Endogeneity

Tom Smith
What is Endogeneity?

• Classic Problem in Econometrics:
  – More police officers might reduce crime but cities with higher crime rates might demand more police officers.
  – More diffuse ownership might affect performance but firms with strong performance might attract diffuse ownership.
  – Banks with adequate capital reserves might have good performance but good performing Banks might have stronger capital reserves.
  – Athletes might develop lactic acid as their performance deteriorates but lactic acid might be facilitating the muscles to aid performance.

• In all cases standard regression analysis might confound these 2 effects.
What is Endogeneity?

• Endogeneity problem is similar to the errors in variables problem.
• Since most of you will be familiar with the errors in variables approach lets review this simpler problem
• Then we can draw out the parts that apply also to the endogeneity problem
Measurement Error Problem

• Lets take the following equation

\[ Y = \alpha + \beta X + \mu \]

where: \( X^* \) = True value

\[ X = X^* + \varepsilon \]

\( X \) Measures \( X^* \) with error

Thus the \( X \) variable is not independent of the error

• OLS minimizes the sum of squared errors

\[ L = E[(Y - \alpha - \beta X)^2] \]

• Which has first order conditions

\[ \frac{\partial L}{\partial \alpha} \rightarrow E[Y - \alpha - \beta X] = 0 \]

\[ \frac{\partial L}{\partial \beta} \rightarrow E[Y - \alpha - \beta X]X = 0 \]

known as the normal equations of OLS
Measurement Error Problem

\[
\frac{\partial L}{\partial \alpha} \rightarrow E[Y - \alpha - \beta X] = 0 \\
\frac{\partial L}{\partial \beta} \rightarrow E[Y - \alpha - \beta X] X = 0 \\
\rightarrow \beta = \frac{Cov(Y, X)}{Var(X)}
\]

Since X is measured with error \( X = X^* + \varepsilon \)

\[
\beta = \frac{Cov(Y, X^* + \varepsilon)}{Var(X^* + \varepsilon)} = \frac{Cov(Y, X^*)}{Var(X^*) + Var(\varepsilon)}
\]

The True Beta is \( \beta^* = \frac{Cov(Y, X^*)}{Var(X^*)} \)

So the calculated Beta is a biased and inconsistent estimator of the true Beta.

This comes about because the X variable is not independent of the error.
Measurement Error Problem

• Solution is to use instrument variable (IV): (z)
• Properties of (z)
  – (1) Uncorrelated with the error (i.e. exogenous)
  – (2) Relevant to $X^*$ (i.e. non zero correlation)

• The IV estimator is: \[ \beta = \frac{\text{Cov}(Y, Z)}{\text{Cov}(X, Z)} \]
Think of this as a two Stage Procedure

- **1st Stage:** \( X = a + b \, z \Rightarrow b = \frac{\text{cov}(X, z)}{\text{var}(z)} \)
  
  Define: \( \hat{X} = a + b \, z \)

- **2nd Stage:** \( Y = \alpha + \beta \hat{X} \)

\[
\Rightarrow \beta = \frac{\text{Cov}(Y, \hat{X})}{\text{Var}(\hat{X})} = \frac{\text{Cov}(Y, a + b \, z)}{\text{Var}(a + b \, z)} = \frac{b \, \text{Cov}(Y, z)}{b^2 \, \text{Var}(z)} = \frac{\text{Cov}(Y, z)}{b \, \text{Var}(z)}
\]

substitute \( b = \frac{\text{cov}(X, z)}{\text{Var}(z)} \)

\[
\Rightarrow \beta = \frac{\text{Cov}(Y, z)}{\frac{\text{Cov}(X, z)}{\text{Var}(z)} \, \text{Var}(z)} = \frac{\text{Cov}(Y, z)}{\text{Cov}(X, z)}
\]

Two Stage Least Squares gives us the familiar measurement error IV estimator
Instrumental Variables (IV)

• Lets think of normal equations of OLS

\[
\begin{align*}
E(Y - \alpha - \beta X) &= 0 \\
E(Y - \alpha - \beta X) X &= 0
\end{align*}
\] \rightarrow \beta = \frac{\text{Cov}(Y, X)}{\text{Var}(X)}

• If we replace \( X \) with \( z \) in the 2\textsuperscript{nd} equation

\[
\begin{align*}
E(Y - \alpha - \beta X) &= 0 \\
E(Y - \alpha - \beta X) z &= 0
\end{align*}
\] \rightarrow \beta = \frac{\text{Cov}(Y, z)}{\text{Cov}(X, z)}

• This gives us another way to think of the IV estimator
Endogeneity

• Think of a specific example

• The Simultaneous Equations are:

\[ Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 + u_1 \quad (1) \]
\[ Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_2 + u_2 \quad (2) \]

Where the Y’s are endogenous variables; the X’s are exogenous variables.

