

Endogeneity & Simultaneous Equation Models

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

Eg

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

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

Endogeneity & Simultaneous Equation Models

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

Eg

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

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

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 endogeneity problems (again violates assumption that $\text{Cov}(X,u) = 0$ and so OLS will give biased estimates)

Eg

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

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

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 endogeneity problems (again violates assumption that $\text{Cov}(X,u) = 0$ and so OLS will give biased estimates)

Eg

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

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

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 endogeneity problems (again violates assumption that $\text{Cov}(X,u) = 0$ and so OLS will give biased estimates)

Eg

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

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

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 endogeneity problems (again violates assumption that $\text{Cov}(X,u) = 0$ and so OLS will give biased estimates)

Eg

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

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

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 endogeneity problems (again violates assumption that $\text{Cov}(X,u) = 0$ and so OLS will give biased estimates)

Eg

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

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

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

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

which given OLS formula implies

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

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

which given OLS formula implies

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

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

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

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

which given OLS formula implies

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

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

which given OLS formula implies

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

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
 $\text{Cov}(X,u) \neq 0$ is said to be **endogenous**

Eg

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

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

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

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

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

which given OLS formula implies

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

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 $\text{Cov}(X,u) \neq 0$ 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)

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} \quad (A)$$

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)

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

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

Example

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

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

Solution: IV estimation

(as with measurement error, since symptom, if not cause, is the same)

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

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

Example

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

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

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)

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

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

Example

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

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

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).

Solution: IV estimation

(as with measurement error, since symptom, if not cause, is the same)

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

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

Example

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

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

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

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

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

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

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

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

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

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

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

- a) $\text{Cov}(\text{Unemployment}, e) = 0$ (by assumption it doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument and
- b) $\text{Cov}(\text{Unemployment}, \text{Wage}) \neq 0$ 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**

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

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

- a) $\text{Cov}(\text{Unemployment}, e) = 0$ (by assumption the variable doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument and
- b) $\text{Cov}(\text{Unemployment}, \text{Wage}) \neq 0$ 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**

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

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

- a) $\text{Cov}(\text{Unemployment}, e) = 0$ (by assumption it doesn't appear in (1)) so uncorrelated with residual, which is one requirement of an instrument and
- b) $\text{Cov}(\text{Unemployment}, \text{Wage}) \neq 0$ 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).

In general we can develop a rule that tells us whether any particular equation will be identified

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

where

K = Total no. of exogenous variables in the system

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

where

K = Total no. of exogenous variables in the system

k = No. of exogenous variables included in the equation

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

where

K = Total no. of exogenous variables in the system

k = No. of exogenous variables included in the equation

m = No. of endogenous variables included in the equation

“In a system of M simultaneous equations, then **any one equation** is identified if the number of **exogenous** variables **excluded** from that equation is greater than or equal to the total number of **endogenous** variables in that equation less one.”

$$K - k \geq m - 1 \quad (B)$$

where

K = Total no. of exogenous variables in the system

k = No. of exogenous variables included in the equation

m = No. of endogenous variables included in the equation

In practice this rule tells us whether we can find an instrument for each and every endogenous RHS variable in the equation

Consider the previous example

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

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

Consider the previous example

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

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

We know Price and Wage are interdependent so can't use OLS

Consider the previous example

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

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

We know Price and Wage are interdependent so can't use OLS

Consider each equation in turn

In (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

Consider the previous example

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

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

We know Price and Wage are interdependent so can't use OLS

Consider each equation in turn

In (1)

- 2 Endogenous variables (Price, Wage) so $m = 2$
- 1 Exogenous variables in the whole system (Unemployment) so $K = 1$
- 0 Exogenous variables in equation (1) so $k = 0$

Consider the previous example

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

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

We know Price and Wage are interdependent so can't use OLS

Consider each equation in turn

In (1)

2 Endogenous variables (Price, Wage) so $m = 2$

1 Exogenous variables in the whole system (Unemployment) so $K = 1$

0 Exogenous variables in equation (1) so $k = 0$

Using the rule $K - k \geq m - 1$

becomes

$$1 - 0 \geq 2 - 1$$

Consider the previous example

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

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

We know Price and Wage are interdependent so can't use OLS

Consider each equation in turn

In (1)

2 Endogenous variables (Price, Wage) so $m = 2$

1 Exogenous variables in the whole system (Unemployment) so $K = 1$

0 Exogenous variables in equation (1) so $k = 0$

Using the rule $K - k \geq m - 1$

becomes

$$1 - 0 \geq 2 - 1$$

so $1 = 1$

so equation (1) is (just) identified. There is an instrument for Wage (which is Unemployment) so can use IV on equation (1)

