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).

Erasmus
School of
Economics

MOOC Econometrics

Lecture 4.2 on Endogeneity: Consequences

Dennis Fok

Erasmus University Rotterdam

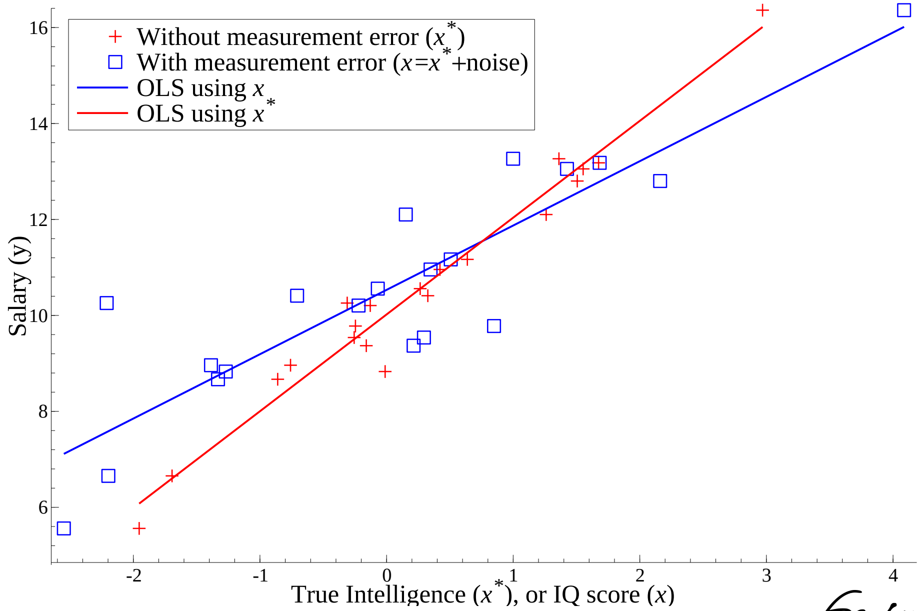
Endogeneity

- Common problem in economics
  1. Omitted variables
  2. Strategic behavior
  3. Measurement errors

  $\rightarrow$ X is correlated with ε
- Endogeneity violates the basic assumptions

$\rightarrow$ How bad is this?

Simulated example, $y = 1 + 2x^* + u$



Measurement error example

Under measurement error (and endogeneity in general):

- we obtain the wrong coefficients!

Test

Can we say anything about the direction of the bias?

Direction of bias in the measurement error case

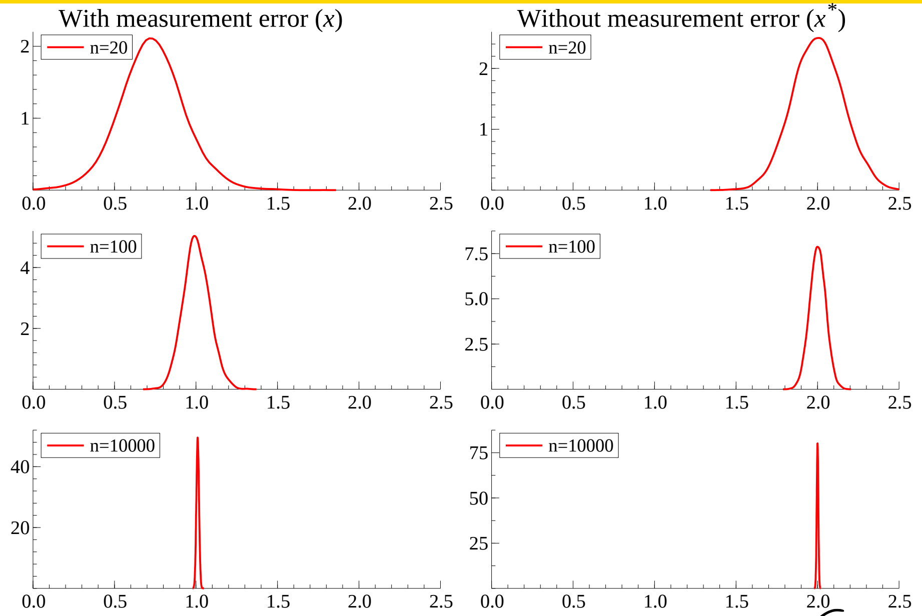
OLS is "biased towards zero"
→ OLS underestimates true effect

