Lecture 6. Natural Experiments and Instrumental Variable

Vinish Shrestha

Towson University

Motivation

Motivation

- So far we've looked at the experimental setting
- RCTs
- Assumptions:
  - independence
  - unconfoundedness
- Treatment is not correlated to potential outcomes
- Treatment is not correlated to potential outcomes conditional upon observed covariates

Motivation

- We can't always rely on RCTs

1. Ethical concerns
2. Costly/time consuming
3. Generalizability (external validity)

- There has been an outburst of observational studies that use the conceptual framework of experimental setting

This is a transitional lecture from experimental to observational setting.

Natural Experiments

Natural Experiments

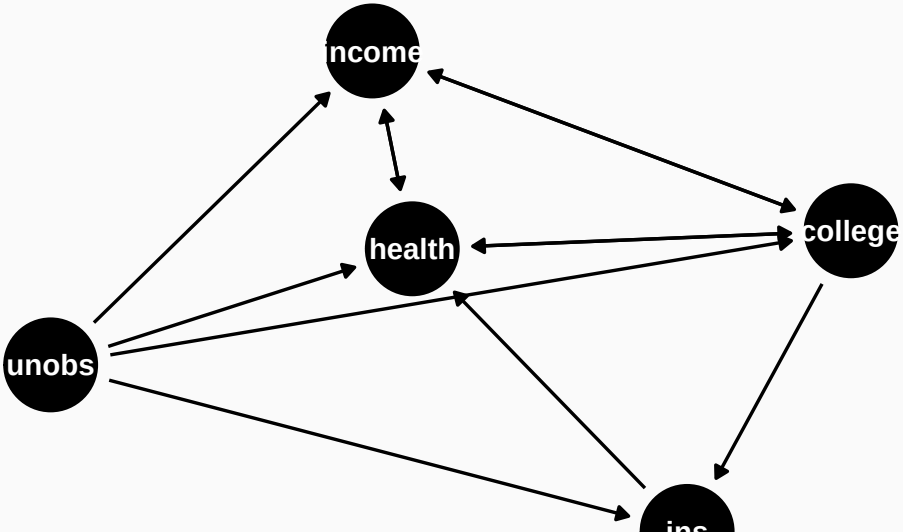
- Analyzes outcomes measures between treatment and control groups in the context of non-randomized setting
- Key: find suitable comparison groups
- Good natural experiment: Transparent source of exogeneous variation in explanatory variables that determine the treatment
- Variation from policy changes, government randomization, nature-related shocks

If one cannot experimentally control the variation one's using, one should understand its source (Meyer (1995)).

Exogeneous variation

Need for natural experiments

- Effects of college education on health



Need for natural experiments

Here, we've got:

1) reverse causality: health causing education
2) omitted variables: unobserved variables affecting both health and college education

- Specification:

$$health_i = \alpha + \beta college_i + \kappa X_i + \epsilon_i \tag{1}$$

- where, X_i is a vector of observed controls (income, race, gender, etc.)
- $\hat{\beta}$ will be biased

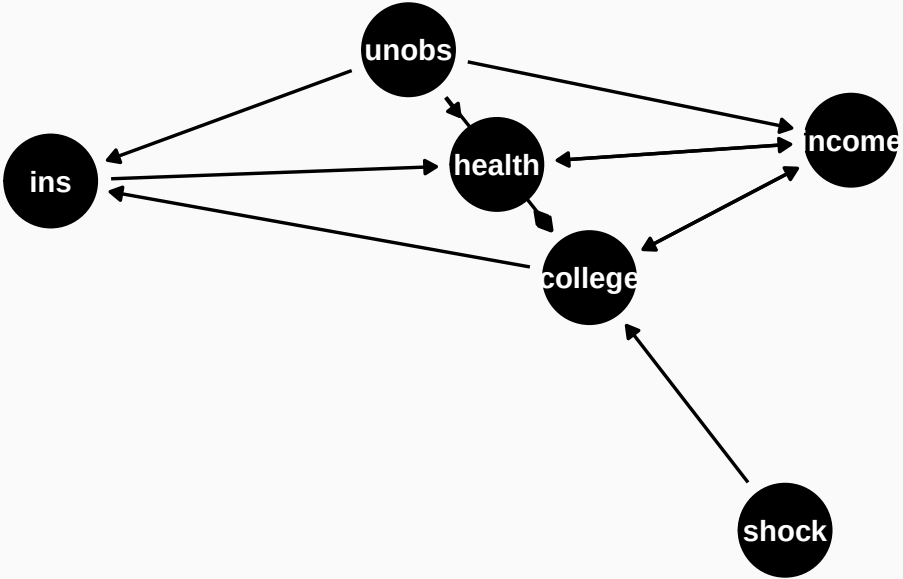
Why?

