Instrumental Variables

Ani Katchova

Outline

- Endogeneity problem
- Instrumental variables
- IV estimation
- 2SLS estimation
- Testing for endogeneity

Endogeneity problem

- Endogeneity problem is when the independent variable is correlated with the error term.
- Endogeneity is a frequent problem in economics and econometrics.
- Sources of endogeneity:
  - Omitted variables - independent variables are not observed and end up in the error term, so the error term is correlated with the independent variables.
  - Measurement error can cause correlation between the mismeasured variable and the error term.
- Solutions for endogeneity:
  - Find and include the unobserved variable in the model.
  - Find and include a proxy variable in the model.
  - Use fixed effects estimator with panel data, by eliminating individual specific effects.
  - Use instrumental variables (IV) method which replaces the endogenous variable with a predicted value that has only exogenous information.

Instrumental variables - definition

- An instrumental variable (or instrument or IV) is a variable that is used in a regression model to correct for the endogeneity problem.
- Dependent variable y
- Endogenous variable x that is correlated with the error term u
- Instrument z is a variable that is related to the endogenous variable x but does not belong in the model for y and is not correlated with the error term.

Regression model – OLS estimation

- Regression model: $y = \beta_0 + \beta_1 x + u$
- If x is exogenous (not correlated with the error term), $cov(x, u) = 0$.
- $cov(x, u) = cov(x, y - \beta_0 - \beta_1 x) = cov(x, y) - \beta_1 var(x) = 0$
- $\beta_1^{OLS} = \beta_1 = \frac{cov(x,y)}{var(x)} = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sum(x-\bar{x})(x-\bar{x})}$
- If x is exogenous, then β_1 will be unbiased and consistent.

Regression model – IV estimation

- Regression model: $y = \beta_0 + \beta_1 x + u$
- If x is endogenous (correlated with the error term), $cov(x, u) \neq 0$.
- Find an instrument z that is not correlated with the error term u, $cov(z, u) = 0$
- $cov(z, u) = cov(z, y - \beta_0 - \beta_1 x) = cov(z, y) - \beta_1 cov(z, x) = 0$
- $\beta_1^{IV} = \beta_1 = \frac{cov(z,y)}{cov(z,x)} = \frac{\sum(z-\bar{z})(y-\bar{y})}{\sum(z-\bar{z})(x-\bar{x})}$
- The coefficient estimated using the above IV formula will be unbiased and consistent.
- If x is exogenous, it can serve as its own instrument $z = x$, and the IV estimate β_1^{IV} would be identical to the OLS estimate β_1^{OLS}.

IV properties

- An instrument z should have three properties:

1) The instrument z does not appear in the original regression model.

$$y = \beta_0 + \beta_1 x + u$$

2) The instrument z is correlated with the endogenous variable x, so $cov(z, x) \neq 0$

$$x = \delta_0 + \delta_1 z + v$$

where $\delta_1 \neq 0$.

3) The instrument z is uncorrelated with the error term u.

$$cov(z, u) = 0$$

2SLS – two stage least squares

- Regression model – OLS estimation: $y = \beta_0 + \beta_1 x + u$
- If x is endogenous, the coefficient $\hat{\beta}_1$ estimated with OLS will be biased.
- 2SLS – first stage: $x = \delta_0 + \delta_1 z + v$ is a regression of the endogenous variable x on the instrument z.
- Get predicted values $\hat{x} = \hat{\delta}_0 + \hat{\delta}_1 z$. The predicted value $\hat{x}$ contains only exogenous information from the instrument z.
- 2SLS – second stage: $y = \beta_0 + \beta_1 \hat{x} + u$. Regression the dependent variable y on the predicted values $\hat{x}$.
- The coefficient $\hat{\beta}_1$ estimated with 2SLS will be unbiased because $\hat{x}$ is exogenous and uncorrelated with the error term u.

2SLS – standard errors

- The standard errors from the second stage regression need to be corrected.
- In OLS, $var(\beta_1) = \frac{\sigma^2}{SST_x}$ In 2SLS, $var(\beta_1) = \frac{\sigma^2}{SST_x R^2_{x,z}}$
- σ^2 is the variance of the error term u. SST_x is the total variation in x.
- $R^2_{x,z}$ is the R^2 from the regression of x on z.
- $var(\hat{\beta}_1^{2SLS}) = \frac{var(\widehat{\beta}_1^{OLS})}{R^2_{x,z}}$ and $se\left(\hat{\beta}_1^{2SLS}\right) = \frac{se(\widehat{\beta}_1^{OLS})}{\sqrt{R^2_{x,z}}}$
- The variance of coefficients using the 2SLS estimation will be higher than the variance of coefficients using the OLS estimation, because the R-squared is less than 1.
- A weaker the relationship between x and z will results in lower $R^2_{x,z}$ and higher variance of the 2SLS coefficients, leading to less significance.