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

In (2) $Wage = d_0 + d_1 Price + d_2 Unemployment + v$

As before

2 Endogenous variables (Price, Wage)

so $m = 2$

1 Exogenous variables in the whole system (Unemployment)

so $K = 1$

but now

In (2) $Wage = d_0 + d_1Price + d_2Unemployment + v$

As before

2 Endogenous variables (Price, Wage)

so $m = 2$

1 Exogenous variables in the whole system (Unemployment)

so $K = 1$

but now

1 Exogenous variables in equation (2)

so $k = 1$

In (2) $Wage = d_0 + d_1 Price + d_2 Unemployment + v$

As before

2 Endogenous variables (Price, Wage)

so $m = 2$

1 Exogenous variables in the whole system (Unemployment)

so $K = 1$

but now

1 Exogenous variables in equation (2)

so $k = 1$

Using (B)

$$K - k \geq m - 1$$

In (2) $Wage = d_0 + d_1 Price + d_2 Unemployment + v$

As before

2 Endogenous variables (Price, Wage)

so $m = 2$

1 Exogenous variables in the whole system (Unemployment)

so $K = 1$

but now

1 Exogenous variables in equation (2)

so $k = 1$

Using (B)

$$K - k \geq m - 1$$

$$1 - 1 \geq 2 - 1$$

In (2) $Wage = d_0 + d_1 Price + d_2 Unemployment + v$

As before

2 Endogenous variables (Price, Wage)

so $m = 2$

1 Exogenous variables in the whole system (Unemployment)

so $K = 1$

but now

1 Exogenous variables in equation (2)

so $k = 1$

Using (B)

$$K - k \geq m - 1$$

$$1 - 1 \geq 2 - 1$$

$$0 < 1$$

In (2) $Wage = d_0 + d_1 Price + d_2 Unemployment + v$

As before

2 Endogenous variables (Price, Wage) so $m = 2$

1 Exogenous variables in the whole system (Unemployment) so $K = 1$

but now

1 Exogenous variables in equation (2) so $k = 1$

Using (B)

$$K - k \geq m - 1$$

$$1 - 1 \quad 2 - 1$$

$$0 < 1$$

so equation (2) is **not** identified. There is no instrument (extra exogenous variable) for Price anywhere else in the system of equations so can't use IV

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1 \text{Price} + d_2 \text{Unemployment} + d_3 \text{Productivity} + v \quad (2)$$

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + d_3\text{Productivity} + v \quad (2)$$

ie added another exogenous variable, Productivity, to (2).

We assume $\text{Cov}(\text{Productivity}, v) = 0$ (ie it is exogenous)

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1\text{Price} + d_2\text{Unemployment} + d_3\text{Productivity} + v \quad (2)$$

ie added another exogenous variable, Productivity, to (2).

We assume $\text{Cov}(\text{Productivity}, v) = 0$ (ie it is exogenous)

Now (2) is still not identified, since

2 Endogenous variables (Price, Wage) and $m = 2$

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1 \text{Price} + d_2 \text{Unemployment} + d_3 \text{Productivity} + v \quad (2)$$

ie added another exogenous variable, Productivity, to (2).

We assume $\text{Cov}(\text{Productivity}, v) = 0$ (ie it is exogenous)

Now (2) is still not identified, since

2 Endogenous variables (Price, Wage) and $m = 2$

and now

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1 \text{Price} + d_2 \text{Unemployment} + d_3 \text{Productivity} + v \quad (2)$$

ie added another exogenous variable, Productivity, to (2).

We assume $\text{Cov}(\text{Productivity}, v) = 0$ (ie it is exogenous)

Now (2) is still not identified, since

2 Endogenous variables (Price, Wage) and $m = 2$

and now

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

2 Exogenous variables in equation (2) (Unemployment, productivity) so $k = 2$

Example 2.

Suppose now that

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

$$\text{Wage} = d_0 + d_1 \text{Price} + d_2 \text{Unemployment} + d_3 \text{Productivity} + v \quad (2)$$

ie added another exogenous variable, Productivity, to (2).

We assume $\text{Cov}(\text{Productivity}, v) = 0$ (ie it is exogenous)

Now (2) is still not identified, since

2 Endogenous variables (Price, Wage) and $m = 2$

and now

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

2 Exogenous variables in equation (2) (Unemployment, productivity) so $k = 2$

and using (B)

$$K - k \geq m - 1$$

$$2 - 2 \geq 2 - 1$$

$$0 < 1$$

(not identified)

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

and

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

and

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

0 Exogenous variables in equation (1)

so $k = 0$

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

and

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

0 Exogenous variables in equation (1)

so $k = 0$

$$K - k \geq m - 1$$

$$2 - 0 \geq 2 - 1$$

$$2 > 1$$

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

and

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

0 Exogenous variables in equation (1)

so $k = 0$

$$K - k \geq m - 1$$

$$2 - 0 \geq 2 - 1$$

$$2 > 1$$

In this case equation (1) is now said to be **over-identified** (more instruments (other exogenous variables) than strictly necessary for IV estimation)

But now in (1)

2 Endogenous variables (Price, Wage)

so $m = 2$

and

2 Exogenous variables in the whole system (Unemployment, productivity)

so $K = 2$

but

0 Exogenous variables in equation (1)

so $k = 0$

$$K - k \geq m - 1$$

$$2 - 0 \geq 2 - 1$$

$$2 > 1$$

In this case equation (1) is now said to be **over-identified** (more instruments (other exogenous variables) than strictly necessary for IV estimation)

So which instrument to use for Wage in (1) ?

