Erasmus
School of
Economics

Econometrics I

Lecture 5: Endogeneity

Annika Camehl

November - December, 2021

Today

1. Endogenous regressors

2. Instrumental variables – IV/2SLS

3. Testing for exogeneity and validity of instruments

Endogenous regressors

Endogenous regressors and instrumental variables

We start with the linear regression model

$$y_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where assumptions 2, 3, 4, 5, and 6 hold.

In matrix notation:

$$y = X\beta + \varepsilon.$$

The regressors x_i are now **stochastic**, assumption 1 (fixed X) is replaced with Assumption 1*: $\operatorname{plim}\left(\frac{1}{n}X'X\right) = Q$ of rank k.

Endogenity

When the orthogonality condition

$$\text{plim}\left(\frac{1}{n}X'\varepsilon\right) = 0$$

holds, the regressors X are called **exogenous**.

There is **endogeneity** or endogenous regressors when this condition does not hold, that is,

$$\text{plim}\left(\frac{1}{n}X'\varepsilon\right) \neq 0,$$

or in other words $\text{cov}(x_{ji}, \varepsilon_i) \neq 0$ for some j.

Examples and intuition

Possible reasons for correlation between X and ε are:

1. Omitted variables that influence X and y (through ε):

Example

Suppose we model the grade for Econometrics I using the number of followed lectures as explanatory factor.
An omitted variable is prior knowledge of the student. This is possibly correlated with the lectures taken and the final grade.
$\rightarrow$ Endogeneity!

Examples and intuition

Possible reasons for correlation between X and ε are:

1. Omitted variables that influence X and y (through ε):

Example

Suppose we model the grade for Econometrics I using the number of followed lectures as explanatory factor.
An omitted variable is prior knowledge of the student. This is possibly correlated with the lectures taken and the final grade.
$\rightarrow$ Endogeneity!

2. Strategic behavior based on unobserved factors:

Example

Model sales vs. prices. If the salesperson determines the price and has extra knowledge (on ε), the price will be correlated with the error.
$\rightarrow$ Endogeneity!

Example 5.1: Endogeneity

Consequences of endogeneity

The OLS estimator equals $b = \beta + (X'X)^{-1}X'\varepsilon = \beta + \left(\frac{1}{n}X'X\right)^{-1}\left(\frac{1}{n}X'\varepsilon\right)$, therefore

$$\text{plim}(b) = \beta + \text{plim}\left(\frac{1}{n}X'X\right)^{-1}\text{plim}\left(\frac{1}{n}X'\varepsilon\right) = \beta + Q^{-1}\text{plim}\left(\frac{1}{n}X'\varepsilon\right).$$

If $\text{plim}\frac{1}{n}X'\varepsilon \neq 0$ then we have endogeneity

- In general we then have correlation between X and ε
  (Note: no correlation does not necessarily mean *independent*)
- When there is endogeneity $\text{plim}(b) \neq \beta$, in other words the OLS estimator is **inconsistent**
- Testing/estimation is no longer useful!

Causes of endogeneity (theoretical)

There are three important causes of endogeneity:

- Omitted variables
- Measurement errors
- Simultaneity

Causes of endogeneity (theoretical)

There are three important causes of endogeneity:

- Omitted variables
- Measurement errors
- Simultaneity

1. Omitted variables

Suppose that in reality y_i depends on two variables, x_{2i} and x_{3i}, but we only use x_{2i} in the model for y_i, that is,

$$\begin{aligned} \text{DGP} &: y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i, \\ \text{Model} &: y_i = \beta_1 + \beta_2 x_{2i} + \varepsilon_i, \end{aligned}$$

such that $\varepsilon_i = \beta_3 x_{3i} + u_i$. When $\beta_3 \neq 0$ and $\text{cov}(x_{2i}, x_{3i}) \neq 0$, it follows that $\text{cov}(x_{2i}, \varepsilon_i) = \beta_3 \text{cov}(x_{2i}, x_{3i}) \neq 0$, or x_{2i} is endogenous.