IV versus 2SLS estimation

- If there is one endogenous variable and one instrument, then the 2SLS estimates (replacing x with $\hat{x}$ based on z) will be the same as the IV estimates $(cov(z, y)/cov(z, x))$.
- The 2SLS estimation can also be used if there is more than one endogenous variable and at least as many instruments.

IV example

- Model for log wages ($lwage$) explained by education ($educ$), which is endogenous. The father's education ($fatheduc$) will serve as an instrument for education.
- $fatheduc$ is a good instrument for $educ$ because it has the three properties:

  1) The instrument $fatheduc$ does not appear in the original regression model.

  $$lwage = \beta_0 + \beta_1 educ + u$$

  2) The instrument $fatheduc$ is correlated with the endogenous variable $educ$, so $cov(fatheduc, educ) \neq 0$

  $$educ = \delta_0 + \delta_1 fatheduc + v$$

  where $\delta_1 \neq 0$.

  3) The instrument $fatheduc$ is uncorrelated with the error term u.

  $$cov(fatheduc, u) = 0$$

- Other potential instruments: number of siblings, college proximity when 16 years old, month of birth.

IV estimation example

- Model for log wages ($lwage$) explained by education ($educ$), which is endogenous. The father's education ($fatheduc$) is an instrument for education.
- Regression model: $lwage = \beta_0 + \beta_1 educ + u$
- $\beta_1^{OLS} = \frac{cov(educ, lwage)}{var(educ)} = \frac{\sum(educ - \overline{educ})(lwage - \overline{lwage})}{\sum(educ - \overline{educ})(educ - \overline{educ})} = 0.109$
- $\beta_1^{IV} = \frac{cov(fatheduc, lwage)}{cov(fatheduc, educ)} = \frac{\sum(fatheduc - \overline{fatheduc})(lwage - \overline{lwage})}{\sum(fatheduc - \overline{fatheduc})(educ - \overline{educ})} = 0.059$
- The coefficient using IV estimation is lower than the coefficient using OLS estimation. One additional year of education is associated with 10.9% increase in wages using OLS but only 5.9% increase in wages using IV.
- The OLS and IV estimates for β_1 appear to be different from each other, so perhaps $educ$ is endogenous.

OLS and 2SLS example

- Regression model – OLS estimation: $lwage = \beta_0 + \beta_1 educ + u$. Get $\hat{\beta}_1^{OLS}$
- Education is an endogenous variable, and father's education is the instrument.
- 2SLS estimation:
- First stage: $educ = \delta_0 + \delta_1 fatheduc + v$, get predicted values $\widehat{educ}$.
- Second stage: $lwage = \beta_0 + \beta_1 \widehat{educ} + u$. Get $\hat{\beta}_1^{2SLS}$.
- 2SLS first stage is regressing education on father's education, getting predicted values for education $\widehat{educ}$. The 2SLS second stage is regressing lwage on $\widehat{educ}$.

OLS and 2SLS estimation

	OLS estimation	2SLS estimation – first stage	2SLS estimation – second stage
VARIABLES	lwage	educ	lwage
educ	0.109***		
	(0.014)		
educ_hat			0.059*
			(0.035)
fatheduc		0.269***	
		(0.029)	
Constant	-0.185	10.237***	0.441
	(0.185)	(0.276)	(0.446)
R-squared	0.12	0.17	0.09

Using the OLS estimation, one additional year of education is associated with 10.9% increase in wages.
Using the 2SLS estimation, one additional year of education is associated with 5.9% increase in wages, which is a lower effect and less significant.
The same IV and 2SLS coefficient of 0.059 are obtained.

2SLS – endogenous variable vs predicted values using instrument

educ	fatheduc	educ_hat $\widehat{educ}$
12	7	12.12
12	7	12.12
12	7	12.12
12	7	12.12
14	14	14.01
12	7	12.12
16	7	12.12
12	3	11.05

- The first few observations for $educ$ and $\widehat{educ}$.
- $\widehat{educ}$ is only based on the exogenous information coming from $fatheduc$.
- $\widehat{educ}$ is not a whole number

- If a variable is binary (0 or 1), the predicted values below 0.5 can be replaced by 0 and above 0.5 can be replaced by 1.