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{\text{Cov}(Z, y)}{\text{Cov}(Z, X)} = \frac{\text{Cov}(\text{Unemp}, \text{Price})}{\text{Cov}(\text{Unemp}, \text{Wage})}$$

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

Both will give unbiased estimate of true value, but likely (especially in small samples) that estimates will be different.

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

Both will give unbiased estimate of true value, but likely (especially in small samples) that estimates will be different.

In practice use **both** at the same time

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

Both will give unbiased estimate of true value, but likely (especially in small samples) that estimates will be different.

In practice use **both** at the same time
- removes possibility of conflicting estimates

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

Both will give unbiased estimate of true value, but likely (especially in small samples) that estimates will be different.

In practice use **both** at the same time

- removes possibility of conflicting estimates
- is more efficient (smaller variance) at least in **large** samples.

If use unemployment then as before, using (A) then the IV estimator

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Unemp, Price)}{Cov(Unemp, Wage)}$$

If instead use Productivity then

$$b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} = \frac{Cov(Prod, Price)}{Cov(Prod, Wage)}$$

Which is best?

Both will give unbiased estimate of true value, but likely (especially in small samples) that estimates will be different.

In practice use **both** at the same time

- removes possibility of conflicting estimates
- is more efficient (smaller variance) at least in **large** samples.

With small samples more efficient to use the **minimum** number of instruments

How in practice is this done (using both instruments at once)?

Given

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

We know

$$\text{Wage} = d_0 + d_1 \text{Price} + d_2 \text{Unemployment} + d_3 \text{Productivity} + v \quad (2)$$

and that Wage is related to the exogenous variables only by

$$\text{Wage} = g_0 + g_1 \text{Unemployment} + g_2 \text{Productivity} + w \quad (3)$$

(the equation which expresses the endogenous right hand side variables solely as a function of exogenous variables is said to be a **reduced form**)

Idea is then to estimate (3) and save the predicted values

$$\hat{wage} = g_0 + g_1 \hat{Unemp} + g_2 \hat{Pr od}$$

and use these as the instrument for Wage in (1)

\hat{wage} satisfies properties of an instrument since clearly correlated with variable of interest and because it is an average only of exogenous variables, is uncorrelated with the residual e in (1) (by assumption)

This strategy is called **Two Stage Least Squares (2SLS)** and in general the formula for the 2SLS estimator is given by

$$\hat{b}_{2sls} = \frac{\overset{\wedge}{Cov}(X, y)}{\overset{\wedge}{Cov}(X, X)} \quad \left(\text{compare with the IV formula } b_{IV}^{\wedge} = \frac{Cov(Z, y)}{Cov(Z, X)} \right)$$

which in the example above becomes

$$\hat{b}_{2sls} = \frac{\overset{\wedge}{Cov}(Wage, Price)}{\overset{\wedge}{Cov}(Wage, Wage)}$$

Note: In many cases you will not have a simultaneous system. More than likely will have just one equation but there may well be endogenous RHS variable(s). In this case the principle is exactly the same.

- Find additional exogenous variables that are correlated with the problem variable but uncorrelated with the error (“as if” there were another equation in the system).

This two stage least squares approach is also useful in helping illuminate whether the instrument(s) is good or not

Sometimes instruments may be statistically significant from zero in the 1st stage and still not good enough.

The t value on the instrument will tell you if, net of the other coefficients, the instrument is a good one (it should be significantly different from zero)

Cannot use the R^2 from the 1st stage since this could be high purely because of the exogenous variables in the model and not the instruments

Instead there is a rule of thumb (at least in the case of a single endogenous variable) that should only proceed with IV estimation if the F value in the test of the goodness of fit of the model on the 1st stage of 2SLS > 10.

Can also look at the **partial R²** which is based on a regression that nets out the effect of the exogenous variable X_2 on both endogenous variable X_1 and the instrument, Z and is obtained from the following regression

1. Regress X_2 on X_1 and save the predicted value \hat{X}_2

2. Regress Z on X_1 and save the predicted value \hat{Z}

3. Regress \hat{X}_2 on \hat{Z}

The R^2 from this regression is the partial R^2 (“partials out” the effect of X_1)

No threshold for the partial R^2 but the higher the value the greater the correlation between instrument and endogenous variable

If there is more than one potentially endogenous rhs variable in your equation the order condition (above) tells us that you will have to find at least one different instrument for **each** endogenous rhs variable (eg 2 endogenous rhs variables requires 2 different instruments).

