

Econometrics

Maqsood Aslam

2024

Wishing you all the prosperous New Year !

May your residuals be minimal,

P-values be small,

Confidence intervals be narrow,

Regression coefficients be significant

R-squared of success be maximized,

Knowledge curve always be upward sloping

and increasing returns to scales of all good things in 2025

Instrumental Variables

Introduction

- ▶ Suppose that the true model is:

$$wage = \beta_0 + \beta_1 education + \beta_2 motivation + u$$

- ▶ But because motivation is impossible to observe, you estimate instead:

$$wage = \beta_0 + \beta_1 education + \varepsilon$$

- ▶ Because motivation and level of schooling are correlated, the estimated effect is biased and inconsistent
- ⇒ Endogeneity: $E(\varepsilon|education) \neq 0$
- ⇒ Endogeneity bias: The estimate of the effect of education on wage partly captures the effect of motivation on education (upward bias if $corr(education, motivation) > 0$)

Introduction

- ▶ The aim of instrumental variables is to try and correct for endogeneity bias
- ▶ How?
- ▶ By using a third variable that will capture only the part of the effect that is due to education
- ▶ The idea is to re-create exogeneity

Definition

Consider the more general regression model:

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

where x is endogenous (correlated with u)

Instrumental variable

An instrumental variable (or instrument) is a variable, denoted z , such that

- ▶ z is correlated with the endogenous variable x : $cov(z, x) \neq 0$
- ▶ z is not correlated with the error term u : $cov(z, u) = 0$

- ▶ $cov(z, u) = 0$ is called the exclusion restriction
- ▶ $cov(z, x) \neq 0$ is called the relevance condition

Instrumental variable

- ▶ What would be a good instrumental variable for education in the wage equation?
- ▶ It would have to be uncorrelated with all unobserved factors that affect wages
- ▶ It would have to be correlated with the level of education
- ▶ The three last digits of individuals' social security number would satisfy the first condition (it is determined randomly)
- ▶ But it is not correlated with the the level of education
- ▶ Individual's IQ (if recorded) is correlated with education
- ▶ But also with other unobserved factors that affect wages

Identification

β_1 is identified by:

$$\text{cov}(z, y) = \beta_1 \text{cov}(z, x) + \text{cov}(z, u)$$

Using the exclusion restriction:

$$\beta_1 = \frac{\text{cov}(z, y)}{\text{cov}(z, x)}$$

The sample analog is the instrumental variable estimator of β_1 :

$$\hat{\beta}_{1IV} = \frac{\sum_{i=1}^N (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^N (z_i - \bar{z})(x_i - \bar{x})}$$

Wald Estimator

When the instrumental variable is binary: $z \in 0, 1$

$$E(y|z = 1) = \beta_0 + \beta_1 E(x|z = 1)$$

$$E(y|z = 0) = \beta_0 + \beta_1 E(x|z = 0)$$

so that

$$\beta_1 = \frac{E(y|z = 1) - E(y|z = 0)}{E(x|z = 1) - E(x|z = 0)}$$

Taking sample analogues gives the Wald estimator:

$$\hat{\beta}_1 = \frac{\bar{y}_{z=1} - \bar{y}_{z=0}}{\bar{x}_{z=1} - \bar{x}_{z=0}}$$

In the MLR setting

This extends to the multiple linear regression case

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$

with x_1 an endogenous variable, and x_2, \dots, x_K exogenous variables

- ▶ Then the OLS estimator of β (vector of parameters) is biased and inconsistent
- ▶ If there exists an instrumental variable z such that
 - ▶ $\text{cov}(z, x_1) \neq 0$
 - ▶ and $\text{cov}(z, u) = 0$
- ▶ The IV estimator of the vector β solves the sample analogs of $\text{cov}(z, u) = 0, \text{cov}(1, u) = 0, \text{cov}(x_2, u) = 0, \dots, \text{cov}(x_K, u) = 0$
- ▶ The IV estimator consistently estimates β