- Because college education is endogeneous in the model
- Note that OLS relies on exogeneity assumption $E(\epsilon|Xs) = 0$.
- The error term has no systematic relationship with the independent variables.
- However, this is not the case here:
  - something is in the error term that is correlated with college education

Goal

- To find **exogeneous variation** that affects college education but isn't correlated with the error term.
- We can think of exogeneous as somthing that's determined outside of the DGP and comes as a "shock"

DAG with exogeneous variable



Exogeneous variable

- Say, there is a variable *shock*

1. Variation in *shock* creates variation in college education

2a. But variation in *shock* does not affect any other variables

2b. The effect of *shock* on health is driven through and only through college education

- Points 1, 2a and 2b are known as the exclusion restrictrion.
- In this case, *shock* is termed as an exogeneous variable

Examples

Currie and Moretti (2003)

- Currie and Moretti (2003) examine the causal relationship between maternal education and infant health outcomes in their paper "Mother's Education and the Intergenerational Transmission of Human Capital." They argue that higher maternal education improves infant health, as measured by birth weight, gestational age, and infant mortality.

Key Research Question: - Does increasing a mother's education causally improve infant health outcomes?

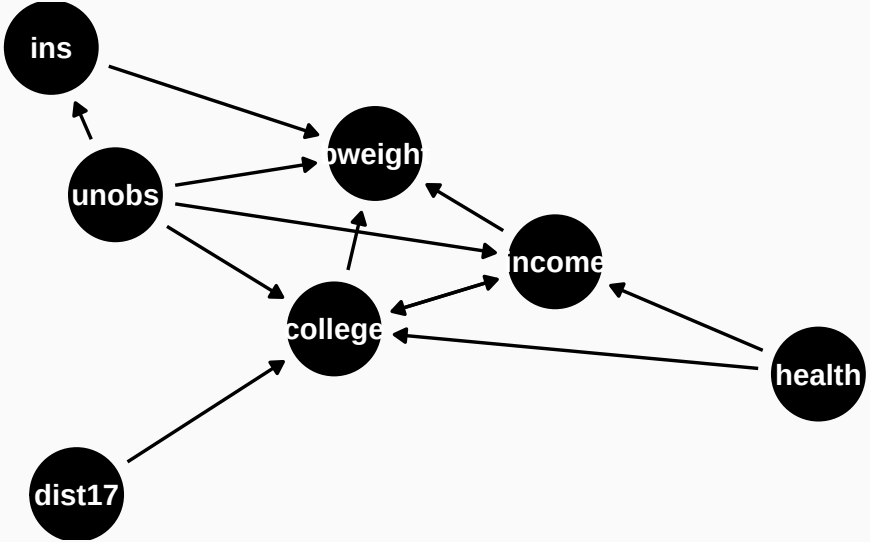
Methodology:

- To address potential endogeneity of education (e.g., reverse causality, omitted variable bias), they use an instrumental variable (IV) approach, leveraging exogenous variation in access to college.

Currie and Moretti (2003): Source of Exogenous Variation

- They use the proximity to a college at age 17 as an instrument for maternal education.
- The idea is that living near a college reduces the cost of obtaining higher education, thereby increasing the likelihood of attending college.
- Findings: They argue that higher maternal education improves infant health, as measured by birth weight, gestational age, and infant mortality.

Currie and Moretti (2003)



Card (1990)

Research Question:

- Does a large influx of low-skilled immigrants negatively affect the wages and employment of native workers, particularly those with similar skill levels?

Card (1990) uses Mariel Boatlift as a natural experiment

- In 1980, the Mariel Boatlift led to the arrival of about 125,000 Cuban immigrants to Miami over a short period, increasing the city's labor force by approximately 7%.
- This event provides a natural experiment to study the effects of immigration on native workers, as the sudden influx of immigrants was largely exogenous.

Card (1990): Methodology

$$native wages_i = \alpha + \beta immigrant\ share_i + \beta X_i + \epsilon_i \tag{2}$$

$$immigrant\ share_i = \alpha + \beta I(Boatlift)_i + \beta X_i + \epsilon_i \tag{3}$$

$$native wages_i = \alpha + \beta \widehat{immigrant}\ share_i + \beta X_i + \epsilon_i \tag{4}$$

Findings:

1. Wages:

- The influx of immigrants had no significant impact on the wages of low-skilled native workers in Miami.
- Even among the most vulnerable group (high school dropouts), wages remained stable.

