + \underset{\downarrow p}{y_i^* \nu_i'} + \underset{\downarrow p}{\nu_i \eta_i'}$$
$$E(x_i^* x_i'^*)\ E(x_i^* \nu_i')\ E(\nu_i x_i'^*)\ E(\nu_i \nu_i') \qquad E(x_i^* y_i'^*)\ E(x_i^* \eta_i')\ E(y_i^* \nu_i')\ E(\nu_i \eta_i')$$

Lemma 13.0.1. Suppose we have sequences of random variables X_n and Y_n such that $X_n \xrightarrow{p} X$ and $Y_n \xrightarrow{p} Y$. Then $X_n Y_n \xrightarrow{p} XY$.

Proof in Appendix

Thus as plim of a sum is the sum of plims, and $1/x$ is a continuous function, we can use the CMT to argue:

$$\xrightarrow{p} (E x_i^* x_i'^* + E \nu_i \nu_i')^{-1} (E(x_i^* (x_i'^* \beta + \varepsilon_i)))$$
$$= (E x_i^* x_i'^* + E \nu_i \nu_i')^{-1} (E x_i^* x_i'^*) \beta \neq \beta$$

We have here an asymmetric bias. In the multiple regression case, the bias will depend on the interaction between explanatory variables.

In the univariate case the above reduces to :

$$\hat{\beta}_{OLS} \xrightarrow{p} \frac{\sigma^2_{x^*}}{\sigma^2_{x^*} + \sigma^2_{\nu}}\beta$$

This represents an attenuation bias of $\hat{\beta}_{OLS}$ toward zero.

13.0.2 Endogeneity with Errors

Any measurement error in the dependent variable is subsumed in the error term as follows:

$$y_i = y_i^* + \eta_i = x_i^*\beta + (\varepsilon_i + \eta_i)$$

The new error term $(\varepsilon_i + \eta_i)$ creates no problem for estimation as η_i is uncorrelated with x_i^*. However if there were measurement errors in x_i^* then we have:

$$y_i = x_i\beta + \underbrace{(\varepsilon_i + \eta_i - \nu_i'\beta)}_{u_i}$$

In this case, u_i is correlated with x_i

$$E(x_i u_i') = E(x_i u_i) = E(x_i(\varepsilon_i + \eta_i - \nu_i'\beta)) = E(x_i\varepsilon_i) + E(x_i\eta_i) - E(x_i\nu_i'\beta_i)$$

$$= -E(x_i\nu_i')\beta = -E(x_i^* + \nu_i)\nu_i'\beta$$

$$= \cancel{E(x_i^*\nu_i')\beta} + E(\nu_i\nu_i')\beta \neq \vec{0}$$

Thus as error term is correlated with x_i, (OLS2') $Ex_iu_i = 0$ does not hold, and OLS is inconsistent. Two solutions:

Solution 1:
If we can estimate $E(\nu_i\nu_i')$ we can undo the error in estimation by using the following:

$$E[x_i x_i'] = E[x_i^* x_i'^*] + E[\nu_i\nu_i']$$

But this is not usually possible.
Solution 2:
Suppose we get another independent measure of $x*$ such that

$$w_i = x_i^* + \tau_i$$

where τ_i is uncorrelated with any of $y_i^*, x_i^*, \eta_i, \nu_i$.

$$E[w_i x_i'] = E[(x_i^* + \tau_i)(x_i^* + \nu_i)'] = E[x_i^* x_i'^*]$$

$$E[w_i y_i] = E[(x_i^* + \tau_i)(y_i^* + \eta_i)] = E[x_i^* y_i^*]$$

Then if $E[w_i x_i']$ is invertible:

$$E[w_i x_i']^{-1}E[w_i y_i] = E[x_i^* x_i'^*]^{-1}E[x_i^* y_i^*] = [x_i^* x_i'^*]^{-1}E[x_i^* x_i'^*]\beta = \beta$$

So $\hat{\beta}_{IV} = (W'X)^{-1}W'Y$ is consistent for β.

This can also be derived directly by multiplying w_i to the estimating equation and taking expectations:

$$w_i y_i = w_i x_i'\beta + w_i e_i$$

$$E[w_i y_i] = E[w_i x_i']\beta + \cancel{E[(x_i^* + \tau_i)(\varepsilon_i + \eta_i - \nu_i'\beta)]}$$

$$\beta = E[w_i x_i']^{-1} E[w_i y_i]$$

$$\Rightarrow \hat{\beta}_{IV} = (\sum_{i=1}^{n} w_i x_i')^{-1} \sum_{i=1}^{n} w_i y_i = (W'X)^{-1} W'Y$$

where

$$w = \begin{pmatrix} w_1' \\ \vdots \\ w_n' \end{pmatrix}, X = \begin{pmatrix} x_1' \\ \vdots \\ x_n' \end{pmatrix}$$

13.1 Endogeneity

Consider the following model

$$y_i = x_i'\beta + \varepsilon_i$$

where $E(\varepsilon_i|x_i) \neq 0$ in violation of OLS2 and OLS2'. To distinguish this from the *regression*, we call the above equation a structural equation and β a structural parameter. A structural equation represents a causal link, rather than just an empirical association.