The problem is similar to the measurement error issue - \( Y \) is correlated with \( u \) since as you can see from eqn (2) \( Y_2 \) contains error which then induces in eqn 1 a problem similar to measurement error with similar consequences: -- ie biased and inconsistent estimators
Two Stage Least Squares (2SLS)

- Consider estimating (1) with 2SLS

  (1) 1st Stage: Regress $Y_2$ on exogenous variables
  
  $Y_2 = a_0 + a_1 X_1 + a_2 X_2 + \varepsilon_2$
  
  Let $\hat{Y}_2 = a_0 + a_1 X_1 + a_2 X_2$

  (2) 2nd Stage: $Y_1 = \alpha_0 + \alpha_1 \hat{Y}_2 + \alpha_2 X_1$

- Consider estimating (2) with 2SLS

  (1) 1st Stage: $Y_1 = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon_1$
  
  Let $\hat{Y}_1 = b_0 + b_1 X_1 + b_2 X_2$

  (2) 2nd Stage:
  
  $Y_2 = \beta_0 + \beta_1 \hat{Y}_1 + \beta_2 X_2$
Let’s go back to the normal equation of OLS:

\[
E \begin{bmatrix}
Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X_1 \\
(Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X_1)(Y_2) \\
(Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X_1)X_1 \\
Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 & X_2 \\
(Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 & X_2)(Y_1) \\
(Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 & X_2)X_2 \\
\end{bmatrix} = 0
\]

Where the endogenous variables are in brackets
Instrumental Variables

• To get IV estimators simply replace the endogenous variables $Y_1$ and $Y_2$ with instruments $(X_1, X_2)$

$$E \begin{bmatrix}
Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)(X_2) \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)X_1 \\
Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_2 \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_2)(X_1) \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_2)X_2
\end{bmatrix} = 0$$

• Doing the above is a good way to see if the system is identified.
• Textbooks spend many pages on rank and order conditions but basically all you need is to be able to write out the IV system as above and have a unique equation for every parameter.
Instrumental Variables - Identification

- Here is an example where you don’t have identification
  
  $Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1$
  
  $Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1$

- Normal equations:

\[
\begin{pmatrix}
Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1 \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)(Y_2) \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)X_1 \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1) \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1)(Y_1) \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1)X_1
\end{pmatrix}
\begin{pmatrix}
E
\end{pmatrix} = 0
\]
Here is an example where you don’t have identification

\[ Y_1 = \alpha_0 + \alpha_1 Y_2 + \alpha_2 X_1 \]
\[ Y_2 = \beta_0 + \beta_1 Y_1 + \beta_2 X_1 \]

Instrumental Variables:

\[
E \begin{bmatrix}
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1) \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)(?) \\
(Y_1 - \alpha_0 - \alpha_1 Y_2 - \alpha_2 X_1)X_1 \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1) \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1)(?) \\
(Y_2 - \beta_0 - \beta_1 Y_1 - \beta_2 X_1)X_1
\end{bmatrix} = 0
\]

System is not identified as these 2 eqns don't have an instrument
Differenced GMM

- Graham (JFE 1996) looked at the difference in leverage as the dependant variable. The logical extension is to look at the whole equation in changes and add a lagged Y variable to take account of the dynamics of the relationship.

\[
Y_{1t} = \alpha_0 + \alpha_1 Y_{t-1} + \alpha_2 Y_{2t} + \alpha_3 X_{1t} + u_{1t} \quad (1)
\]

\[
Y_{2t} = \beta_0 + \beta_1 Y_{2t-1} + \beta_2 Y_{1t} + \beta_3 X_{2t} + u_{2t} \quad (2)
\]

- Taking first difference of equation (1) gives:

\[
\Delta Y_{1t} = \alpha_1 \Delta Y_{1t-1} + \alpha_2 \Delta Y_{2t} + \alpha_3 \Delta X_{1t} + \Delta u_{1t}
\]

- Note that the intercept term cancels and also that the error term will now be correlated.
Differenced GMM

\[ \Delta Y_{1t} = \alpha_1 \Delta Y_{1t-1} + \alpha_2 \Delta Y_{2t} + \alpha_3 \Delta X_{1t} + \Delta u_{1t} \]

• The estimated moment conditions are:

\[
\begin{bmatrix}
\Delta Y_{1t} - \alpha_1 \Delta Y_{1t-1} - \alpha_2 \Delta Y_{2t} - \alpha_3 \Delta X_{1t} \\
E\left( ( \Delta Y_{1t} - \alpha_1 \Delta Y_{1t-1} - \alpha_2 \Delta Y_{2t} - \alpha_3 \Delta X_{1t} )Y_{2t-2} \right) \\
E\left( ( \Delta Y_{1t} - \alpha_1 \Delta Y_{1t-1} - \alpha_2 \Delta Y_{2t} - \alpha_3 \Delta X_{1t} )X_{1t-2} \right)
\end{bmatrix} = 0
\]

The instruments are at t-2 as the change in the X variables are from t-1 to t
System GMM