2SLS – standard errors

- The R-squared of the 2SLS first stage regression of $educ$ on $fatheduc$ is $R^2_{x,z}$=0.17. The R-squared is not very high.
- From the regression output, $se\left(\hat{\beta}_1^{OLS}\right)$=0.014 and $se\left(\hat{\beta}_1^{2SLS}\right)$=0.034. The 2SLS coefficient has a higher standard error and is less significant.
- The exact relationship for the standard errors is:

$$se\left(\hat{\beta}_1^{2SLS}\right) = \frac{se\left(\hat{\beta}_1^{OLS}\right)}{\sqrt{R^2_{x,z}}} = \frac{0.014}{\sqrt{0.17}} = 0.034$$

Multiple regression model – IV estimation

- Multiple regression model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

- here y_2 is the endogenous variable that is correlated with the error term u_1, and z_1 and z_2 are exogenous variables.
- Find two instruments z_3 and z_4 for the endogenous variable y_2, that are uncorrelated with the error term.
- The exogeneity conditions for the instruments are:
    - $cov(z_3, u_1) = cov(z_3, y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1 - \beta_3 z_2) = 0$
    - $cov(z_4, u_1) = cov(z_4, y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1 - \beta_3 z_2) = 0$
    - $E(u_1) = E(y_1 - \beta_0 - \beta_1 y_2 - \beta_2 z_1 - \beta_3 z_2) = 0.$
    - These equations are solved to obtain the IV coefficients $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$, and $\hat{\beta}_3$.

Multiple regression model – 2SLS

- Multiple regression model

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

- here y_2 is the endogenous variable that is correlated with the error term u_1, and z_1 and z_2 are exogenous variables.
- Find two instruments z_3 and z_4 for the endogenous variable y_2.
- The 2SLS first stage reduced form equation is:

$$y_2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \delta_4 z_4 + v_2$$

- Obtain fitted values: $\hat{y}_2 = \hat{\delta}_0 + \hat{\delta}_1 z_1 + \hat{\delta}_2 z_2 + \hat{\delta}_3 z_3 + \hat{\delta}_4 z_4$
- The 2SLS second stage is to estimate the structural model where the endogenous variable y_2 is replaced by $\hat{y}_2$:

$$y_1 = \beta_0 + \beta_1 \hat{y}_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

IV properties

- The instruments z_3 and z_4 should have three properties:

1) The instruments do not appear in the original regression model.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$$

2) The instruments are correlated with the endogenous variable y_2, so $cov(z_3, y_2) \neq 0$ and $cov(z_4, y_2) \neq 0$

$$y_2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \delta_4 z_4 + v_2$$

where $\delta_3 \neq 0$ and $\delta_4 \neq 0$.

3) The instruments are uncorrelated with the error term u_1.

$cov(z_3, u_1) = 0$ and $cov(z_4, u_1) = 0$.

IV and 2SLS discussion

- The IV estimation is equivalent to the 2SLS estimation.
- The 2SLS estimation works because the endogenous variable y_2 is replaced in the second stage by $\hat{y}_2$ that contains only exogenous information from instruments and exogenous variables, but not the endogenous part that is correlated with the error term.

2SLS example

- Structural equation model:
$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1$$
- here $educ$ is endogenous and $exper$ and $exper^2$ are exogenous.
- Find two instruments $fatheduc$ and $motheduc$ for $educ$.
- 2SLS first stage - estimate the reduced form equation:
- $educ = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 fatheduc + \delta_4 motheduc + v_2$
- Obtain the predicted values $\widehat{educ}$, which contain only exogenous information.
- 2SLS second stage - estimate the structural equation replacing $educ$ with $\widehat{educ}$:
$$lwage = \beta_0 + \beta_1 \widehat{educ} + \beta_2 exper + \beta_3 exper^2 + u_1$$

2SLS example

VARIABLES	OLS lwage	2SLS – first stage educ	2SLS – second stage lwage
educ	0.108***		
	(0.014)		
educ_hat			0.061*
			(0.031)
exper	0.042***	0.045	0.044***
	(0.013)	(0.040)	(0.013)
expersq	-0.0008**	-0.001	-0.0009**
	(0.0004)	(0.001)	(0.0004)
fatheduc		0.190***	
		(0.034)	
motheduc		0.158***	
		(0.036)	
Constant	-0.522***	9.103***	0.048
	(0.199)	(0.427)	(0.400)

- 2SLS estimation – estimate 2SLS first stage for education, get predicted values educ_hat and use them instead of educ in the 2SLS second stage.
- The coefficient on education goes down from 0.108 using OLS to 0.061 using 2SLS.
- One additional year of education is associated with 10.8% increase in wages using OLS, and with 6.1% increase in wages using 2SLS. The effect is smaller and less significant using the 2SLS after correcting for the endogeneity.