Intuitively:

- x-values on the *left* likely have negative measurement errors
- x-values on the *right* likely have positive measurement errors

Measurement errors "stretch" the scatter in the horizontal direction
→ a flatter regression line

Distribution of estimator for different n, true value= 2



Consistency: formal argumentation

If X is endogenous:

- If n grows the OLS estimator converges to the wrong value.
  → OLS is inconsistent

Consider the standard model $y = X\beta + \varepsilon$ and the OLS estimator

$$
\begin{aligned}
b &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + \varepsilon) \\
&= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'\varepsilon \\
&= \beta + (X'X)^{-1}X'\varepsilon
\end{aligned}
$$

So, b can be split into

1. True parameter value β
2. Random deviation $(X'X)^{-1}X'\varepsilon$

Asymptotic properties

What happens to b as $n \to \infty$?

Recall: $b = \beta + (X'X)^{-1}X'\varepsilon$

- β is constant
- Elements of $(X'X)$ and $X'\varepsilon$ are sums over observations:

$$
X'X = \begin{pmatrix} \sum_{i=1}^n x_{1i}^2 & \sum_{i=1}^n x_{1i}x_{2i} & \cdots & \sum_{i=1}^n x_{1i}x_{ki} \\ \sum_{i=1}^n x_{1i}x_{2i} & \sum_{i=1}^n x_{2i}^2 & \cdots & \sum_{i=1}^n x_{2i}x_{ki} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n x_{ki}x_{1i} & \sum_{i=1}^n x_{ki}x_{2i} & \cdots & \sum_{i=1}^n x_{ki}^2 \end{pmatrix}, X'\varepsilon = \begin{pmatrix} \sum_{i=1}^n x_{1i}\varepsilon_i \\ \sum_{i=1}^n x_{2i}\varepsilon_i \\ \vdots \\ \sum_{i=1}^n x_{ki}\varepsilon_i \end{pmatrix}
$$

→ these diverge as $n \to \infty$

Asymptotic properties

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

- $\left(\frac{1}{n}X'X\right)$ is an average
  $\rightarrow$ in general converges to, say, Q
- $\left(\frac{1}{n}X'\varepsilon\right)$ also converges in general

Consistency result:
b converges to β as $n \rightarrow \infty$ if

1. $\frac{1}{n}X'X$ converges to Q, *and*
2. Q^{-1} exists, *and*
3. $\frac{1}{n}X'\varepsilon$ converges to 0
   - No correlation between X and ε (for large n)
   - X is exogenous

X endogenous: b does not converge to β!

Small sample properties

So far we discussed what happens for $n \rightarrow \infty$

Test
Why can't we derive the bias?

To obtain the bias

- need to evaluate

$$\begin{aligned} \mathrm{E}[b] &= \mathrm{E}[(X'X)^{-1}X'y] = \mathrm{E}[(X'X)^{-1}X'(X\beta+\varepsilon)] \\ &= \mathrm{E}[\beta + (X'X)^{-1}X'\varepsilon] = \beta + \underbrace{\mathrm{E}[(X'X)^{-1}X'\varepsilon]}_{=?}. \end{aligned}$$

- X is stochastic
- cannot simplify final expectation (without further assumptions)

OLS in presence of endogeneity

If X endogenous

- X correlated with ε
- OLS estimator for β is not consistent
- Even with in infinite amount of data: OLS does not give useful estimates

TRAINING EXERCISE 4.2

- 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).