Causes of endogeneity (theoretical)

2. Measurement error

- Suppose that in reality y_i depends on x_{2i}^*

$$\text{DGP} : y_i = \beta_1 + \beta_2 x_{2i}^* + u_i.$$

- x_{2i}^* is observed with an error e_i (with mean 0 and variance σ_e^2). We observe $x_{2i} = x_{2i}^* + e_i$.
- x_{2i} is endogenous in the regression of x_{2i} on y_i

Causes of endogeneity (theoretical)

3. Simultaneity

DGP:

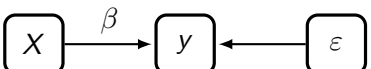
$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$
$$x_i = \gamma_1 + \gamma_2 \varepsilon_i + \eta_i$$

Model: $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$

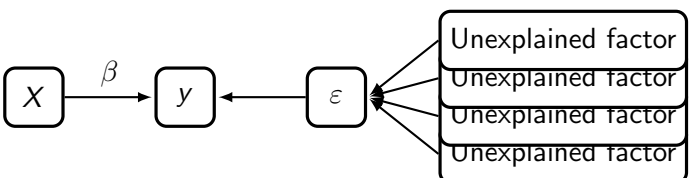
Now it holds "by definition" that x_i is correlated with ε_i.
$\rightarrow$ Both variables y_i and x_i are determined at the same time (simultaneously)

Instrumental variables

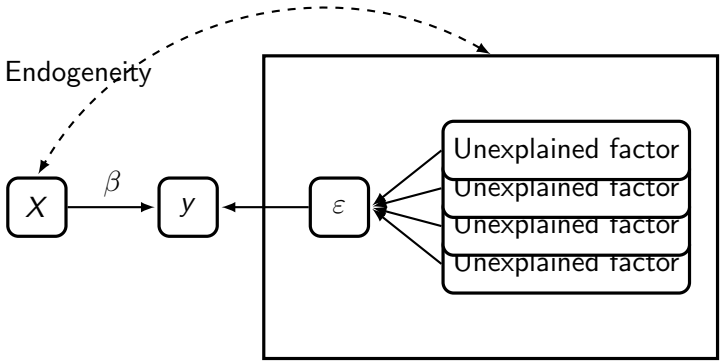
"Solving endogeneity": Graphical representation



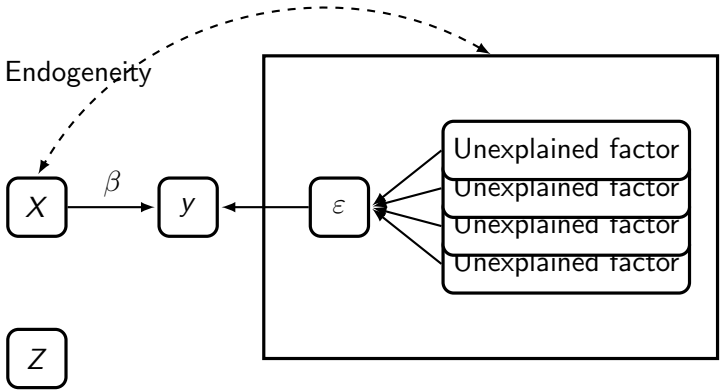
"Solving endogeneity": Graphical representation



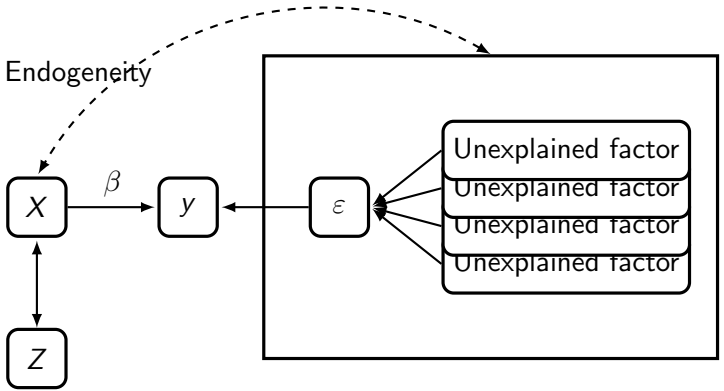
"Solving endogeneity": Graphical representation



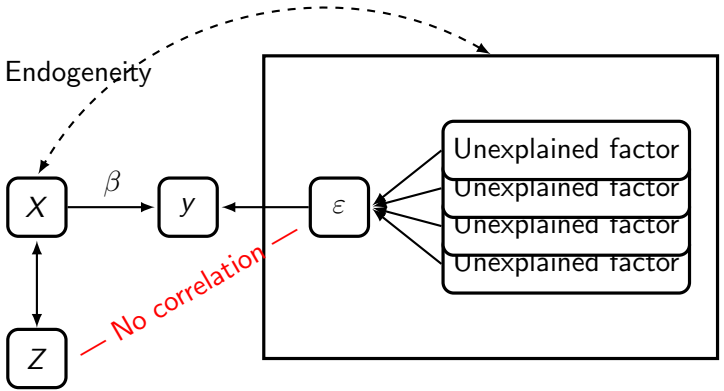
"Solving endogeneity": Graphical representation



