Instrumental Variable

Econ 258 Data Analytics with R

Prepared by Ardina Hasanbasri

Outline

1) Introduction to IV: How is cholera disease transmitted?
2) Instrumental Variable
    1) IV and Two stage least square
3) Coding Example: Cigarettes
    - Demonstration of stargazer package for publication ready tables
4) DIY Coding: The Returns to Schooling

Causality Models So Far

- Propensity Score Matching
  - Comparing apples to apples

- Regression Discontinuity Design
  - Looking at similar individuals around a cutoff

- Instrumental Variable
  - Is there another variable that can create randomness?

Questions Preview

How much do prices affect cigarette consumption?

What are the returns to going to collage?

1.1 Instrument Variable Illustration: Cholera Outbreak

- John Snow is a 19th century physicians who wanted to understand how cholera was transmitted. At the time, London experienced around three major cholera outbreaks.
    - Current theory popular at the time was *miasmis,* transmission was through air.
    - John Snow alternatively thought that transmission was through water.
- ***Q**: **Does unhygienic water transmit the cholera disease much faster?***

1.2 Find a Natural Experiment

Are households with clean water versus non-clean water the same?

John Snow looked at water companies.

- **The Lambeth Company** moved its source of water to a purer water source, while two other companies kept the same water source.

- The households these company were serving are similar.

1.3 Cholera death conditional on water company used

Snow's Table IX

Company name	Number of houses	Cholera deaths	Deaths per 10,000 houses
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37

2.1 Instrument Variable

Under homogeneous treatment effect:

An ***instrumental variable*** approach assumes that there is a variable ***z*** such that:

$$P(d|z) \neq P(d)$$

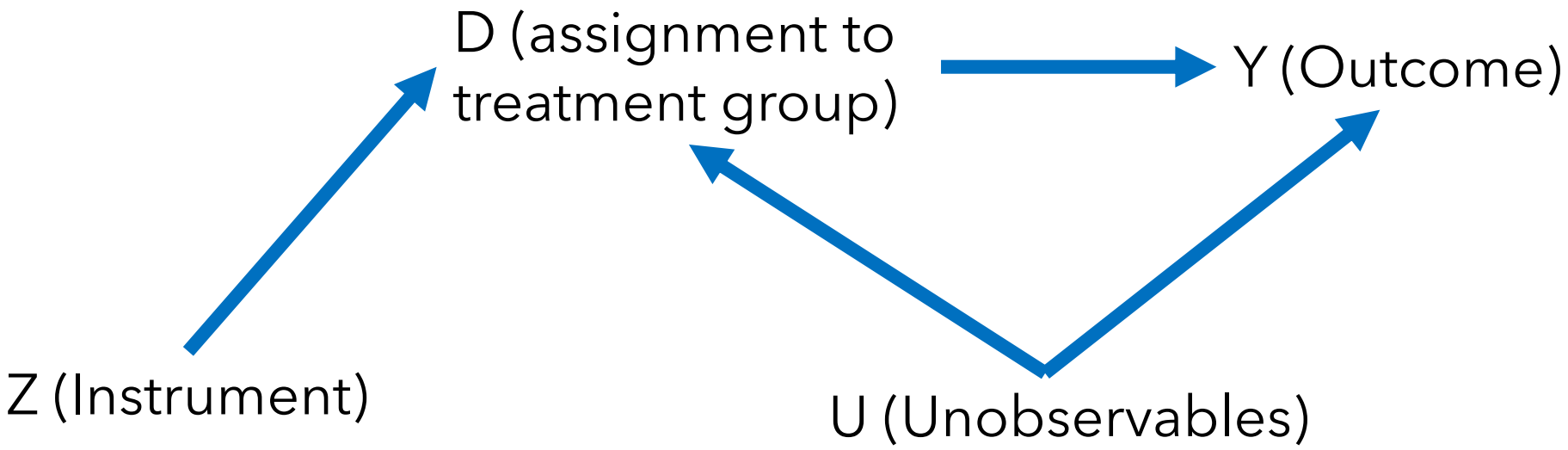
and

$$E[u|z] = E[u]$$

This is actually the conditional independence assumption for IV.