Erasmus
School of
Economics

MOOC Econometrics

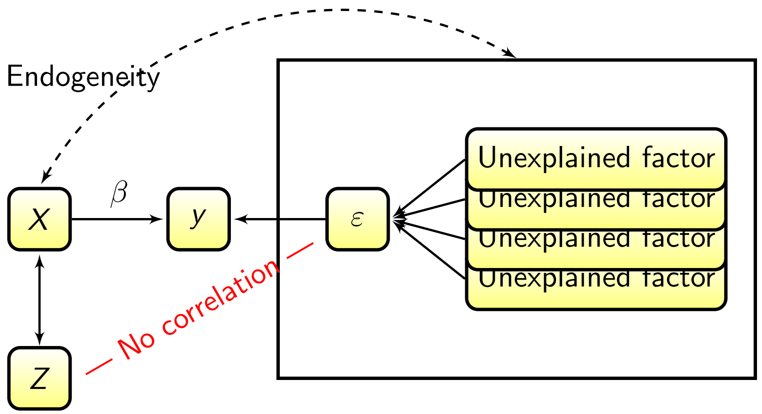
Lecture 4.3 on Endogeneity: Estimation under endogeneity

Dennis Fok

What have we so far?

- Endogeneity is a common problem
- Endogeneity causes OLS to be inconsistent
- Estimation requires another estimation technique

"Solving endogeneity": Graphical representation



Instrumental variable estimation

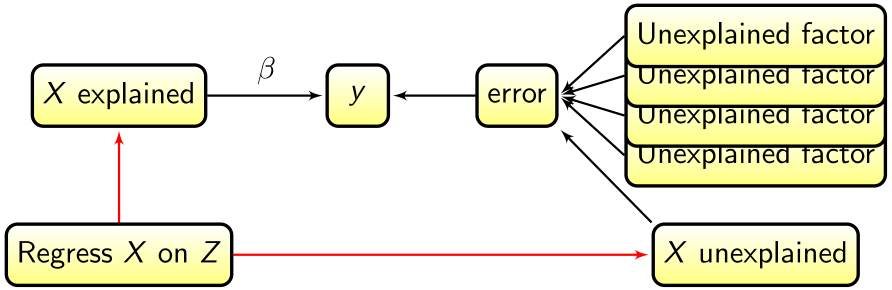
- Z variables are *instruments* if
  - Z and X are correlated
  - Z does not correlate with ε
- Correlation between instruments and y is only due to X

$$\begin{aligned}\mathrm{Cov}(Z, y) &= \mathrm{Cov}(Z, X\beta + \varepsilon) = \mathrm{Cov}(Z, X\beta) + \underbrace{\mathrm{Cov}(Z, \varepsilon)}_{=0} \\ &= \mathrm{Cov}(Z, X)\beta\end{aligned}$$

- Use instruments to estimate 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



Endogeneity is solved as

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

2SLS in matrix notation

Given model

$$y = X\beta + \varepsilon, \quad \text{Var}[\varepsilon] = \sigma^2 I$$

and instruments Z

1. Regress X on Z to get explained part:
   - Model: $X = Z\gamma + \eta$
   - OLS estimate: $(Z'Z)^{-1}Z'X$
   - Fitted value: $\hat{X} = \underbrace{Z(Z'Z)^{-1}Z'}_{H_Z}X = H_Z X$
2. Regress y on $\hat{X}$:

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

Use: $H_Z = H_Z' = H_Z' H_Z$

Properties 2SLS

- Variance of b_{2SLS}: $\text{Var}[b_{2SLS}] = \sigma^2(X'H_Z X)^{-1}$
- Estimating σ^2:
  - $\hat{\sigma}^2 = \frac{1}{n-k}(y - Xb_{2SLS})'(y - Xb_{2SLS})$
  - Do **not** use residuals (or reported standard errors) of second stage regression!

