Causal Inference Workshop

Week 4 - Instrumental Variables and Regression Discontinuity
Application and Implementation

Causal Inference Workshop

February 9, 2024

Workshop outline

A. Causal inference fundamentals
  - Modeling assumptions matter too
  - Conceptual framework (potential outcomes framework)

B. Design stage: common identification strategies
  - IV + RDD [coding]
  - DiD, DiDiD, Event Studies, New TWFE Lit [coding]
  - Synthetic Control / Synthetic DiD [coding]

C. Analysis stage: strengthening inferences
  - Limitations of identification strategies, pre-estimation steps
  - Estimation [controls] and post-estimation steps [supporting assumptions]

D. Other topics in causal inference and sustainable development
  - Inference (randomization inference, bootstrapping)
  - Weather data regressions, other common/fun SDev topics [coding]
  - Remote sensing data, other common/fun SDev topics

Causal inference roadmap

- *Potential outcomes* [framework]
  - Causal effect is the difference between two potential outcomes
  - We can't observe this difference, but can see differences in average observed outcomes
  - If **(conditional) independence assumption** holds, can estimate unbiased ATT
- *Identification* [application/implementation] [last week, and today, ... and next week!]
  - In most empirical settings, IA and CIA do not hold, which is why we need an **identification strategy**
  - Want to eliminate selection bias (identification problem)
- *Estimation* [application/implementation]
  - (Usually) use linear regression model
  - $\hat{\beta}_{OLS}$ unbiased estimator for ATT if e is uncorrelated with treatment (regression problem)

Outline

Workshop outline

Instrumental variables

Outline

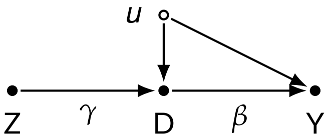
Workshop outline

Instrumental variables

- IV examples
- IV coding
- RDD examples
- RDD coding

IV recap

$$D_i = \delta + \gamma Z_i + v_i$$
$$Y_i = \alpha + \beta D_i + u_i, \quad cov[D_i, u_i] \neq 0$$



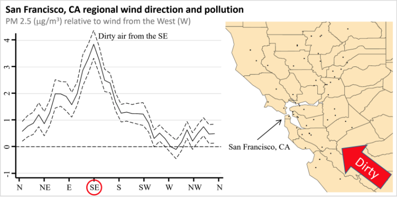
- D_i is endogenous; but there exists a binary instrument Z_i that is a random source of variation in D_i, it "assigns" or changes the probability of treatment
  → We use the instrument to isolate variation in D that is unrelated to u and recover β

- Identifying assumptions:

A1. independence (of Z)	no unmeasured confounder affecting instr. & outcome
A2. exclusion restriction	no direct effect of Z on Y; Z affects Y only through D
A3. relevance (of Z)	Z does affect D
A4. monotonicity (of Z on D)	no defiers, Z is an incentive, doesn't discourage treat.

An SDev-y IV example

- Deryugina et al. (2019), AER
  → instrument for air pollution using changes in local wind direction; estimate the causal effects of acute PM exposure on mortality, health care use, and medical costs among the elderly



San Francisco, CA regional wind direction and pollution

PM 2.5 (μg/m³) relative to wind from the West (W)

Figure 2. Relationship between daily average wind direction and PM 2.5 concentrations for counties in and around the Bay Area, CA. The left panel shows regression estimates of equation (A1) from the Online Appendix, where the dependent variable is the county average daily PM 2.5 concentration and the key independent variables are a set of indicators for the daily wind direction falling into a particular 10-degree angle bin. Controls include county, month-by-year, and state-by-month fixed effects, as well as a flexible function of maximum and minimum temperatures, precipitation, wind speed, and the interactions between them. The dashed lines represent 95 percent confidence intervals based on robust standard errors. The right panel shows the location of the PM 2.5 pollution monitors (black dots) in the Bay Area that provided the pollution measures for this regression.

