Data Processing
– Causal Inference using Instrument Variables

Jung PARK, PhD
Research Fellow in Data Science

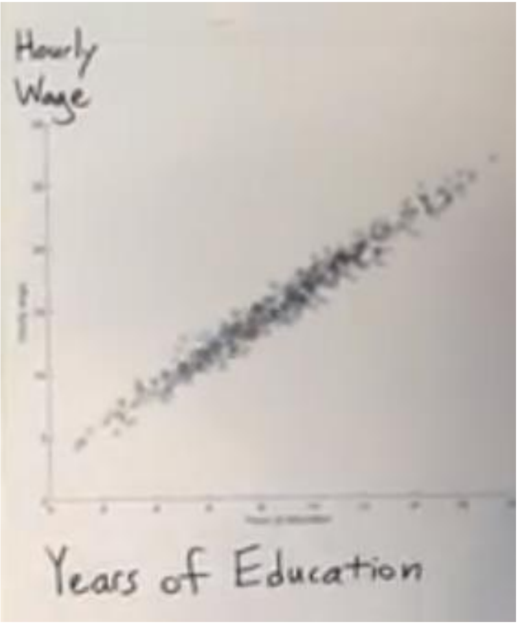
First uploaded 2019 Aug 14
Last modified 2019 Aug 16

Table of contents
: Causal inference using instrument variables

Correlation is not causality

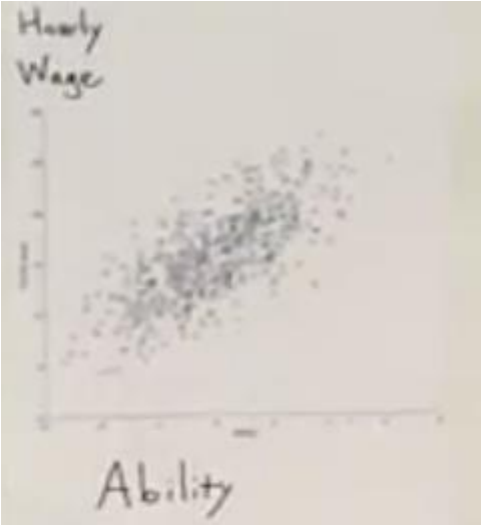
- One of the basics of scientific experiments is to change a parameter of our interest while keeping everything else the same.
- However, in social science, researchers often collect observation data instead of designing an experiment. The approach is equivalently valid only if the samples are randomly collected so that the values of everything else are equivalent.
- Machine learning easily do over-fitting of data so that many variables can show high correlations; we need to be more careful to prove the causality

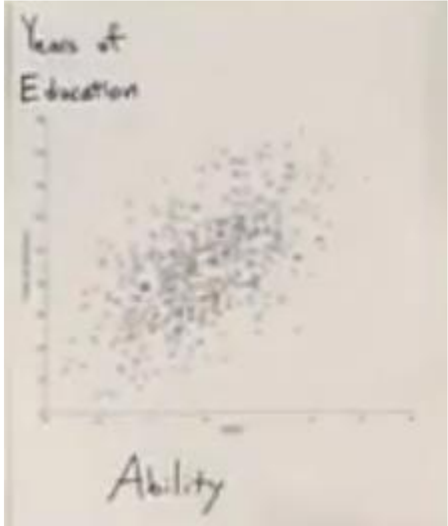
Correlation is not causality - example



- Let’s assume that we found a correlation between years of education and hourly wage in a dataset
- This doesn’t mean that if you have more education, you will get higher hourly wage