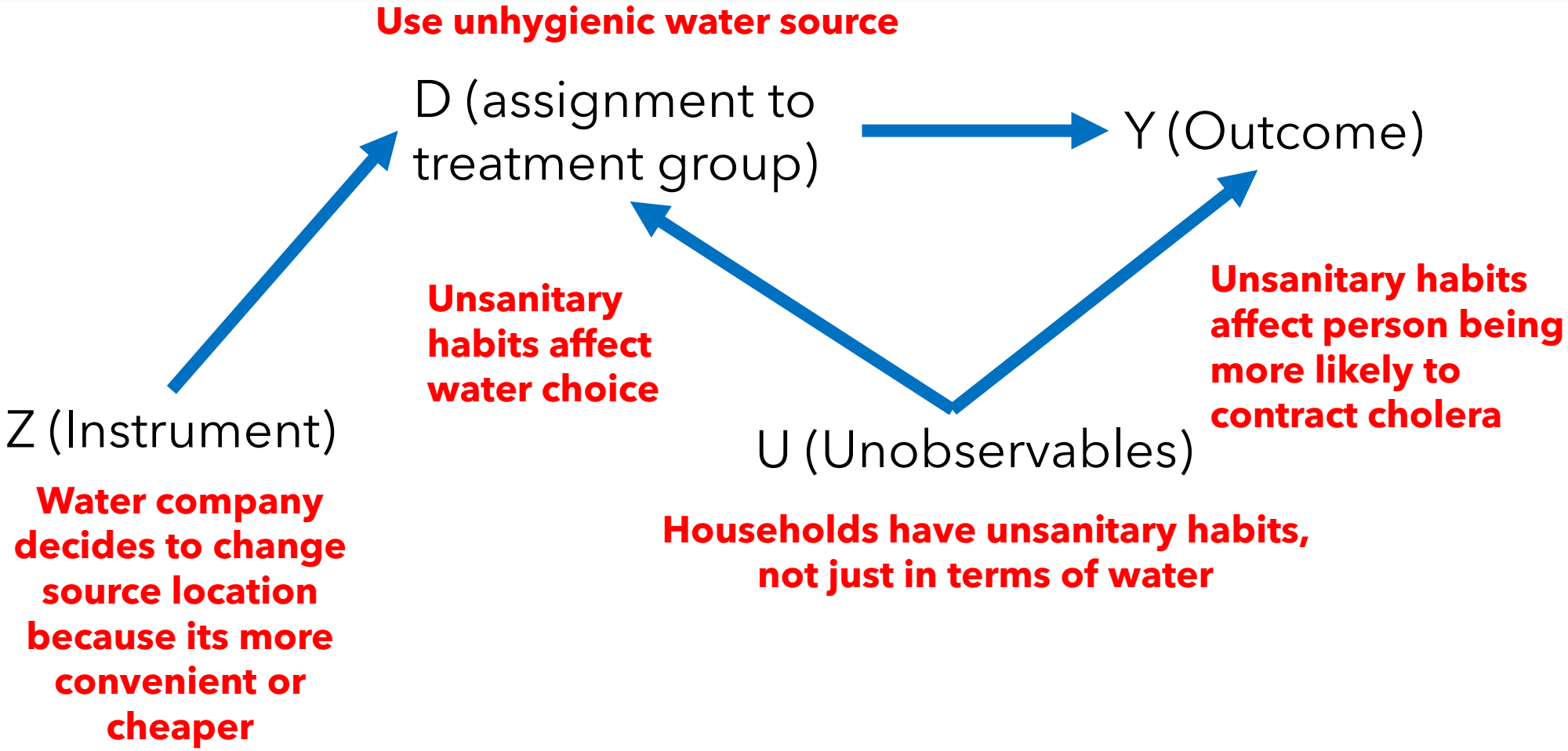
Intuition:
First equation ⇒ the variable z affects the probability of treatment
Second equation ⇒ Ensure randomization. On average, people's "shocks" are not different when you control for z. For example: the choice of water company you choose is quite random.