In the MLR setting

- ▶ Note that the exclusion restriction cannot be tested, since u is not observed
- ▶ The relevance condition can and should always be tested
- ▶ By estimating:

$$x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 \cdots + \pi_K x_K + v$$

- ▶ π_1 has to be different from zero for the relevance condition to hold
- ⇒ One need to test: $H_0 : \pi_1 = 0$

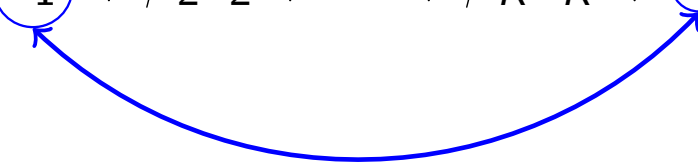
Properties of the IV estimator

- ▶ The IV estimator is consistent, but not unbiased
- ▶ The IV estimator's variance is always larger than the OLS estimator's variance
- ▶ Under H_1 to H_5 , the IV estimator is normally asymptotically distributed
- ▶ One can thus define and use t-statistics

NB The R-squared from IV estimation cannot be interpreted as the fraction of the sample variation in y that is explained by the regressors, because SST cannot be decomposed as the sum of SSR and SSE

The R-squared can even be negative (the SSR can be larger than the SST)

Estimation: principle

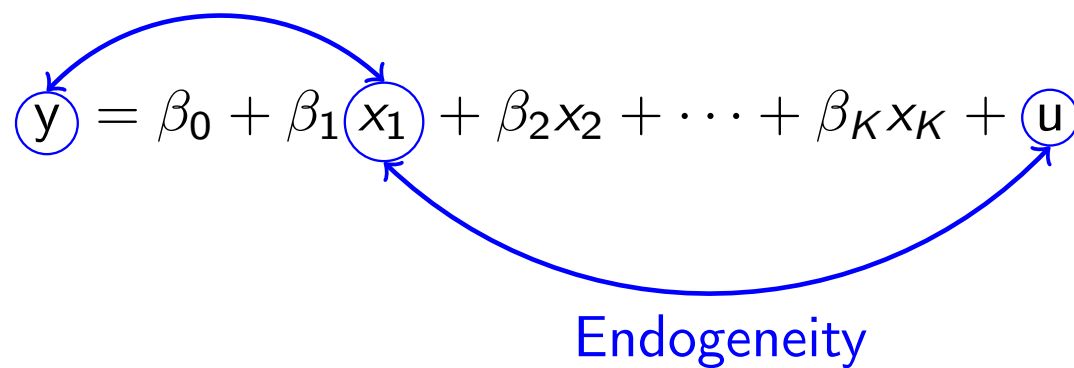
$$\textcircled{y} = \beta_0 + \beta_1 \textcircled{x_1} + \beta_2 x_2 + \cdots + \beta_K x_K + \textcircled{u}$$


The diagram illustrates the concept of endogeneity in a regression model. It shows the equation $\textcircled{y} = \beta_0 + \beta_1 \textcircled{x_1} + \beta_2 x_2 + \cdots + \beta_K x_K + \textcircled{u}$ where the dependent variable y , the independent variable x_1 , and the error term u are each enclosed in a blue circle. A blue curved double-headed arrow connects the circle around x_1 and the circle around u , indicating a causal relationship between them. Below this arrow, the word "Endogeneity" is written in blue.