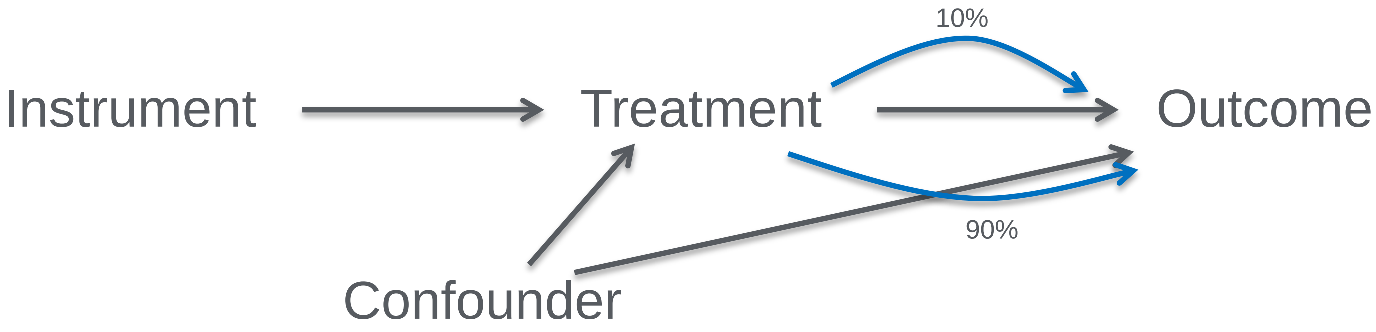
Correlation is not causality - example





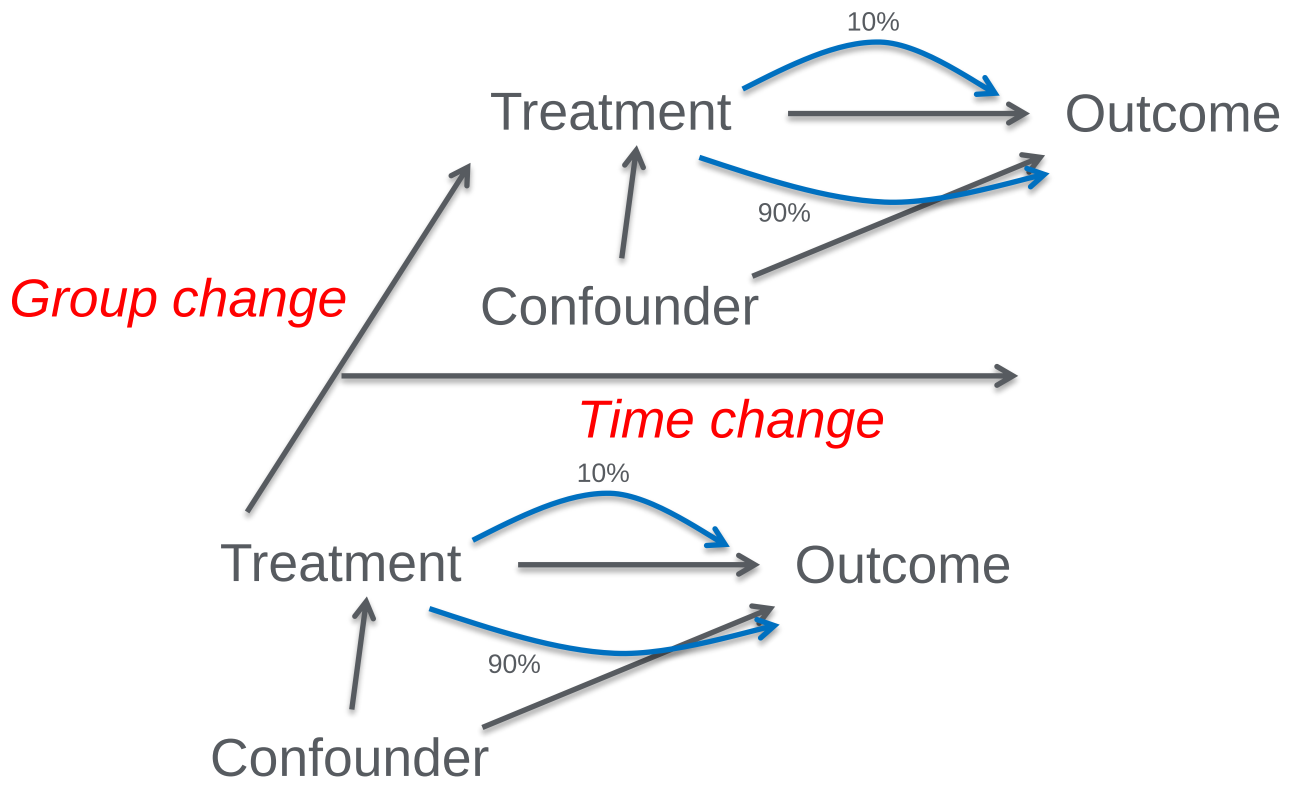
- The correlation may exists simply because the two variables are correlated by another variable, namely a confounder.
- For example, ability (can be instead considered as self-confidence, inherited intelligence, emotional intelligence, etc) can be correlated to both variables. A confounder can be conceptual and unmeasureable.
- We say that the year of education is endogenous as it is related to hourly wage through ability