- The above treats all of the explanatory variables as endogenous and is known as Differenced GMM. It can be run in STATA as a Dynamic Panel. STATA also lets you run System GMM which is the equation in difference form, instrumented in lagged levels and the equation in levels, instrumented with lagged differences. So this second equation would appear as follows:

\[
\begin{align*}
E \left[ Y_{1t} - \alpha_1 Y_{1t-1} - \alpha_2 Y_{2t} - \alpha_3 X_{1t} \\
( Y_{1t} - \alpha_1 Y_{1t-1} - \alpha_2 Y_{2t} - \alpha_3 X_{1t}) \Delta Y_{2t-2} \right] = 0 \\
( Y_{1t} - \alpha_1 Y_{1t-1} - \alpha_2 Y_{2t} - \alpha_3 X_{1t}) \Delta X_{1t-2} \right]
\]

- The above are known as the Arellano and Bond estimators. Stephen Bond has an excellent site on the use of Differenced GMM and System GMM. See his course on panel data econometrics at [www.nuffield.ox.ac.uk/users/bond](http://www.nuffield.ox.ac.uk/users/bond)

- Good Example of application in Corporate Governance is Wintoki et al “Endogeneity and the Dynamics of Corporate Governance” SSRN
Specification Tests

- Endogeneity Test
  - Hausman-Wu
- Instruments correlated with Endogeneous Variables
  - F test
- Instruments uncorrelated with error
  - Overidentifying Chi-squared test
How to Test for Endogeneity

• Perform Hausman-Wu test of Endogeneity using the following test:
  \[ H_0 : \beta_{OLS} \text{ is efficient and consistent} \]
  \[ H_1 : \beta_{IV} \text{ is consistent} \]

• This hypothesis is tested by the following statistics which has a \( \chi^2 \) distribution with degrees of freedom equal to the number of elements in \( \beta \)

\[
(\beta_{OLS} - \beta_{IV})^T [V_{IV} - V_{OLS}]^{-1} (\beta_{OLS} - \beta_{IV}) \mathrel{\pmod{\#\beta\text{ elements}}} \chi^2
\]
How to Test for Instruments correlated with endogenous variable

• Perform 1\textsuperscript{st} Stage Regression of Endogenous variable on Instruments:

\[ Y_1 = b_0 + b_1 X_1 + b_2 X_2 + \varepsilon_1 \]

• Check \( R^2 \) and F statistic of this regression
  – Rule of Thumb F > 10 Staiger Stock (1997)
How to Test for Instruments uncorrelated with error

• Overidentifying J test is a Chi-squared test of whether the instruments are uncorrelated with the error

\[
\begin{bmatrix}
Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X \\
(Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X_1)(X_2) \\
(Y_1 - \alpha_0 - \alpha_1 & Y_2 - \alpha_2 & X_1)X_1 \\
Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 & X_2 \\
(Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 X_2)(X_1) \\
(Y_2 - \beta_0 - \beta_1 & Y_1 - \beta_2 X_2)X_2
\end{bmatrix} = 0
\]
Coming up with Instruments

- Use exogenous variables that appear elsewhere in the system
- Used lagged values of the endogenous variable
- Come up with economically meaningful instruments
  - Random Lottery draw used in determining effect of serving in Vietnam war on future earning prospects – see Angrist (AER, 1990)
  - Sex of first born used in determining effect of family successions on firm performance – see Bennedsen, Nielsen, et al (QJE, 2007)
  - Top Managers prior Military service as a proxy for Disclosure Quality – see Smith Bamber (Accounting Review 2010)
  - Marginal Tax rate as a proxy for leverage in an examination of the effects of default probability -- see Molina (JF 2005)
  - Education as an instrument in determining the effect of family successions on firm value – see Perez-Gonzalez (AER 2006)
Endogeneity Example

Each Firm optimises Ownership/performance relation

Performance

Insider Ownership

A
B
C

3% 5% 7%
Endogeneity Example

Performance

Insider Ownership

Equilibrium for each firm

A

3%

5%

7%

B

C
A Senate committee requires a study to examine the ownership-performance relation.
Endogeneity Example

A Senate committee requires a study to examine ownership performance relation.
Endogeneity Example

Performance

Insider Ownership

A 3% 5% 7% B

Senate committee
Conclude 5% is optimum
Passes a law saying
Must have 5% ownership
Endogeneity Example

However at 5% ownership A and B are no longer at Optimum
Endogeneity Example

Comparison with Optimum
For A and B
Endogeneity Example

Arrows show the loss in Performance to A and B If they have 5% Ownership
**Endogeneity Example**

In contrast, IV estimation finds that there is no relation.
• Good References here:
  • Demsetz (Journal of Law and Economics, 1983)
  • Demsetz and Lehn (Journal of Political Economy, 1985)
  • Coles, Lemmon and Meschke (SSRN, 2007)
• Firms know that by having less ownership they increase agency issues but do so because there are additional benefits
• Cross sectional regression will find no relation because of endogeneity
• Much better to model the ownership/performance relation at the firm level
• Observe the managerial pay-performance sensitivity that maximizes performance
• Coles et al look at the productivity parameters for managerial input and investment that would give rise to observed level of ownership and investment
• Get very sensible results
  • Productivity of managerial input high in personal and business services and equipment industries (education, software, networking, computers)
  • Productivity of managerial input low in metal mining and utilities
  • Model Q values highly correlated with observed Q