2.2 Instrumental Variable in Graph Form

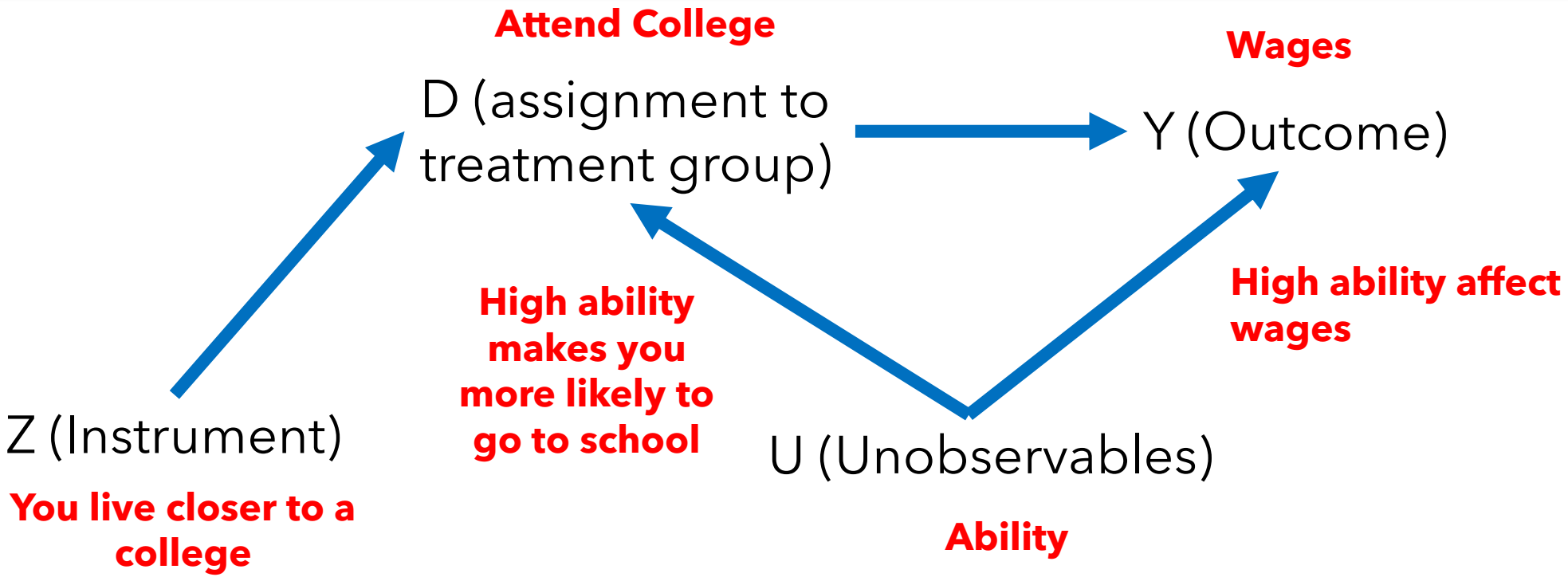


Important assumption: Z affects Y only through D!!!

2.3 Instrumental Variable in Graph Form



2.4 Schooling Example



2.5 Two Stage Least Square

2SLS is one type of IV estimator.

causal when run with $\hat{S}_i$ not S_i

$$Y_i = \alpha + \delta S_i + \varepsilon_i$$ → second stage

run regression not with S_i but $\hat{S}_i$

$$S_i = \gamma + \beta Z_i + \epsilon_i$$ → first stage

characteristics

$$\hat{S}_i = 0.7 + 0.2 Z_i$$

regress school X Z

school: 0 1

Z: being near school

S_i is an endogenous variable. People can select themselves into S_i. For example, schooling or hygienic practices. Z_i is an instrument or a vector of instruments that affects S_i, for example distance to schooling or water companies.

2SLS estimates a first stage (second equation), then uses the fitted values in the first equation (second stage). Intuition: The new fitted value $\widehat{S_i}$ are now exogenous

3.1 How much do prices affect cigarette consumption?

Please see "Lesson 12 Cigarette Example.R" for complete code.

