Lecture 15. Endogeneity & Instrumental Variable Estimation

Saw that measurement error (on right hand side) means that OLS will be biased (biased toward zero)

Potential solution to endogeneity – **instrumental variable** estimation

- A variable that is correlated with the problem variable but which does not suffer from measurement error

Tests for endogeneity

Other sources of endogeneity

Problems with weak instruments

Idea of Instrumental Variables attributed to

Philip Wright 1861-1934

interested in working out whether price of butter was demand or supply driven

More formally, an instrument Z for the variable of concern X satisfies

1) $Cov(X,Z) \neq 0$

More formally, an instrument Z for the variable of concern X satisfies

1) $Cov(X,Z) \neq 0$

correlated with the problem variable

More formally, an instrument Z for the variable of concern X satisfies

1) $Cov(X,Z) \neq 0$

correlated with the problem variable

2) $Cov(Z,u) = 0$

More formally, an instrument Z for the variable of concern X satisfies

1) $Cov(X,Z) \neq 0$

correlated with the problem variable

2) $Cov(Z,u) = 0$

but uncorrelated with the residual (so does not suffer from measurement error and also is not correlated with any unobservable factors influencing the dependent variable)

Instrumental variable (IV) estimation proceeds as follows:

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1 X + u \qquad (1)$$

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply (1) by the instrument Z

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Follows that

$$Cov(Z,y) \quad = Cov[Zb_0 + b_1ZX + Zu]$$

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Follows that

$$\begin{aligned} Cov(Z,y) &= Cov[Zb_0 + b_1ZX + Zu] \\ &= Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u) \end{aligned}$$

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Follows that

$$\begin{aligned} Cov(Z,y) &= Cov[Zb_0 + b_1ZX + Zu] \\ &= Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u) \end{aligned}$$

since $Cov(Zb_0) = 0$ (using rules on covariance of a constant)

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Follows that

$$\begin{aligned} Cov(Z,y) &= Cov[Zb_0 + b_1ZX + Zu] \\ &= Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u) \end{aligned}$$

since $Cov(Zb_0) = 0$ (using rules on covariance of a constant)

and $Cov(Z,u) = 0$

(if assumption above about the properties of instruments is correct)

Instrumental variable (IV) estimation proceeds as follows:

Given a model

$$y = b_0 + b_1X + u \qquad (1)$$

Multiply by the instrument Z

$$Zy = Zb_0 + b_1ZX + Zu$$

Follows that

$$\begin{aligned} Cov(Z,y) &= Cov[Zb_0 + b_1ZX + Zu] \\ &= Cov(Zb_0) + Cov(b_1Z,X) + Cov(Z,u) \end{aligned}$$

since $Cov(Zb_0) = 0$ (using rules on covariance of a constant)

and $Cov(Z,u) = 0$

(if assumption above about the properties of instruments is correct) then $\quad Cov(Z,y) = 0 + b_1 Cov(Z,X) + 0$

Solving $Cov(Z,y) \quad = 0 + b_1 Cov(Z,X) + 0$ for b_1

gives the formula to calculate the instrumental variable estimator

Solving $Cov(Z,y) \quad = 0 + b_1 Cov(Z,X) + 0$

for b_1

gives the formula to calculate the instrumental variable estimator

So $b_1^{IV} = \dfrac{Cov(Z,y)}{Cov(Z,X)}$

Solving $Cov(Z,y) = 0 + b_1 Cov(Z,X) + 0$

for b_1

gives the formula to calculate the instrumental variable estimator

So $b_1^{IV} = \dfrac{Cov(Z,y)}{Cov(Z,X)}$ (compare with $b_1^{OLS} = \dfrac{Cov(X,y)}{Var(X)}$)

Solving Cov(Z,y) $= 0 + b_1 Cov(Z,X) + 0$

for b_1

gives the formula to calculate the instrumental variable estimator

So $b_1^{IV} = \dfrac{Cov(Z,y)}{Cov(Z,X)}$ (compare with $b_1^{OLS} = \dfrac{Cov(X,y)}{Var(X)}$)

In the presence of measurement error (or endogeneity in general) the IV estimate is **unbiased** in large samples (but may be biased in small samples)