Derivation of variance (use $\text{Var}[\varepsilon] = \sigma^2 I$):

$$\begin{aligned} b_{2SLS} &= (X'H_Z X)^{-1}X'H_Z y = (X'H_Z X)^{-1}X'H_Z(X\beta + \varepsilon) \\ &= \beta + (X'H_Z X)^{-1}X'H_Z\varepsilon \\ \text{Var}[b_{2SLS}] &= \text{Var}[(X'H_Z X)^{-1}X'H_Z\varepsilon] \\ &= (X'H_Z X)^{-1}X'H_Z \text{Var}[\varepsilon]\left((X'H_Z X)^{-1}X'H_Z\right)' \\ &= (X'H_Z X)^{-1}X'H_Z(\sigma^2 I)H_Z'X(X'H_Z X)^{-1} \\ &= \sigma^2(X'H_Z X)^{-1}\underbrace{X'\underbrace{H_Z H_Z'}_{H_Z}X(X'H_Z X)^{-1}}_{I} = \sigma^2(X'H_Z X)^{-1} \end{aligned}$$

Properties of 2SLS

- 2SLS is consistent if (when $n \to \infty$)
  - Z and ε not correlated: $\frac{1}{n}Z'\varepsilon \to 0$
  - Z not multicollinear: $\frac{1}{n}Z'Z \to Q_{ZZ}$, and Q_{ZZ} invertible
  - X and Z sufficiently correlated: $\frac{1}{n}X'Z \to Q_{XZ}$, and Q_{ZZ} rank k

Sketch of proof:

$$\begin{aligned} b_{2SLS} &= \beta + (X'H_Z X)^{-1}X'H_Z\varepsilon = \beta + (X'Z(Z'Z)^{-1}Z'X)^{-1}X'Z(Z'Z)^{-1}Z'\varepsilon \\ &= \beta + \underbrace{\underbrace{\left(\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}\frac{1}{n}Z'X\right)^{-1}}_{\downarrow \atop (Q_{XZ}Q_{ZZ}^{-1}Q_{XZ}')^{-1}}\underbrace{\frac{1}{n}X'Z\left(\frac{1}{n}Z'Z\right)^{-1}}_{\downarrow \atop Q_{XZ}Q_{ZZ}^{-1}}\underbrace{\frac{1}{n}Z'\varepsilon}_{\downarrow \atop 0}}_{=\beta+0} \end{aligned}$$

Finding instruments

What are good instruments?

- All exogenous variables in X (incl. constant)
- Other instruments are always needed:
  - At least one for every endogenous variable
  - Want: strong correlation between Z and X
  - Need: no correlation between Z and ε

Examples of instruments

Explain obtained grade using attendance:

Potential instruments:

- Travel time home to university
- Policy change to obligatory attendance

Test

What variable would be an instrument for price when modeling consumer sales of ice cream using sales $= \alpha + \beta$price $+ \varepsilon$?

Potential instruments?

1. Prices of raw materials (valid)
2. ~~Competitor prices~~ (direct influence on sales, so part of ε)
3. ~~Outside temperature~~ (direct influence on sales, so part of ε)

Summary

If X is in fact exogenous

- OLS and 2SLS both consistent
- Variance OLS smaller than variance 2SLS!

$\rightarrow$ Use OLS

If X is endogenous

- 2SLS is consistent
- OLS inconsistent

$\rightarrow$ Use 2SLS

TRAINING EXERCISE 4.3

- 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).

Erasmus School of Economics

MOOC Econometrics

Lecture 4.4 on Endogeneity: Testing for endogeneity

Dennis Fok

Erasmus University Rotterdam

Outline

Given

- Model $y = X\beta + \varepsilon$
- Instruments Z

Two important things to test

1. Z satisfies assumptions for instruments?
2. X exogenous or endogenous?

Testing the validity of instruments

Valid instruments satisfy three conditions

1. There are enough instruments
   $\rightarrow$ Easy! Just count.
2. Instruments are correlated (enough) with X
   $\rightarrow$ Check significance of instruments in first stage regression
3. Instruments are not correlated with ε
   $\rightarrow$ Perform *Sargan test*

Test correlation Z vs. X

- X_1 potentially endogenous variables
- X_2 exogenous variables
- $Z = (Z^*, X_2)$ instruments

First-stage regression: apply OLS to $X_1 = Z^*\gamma_1 + X_2\gamma_2 + \eta$

Test

Why does 2SLS require $\gamma_1 \neq 0$?