A method to separate a causal part from confounding – Instrument variable

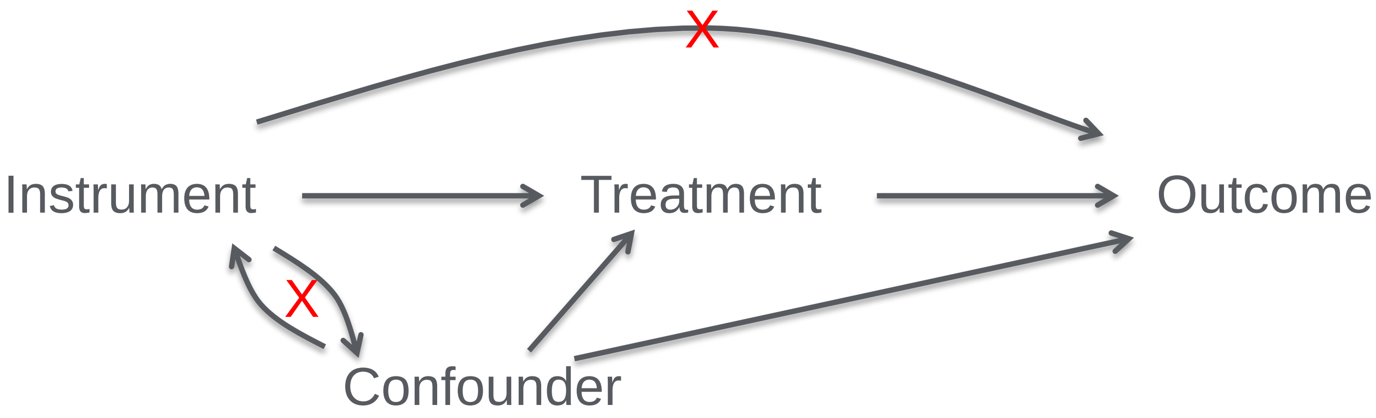


- By using an instrument, we can separate the direct causal effect of treatment to outcome from the indirect correlation caused by confounders

A method to separate a causal part from confounding – Instrument variable



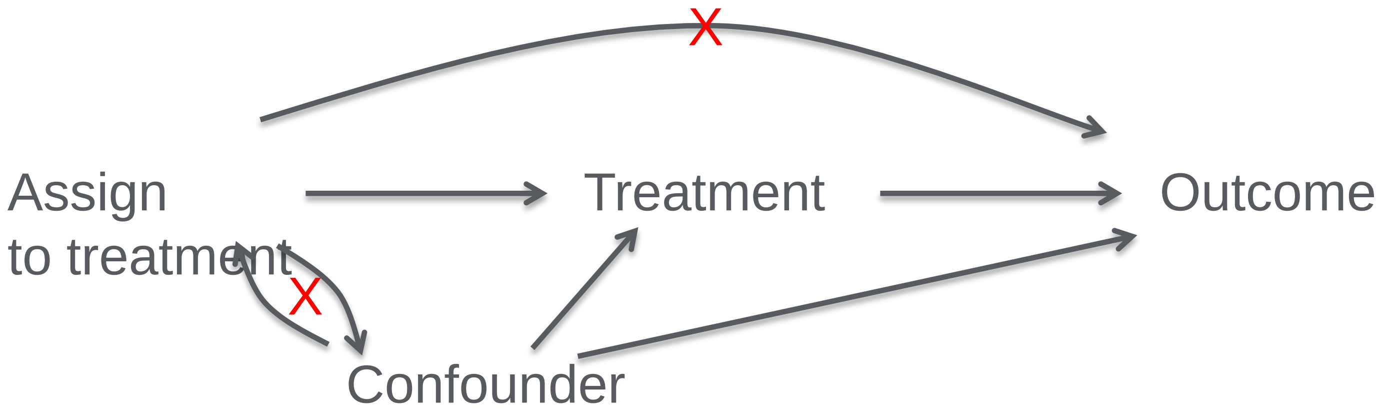
How to find a good instrument variable - three assumptions