"Solving endogeneity": Graphical representation



"Solving endogeneity": Graphical representation



Important: $\text{corr}(Z, \varepsilon) = 0$ does not imply $\text{corr}(y, Z) = 0$

"Solving endogeneity": Instrumental variables

Given the model

$$y_i = x_i'\beta + \varepsilon_i, \qquad \text{with } \operatorname{cov}(x_i, \varepsilon_i) \neq 0$$

and with k explanatory variables

Idea:

- Find m variables (z_i) that are correlated with x_i, but not with ε_i
- Part of x_i that is explained by z_i is by definition not correlated with ε_i

Procedure:

- Explain X with $Z \rightarrow$ explained part uncorrelated with ε.
- Use explained part to explain y.

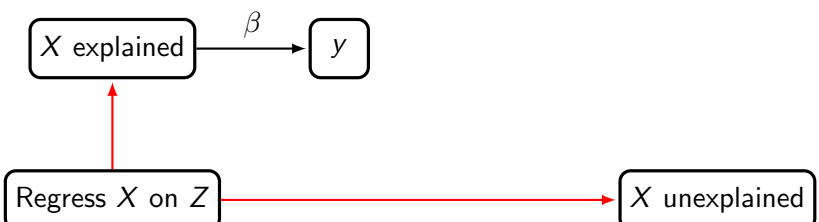
$\rightarrow$ 2-stage least squares (2SLS)

"Solving endogeneity": Graphical representation

1. Use Z to decompose X in explained and unexplained part

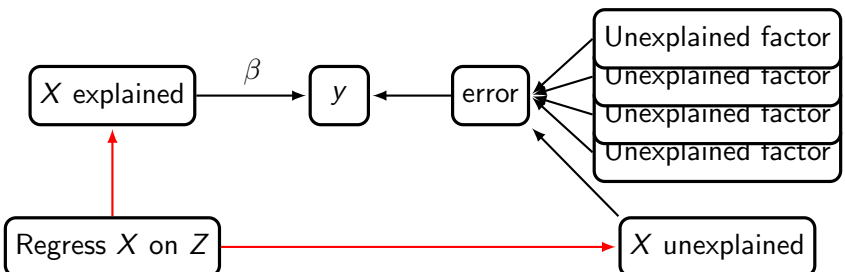
"Solving endogeneity": Graphical representation

1. Use Z to decompose X in explained and unexplained part
2. Effect size of explained part on y equals β



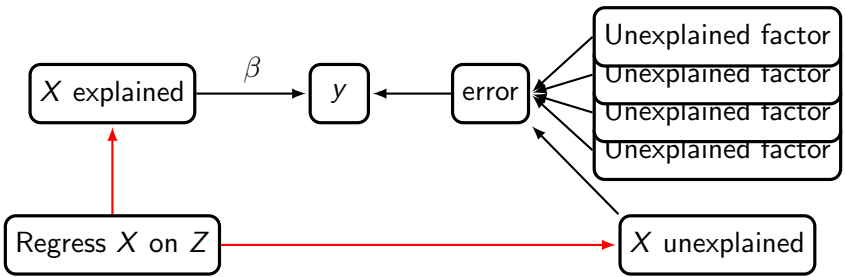
"Solving endogeneity": Graphical representation

1. Use Z to decompose X in explained and unexplained part
2. Effect size of explained part on y equals β
3. Unexplained part is added to error term



"Solving endogeneity": Graphical representation

1. Use Z to decompose X in explained and unexplained part
2. Effect size of explained part on y equals β
3. Unexplained part is added to error term



Endogeneity is solved as

- X unexplained not correlated with X explained
- X explained is exogenous

Two conditions for the instruments

(1) Z and ε uncorrelated

$$\text{plim}\,\frac{1}{n}Z'\varepsilon = 0$$

validity of the instrument

(2) Z (sufficiently) correlated with X

$$\text{plim}\,\frac{1}{n}Z'X = Q_{ZX}$$

with $\text{rank}(Q_{ZX}) = k$ (rank condition)
relevance of the instrument

$\rightarrow$ If the instrument fails condition 1 (Z and ε uncorrelated) it is invalid, if it fails condition 2 (Z correlated with X) it is irrelevant.

Finding instruments

What are good instruments?