2SLS example

	2SLS – second stage correct standard errors	2SLS – second stage incorrect standard errors
VARIABLES	lwage	lwage
educ		
educ_hat	0.061*	0.061*
	(0.031)	(0.033)
exper	0.044***	0.044***
	(0.013)	(0.014)
expersq	-0.0009**	-0.0009**
	(0.0004)	(0.0004)
fatheduc		
motheduc		
Constant	0.048	0.048
	(0.400)	(0.420)

If estimating the second stage of 2SLS, the standard errors need to be corrected.

In OLS, $var(\beta) = \frac{\sigma^2}{SST_x}$

In 2SLS, $var(\beta) = \frac{\sigma^2}{SST_x R^2_{x,z}}$

The standard error on the coefficient on education is higher when corrected (0.033 vs 0.031).

Many software packages provide the corrected standard errors.

Testing for endogeneity

- Structural equation model: $y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1$
- Testing for endogeneity of y_2.
- Find two instruments z_3 and z_4 for y_2.
- Estimate the reduced form equation:
$$y_2 = \delta_0 + \delta_1 z_1 + \delta_2 z_2 + \delta_3 z_3 + \delta_4 z_4 + v_2$$
- Obtain the residuals $\hat{v}_2$, which would contain the endogenous information.
- The predicted values $\hat{y}_2$ only contains the exogenous information.
- So the endogenous variable is broken down in exogenous part $\hat{y}_2$ and endogenous part $\hat{v}_2$, $y_2 = \hat{y}_2 + \hat{v}_2$.
- Estimate the structural equation with the residuals $\hat{v}_2$ included:
$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \gamma_1 \hat{v}_2 + u_1$$
- H_0: $\gamma_1 = 0$ (exogeneity)
- H_a: $\gamma_1 \neq 0$ (endogeneity)

Testing for endogeneity example

- Structural equation model: $lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + u_1$
- Testing for endogeneity of $educ$.
- Find two instruments $fatheduc$ and $motheduc$ for $educ$.
- Estimate the reduced form equation:
- $educ = \delta_0 + \delta_1 exper + \delta_2 exper^2 + \delta_3 fatheduc + \delta_4 motheduc + v_2$
- Obtain the residuals $\hat{v}_2$, which would contain the endogenous information.
- The predicted values $\widehat{educ}$ only contains the exogenous information.
- Estimate the structural equation with the residuals $\hat{v}_2$ included:
$$lwage = \beta_0 + \beta_1 educ + \beta_2 exper + \beta_3 exper^2 + \gamma_1 \hat{v}_2 + u_1$$
- H$_0$: $\gamma_1 = 0$ (exogeneity)
- H$_a$: $\gamma_1 \neq 0$ (endogeneity)

Testing for endogeneity

	Structural model	Reduced form model	Structural model with residuals
VARIABLES	lwage	educ	lwage
educ	0.107***		0.061**
	(0.014)		(0.031)
exper	0.042***	0.045	0.044***
	(0.013)	(0.040)	(0.013)
expersq	-0.001**	-0.001	-0.001**
	(0.0004)	(0.001)	(0.0003)
fatheduc		0.190***	
		(0.033)	
motheduc		0.157***	
		(0.035)	
vhat			0.058*
			(0.034)
Constant	-0.522***	9.103***	0.048
	(0.199)	(0.427)	(0.395)

Estimate the reduced form model for education, obtain the residuals, and include them in the structural model for lwage.

The coefficient on the residual vhat is significant at 10%, so the variable education is endogenous. Instrumental variables need to be used to correct for the endogeneity.

Review questions

- Describe the three properties of a good instrument.
- Describe the IV estimator.
- Describe the 2SLS procedure with first and second stage estimation.
- Describe the test for endogeneity of an independent variable.