- technically the IV estimator is said to be **consistent** –

Solving Cov(Z,y) $= 0 + b_1 Cov(Z,X) + 0$

for b_1

gives the formula to calculate the instrumental variable estimator

So $b_1^{IV} = \frac{Cov(Z,y)}{Cov(Z,X)}$ (compare with $b_1^{OLS} = \frac{Cov(X,y)}{Var(X)}$)

In the presence of measurement error (or endogeneity in general) the IV estimate is **unbiased** in large samples (but may be biased in small samples)

- technically the IV estimator is said to be **consistent** – while the OLS estimator is inconsistent *IN THE PRESENCE OF ENDOGENEITY*

which makes IV a useful estimation technique to employ

However can show that (in the 2 variable case) the variance of the IV estimator is given by

$$Var(\hat{\beta}_1^{IV}) = \frac{s^2}{N * Var(X)} * \frac{1}{r_{X\ Z}^2}$$

where r_{xz}^2 is the square of the correlation coefficient between endogenous variable and instrument

However can show that (in the 2 variable case) the variance of the IV estimator is given by

$$Var(\hat{\beta}_1^{IV}) = \frac{s^2}{N * Var(X)} * \frac{1}{r_{X\ Z}^2}$$

where r_{xz}^2 is the square of the correlation coefficient between endogenous variable and instrument

(compared with OLS $Var(\hat{\beta}_1^{OLS}) = \frac{s^2}{N * Var(X)}$)

However can show that (in the 2 variable case) the variance of the IV estimator is given by

$$Var(\hat{\beta}_1^{IV}) = \frac{s^2}{N * Var(X)} * \frac{1}{r^2_{X\ Z}}$$

where r_{xz}^2 is the square of the correlation coefficient between endogenous variable and instrument

(compared with OLS $Var(\hat{\beta}_1^{OLS}) = \frac{s^2}{N * Var(X)}$)

Since $r^2 > 0$

So IV estimation is less precise (efficient) than OLS estimation

May sometimes want to trade off bias against efficiency

So why not ensure that the correlation between X and the instrument Z is as high as possible?

So why not ensure that the correlation between X and the instrument Z is as high as possible?

 - if X and Z are perfectly correlated then Z must also be correlated with u and so suffer the same problems as X – the initial problem is not solved.

So why not ensure that the correlation between X and the instrument Z is as high as possible?

- if X and Z are perfectly correlated then Z must also be correlated with u and so suffer the same problems as X – the initial problem is not solved.

Conversely if the correlation between the endogenous variable and the instrument is small there are also problems

Since can always write the IV estimator as

$$b_1^{IV} = \frac{Cov(Z, y)}{Cov(Z, X)}$$

Since can always write the IV estimator as

$$b_1^{IV} = \frac{Cov(Z, y)}{Cov(Z, X)}$$

sub. in for $y = b_0 + b_1X + u$

Since can always write the IV estimator as

$$b_1^{IV} = \frac{Cov(Z,y)}{Cov(Z,X)}$$

sub. in for $y = b_0 + b_1X + u$

$$\frac{Cov(Z, b_0 + b_1X + u)}{Cov(Z,X)}$$

Since can always write the IV estimator as

$b_1^{IV} = \dfrac{Cov(Z,y)}{Cov(Z,X)}$

sub. in for $y = b_0 + b_1X + u$

$$b_1^{IV} = \frac{Cov(Z, b_0 + b_1X + u)}{Cov(Z,X)}$$

$$= \frac{Cov(Z,b_0) + b_1Cov(Z,X) + Cov(Z,u)}{Cov(Z,X)}$$

Since can always write the IV estimator as

$$b_1^{IV} = \frac{Cov(Z,y)}{Cov(Z,X)}$$

sub. in for $y = b_0 + b_1X + u$

$$b_1^{IV} = \frac{Cov(Z,b_0 + b_1X + u)}{Cov(Z,X)}$$

$$= \frac{Cov(Z,b_0) + b_1Cov(Z,X) + Cov(Z,u)}{Cov(Z,X)}$$

$$b_1^{IV} = \frac{0 + b_1Cov(Z,X) + Cov(Z,u)}{Cov(Z,X)}$$

So $b_1^{IV} = b_1 + \dfrac{Cov(Z,u)}{Cov(Z,X)}$

Since can always write the IV estimator as

$$b_1^{IV} = \frac{Cov(Z, y)}{Cov(Z, X)}$$

sub. in for $y = b_0 + b_1X + u$

$$b_1^{IV} = \frac{Cov(Z, b_0 + b_1X + u)}{Cov(Z, X)}$$

$$= \frac{Cov(Z, b_0) + b_1Cov(Z, X) + Cov(Z, u)}{Cov(Z, X)}$$

$$b_1^{IV} = \frac{0 + b_1Cov(Z, X) + Cov(Z, u)}{Cov(Z, X)}$$

So $b_1^{IV} = b_1 + \dfrac{Cov(Z, u)}{Cov(Z, X)}$

So if Cov(X,Z) is small then the IV estimate can be a long way from the true value b_1