1. Relevance: instrument is causing treatment
2. Exclusion restriction: instrument is related to outcome but without cauality
3. Exogenous assumption: instrument is randomly distributed regardless to confounders

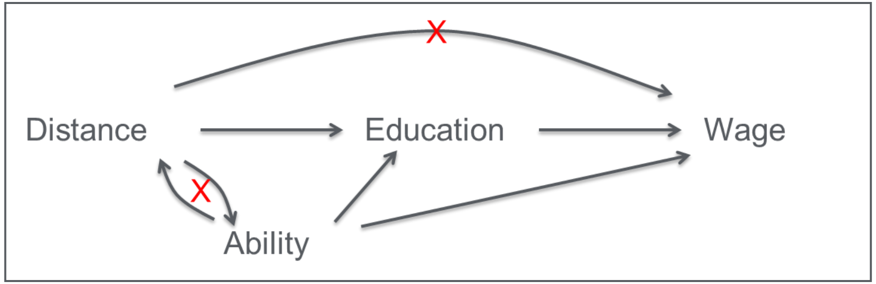
Another usage of IV: handling noncompliers

- A human being doesn’t always follow the intention of experiment
- “Assign to treatment” is an excellent example of instrument variable satisfying the three assumptions of relevance, exclusion and exogeneity
- Because of noncompliers, we need to use 2SLS method to segregate the actual effect of treatment to outcome

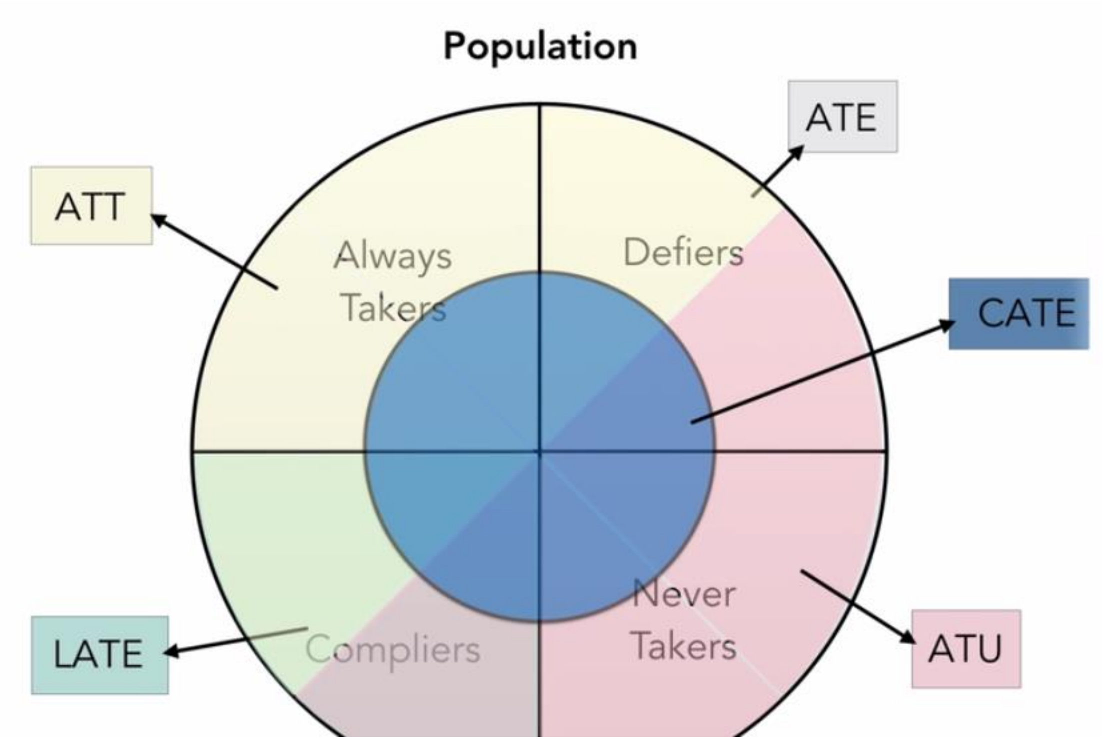


Actual calculation of causality using R
- Two Stage Least Square (2SLS) method

```
# attach the package Applied Econometrics with R (AER)
library(AER)

# load the `CollegeDistance` data set
data(CollegeDistance)

# first stage: regress education on distance
first <- lm(education ~ distance, data = CollegeDistance)

# generate predicted education
CollegeDistance$ed.pred<- predict(first)

# second stage: regress log(wage) on predicted education
second <- lm(log(wage) ~ ed.pred, data = CollegeDistance)

#  the same 2SLS using ivreg
TwoStage <- ivreg(log(wage) ~ education | distance, data = CollegeDistance)

# modified from https://www.econometrics-with-r.org/12-6-exercises-10.html
```



