Erasmus School of Economics

MOOC Econometrics

Lecture 4.1 on Endogeneity: Motivation

Dennis Fok

Erasmus University Rotterdam

Motivating example

We want to explain

- *Number of flights at an airport per month* (y) using
- *Number of travel insurances made in previous month* (x)

Suppose OLS yields

$$y = 10,000 + .25x + e$$

Test

How should we interpret the obtained coefficients?
What does the estimate .25 really mean?

Interpretation of parameters

Given the estimates (y: flights, x: insurances)

$$y = 10,000 + .25x + e$$

Correct: $4,000$ *insurances sold* $\rightarrow$ *expected number of flights*
$= 10,000 + .25 \times 4,000 = 11,000$

- High x tends to go together with high y.
- The identified correlation yields adequate predictions.

Incorrect: *Selling* $4,000$ *additional insurances causes*
$.25 \times 4,000 = 1,000$ *additional flights*

- The regression does not identify a *causal* impact!
- A third variable (*travel demand*) affects y (*flights*) and x (*insurances*).

Endogeneity

OLS requires some assumptions:

- explanatory variables should be exogenous
- violation of this: *endogeneity*.

In this set of lectures, you will learn to:

1. Understand/recognize endogeneity.
2. Know the consequences of endogeneity.
3. Estimate parameters under endogeneity.
4. Know the intuition of the new estimator.
5. Argue/test assumptions underlying this new estimator.

Stochastic vs. non-stochastic regressors

Standard assumptions for linear model ($y = X\beta + \varepsilon$) include

A2 Explanatory variables are *non-stochastic*

Implications:

- Obtain new data: X stays constant (and y changes)
- Need "controlled experiment"
- OLS estimator b converges to true coefficient β for $n \to \infty$ (OLS is *consistent*)

Economic models

In economics:

- Controlled (or natural) experiments are rare
- New data with same X cannot be obtained
- Explanatory variables are *stochastic*!

If X stochastic:

- new data set $\to$ new X values
- X can be correlated with other variables
- If X correlated with ε
  - X is endogenous
  - There is another variable that affects y and X
  - OLS does not properly estimate β (inconsistent)
- If X uncorrelated with ε
  - X is exogenous
  - OLS consistent

Other examples of endogeneity – Omitted variables

- True model is
$$y = X_1\beta_1 + X_2\beta_2 + \eta$$
but we ignore X_2 and perform OLS on
$$y = X_1\beta_1 + \varepsilon$$
- We have: $\varepsilon = X_2\beta_2 + \eta$
- X_1 correlated with ε (X_1 is endogenous) if
  - X_1 correlated with X_2 *and*
  - $\beta_2 \neq 0$

  Derivation:
$$\begin{aligned}\text{Cov}(X_1, \varepsilon) &= \text{Cov}(X_1, X_2\beta_2 + \eta) \\ &= \text{Cov}(X_1, X_2)\beta_2 + \underbrace{\text{Cov}(X_1, \eta)}_{=0}\end{aligned}$$

Omitted variable – Example

Model *student's grade* using *attendance* at lectures.

Test

Which omitted factor would lead to endogeneity of attendance?

Three possible omitted factors:

1. Difficulty of exam
   NO: not correlated with attendance.
2. Motivation of the students?
   YES: correlates with attendance and affects grade.
3. Compulsory attendance yes/no?
   NO: does not directly impact the grade

Other examples – Strategic behavior

Consider a model explaining demand using price.

Strategic price setting:

1. Sets high price when high demand is expected
2. Price and sales positively correlated
3. Price will be endogenous in regression of demand on price.

Other examples – Measurement errors

- y (eg. salary) depends on x^* (eg. intelligence)
- x^* (intelligence) difficult to observe
- $x = x^* +$ measurement error: noisy measurement (eg. IQ score)
- measurement error: x is endogenous in $y = \alpha + \beta x + \varepsilon$

Summary & what's next?

- Endogeneity is a common problem
- OLS is not useful under endogeneity

Upcoming topics:

- How to solve for endogeneity?
- How to test for endogeneity?

TRAINING EXERCISE 4.1

- Train yourself by making the training exercise (see the website).
- After making this exercise, check your answers by studying the webcast solution (also available on the website).