2. Employment:

- The employment rates of native workers were also not negatively affected by the immigration shock.

3. Labor Market Absorption:

- The labor market adjusted to the sudden increase in the labor force without significant displacement or wage reductions for native workers.

Instrumental Variable

Use moment condition (case of one endogeneous variable and one instrument)

$$E(Z_i(Y_i - X_i\beta)) = 0$$

sample analog:

$$\frac{1}{n}\sum Z_i(Y_i - X_i\beta) = 0$$

solution:

$$\hat{\beta_{IV}} = \frac{\sum(ZY)}{\sum(Z_iX_i)}$$

Instrumental Variable

- Using the examples, we established a ground-work for instrumental variable (IV)
- Instrumental Variable (Z):

1. Correlated with the variable of interest X_1 *relevance*
2. Affects the outcome variable Y only through X_1 (exclusion restriction)

- You can identify the IV effect using the 2 stage least squares (2SLS).

Implementation

2SLS (widely conducted approach)

- Say you are interested in evaluating the causal effect of X_1 on Y in the following specification.

$$Y_i = \alpha + \beta_1 X1_i + \beta_2 X2_i + \epsilon_i$$

- here, X_1 is endogenous and you want to instrument it using Z.
- Assume that relevance and exclusion restriction conditions are satisfied

2SLS implementation

1. First stage: Use Z to isolate variation in $X1$ pertaining to Z

$$X1_i = \alpha + \kappa Z_i + v_i$$

- generate the predicted values of $X1$ using $\hat{\alpha}$ and $\hat{\beta}$.

2. Use $\hat{X1}$ from the FS and estimate:

$$Y_i = \alpha + \beta_1 \hat{X1}_i + \beta_2 X2_i + \epsilon_i$$

Example

- Cai, Janvry, and Sadoulet (2015) : *Social networks and the decision to insure*
- Studies the decision that farmers make whether to buy insurance against weather events.
  - How much does a friend's decision infuences your decision?
- In the first round friends are put in two groups
  - 1. Default "buy" informational session
  - 2. Default "don't buy" informational session
- In the first round, observe people's decision
  - if people abide to default then those who were in *default buy* option should have increased probability of getting insurance.
- After sometime, in the second round, people look at what their friends did in the first-round and make the decision whether to purchase insurance.

Run 2SLS (manually)

```r
# load data
data <-load(file.path(datapath, "social_insure.rda"))
data  <- na.omit(social_insure)

# FIRST STAGE
mod_FS  <- lm(pre_takeup_rate ~ default + male + age + agpop +
              ricearea_2010 +
              literacy + intensive + risk_averse + disaster_prob +
              factor(village), data)

data$FS_predict  <- predict(mod_FS)
```

Run 2SLS (manually)

```
# SECOND STAGE
mod_SS  <- lm(takeup_survey ~ FS_predict + male + age + agpop +
              ricearea_2010 +
              literacy + intensive + risk_averse + disaster_prob +
              factor(village), data)
```

First Stage Results

```
##
## Call:
## lm(formula = pre_takeup_rate ~ default + male + age + agpop +
##     ricearea_2010 + literacy + intensive + risk_averse + disaster_prob +
##     factor(village), data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49136 -0.09949 -0.00761  0.08216  0.60155
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1.248e-01  6.064e-02   2.059 0.039714 *
## default                    1.180e-01  1.094e-02  10.786  < 2e-16 ***
```

Second Stage Results

```
##
## Call:
## lm(formula = takeup_survey ~ FS_predict + male + age + agpop +
##     ricearea_2010 + literacy + intensive + risk_averse + disaster_prob +
##     factor(village), data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0611 -0.4123 -0.1972  0.4725  0.9098
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.0204836  0.1677946   0.122 0.902858
## FS_predict                0.7910970  0.2465491   3.209 0.001365 **
```

References

References i

Cai, Jing, Alain De Janvry, and Elisabeth Sadoulet. 2015. "Social Networks and the Decision to Insure." *American Economic Journal: Applied Economics* 7 (2): 81–108.

Card, David. 1990. "The Impact of the Mariel Boatlift on the Miami Labor Market." *Ilr Review* 43 (2): 245–57.

Currie, Janet, and Enrico Moretti. 2003. "Mother's Education and the Intergenerational Transmission of Human Capital: Evidence from College Openings." *The Quarterly Journal of Economics* 118 (4): 1495–1532.

Meyer, Breed D. 1995. "Natural and Quasi-Experiments in Economics." *Journal of Business & Economic Statistics* 13 (2): 151–61.