When $E(\varepsilon_i|x_i) \neq 0$, we say that x_i is endogenous. When this occurs, usually this is only by a few components of x_i being correlated with ε_i. The components causing this are referred to as endogenous and the rest exogenous. We then can partition x_i into the exogenous part x_{1i} and the endogenous part x_{2i} by rearranging.

The endogeneity problem may not only be caused through measurement error, but also joint determination, reverse causality or omitted variable bias.

13.1.1 Joint Determination: Supply and Demand

Consider the following model where q_i and p_i are determined jointly by the demand equation

$$\text{Demand: } q_i = -\beta_d p_i + \varepsilon_{di}$$

$$\text{Supply: } q_i = -\beta_s p_i + \varepsilon_{si}$$

In matrix notation:

$$\begin{pmatrix} 1 & \beta_d \\ 1 & \beta_s \end{pmatrix} \begin{pmatrix} q_i \\ p_i \end{pmatrix} = \begin{pmatrix} \varepsilon_{di} \\ \varepsilon_{si} \end{pmatrix}$$

$$\begin{pmatrix} q_i \\ p_i \end{pmatrix} = \frac{1}{-\beta_s - \beta_d} \begin{pmatrix} -\beta_s & -\beta_d \\ -1 & 1 \end{pmatrix} \begin{pmatrix} \varepsilon_{di} \\ \varepsilon_{si} \end{pmatrix}$$

$$= \begin{pmatrix} (\beta_s \varepsilon_{di} + \beta_d \varepsilon_{si})/(\beta_s + \beta_d) \\ (\varepsilon_{si} - \varepsilon_{di})/(\beta_s + \beta_d) \end{pmatrix}$$

Thus neither $E[p_i \varepsilon_{di}]$ nor $E[p_i \varepsilon_{si}]$ are zero, so p_i is endogenous.

Running an OLS of q_i on p_i will give a biased estimate of β_d. We estimate $Cov(q_i, p_i)/Var(p_i)$, and assuming demanad and supply shocks uncorrelated:

$$Cov(q_i, p_i)/Var(p_i) = \frac{\frac{\beta_s}{(\beta_s+\beta_d)^2} Var(\varepsilon_{di}) - \frac{\beta_d}{(\beta_s+\beta_d)^2} Var(\varepsilon_{si})}{\frac{1}{(\beta_s+\beta_d)^2} Var(\varepsilon_{di}) - \frac{1}{(\beta_s+\beta_d)^2} Var(\varepsilon_{si})}$$

$$= \beta_s \frac{Var(\varepsilon_{di})}{Var(\varepsilon_{di}) - Var(\varepsilon_{si})} - \beta_d \frac{Var(\varepsilon_{si})}{Var(\varepsilon_{di}) - Var(\varepsilon_{si})}$$

That is, some linear combination of the slopes of the demand and supply curves.

13.1.2 Omitted Variables

Another example of endogeneity would be a structural equation connecting two variables that are both chosen by economics agents, say, wage and educaiton

$$wage_i = \beta_1 + \beta_2 educ_i + \varepsilon_i$$

Both $wage_i$ and $educ_i$ may be affected by person i's ability or some other factor belonging to ε_i. Here the structural equation is thought of reflecting a causal relationship, which would be observed if we could randomly assign education levels to people independent of ability or anything else. In reality as this does not occur we cannot rule out that this choice may have been affected by other factors influencing wage.

13.2 Instrumental Variables

We formalise the device used in Solution 2 of the measurement error. Consider a linear regression model:

$$y_i = x_i'\beta + \varepsilon_i$$

where $E(\varepsilon_i|x_i) = 0$ and so x_i is endogenous. Suppose we have a variable w_i such that:

$$E[w_i\varepsilon_i] = 0 \text{ (exogeneity)}$$

$$E[w_i w_i'] > 0 \text{ (no redundant instruments, =non singular?)}$$

$$E[w_i x_i'] \text{ has full column rank (relevance)}$$

Then

$$0 = E[w_i\varepsilon_i] = E[w_i(y_i - x_i'\beta)] = E[w_i y_i] - E[w_i x_i']\beta$$
$$\Rightarrow \beta = E[w_i x_i']^{-1} E[w_i y_i]$$

