Erasmus School of Economics

MOOC Econometrics

Lecture 4.5 on Endogeneity: Application

Dennis Fok

Erasmus University Rotterdam

Application

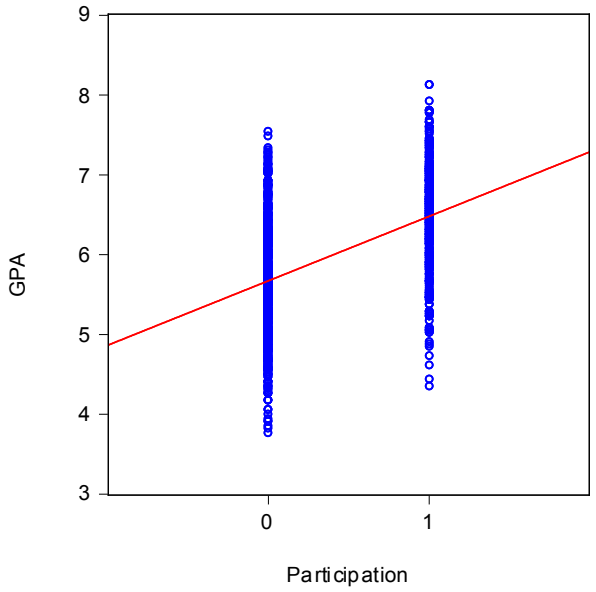
Setting:

- Online learning platform
- Grade Point Average (GPA) in MOOC on engineering
- Impact of preparatory mathematics course
  $\rightarrow$ participation is voluntary!

Data statistics:

- 1000 learners
- 48.8% male
- 33.7% participated in prep course
- Average GPA 5.94 (on 10 point scale)

Correlation of GPA with participation



Correlation vs. regression

Seems positive impact

- How large?
- Significant?
- Correction for male vs. female?

$\rightarrow$ Need econometric model!

OLS estimation

Regress GPA on

1. Constant
2. Gender: dummy variable (male=1, female=0)
3. Participation: dummy variable (yes=1, no=0)

Dependent variable: GPA Sample size: 1000			
	Coefficient	Standard error	t-statistic
Constant	5.77	0.034	169.87
Gender	−0.21	0.044	−4.82
Participation	0.82	0.047	17.59
R^2	0.24		

Discussion of OLS

Should we trust the OLS estimates?

→ No, participation likely endogenous!

- Learners self-select for prep course
- Omitted factors (characteristics of learners) relate to this selection
- Same characteristics may relate to GPA

Over- or underestimation by OLS?

If prep course participation is endogenous

- OLS is inconsistent
- OLS does not estimate causal effect of prep course

Test

What omitted factor would lead OLS to overestimate the impact of the preparatory course?

Over- or underestimation by OLS?

Overestimation

- Omitted factor: Motivation
  High motivation → Get high GPA & Take course

Underestimation

- Omitted factor: Mathematics level
  High level → Get high GPA & Do not take course

Net effect:

- Difficult to judge
- Depends on importance of effects
- Also depends on other variables (age?)

Consistent estimation

- Use two-stage least squares (2SLS)
- Need instruments!

Test

What variable can you think of that qualifies as instrument for participation?

Instruments

Instruments should...

- relate to prep course participation
- not affect GPA

Many learner specific variables, such as

- Intelligence (IQ-score)
- Number of MOOCs followed before
- Age of learner

are likely not valid!
→ All will impact performance directly!

Instruments

Finding instruments

- be creative! ... and lucky

Here

- Learners get email invitation for prep course
- Platform email problem: some did not get email
- Variable

$$\text{Email} = \begin{cases} 0 & \text{if email not received} \\ 1 & \text{if email received} \end{cases}$$

  is perfect instrument if
  - Email problem is random
  - Invitation affects participation

First-stage regression

Explain participation using all instruments (constant, gender, email)

Dependent variable: Participation Sample size: 1000			
	Coefficient	Standard error	t-statistic
Constant	0.10	0.023	4.41
Gender	0.05	0.027	1.80
Email	0.41	0.027	15.35
R^2	0.20		

→ Email affects participation significantly

2SLS estimation

Dependent variable: GPA Sample size: 1000 Instruments used: Constant, Gender, Email			
	Coefficient	Standard error	t-statistic
Constant	5.95	0.048	123.54
Gender	−0.17	0.048	−3.59
Participation	0.24	0.115	2.09
R^2	0.13		

- Prep course still has significant positive impact
- Effect size decreased (from 0.82 (OLS) to 0.24 (2SLS))
- 2SLS increases variance
  - Only acceptable when Participation is endogenous
  - Perform Hausman test

Hausman test (H_0: Participation is exogenous)

Dependent variable: Residuals from OLS Sample size: 1000			
	Coefficient	Standard error	t-statistic
Constant	0.18	0.044	4.02
Gender	0.04	0.044	0.93
Participation	− 0.58	0.105	−5.55
First-stage residuals (v)	0.72	0.117	6.17
R^2	0.0368		

- Test-statistic: $nR^2 = 1000 \times 0.0368 = 36.8$
- Reject H_0 (critical value from $\chi^2(1)$: 3.8)
- Participation is endogenous
- 2SLS is needed

TRAINING EXERCISE 4.5

- 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).