

# Class 16 Instrumental Variables and Two-Stage Least Squares

Dr. Wei Miao

UCL School of Management

November 20, 2024

## Section 1

# Instrumental Variable

# Class Objectives

- The requirements of a valid instrumental variable and how to find good instruments
- Intuition of why instrumental variables solve endogeneity problems
- Apply two-stage least square method to estimate the causal effects using instrumental variables

# Causal Inference from OLS

- From non-experimental secondary data, it is impossible to control all confounding factors, which means we can never obtain causal effects from OLS regressions.
- Can we still obtain causal inference from secondary data?

# What is an Instrumental Variable

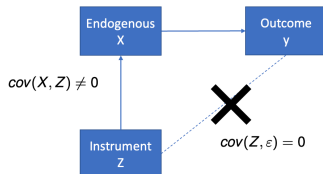
## Instrumental Variable

An instrumental variable is a set of variables  $Z$  that satisfies the following requirements:

- ①  $z$  is exogenous and uncorrelated with  $\epsilon$ ; that is,  $cov(Z, \epsilon) = 0$
- ②  $z$  only affects  $Y$  through  $X$ , but not directly affect  $Y$
- ③  $z$  affects  $x$  to some extent, that is,  $cov(Z, x) \neq 0$

- Point 1 is called **exogeneity** requirement: the instrumental variable should be beyond individual's control, such that the instrumental variables are uncorrelated with any individual's unobserved confounding factors.
  - Potential IVs: government policy; natural disasters; randomized experiment; birthdays; etc.
- Point 2 is called **exclusion restriction**: the instrumental variable should only affect  $Y$  through  $X$ , but not directly affect  $Y$ .
- Point 3 is called **relevance requirement**: though beyond an individual's control, the instrumental variable should still affect the individual's  $X$ , causing some exogenous changes in  $X$  that is beyond individual control.
  - If the correlation between  $z$  and  $x$  is too small, we have a **weak IV problem**.

## Graphical Illustration of IV



# A Classic Example of Instrumental Variable

## Return of Military Service to Lifetime Income<sup>1</sup>

$$Income = \beta_0 + \beta_1 MilitaryService + \epsilon$$

- OLS suffers from endogeneity problems. What are the potential endogeneity issues?
- A lottery was used to determine if soldiers with certain birthdays are drafted to the frontline.

---

<sup>1</sup>Angrist, Joshua D., Stacey H. Chen, and Jae Song. "Long-term consequences of Vietnam-era conscription: New estimates using social security data." *American Economic Review* 101, no. 3 (2011): 334-38.

- The date of birth ( $z$ ) or zodiacs can be an **instrumental variable** for military service ( $x$ ) in this case.
  - Relevance requirement: Affects years of military service:  $cov(z, x) \neq 0$
  - Exogeneity requirement: Randomly drawn and thus uncorrelated with any confounders:  $cov(z, \epsilon) = 0$
  - Exclusion restriction:  $z$  only affects  $Y$  through  $X$ , but not directly affect  $Y$ .

[illegible]



## More Examples of IVs

Can you come up with IV candidates for the following causal questions?<sup>2</sup>

- Number of restaurants on UberEat  $\Rightarrow$  Number of orders on UberEat
- Retail price  $\Rightarrow$  Sales

---

<sup>2</sup>See html version for answers.

## Section 2

# Two-Stage Least Squares

# Solving Endogeneity Using IV

- Given an endogenous OLS regression,

$$y_i = X_i\beta + \varepsilon_i, \quad \text{cov}(X_i, \varepsilon_i) \neq 0$$

- Find instrumental variables  $Z_i$  that do not (directly) influence  $y_i$ , but are correlated with  $X_i$

# Two-Stage Least Squares: Stage 1

- ① Run a regression with  $X \sim Z$ . The predicted  $\hat{X}$  is predicted by  $Z$ , which should be uncorrelated with the error term  $\epsilon$ .
- $\hat{X}$  (the part of changes in  $X$  due to  $Z$ ) is exogenous, because  $Z$  is exogenous
  - All endogenous parts are now left over in the error term in the first-stage regression  $\epsilon_i$

$$X_i = Z_i\eta + \epsilon_i$$

## Two-Stage Least Squares: Stage 2

- ② Run a regression with  $Y \sim \hat{X}$ : now  $\hat{X}$  is uncorrelated with the error term and thus we can get causal inference from the second stage regression.

$$y_i = \hat{X}\beta + \varepsilon_i, \quad \text{cov}(\hat{X}_i, \varepsilon_i) = 0$$

## After-Class Readings

- Next week, we are going to discuss a case study using IV and 2SLS. Please read the case study before the next class.
- (optional) [Econometrics with R: Instrumental Variables Regression](#)