So: always check extent of correlation between X and Z before any IV estimation (see later)

So: always check extent of correlation between X and Z before any IV estimation (see later)

In large samples you can have as many instruments as you like – though finding good ones is a different matter.

In small samples a minimum number of instruments is better (bias in small samples increases with no. of instruments).

Where to find good instruments?

Where to find good instruments?
- difficult

Where to find good instruments?

- difficult
- The appropriate instrument will vary depending on the issue under study.

In the case of measurement error, could use the *rank* of X as an instrument (ie order the variable X by size and use the number of the order rather than the actual vale.

In the case of measurement error, could use the *rank* of X as an instrument (ie order the variable X by size and use the number of the order rather than the actual vale.

Clearly correlated with the original value but because it is a rank should not be affected with measurement error

In the case of measurement error, could use the *rank* of X as an instrument (ie order the variable X by size and use the number of the order rather than the actual vale.

Clearly correlated with the original value but because it is a rank should not be affected with measurement error

- Though this assumes that the measurement error is not so large as to affect the (true) ordering of the X variable

```
   egen rankx=rank(x_obs)     /* stata command to create the ranking of x_observ  */

. list x_obs rankx
      x_observ      rankx
  1.        60          1
  2.        80          2
  3.       100          3
  4.       120          4
  5.       140          5
  6.       200          6
  7.       220          7
  8.       240          8
  9.       260          9
 10.       280         10
```

ranks from smallest observed x to largest
Now do instrumental variable estimates using rankx as the instrument for x_obs

```
ivreg y_t (x_ob=rankx)

Instrumental variables (2SLS) regression

      Source |       SS       df       MS              Number of obs =      10
-------------+------------------------------           F(  1,     8) =   84.44
       Model |  11654.5184     1  11654.5184           Prob > F      =  0.0000
    Residual |  1125.47895     8  140.684869           R-squared     =  0.9119
-------------+------------------------------           Adj R-squared =  0.9009
       Total |  12779.9974     9  1419.99971           Root MSE      =  11.861
------------------------------------------------------------------------------
      y_true |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    x_observ |    .460465   .0501086     9.19   0.000     .3449144    .5760156
       _cons |   48.72095   9.307667     5.23   0.001     27.25743    70.18447
------------------------------------------------------------------------------
Instrumented:  x_observ
Instruments:   rankx
------------------------------------------------------------------------------
```

Can see both estimated coefficients are a little closer to their true values than estimates from regression with measurement error (but not much)In this case the rank of X is not a very good instrumentNote that standard error in instrumented regression is larger than standard error in regression of y_true on x_observed as expected with IV estimation

Testing for Endogeneity

It is good practice to compare OLS and IV estimates. If estimates are very different this may be a sign that things are amiss.

Testing for Endogeneity

It is good practice to compare OLS and IV estimates. If estimates are very different this may be a sign that things are amiss.

Using the idea that IV estimation will always be (asymptotically) unbiased whereas OLS will only be unbiased if Cov(X,u) = 0 then can do the following:

Wu-Hausman Test for Endogeneity

Testing for Endogeneity

It is good practice to compare OLS and IV estimates. If estimates are very different this may be a sign that things are amiss.

Using the idea that IV estimation will always be (asymptotically) unbiased whereas OLS will only be unbiased if Cov(X,u) = 0 then can do the following:

Wu-Hausman Test for Endogeneity

1. Given $y = b_0 + b_1X + u$ (A)

   Regress the endogenous variable X on the instrument(s) Z

   $$X = d_0 + d_1Z + v \qquad (B)$$

Testing for Endogeneity

It is good practice to compare OLS and IV estimates. If estimates are very different this may be a sign that things are amiss.