- All exogenous variables in x_i (incl. constant)
  (These have to be included in Z)
- Other instruments are always needed. Use the properties of the problem at hand.
  $\rightarrow$ These additional instruments should not directly influence y.
- The stronger the correlation between Z and X the better (as long as there is no correlation between Z and ε)

$\rightarrow$ For every endogenous variable we need to find at least one instrument

Examples of instruments

- In example attendance/grade: policy change to obligatory attendance
- In example price/sales: prices of raw materials

Requirements IV/2SLS

To perform 2SLS we need:

1. Z and ε uncorrelated
$$\text{plim}\,\frac{1}{n}Z'\varepsilon = 0$$
2. Z (sufficiently) correlated with X
$$\text{plim}\,\frac{1}{n}Z'X = Q_{ZX}$$
with $\text{rank}(Q_{ZX}) = k$ (rank condition)
3. Z "stable" and not "multi collinear"
$$\text{plim}\,\frac{1}{n}Z'Z = Q_{ZZ}$$
with $\text{rank}(Q_{ZZ}) = m$
4. Enough instruments (Z) (order condition)
$$m \geq k$$

IV estimator

case: $m = k$

Given model

$$y = X\beta + \varepsilon$$

with all assumptions for IV (see earlier slide).

IV estimator:

$$b_{IV} = (Z'X)^{-1}Z'y$$

Intuitively for single regression (one variable and one instrument):

$$\begin{aligned} b_{IV} &= \frac{\delta y/\delta Z}{\delta X/\delta Z} \\ &= \frac{(Z'Z)^{-1}Z'y}{(Z'Z)^{-1}Z'X} \\ &= (Z'X)^{-1}Z'y \end{aligned}$$

2-stage least squares

case: $m \geq k$

Given model

$$y = X\beta + \varepsilon$$

with all assumptions for IV (see earlier slide).

Steps:

1. regress X on Z
   $\rightarrow \hat{X} = Z(Z'Z)^{-1}Z'X = P_Z X$
2. regress y on $\hat{X}$
   2SLS estimator:

$$\begin{aligned} b_{2SLS} &= (\hat{X}'\hat{X})^{-1}\hat{X}'y \\ &= (X'P_Z'P_Z X)^{-1}X'P_Z'y \\ &= (X'P_Z X)^{-1}X'P_Z y \end{aligned}$$

If $m = k$: $b_{IV} = (Z'X)^{-1}Z'y$ the 2SLS estimator simplifies to the IV estimator

Instrumental variables as GMM estimator

The assumptions on ε_i, x_i and z_i can be summarized in m moment conditions

$$E[z_i\varepsilon_i] = E\begin{pmatrix} z_{1i}\varepsilon_i \\ z_{2i}\varepsilon_i \\ \vdots \\ z_{mi}\varepsilon_i \end{pmatrix} = 0$$

As $\varepsilon_i = y_i - x_i'\beta$ we have $E[z_i\varepsilon_i] = E[z_i(y_i - x_i'\beta)]$
If $m \geq k$ the m conditions provide enough information to estimate the k parameters (if $m = k$ exactly enough info).

For a specific dataset we need to solve

$$G_n(\beta) = \sum_{i=1}^{n} g_i(\beta) = \sum_{i=1}^{n} z_i(y_i - x_i'\beta) = Z'(y - X\beta) = 0$$

If $m > k$ this cannot be done exactly

IV as GMM estimator

"Solving": $G_n(\beta) = Z'(y - X\beta) = 0$
$\rightarrow$ Try to come as close as possible to 0

Solve: $\min_\beta G_n' W G_n$

The optimal weight matrix is

$$W = J_0^{-1} = (\mathrm{E}[g_i(\beta) g_i(\beta)'])^{-1} = (\mathrm{E}[(z_i \varepsilon_i)(z_i \varepsilon_i)'])^{-1} = (\mathrm{E}[\varepsilon_i^2 z_i z_i'])^{-1}$$

For a specific dataset we estimate W with $(\hat{\sigma}^2 \frac{1}{n} Z'Z)^{-1}$

So

$$\min(y - X\beta)' Z (Z'Z)^{-1} Z' (y - X\beta) = \min(y - X\beta)' P_Z (y - X\beta)$$

$\rightarrow b_{IV} = (X' P_Z X)^{-1} X' P_Z y$

Properties of two stage least squares

The IV/2SLS estimator b_{IV}/b_{2SLS} has the following properties:

1. consistency
2. asymptotic normality:

$$b_{2SLS} \approx N\left(\beta, \sigma^2(\hat{X}'\hat{X})^{-1}\right).$$

Important remarks

1. In the "first-stage" regression you have to include all exogenous variables
2. The variance of the residuals ε_i, σ^2 (necessary for the standard errors of b_{2SLS}), is to be estimated using the "2SLS"-residuals: $e_{2SLS} = y - Xb_{2SLS}$.