Again, each instrument should be correlated with the endogenous rhs variable it replaces **net** of the other existing exogenous rhs variables. In this case two stage least squares estimation means predicting a value for each of the endogenous variables based on the instruments.

The examples above are based on models where there is only an endogenous variable on the right hand side of the model

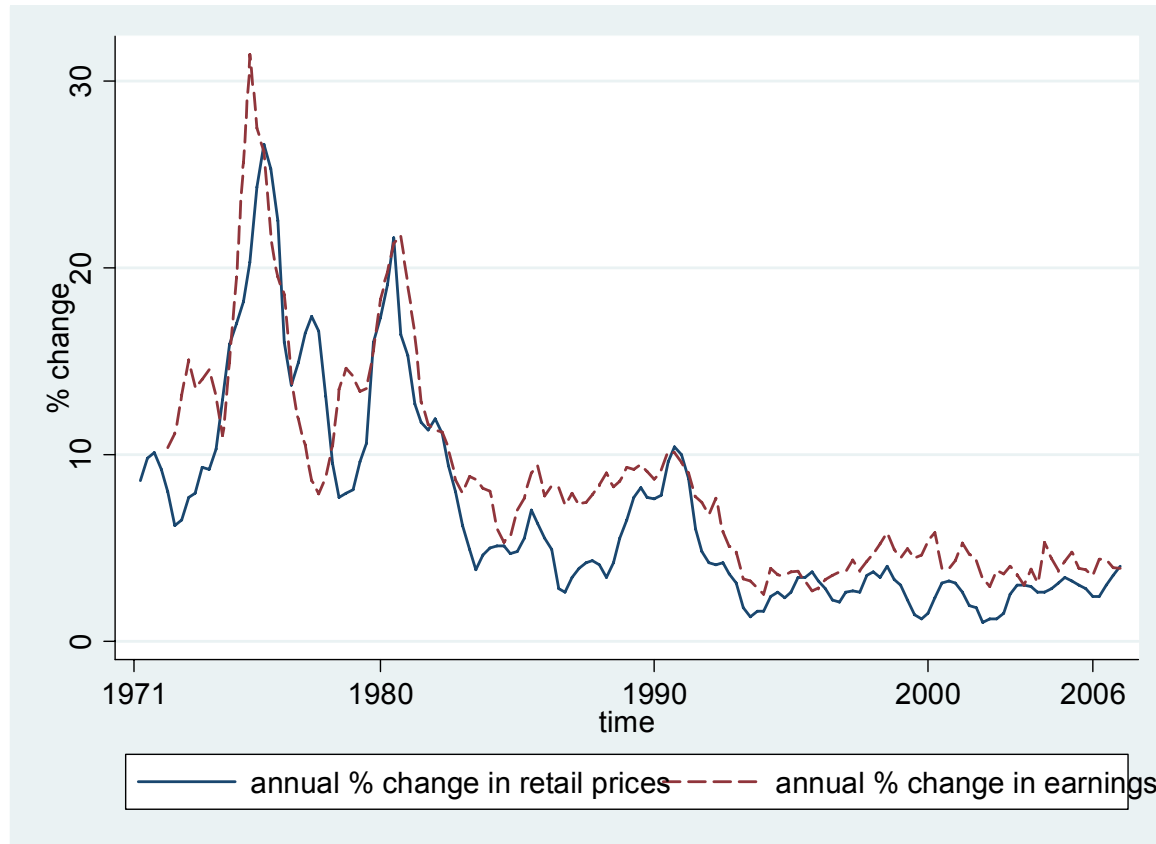
In many cases you will have a combination of exogenous variables (X_1) and endogenous variables (X_2) on the right hand side

$$Y = b_0 + b_1X_1 + b_2X_2 + u$$

In this case the only difference in estimation procedures is to make sure that you include the exogenous variables X_1 at both stages of the two stage estimation process

Example

```
two (line inflation time) (line avearn time, xlabel(1971 1980 1990 2000 2006) ytitle(% change) clpattern(dash))
```



The data set *prod.dta* contains quarterly time series data on wage, price, unemployment and productivity changes

The graph suggests that wages and prices move together over time and suggests may want to run a regression of inflation on wage changes where by assumption (and nothing else) the direction of causality runs from changes in wages to changes in inflation

```
. reg inf avearn
```

Source	SS	df	MS	Number of obs =	140
--------	----	----	----	-----------------	-----

	Model	Residual	Total		F(1, 138) = 569.25	Prob > F = 0.0000	R-squared = 0.8049	Adj R-squared = 0.8035	Root MSE = 2.5322
	3650.00027	884.851892	4534.85216	1	138	32.6248357			
inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]				
avearn	.8984636	.0376574	23.86	0.000	.8240036	.9729237			
_cons	-.8945061	.3897212	-2.30	0.023	-1.665103	-.123909			

which suggests an almost one-for-one relation between inflation and wage changes over this period.