Using the idea that IV estimation will always be (asymptotically) unbiased whereas OLS will only be unbiased if Cov(X,u) = 0 then can do the following:

Wu-Hausman Test for Endogeneity

1. Given $y = b_0 + b_1X + u$ (A)

   Regress the endogenous variable X on the instrument(s) Z

   $$X = d_0 + d_1Z + v \qquad \text{(B)}$$

   Save the residuals $\hat{v}$

2. Include this residual as an extra term in the original model

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Why ?

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$

Endogenous X = instrument + something else

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$ (A)

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$

Endogenous X = instrument + something else

and so only way X could be correlated with u in (A) is through v

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$ (A)

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u
If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$
Endogenous X = instrument + something else

and so only way X could be correlated with u in (A) is through v
(since Z is not correlated with u by assumption)

This means the residual u in (A) depends on v + some other residual

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$ (A)

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u
If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$

and so only way X could be correlated with u is through v

This means the residual in (A) depends on v + some other residual

$$u = b_2v + e$$

Include this residual as an extra term in the original model

ie given $y = b_0 + b_1X + u$ (A)

estimate

$$y = b_0 + b_1X + b_2\hat{v} + e \qquad \text{(B)}$$

and test whether $b_2 = 0$ (using a t test)

If $b_2 = 0$ conclude there is no correlation between X and u

If $b_2 \neq 0$ conclude there is correlation between X and u

Why ? because $X = d_0 + d_1Z + v$

and so only way X could be correlated with u is through v

This means the residual in (A) depends on v + some residual

$$u = b_2v + e$$

So estimate (B) instead and test whether coefficient on v is significant

$$y = b_0 + b_1 X + b_2 \hat{v} + e \qquad \text{(B)}$$

If it is, conclude that X and error term are indeed correlated;

there is endogeneity

N.B. This test is only as good as the instruments used and **is only valid asymptotically**. This may be a problem in small samples and so you should generally use this test only with sample sizes well above 100.

Example:

The data set *ivdat.dta* contains information on the number of GCSE passes of a sample of 16 year olds and the total income of the household in which they live.

Income tends to be measured with error. Individuals tend to mis-report incomes, particularly third-party incomes and non-labour income. The following regression may therefore be subject to measurement error in one of the right hand side variables, (the gender dummy variable is less subject to error).

```
. reg nqfede inc1 female

      Source |       SS       df       MS              Number of obs =     252
-------------+------------------------------           F(  2,   249) =   14.55
       Model |  274.029395     2  137.014698           Prob > F      =  0.0000
    Residual |   2344.9706   249  9.41755263           R-squared     =  0.1046
-------------+------------------------------           Adj R-squared =  0.0974
       Total |     2619.00   251  10.4342629           Root MSE      =  3.0688
------------------------------------------------------------------------------
      nqfede |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        inc1 |   .0396859   .0087786     4.52   0.000      .022396    .0569758
      female |   1.172351    .387686     3.02   0.003     .4087896    1.935913
       _cons |   4.929297   .4028493    12.24   0.000      4.13587    5.722723
```

To test endogeneity first regress the suspect variable on the instrument and any exogenous variables in the original regression

```
reg inc1 ranki female

       Source |       SS       df       MS             Number of obs =     252
-------------+------------------------------           F(  2,   249) =  247.94
       Model |  81379.4112     2  40689.7056           Prob > F      =  0.0000
    Residual |   40863.626   249  164.110948           R-squared     =  0.6657
-------------+------------------------------           Adj R-squared =  0.6630
       Total |  122243.037   251  487.024053           Root MSE      =  12.811
------------------------------------------------------------------------------
        inc1 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       ranki |   .2470712   .0110979    22.26   0.000     .2252136    .2689289
      female |   .2342779   1.618777     0.14   0.885    -2.953962    3.422518
       _cons |   .7722511   1.855748     0.42   0.678    -2.882712    4.427214
```