This motivates $\hat{\beta}_{IV}$

$$\hat{\beta}_{IV} = (W'X)^{-1}W'Y = \left(\frac{1}{n}\sum_{i=1}^{n} w_i x_i'\right)^{-1} \frac{1}{n}\sum_{i=1}^{n} w_i y_i$$

$$\xrightarrow{p} E[w_i x_i']^{-1} E[w_i y_i] = E[w_i x_i']^{-1} E[w_i x_i']\beta + E[w_i\varepsilon_i] = \beta$$

Note:-

Understanding the Relevance Condition
Suppose we start with the regression model:

$$y_i = x_{1i}'\beta_1 + x_{2i}'\beta_2 + \varepsilon_i$$

where x_{2i} is a scalar endogenous variable, and x_{1i} is a $(k-1)\times 1$ vector of exogenous variables.

We define w_i as a vector of instruments, where exogenous variables instrument themselves and z_i is a scalar IV for x_{2i}

Given $E[w_i w_i']$ nonsingular

$$E[w_i x_i'] \text{ full column rank} \Leftrightarrow E(w_i w_i')^{-1} E[w_i x_i'] \text{ full column rank}$$

But this represents the set of population coefficients in a regression model of x_i on w_i.

$$\because E(w_i w_i')^{-1} E[w_i x_i'] = E(w_i w_i')^{-1} E[w_i'(x_{(1,1i)} \dots x_{(k-1,1i)}\, x_{2i})]$$

$$= (\vec{\beta}_{1,1} \dots \vec{\beta}_{k-1,1} \dots \vec{\beta}_2)$$

Consider $\vec{\beta}_{1,1}$:
This represents the coefficient of $x_{1,1i}$ in the regression of $x_{1,1i}$ on w_i. But clearly as $x_{1,1i}$ is included as a regressor in w_i, $\vec{\beta}_{1,1} = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$.

For $\vec{\beta}_2$:
This represents β in the regression model:

$$x_{2i} = w_i'\vec{\beta} + \nu_i = \pi_1 x_{1,1i} + \dots + \pi_{k-1} x_{k-1,1i} + \pi_k z_i + \nu_i$$

Therefore:

$$E(w_i w_i')^{-1} E[w_i x_i'] = \begin{pmatrix} I_{k-1} & \vec{\pi} \\ \vec{0}' & \pi_k \end{pmatrix}$$

where $\vec{\pi}$ is $(\pi_1 \dots \pi_{k-1})'$

Clearly then for this to be full rank (implying the relevance condition), **we need** $\pi_k \neq 0$.
This means that the instrument z_i is correlated with x_{2i}, even after the effects of all the other exogenous variables have been controlled for.

Example. Demand and Supply shifters
Suppose the market is a local fish market as in Graddy(1985). We may think supply would be affected by weather offshore w_i, so that:

$$q_i = -\beta_s p_i + \gamma w_i + \varepsilon_{si}$$

whereas demand will not be directly affected by w_i so that:

$$q_i = \beta_d p_i + \varepsilon_{di}$$

Then we can use w_i as an instrument for p_i in the estimation of the demand (but not supply) equation. As these two equations need to equal, we can guarantee the relevance condition for $\gamma \neq 0$, as we can express p_i as a function of w_i.

Example. Education and Wage
Angrist and Krueger (1991) propose the quarter of birth indicator as the instrument for education. Due to compulsory education laws in the United States, you cannot drop out from school until you are 16, so people who are born in the first quarter of the year, being oldest in their class, may drop out more often than those born in the other quarters. This insures that $E[w_i x_i] \neq 0$, whereas, arguably, the quarter of birth should not be related to any other determinants of your wage, so that $E[w_i \varepsilon_i] = 0$.

13.3 Appendix

Proof. Approach 1 (CMT only):

$$X_n \xrightarrow{p} X \text{ and } Y_n \xrightarrow{p} Y \Leftrightarrow (X_n, Y_n) \xrightarrow{p} (X, Y)$$

Define a continuous function:

$$f(x, y) = xy$$

Then by the CMT (applied to vector):

$$f(X_n, Y_n) \xrightarrow{p} f(X, Y)$$

$$\Rightarrow X_n Y_n \xrightarrow{p} XY$$

Approach 2 (Slutsky + CMT):

$$X_n \xrightarrow{p} X \Rightarrow X_n \xrightarrow{d} X$$

$$Y_n \xrightarrow{p} X$$

By Slutsky's theorem:

$$X_n Y_n \xrightarrow{d} XY$$

$$\Rightarrow X_n Y_n \xrightarrow{p} XY$$

(this implication only holds for RHS constant) □