However you might equally run a regression of wage changes on prices (where now the implied direction of causality is from changes in the inflation rate to changes in wages)

```
. reg avearn inf
```

Source	SS	df	MS	Number of obs = 140	F(1, 138) = 569.25	Prob > F = 0.0000	R-squared = 0.8049	Adj R-squared = 0.8035	Root MSE = 2.5285
Model	3639.3317	1	3639.3317						
Residual	882.265564	138	6.39322872						
Total	4521.59727	139	32.5294767						
avearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]				
inflation	.8958375	.0375473	23.86	0.000	.8215951	.9700799			
_cons	2.488974	.3351546	7.43	0.000	1.826271	3.151676			

this suggests wages grow at constant rate of around 2.5% a year (the coefficient on the constant) and then each 1 percentage point increase in the inflation rate adds a .89 percentage point increase in wages.

Which is right specification?

In a sense both, since wages affect prices but prices also affect wages. The 2 variables are interdependent and said to be **endogenous**. This means that $Cov(X,u) \neq 0$ ie a correlation between right hand side variables and the residuals which makes OLS estimates biased and inconsistent.

Need to **instrument** the endogenous right hand side variables. ie find a variable that is correlated with the suspect right hand side variable but uncorrelated with the error term.

Now it is not easy to come up with good instruments in this example since many macro-economic variables are all interrelated, but one possible solution with time series data is to use **lags** of the endogenous variable. The idea is that while inflation may affect wages and vice versa it is less likely that inflation can influence past values of wages and so they might be used as instruments for wages

Suppose decide to use the 3 and 4 year lag of wages as instruments for wages in the inflation regression

Which instrument to use? Both should give same estimate if sample size is large enough but in finite (small) samples the two IV estimates can be quite different.

```
sort year q          /* important to sort the data before taking lags */
g wlag3=avearn[_n-12]
(16 missing values generated)
g wlag4=avearn[_n-16]
(20 missing values generated)
/* note lose observations when take lags - cant calculate lags of values toward the start of the time period */
```

Using the 3 year lag as an instrument

```
ivreg inf (avear=wlag3)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS			
Model	3357.2168	1	3357.2168	Number of obs =	128	
Residual	785.683138	126	6.23558046	F(1, 126) =	145.52	
				Prob > F =	0.0000	
				R-squared =	0.8104	
				Adj R-squared =	0.8088	
Total	4142.89994	127	32.6212594	Root MSE =	2.4971	

inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
avearn	1.026884	.085127	12.06	0.000	.8584203	1.195348
_cons	-1.790507	.7228275	-2.48	0.015	-3.220961	-.3600522

```
Instrumented:  avearn
Instruments:   wlag3
```

Using the 4 year lag as an instrument

```
ivreg inf (avear=wlag4)
```

Instrumental variables (2SLS) regression

Source	SS	df	MS			
Model	2192.9109	1	2192.9109	Number of obs =	124	
Residual	646.995151	122	5.30323894	F(1, 122) =	166.23	
				Prob > F =	0.0000	
				R-squared =	0.7722	
				Adj R-squared =	0.7703	
Total	2839.90605	123	23.088667	Root MSE =	2.3029	

inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
avearn	1.005013	.0779507	12.89	0.000	.8507014	1.159324
_cons	-1.57803	.6190175	-2.55	0.012	-2.803437	-.3526229

```
Instrumented:  avearn
Instruments:   wlag4
```

Both estimates are similar and higher than original OLS estimate So which one?

Best idea (which also gives more efficient estimates ie ones with lower standard errors) is to use **all the instruments** at the same time – at least in large samples

1. Regress endogenous variables (wages) on both instruments (wage_{t-3} and wage_{t-4})

```
. reg avearn wlag4 wlag3
```

Source	SS	df	MS			
Model	915.451476	2	457.725738	Number of obs =	124	
Residual	1457.40163	121	12.0446416	F(2, 121) =	38.00	
				Prob > F =	0.0000	
				R-squared =	0.3858	
				Adj R-squared =	0.3756	
Total	2372.85311	123	19.2914887	Root MSE =	3.4705	

avearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wlag4	.3527246	.078543	4.49	0.000	.1972279	.5082213
wlag3	.146539	.0778473	1.88	0.062	-.0075803	.3006584
_cons	2.909309	.6131674	4.74	0.000	1.695382	4.123235

Save predicted wage

```
. predict wagehat /* stata command to save predicted value of dep. var. */
```

2. Include this instead of wages on the right hand side of the inflation regression

```
. reg inf wagehat
```

Source	SS	df	MS			
Model	930.575733	1	930.575733	Number of obs =	124	
Residual	1909.33031	122	15.6502485	F(1, 122) =	59.46	
				Prob > F =	0.0000	
				R-squared =	0.3277	
				Adj R-squared =	0.3222	
Total	2839.90605	123	23.088667	Root MSE =	3.956	

inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wagehat	1.008227	.1307504	7.71	0.000	.7493933	1.26706
_cons	-1.602087	1.041137	-1.54	0.126	-3.663121	.4589473

This gives unbiased estimate of effect of wages on prices.

Compare with original (biased) estimate,

can see wage effect is a little larger, (though standard error of IV estimate is larger than in OLS)