---

1. save the residuals

```
. predict uhat, resid
```

2. include residuals as additional regressor in the original equation

```
.  reg nqfede inc1 female uhat

      Source |       SS       df       MS              Number of obs =     252
-------------+------------------------------           F(  3,   248) =    9.94
       Model |  281.121189     3  93.7070629           Prob > F      =  0.0000
    Residual |  2337.87881   248  9.42693069           R-squared     =  0.1073
-------------+------------------------------           Adj R-squared =  0.0965
       Total |     2619.00   251  10.4342629           Root MSE      =  3.0703
------------------------------------------------------------------------------
      nqfede |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        inc1 |   .0450854   .0107655     4.19   0.000     .0238819    .0662888
      female |   1.176652   .3879107     3.03   0.003     .4126329    1.940672
        uhat |  -.0161473   .0186169    -0.87   0.387    -.0528147    .0205201
       _cons |   4.753386   .4512015    10.53   0.000     3.864711    5.642062
------------------------------------------------------------------------------
```

Now added residual is not statistically significantly different from zero, so conclude that there is no endogeneity bias in the OLS estimates. Hence no need to instrument.

Note you can also get this result by typing the following command after the ivreg command

```
ivendog

Tests of endogeneity of: inc1
H0: Regressor is exogenous
    Wu-Hausman F test:                 0.75229  F(1,248)    P-value = 0.38659
    Durbin-Wu-Hausman chi-sq test:     0.76211  Chi-sq(1)   P-value = 0.38267
```

the first test is simply the square of the t value on uhat in the last regression (since $t^2 = F$)

N.B. This test is only as good as the instruments used and is only valid asymptotically. This may be a problem in small samples and so you should generally use this test only with sample sizes well above 100.

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Any shock, represented by Δe

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that Cov(X,u) = 0 and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Any shock, represented by $\Delta e \rightarrow \Delta C$ in (1)

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Any shock, represented by $\Delta e \rightarrow \Delta C$ in (1)

but then this $\Delta C \rightarrow \Delta Y$ from (2)

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that $Cov(X,u) = 0$ and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Any shock, represented by $\Delta e \rightarrow \Delta C$ in (1)

but then this $\Delta C \rightarrow \Delta Y$ from (2)

and then this $\Delta Y \rightarrow \Delta C$ from (1)

Endogeneity & Simultaneous Equation Models

Often failure to establish a one-way causal relationship in an econometric model also leads to to endogeneity problems (again violates assumption that Cov(X,u) = 0 and so OLS will give biased estimates)

Eg

$$C = a + bY + e \qquad (1)$$

$$Y = C + I + G + v \qquad (2)$$

This is a 2 equation simultaneous equation system. C and Y appear on both sides of respective equations and are **interdependent** since

Any shock, represented by $\Delta e \rightarrow \Delta C$ in (1)
but then this $\Delta C \rightarrow \Delta Y$ from (2)

and then this $\Delta Y \rightarrow \Delta C$ from (1) so changes in C lead to changes in Y **and** changes in Y lead to changes in C

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$ means Cov(X,u) (or in this case Cov(Y,e)) $\neq$ □0 in (1)

$$C = a + bY + e \qquad (1)$$

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$

means Cov(X,u) (or in this case Cov(Y,e)) $\neq$ □0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)}$$

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$

means Cov(X,u) (or in this case Cov(Y,e)) $\neq$ □0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(Y,C)}{Var(Y)}$$ (in this example)

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$

means Cov(X,u) (or in this case Cov(Y,e)) $\neq$ 0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(Y,C)}{Var(Y)} = b + \frac{Cov(Y,e)}{Var(Y)}$$ (sub in for C from (1))

but the fact that Δe → ΔC → ΔY

means Cov(X,u) (or in this case Cov(Y,e)) ≠ □0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(Y,C)}{Var(Y)} = b + \frac{Cov(Y,e)}{Var(Y)}$$ (sub in for C from (1))

means $E(\hat{b}) \neq b$

but the fact that Δe → ΔC → ΔY

means Cov(X,u) (or in this case Cov(Y,e)) ≠ □0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(Y,C)}{Var(Y)} = b + \frac{Cov(Y,e)}{Var(Y)}$$

means $E(\hat{b}) \neq b$

So OLS in the presence of interdependent variables gives biased estimates.

but the fact that $\Delta e \rightarrow \Delta C \rightarrow \Delta Y$

means Cov(X,u) (or in this case Cov(Y,e)) $\neq$ 0 in (1)

$$C = a + bY + e \qquad (1)$$

which given OLS formula implies

$$\hat{b} = \frac{Cov(X,Y)}{Var(X)} = \frac{Cov(Y,C)}{Var(Y)} = b + \frac{Cov(Y,e)}{Var(Y)}$$

means $E(\hat{b}) \neq b$

So OLS in the presence of interdependent variables gives biased estimates.