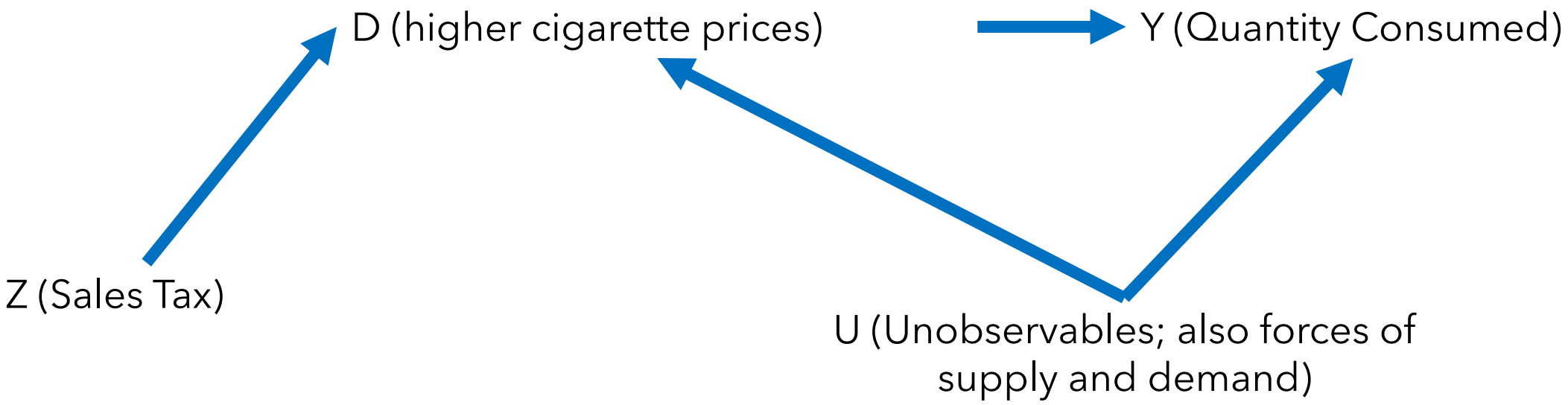
- Prices are endogenous, determined by the supply and demand of cigarettes.
- We cannot regress quantity of cigarettes consumed on cigarettes prices. Why?

$$Q_{cigarettes\_i} = \alpha_0 + \alpha_1 prices_i + \epsilon_i$$

Why does α_i not answer our question?

3.2 Using Sales Tax as an Instrument

- Different states have different sales tax.
- Higher sales tax affects higher prices of cigarettes (states belong to treatment group).
- However, sales tax is not affected by cigarettes consumption.



3.3 What is 2SLS doing?

```
stage1 <- lm(log(rprice) ~ log(rincome) + tdiff + I(tax/cpi), data = CigarettesSW, subset = year == "1995")
CigarettesSW$predict <- predict(stage1)

stage2 <- lm(log(packs) ~ predict + log(rincome), data = CigarettesSW, subset = year == "1995")
```

- **Stage1:** Estimate a model of prices with respect to the instrument.

This model thus can be used to create a new price variable that is exogenous.

- **Stage2:** Instead of regressing quantity of cigarettes on original prices, regress it on the newly created instrument instead.

3.4 Create Publication Ready Tables using Stargazer Package

Table 1: **Do prices affect cigarette consumption?**

	Dependent variable:		
	log(packs)		
	OLS		*instrumental variable*
	(1)	(2)	(3)
log(rprice)	−1.407***		−1.277***
	(0.251)		(0.263)
predict		−1.277***	
		(0.284)	
log(rincome)	0.344	0.280	0.280
	(0.235)	(0.257)	(0.239)
Constant	10.342***	9.895***	9.895***
	(1.023)	(1.141)	(1.059)
Observations	48	48	48
R^2	0.433	0.337	0.429
Adjusted R^2	0.408	0.307	0.404
Residual Std. Error (df = 45)	0.187	0.203	0.188
F Statistic (df = 2; 45)	17.165***	11.426***	

Note: An exercise done in Econ 258. *p<0.1; **p<0.05; ***p<0.01

3.5 Be Careful with Weak Instruments

- Instrumental Variables are clever ways to create randomization.
- Critics of IV argue that it is unclear sometimes what the IV estimate is estimating.
- Another problem with IV is that you have to have a good instrument.

Weak Instrument

- A weak instrument means the correlation between the instrument and our endogenous variable of interest is not significant.

4 Coding Exercise: Return to Schooling

Homework 3 Part A Instructions!

1) What is the coefficient of years of schooling (educ) on log wages in an OLS? Interpret the coefficient.

- Run the regression and control also for variables: exper, black, south, married, smsa

2) What is the correlation between educ and nearc4?

- Control also for variables: exper, black, south, married, smsa
- Is nearc4 a weak instrument for educ? (Run the ols and see if its significant)

3) Do an IV regression (2SLS) using the command "ivreg" to get the effect of schooling.

- Check the help file for ivreg to understand the command inputs.
- After running ivreg, use command "summary(ivreg_name)" to get results.

4) Do 2SLS without the "ivreg" command, just regressions. (Remember the standard errors would not be correct if you do this just with regressions, but you can check the estimates are the same).

5) Create a table with multiple regression, use stargazer to create a nice publication ready table in overleaf.

- Answer: Cunningham Textbook section 7.7.1.

Will need to submit this code as part of **Homework 3!**