```
. reg inf avearn if e(sample)
```

Source	SS	df	MS	Number of obs = 124		
Model	2197.24294	1	2197.24294	F(1, 122)	=	417.11
Residual	642.663105	122	5.26773037	Prob > F	=	0.0000
				R-squared	=	0.7737
				Adj R-squared	=	0.7718
Total	2839.90605	123	23.088667	Root MSE	=	2.2952

inflation	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
avearn	.9622848	.0471169	20.42	0.000	.8690122	1.055557
_cons	-1.258218	.4084766	-3.08	0.003	-2.066838	-.4495972

Stata does all this automatically using the `ivreg2` command. Adding “first” to the command will also give the first stage of the two stage least squares regression which will help you decide whether the instruments are weak or not.

```
ivreg2 inf (avearn=wlag3 wlag4), first
```

First-stage regression of avearn:

Ordinary Least Squares (OLS) regression

Total (centered) SS	=	2372.853108	Number of obs	=	124
Total (uncentered) SS	=	9319.731851	F(2, 121)	=	38.00
Residual SS	=	1457.401632	Prob > F	=	0.0000
			Centered R2	=	0.3858
			Uncentered R2	=	0.8436
			Root MSE	=	3.5

avearn	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wlag3	.146539	.0778473	1.88	0.062	-.0075803	.3006584

wlag4	.3527246	.078543	4.49	0.000	.1972279	.5082213
_cons	2.909309	.6131674	4.74	0.000	1.695382	4.123235

Partial R-squared of excluded instruments: 0.3858

Test of excluded instruments:

F(2, 121) = 38.00
Prob > F = 0.0000

Summary results for first-stage regressions:

Variable	Partial R2	Partial R2	F(2, 121)	P-value
avearn	0.3858	0.3858	38.00	0.0000

Instrumental variables (2SLS) regression

Total (centered) SS	=	2839.906047	Number of obs	=	124
Total (uncentered) SS	=	7221.48998	F(1, 122)	=	175.29
Residual SS	=	647.6713926	Prob > F	=	0.0000
			Centered R2	=	0.7719
			Uncentered R2	=	0.9103
			Root MSE	=	2.3

inflation	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
avearn	1.008227	.0755351	13.35	0.000	.8601806 1.156273
_cons	-1.602087	.6014696	-2.66	0.008	-2.780945 -.423228

Sargan statistic (overidentification test of all instruments): 0.037
Chi-sq(1) P-val = 0.84741

Instrumented: avearn
Instruments: wlag3 wlag4

Note that the instruments are jointly significant in the first stage (as suggested by the F value and the R²)

Example 2: Poor Instruments & IV Estimation

Often a poor choice of instrument can make things much worse than the original OLS estimates.

Consider the example of the effect of education on wages (taken from the data set *video.dta*). Policy makers are often interested in the costs and benefits of education. Some people argue that education is endogenous (because it also picks up the effects of omitted variables like ability or motivation and so is correlated with the error term).

The OLS estimates from a regression of log hourly wages on years of education suggest that

```
. reg lhw yearsed
```

Source	SS	df	MS			
Model	246.491385	1	246.491385	Number of obs =	6076	
Residual	2540.90296	6074	.41832449	F(1, 6074) =	589.23	
				Prob > F =	0.0000	
				R-squared =	0.0884	
				Adj R-squared =	0.0883	
Total	2787.39434	6075	.458830344	Root MSE =	.64678	

lhw	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
yearsed	.0759492	.0031288	24.27	0.000	.0698156	.0820828
_cons	5.468133	.0400159	136.65	0.000	5.389688	5.546579

1 extra year of education is associated with 7.6% increase in earnings.

If endogeneity is a problem, then these estimates are biased, (upward if ability and education are positively correlated – see lecture notes on omitted variable bias).

So try to instrument instead.

For some reason you choose whether the individual owns a video recorder.

To be a good instrument the variable should be a) uncorrelated with the residual and by extension the dependent variable (wages) but b) correlated with the endogenous right hand side variable (education).

To test assumption a) you first examine the correlation between wages and videos in a regression.