Any right hand side variable which has the property $Cov(X,u) \neq 0$ is said to be **endogenous**

Solution: IV estimation
(as with measurement error, since symptom, if not cause, is the same)

Solution: IV estimation
(as with measurement error, since symptom, if not cause, is the same)

$$\hat{b}_{IV} = \frac{Cov(Z, y)}{Cov(Z, X)} \qquad \text{(A)}$$

Again, problem is where to find instruments. In a simultaneous equation model, the answer may often be in the system itself

Solution: IV estimation
(as with measurement error, since symptom, if not cause, is the same)

$$\hat{b}_{IV} = \frac{Cov(Z, y)}{Cov(Z, X)} \qquad \text{(A)}$$

Again, problem is where to find instruments. In a simultaneous equation model, the answer may often be in the system itself

Example

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

Solution: IV estimation
(as with measurement error, since symptom, if not cause, is the same)

$$\hat{b}_{IV} = \frac{Cov(Z, y)}{Cov(Z, X)} \qquad \text{(A)}$$

Again, problem is where to find instruments. In a simultaneous equation model, the answer may often be in the system itself

Example

$$Price = b_0 + b_1 Wage + e \qquad (1)$$
$$Wage = d_0 + d_1 Price + d_2 Unemployment + v \qquad (2)$$

This time wages and prices are interdependent so OLS on either (1) or (2) will give biased estimates….. but

Solution: IV estimation
(as with measurement error, since symptom, if not cause, is the same)

$$\hat{b}_{IV} = \frac{Cov(Z, y\ \ )}{Cov(Z, X)} \qquad \text{(A)}$$

Again, problem is where to find instruments. In a simultaneous equation model, the answer may often be in the system itself

Example

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$
$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

This time wages and prices are interdependent so OLS on either (1) or (2) will give biased estimates….. but

unemployment does not appear in (1) – by assumption (can this be justified?) but is correlated with wages through (2).

This means unemployment can be used as an instrument for wages in (1) since

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

a) Cov(Unemployment, e) = 0 (by assumption it doesn’t appear in (1)) so uncorrelated with residual, which is one requirement of an instrument

This means unemployment can be used as an instrument for wages in (1) since

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

a) Cov(Unemployment, e) = 0 (by assumption it doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument

and

b) Cov(Unemployment, Wage) ≠ 0 so correlated with endogenous RHS variable, which is the other requirement of an instrument

This means unemployment can be used as an instrument for wages in (1) since

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

a) Cov(Unemployment, e) = 0 (by assumption it doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument

and

b) Cov(Unemployment, Wage) ≠ 0 so correlated with endogenous RHS variable, which is the other requirement of an instrument

This process of using extra exogenous variables as instruments for endogenous RHS variables is known as **identification**

This means unemployment can be used as an instrument for wages in (1) since

$$\text{Price} = b_0 + b_1\text{Wage} + e \qquad (1)$$

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + v \qquad (2)$$

a) Cov(Unemployment, e) = 0 (by assumption the variable doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument

and

b) Cov(Unemployment, Wage) ≠ 0 so correlated with endogenous RHS variable, which is the other requirement of an instrument

This process of using extra exogenous variables as instruments for endogenous RHS variables is known as **identification**

If there are no additional exogenous variables outside the original equation that can be used as instruments for the endogenous RHS variables then the equation is said to be **unidentified**

This means unemployment can be used as an instrument for wages in (1) since

$$Price = b_0 + b_1 Wage + e \qquad (1)$$
$$Wage = d_0 + d_1 Price + d_2 Unemployment + v \qquad (2)$$

a) Cov(Unemployment, e) = 0 (by assumption it doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument and
b) Cov(Unemployment, Wage) ≠ 0 so correlated with endogenous RHS variable, which is the other requirement of an instrument

This process of using extra exogenous variables as instruments for endogenous RHS variables is known as **identification**

If there are no additional exogenous variables outside the original equation that can be used as instruments for the endogenous RHS variables then the equation is said to be **unidentified**
(In the example above (2) is unidentified because despite Price being endogenous , there are no other exogenous variables not already in (2) that can be used as instruments for Price).