Endogeneity

Estimation: principle

Effect of x_1 + effect of $\text{corr}(x_1, u)$

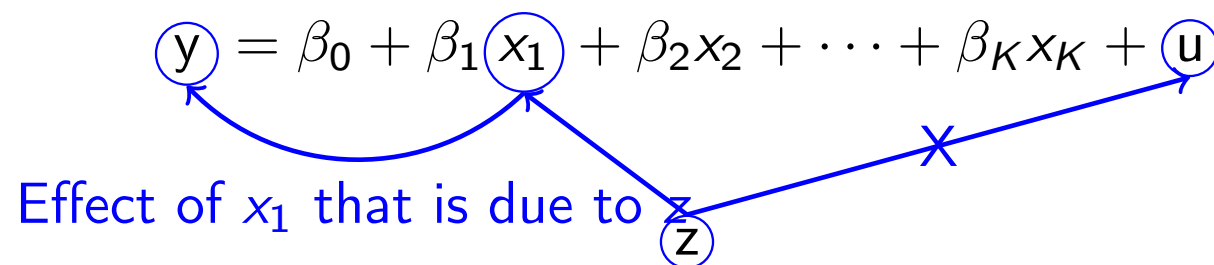


Estimation: principle

$$\textcircled{y} = \beta_0 + \beta_1 \textcircled{x_1} + \beta_2 x_2 + \cdots + \beta_K x_K + \textcircled{u}$$

A diagram illustrating a path from a node z to a node x_1 and then to a node u . The nodes z , x_1 , and u are represented by blue circles. A blue arrow points from z to x_1 . A blue arrow points from z to u , with a blue 'X' mark on the path between z and u .

Estimation: principle



Two-stage least square

The IV estimator is equivalent to the two-stage least square estimator:

- 1) 1st stage: regress the endogenous variable on the instrument and all the exogenous variables:

$$x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 \cdots + \pi_K x_K + v$$

⇒ get the OLS estimation

$$\hat{x}_1 = \hat{\pi}_0 + \hat{\pi}_1 z + \hat{\pi}_2 x_2 \cdots + \hat{\pi}_K x_K$$

- 2) 2nd stage: estimate the model by OLS, replacing x_1 endogenous by \hat{x}_1 exogenous:

$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$

⇒ The OLS estimator $\hat{\beta}_{2SLS} = \hat{\beta}_{IV}$ is consistent

Testing for endogeneity

- ▶ Suppose that all explanatory variables are exogenous
 - ▶ Then both the OLS and IV/2SLS estimator are consistent
 - ▶ But the IV estimator is less efficient than the OLS estimator (higher variance)
 - ▶ Thus when the variable that is supposed to be endogenous is not, the use of IV comes at a price: the variance of the IV estimator is larger than the variance of the OLS estimator
-
- ⇒ You may want to test for the endogeneity of the variable
 - ⇒ To know whether IV is necessary

Hausman test

One want to test:

$$H_0 : cov(x_1 u) = 0$$

$$H_1 : cov(x_1, u) \neq 0$$

Hausman test

1) Estimate the first stage equation by OLS:

$$x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 \cdots + \pi_K x_K + v$$

⇒ get the estimated residuals \hat{v}

2) Estimate the model, adding \hat{v} as a covariate:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + \gamma \hat{v} + w$$

⇒ Test for $H_0 : \gamma = 0$ with a t-test

Multiple instruments

- ▶ It is possible to use more than one instrument
- ▶ You need to have at least one instrument for one endogenous variable
- ▶ Suppose that you have two instruments z_1 and z_2 for x_1
- ▶ In this case, you can test that:

$$H_0 : E(z_1 u) = E(z_2 u) = 0$$

NB You can never formally test the exclusion restriction

- ▶ This is an over-identification test

Sargan test

Sargan test

- 1) Estimate the model by 2SLS

$$x_1 = \pi_0 + \pi_1 z + \pi_2 x_2 + \cdots + \pi_K x_K + v$$
$$y = \beta_0 + \beta_1 \hat{x}_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$

⇒ get the 2SLS residuals \hat{u}

- 2) Regress \hat{u} on a constant and all exogenous variables:

$$\hat{u} = \lambda_0 + \lambda_1 z + \lambda_2 x_2 + \cdots + \lambda_K x_K + \omega$$

⇒ get the R-squared from the regression $R_{\hat{u}}^2$

$$NR_{\hat{u}}^2 \underset{H_0}{\sim} \chi^2(1)$$

⇒ If the null is rejected, then at least one of the instrument is not exogenous