Local Average Treatment Effect (LATE)



https://modu.ssri.duke.edu/module/your-guide-instrumental-variables/ates-cates-and-lates-whats-difference

References

YOUR GUIDE TO INSTRUMENTAL VARIABLES MODULE by Matt Masten
https://modu.ssri.duke.edu/module/your-guide-instrumental-variables

Introduction to Econometrics with R, by Christoph Hanck, Martin Arnold, Alexander Gerber and Martin Schmelzer
https://www.econometrics-with-r.org/12-6-exercises-10.html

Panel Data Analysis

Assumption + Data = Conclusion

Losen assumption + Strong Data = As Good Conclusion

Panel data is stronger than cross-sectional data

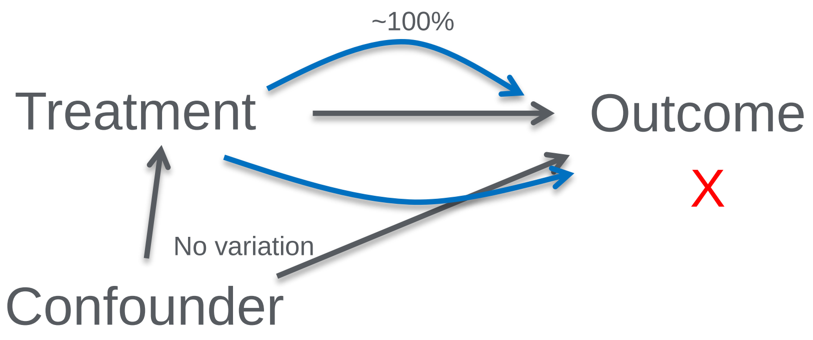
Comparison in Cross-sectional data requires a strong assumption that the samples are randomized; or the variation of the treatment variable is exogenous; or all other variables between the treatment group and control groups are equivalent

With pandel data, the assumption is losen: it allows to have variation between the groups. Except the common trend assumption - the time trends of both groups have to be the same

Then, we can calculate the caulity using Differences-in-Differences method

V_treatment (delta t) – V_control (delta t) : (differences between groups) of (the time differences in each group)

Regression Discontinuity Design



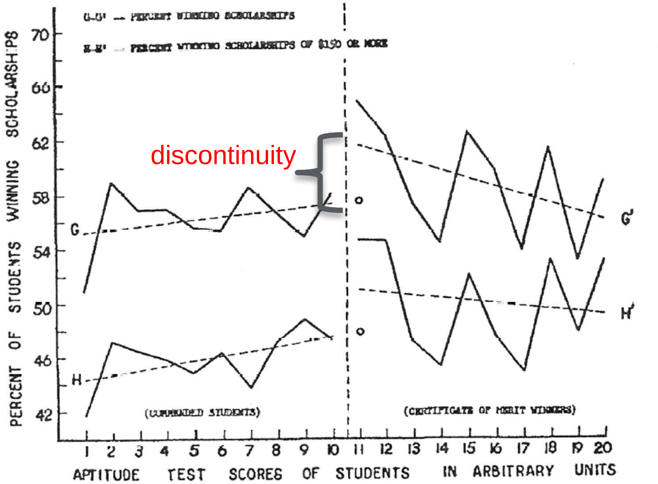


Fig. 2. Regression of success in winning scholarships on exposure determiner.

- By limiting samples with ignorable variations in confounders, we can calculate the causality
- Regression Discontinuity Design selects samples just below and above thresholds; as they have all other characteristics equivalent regardless the treatment

References

Regression Discontinuity by Matt Masten
https://modu.ssri.duke.edu/module/your-guide-regression-discontinuity/regression-discontinuity-more-analysis-thistlethwaite-campbell

Thistlethwaite and Campbell (1960). “Regression-discontinuity analysis: an alternative to the ex post facto experiment.” *Journal of Educational Psychology*, 51(6), 309-317.