If $\gamma_1 \approx 0$:

- $\hat{X}_1 \approx X_2\hat{\gamma}_2$
  $\rightarrow \hat{X}_1$ almost perfectly correlated with X_2
- (Extremely) large estimation uncertainty

Test for sufficient correlation:

- Test $H_0 : \gamma_1 = 0$ in first-stage regression.

Sargan test

Ingredients:

- Model: $y = X\beta + \varepsilon$
- Explanatory variables: $X = (X_1, X_2)$
  X_1 (endogenous), X_2 (exogenous)
- Instruments: $Z = (Z^*, X_2)$

Null hypothesis (H_0): Correlation Z and ε equals 0

Test procedure:

- Rewrite to $H_0 : \delta = 0$ in
$$\varepsilon = Z\delta + \xi$$
- ε cannot be observed
  $\rightarrow$ Estimate ε using 2SLS

Sargan test

Procedure:

1. Use Z to obtain 2SLS estimator b_{2SLS} for β
2. Calculate $e_{2SLS} = y - Xb_{2SLS}$
3. Regress e_{2SLS} on Z
4. $nR^2 \approx \chi^2(m-k)$ under H_0 (valid instruments)
   - m instruments in Z
   - k explanatory variables in X

Test

The Sargan test requires $m > k$. What happens when $m = k$?

Notes on the Sargan test

- Test only works when there are “too many” instruments ($m > k$)
- At least k of the instruments should be valid
- Test cannot indicate which instruments are invalid!

Testing for exogeneity of variables – Hausman test

Intuition:

- Use the instruments to split potentially endogenous variables into
  1. a guaranteed exogenous part
  2. a potentially endogenous part
- Check whether the endogenous and exogenous part affect y differently.

Hausman test – procedure

Ingredients:

- Explanatory variables: $X = (X_1, X_2)$
- Potentially endogenous: X_1 (k_1 variables)
- Exogenous variables: X_2 (k_2 variables)
- Instruments: Z

Null hypothesis (H_0): X_1 is exogenous

Formal procedure:

1. Regress y on X $\rightarrow$ calculate $e = y - Xb$
2. Regress X_1 on Z $\rightarrow$ calculate residuals V
3. Regress e on X and V
4. $nR^2 \approx \chi^2(k_1)$ under H_0 of exogeneity

TRAINING EXERCISE 4.4

- 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).

Erasmus School of Economics

MOOC Econometrics

Lecture 4.5 on Endogeneity: Application

Dennis Fok

Erasmus University Rotterdam

Application

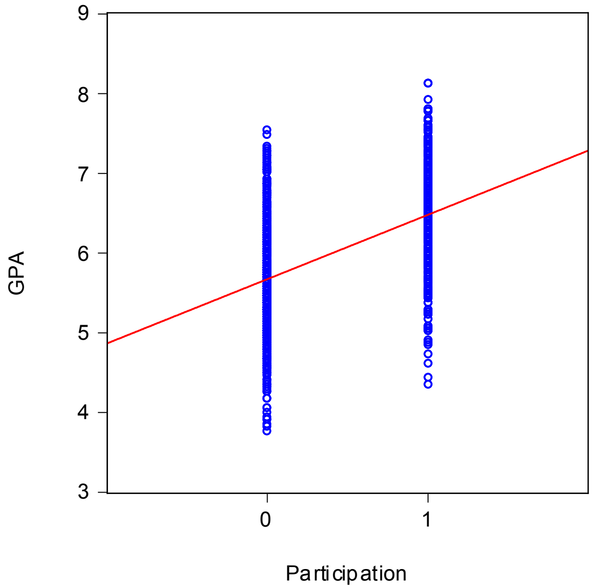
Setting:

- 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!
$\rightarrow$ 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		

$\rightarrow$ 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).