```
. reg lhw video
```

Source	SS	df	MS			
				Number of obs =	6706	

-----+-----				F(1, 6704) = 0.20	
Model		.089653333	1	.089653333	Prob > F = 0.6576
Residual		3059.15176	6704	.456317387	R-squared = 0.0000
-----+-----				Adj R-squared = -0.0001	
Total		3059.24142	6705	.456262702	Root MSE = .67551
-----+-----					
lhw		Coef.	Std. Err.	t	P> t [95% Conf. Interval]
video		-.0178585	.0402898	-0.44	0.658 [-.0968393 .0611223]
_cons		6.444377	.0428575	150.37	0.000 [6.360363 6.528391]

The regression shows that it satisfies the first requirement since it is uncorrelated with wages (videos are relatively cheap now so that ownership is more a matter of taste than income).

However when you instrument years of education using video ownership

```
. ivreg lhw (years=video)
```

Instrumental variables (2SLS) regression				Number of obs = 6076	
Source		SS	df	MS	F(1, 6074) = 0.03
Model		-57.4179219	1	-57.4179219	Prob > F = 0.8657
Residual		2844.81226	6074	.46835895	R-squared = .
-----+-----				Adj R-squared = .	
Total		2787.39434	6075	.458830344	Root MSE = .68437
-----+-----					
lhw		Coef.	Std. Err.	t	P> t [95% Conf. Interval]
years		-.0083832	.0495839	-0.17	0.866 [-.1055853 .0888189]
_cons		6.52326	.6204326	10.51	0.000 [5.306992 7.739528]
-----+-----					
Instrumented:		years			
Instruments:		video			

The IV estimates are now negative (and insignificant). This does not appear very sensible. The reason is that video ownership is hardly correlated with education as the regression below shows.

```
. reg years video
```

Source	SS	df	MS			
Model	1.91564494	1	1.91564494	Number of obs =	12334	
Residual	83278.6218	12332	6.75305075	F(1, 12332) =	0.28	
				Prob > F =	0.5943	
				R-squared =	0.0000	
				Adj R-squared =	-0.0001	
Total	83280.5375	12333	6.75265851	Root MSE =	2.5987	

yearsed	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
video	-.0500536	.0939784	-0.53	0.594	-.234266	.1341587
_cons	12.15383	.1029141	118.10	0.000	11.9521	12.35556

Moral: Always check the correlation of your instrument with the endogenous right hand variable. The “first” option on stata’s ivreg command will always print the 1st stage of the 2SLS regression – ie the regression above – in which you can tell if the proposed instrument is correlated with the endogenous variable by simply looking at the t value.

Using the rule of thumb (at least in the case of a single endogenous variable) that should only proceed with IV estimation if the F value on the 1st stage of 2SLS > 10.

In the example above this is clearly not the case.

In practice you will often have a model where some but not all of the right hand side variables are endogenous.

$$Y = b_0 + b_1X_1 + b_2X_2 + u$$

where X_1 is exogenous and X_2 is endogenous

The only difference between this situation and the one described above is that you must include the exogenous variables X_1 in **both** stages of the 2SLS estimation process

Example 3. 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.