Example 5.1: IV estimation

Testing for exogeneity and validity of instruments

Testing for exogeneity and validity of instruments

Testing for exogeneity

1. Hausman test (Durbin, Wu, & Hausman): comparison of covariance matrices of OLS and IV estimator under H_0 of exogeneity

Testing the validity of instruments

1. Sargan test (look at correlation of residuals with Z)
2. Look at correlation of endogenous variables and Z

Hausman test

$H_0 : plim \frac{1}{n} X'\varepsilon = 0$

- under H_0 : OLS is consistent + efficient and IV is consistent
- under H_0 (exogeneity) OLS and IV give comparable results (both consistent), under H_1 we expect a difference
- Let $d = b_{IV} - b$. Under H_0 : $\mathrm{E}(d) \approx 0$ and, if errors are *iid*, $\mathrm{Var}(d) \approx \mathrm{Var}(b_{IV}) - \mathrm{Var}(b)$
- Test statistic:
  $(b_{IV} - b)'(\mathrm{Var}(b_{IV}) - \mathrm{Var}(b))^{-1}(b_{IV} - b) \approx \chi^2(k)$ under H_0
- Problem: *iid* is a strong assumption and in small samples one can have problems as $\mathrm{Var}(b_{IV}) - \mathrm{Var}(b)$ is not always positive definite.

Alternative testing procedure

Use $\frac{1}{n}\sum_{i=1}^{n} x_{ji}\varepsilon_i$ with an estimate of ε_i, and next test whether this significantly differs from 0.

- Note that we cannot do this using OLS residuals $e_{OLS,i} = y_i - x_i' b_{OLS}$, as it holds that

$$\frac{1}{n}\sum_{i=1}^{n} x_{ji} e_{OLS,i} = 0.$$

- Split x_{ji} in an exogenous and an endogenous part, by performing the "first-stage" regression:

$$x_{ji} = z_i'\gamma_j + v_{ji}.$$

$\Rightarrow$ It holds $E[x_{ji}\varepsilon_i] = E[v_{ji}\varepsilon_i]$ as we assume that $E[z_i\varepsilon_i] = 0$.

Testing exogeneity

- We can now test if $E[v_{ji}\varepsilon_i] = 0$ by testing $H_0 : \alpha = 0$ in

$$\varepsilon_i = \alpha v_{ji} + w_i,$$

or

$$y_i = x_i'\beta + \alpha(x_{ji} - z_i'\gamma) + w_i,$$

Procedure (for k_0 endogenous variables)

1. Regress y on $X \rightarrow e = y - Xb$
2. Regress all endogenous variables x_j on $Z \rightarrow v_j = x_j - Z\hat{\gamma}_j$
3. Regress e on X and v_j, $j = 1, \ldots, k_0$
4. $nR^2 \approx \chi^2(k_0)$ under H_0 (exogeneity)

Testing the validity of instruments

A crucial assumption for IV/2SLS is exogeneity of the instruments $z_{1i}, \ldots, z_{mi}$, that is, $E[z_{li}\varepsilon_i] = 0, \quad l = 1, \ldots, m.$

If this does not hold, then IV is not consistent and the Hausman test is not valid.

We can test this assumption by testing $H_0 : \gamma = 0$ in

$$\varepsilon_i = z_i'\gamma + \eta_i,$$

where for implementation it is important that we use a consistent estimate of ε_i. This is given by the "IV"- or "2SLS"-residuals

$$e_{2SLS,i} = y_i - x_i' b_{2SLS}.$$

This leads to the so-called Sargan-test.

Testing the validity of instruments

Procedure Sargan test

1. Perform IV for $y = X\beta + \varepsilon$ with instruments $Z \rightarrow e_{IV} = y - Xb_{IV}$
2. Regress e_{IV} on Z
3. $nR^2 \approx \chi^2(m-k)$ under H_0 (valid instruments)

Notes:

- this test only works if the model is overidentified ($m > k$)
- "Identifying assumptions" must still hold!

Example 5.1: Testing for exogeneity and validity

Let's do a small quiz