```
ivreg2 nqfede (incl=ranki) female, first
```

First-stage regressions

		Number of obs =	252
		F(2, 249) =	247.94
		Prob > F =	0.0000
Total (centered) SS	=	122243.0372	
Total (uncentered) SS	=	382752.6464	
Residual SS	=	40863.62602	
		Centered R2 =	0.6657
		Uncentered R2 =	0.8932
		Root MSE =	13

incl	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
female	.2342779	1.618777	0.14	0.885	-2.953963	3.422518
ranki	.2470712	.0110979	22.26	0.000	.2252136	.2689289
_cons	.7722511	1.855748	0.42	0.678	-2.882712	4.427215

Partial R-squared of excluded instruments: 0.6656
 Test of excluded instruments:
 F(1, 249) = 495.64
 Prob > F = 0.0000

IV (2SLS) regression with robust standard errors	Number of obs =	252
	F(2, 249) =	14.57
	Prob > F =	0.0000
	R-squared =	0.1033
	Root MSE =	3.0711

nqfede	Coef.	Robust Std. Err.	t	P> t	[95% Conf. Interval]	
incl	.0450854	.0101681	4.43	0.000	.0250589	.0651119
female	1.176652	.3883785	3.03	0.003	.4117266	1.941578
_cons	4.753386	.448987	10.59	0.000	3.86909	5.637683

Instrumented: incl
 Instruments: female ranki

Note that the exogenous variable “female” appears in both stages

Testing Instrumental Validity (Overidentifying Restrictions)

If you have more instruments than endogenous right hand side variables (the equation is overidentified – hence the name for the test) then it is possible to test whether (some of the) instruments are valid – in the sense that they satisfy the assumption of being uncorrelated with the residual in the original model.

Testing Instrumental Validity (Overidentifying Restrictions)

If you have more instruments than endogenous right hand side variables (the equation is overidentified – hence the name for the test) then it is possible to test whether (some of the) instruments are valid – in the sense that they satisfy the assumption of being uncorrelated with the residual in the original model.

One way to do this would be, as in the example above, to compute two different 2SLS estimates, one using one instrument and another using the other instrument. If these estimates are radically different you might conclude that one (or both) of the instruments was invalid (not exogenous). If these estimates were similar you might conclude that both instruments were exogenous.

Testing Instrumental Validity (Overidentifying Restrictions)

If you have more instruments than endogenous right hand side variables (the equation is overidentified – hence the name for the test) then it is possible to test whether (some of the) instruments are valid – in the sense that they satisfy the assumption of being uncorrelated with the residual in the original model.

One way to do this would be, as in the example above, to compute two different 2SLS estimates, one using one instrument and another using the other instrument. If these estimates are radically different you might conclude that one (or both) of the instruments was invalid (not exogenous). If these estimates were similar you might conclude that both instruments were exogenous.

An implicit test of this – that avoids having to compute all of the possible IV estimates - is based on the following idea

Given $y = b_0 + b_1X + u$ and $\text{Cov}(X,u) \neq 0$

If an instrument Z is valid (exogenous) it is uncorrelated with u

To test this simply regress u on **all** the possible instruments.

$$u = d_0 + d_1Z_1 + d_2Z_2 + \dots + d_lZ_l + v$$

If the instruments are exogenous they should be uncorrelated with u and so the coefficients $d_1 \dots d_l$ should all be zero (ie the Z variables have no explanatory power)

To test this simply regress u on **all** the possible instruments.

$$u = d_0 + d_1 Z_1 + d_2 Z_2 + \dots + d_l Z_l + v$$

If the instruments are exogenous they should be uncorrelated with u and so the coefficients $d_1 \dots d_l$ should all be zero (ie the Z variables have no explanatory power)

Since u is never observed have to use a proxy for this. This turns out to be the residual from the 2SLS estimation estimated using all the possible instruments

$$\hat{u}^{2sls} = y - \hat{b}_0^{2sls} - \hat{b}_1^{2sls} X$$

(since this is a consistent estimate of the true unknown residuals)

So to Test Overidentifying Restrictions

1. Estimate model by 2SLS and save the residuals
2. Regress these residuals on *all* the exogenous variables (including those X_1 variables in the original equation that are not suspect)

$$\hat{u}^{2sls} = d_0 + b_1 X_1 + d_1 Z_1 + d_2 Z_2 + \dots + d_l Z_l + v$$

and save the R^2

3. Compute $N \cdot R^2$

4. Under the null that all the instruments are uncorrelated then

$$N \cdot R^2 \sim \chi^2 \quad \text{with } L-k \text{ degrees of freedom}$$

(L is the number of instruments and k is the number of endogenous right hand side variables in the original equation)

So to Test Overidentifying Restrictions

1. Estimate model by 2SLS and save the residuals
2. Regress these residuals on *all* the exogenous variables (including those X_1 variables in the original equation that are not suspect)

$$\hat{u}^{2sls} = d_0 + b_1 X_1 + d_1 Z_1 + d_2 Z_2 + \dots + d_l Z_l + v$$

and save the R^2

3. Compute $N \cdot R^2$

4. Under the null that all the instruments are uncorrelated then

$$N \cdot R^2 \sim \chi^2 \quad \text{with } L-k \text{ degrees of freedom}$$

(L is the number of instruments and k is the number of endogenous right hand side variables in the original equation)

Note that can only do this test if there are more instruments than endogenous right hand side variables (in just identified case the residuals and right hand side variables are uncorrelated by construction)

Also this test is again only valid in large samples

Example: using the *prod.dta* file we can test whether some of the instruments ($wage_{t-3}$ and $wage_{t-4}$) are valid instruments for wages

1st do the two stage least squares regression using all the possible instruments

```
ivreg inf (avearn=wlag3 wlag4)
```

Now save these 2sls residuals

```
. predict ivres, resid
```

and regress these on all the exogenous variables in the system (remember there may be situations where the original equation contained other exogenous variables, in which case include them here also)

```
. reg ivres wlag3 wlag4
```

Source	SS	df	MS			
Model	.193389811	2	.096694905	Number of obs =	124	
Residual	647.478005	121	5.3510579	F(2, 121) =	0.02	
Total	647.671395	123	5.2656211	Prob > F =	0.9821	
				R-squared =	0.0003	
				Adj R-squared =	-0.0162	
				Root MSE =	2.3132	

ivres	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
wlag3	.0096316	.051888	0.19	0.853	-.0930943	.1123575
wlag4	-.0085304	.0523517	-0.16	0.871	-.1121743	.0951136
_cons	-.0073752	.4086975	-0.02	0.986	-.8164997	.8017493

The test is $N \cdot R^2 = 124 \cdot 0.0003 = 0.04$ which is $\sim \chi^2$ ($L-k = 2-1$)

($L=2$ instruments and $k=1$ endogenous right hand side variable)

Since $\chi^2_{(1)}^{\text{critical}} = 3.84$, estimated value is below critical value so **accept** null hypothesis that some of the instruments are valid

Can obtain these results automatically using the command:

overid

Tests of overidentifying restrictions:

Sargan N*R-sq test	0.037	Chi-sq(1)	P-value = 0.8474
Basman test	0.036	Chi-sq(1)	P-value = 0.8492