Outline

Workshop outline

Instrumental variables

- IV examples
- IV coding
- RDD examples
- RDD coding

IV coding, part I

Use: `01a_iv_simulated`

- Simulated data (DGP)
- Run code step-by-step first
  - DGP
  - OLS estimate
  - 2SLS manually (and bootstrapped SEs)
  - 2SLS using package
- To-do:
  - Modify the strength of the instrument - what happens to 2SLS estimates?
  - Modify correlation between D and e - what happens to OLS vs. 2SLS estimates?
  - (Bonus) modify the DGP to include another variable affected by the instrument that then affects the outcome (e.g., rainfall example) - how does this change estimates?

IV coding, part II

Use: `01b_iv_card1995`

- From Card (1993) (link to WP version, published in 1995)
  → use college proximity as an IV for schooling; use NLS Young Men Cohort data; finds returns to schooling higher than OLS estimates
- Run step-by-step first
  - OLS estimate
  - 2SLS manually (and bootstrapped SEs)
  - 2SLS using package
- To-do:
  - Play around with control variables or anything else

Outline

Workshop outline

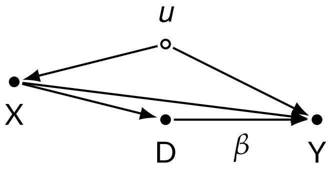
Instrumental variables

- IV examples
- IV coding
- RDD examples
- RDD coding

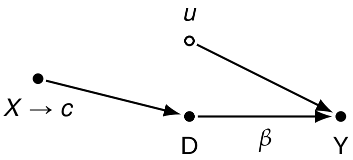
RDD recap

$$Y_i = \alpha + \beta D_i + f(X_i, \phi) + u_i$$

- Treatment D_i is not randomly assigned, it is deterministic, but *discontinuous* along a continuous pretreatment running variable X_i, such that there is “local randomization” in a small neighborhood ((bandwidth) around the cutoff c (e.g., $D_i = \mathbb{1}\{X_i \geqslant c\}$)



As $X \to c$:



- Identifying assumptions

A1. *local* continuity	other determinants of Y don’t jump at c
A2. relevance	discontinuity in the dependence of D_i on X_i

→ We can attribute jump in Y_i at c to D_i’s causal effect

Hannah’s SDev-y RDD example

Figure: Thank you!

Outline

Workshop outline

Instrumental variables

- IV examples
- IV coding
- RDD examples
- RDD coding

RDD coding, part I

Use: `01c_rdd_simulated`

- Simulated data (DGP)
- Run code step-by-step first
  - Part 1:
    - Linear DGP
    - Plot data using standard plotting (e.g., `ggplot2`) and `rdrobust` package (`rdplot`)
    - Same / different slope regressions both using standard regressions and `rddtools` package
  - Part 2:
    - Nonlinear DGP, no discontinuity
    - Same / different slope linear $f()$; quadratic $f()$
- To-do:
  - Modify some of the arguments of `rdplot`
  - Change DGP in an example and see what happens to estimate

RDD coding, part II

Use: `01b_iv_card1995`

- From Carpenter and Dobkin (2009) (link)
  → use minimum drinking age in RDD to estimate the effect of alcohol consumption on mortality; 9% increase in mortality rate at age 21 (motor vehicle accidents, alcohol-related deaths, and suicides)
- Run step-by-step first
  - Load data (save from folder or from here)
  - Same / different slope linear $f()$ regressions
  - Quadratic $f()$ regression
- To-do:
  - Run a couple of sensitivity checks (bandwidth, functional form)

Questions? Comments?

Thank you!

References

Heavily based on Claire Palandri's 2022 version of the Causal Inference Workshop.

Card, David. 1993. *Using Geographic Variation in College Proximity to Estimate the Return to Schooling.* Working Paper, Working Paper Series 4483. National Bureau of Economic Research. https://doi.org/10.3386/w4483. http://www.nber.org/papers/w4483.

Carpenter, Christopher, and Carlos Dobkin. 2009. "The Effect of Alcohol Consumption on Mortality: Regression Discontinuity Evidence from the Minimum Drinking Age." *American Economic Journal: Applied Economics* 1 (1): 164–82. https://doi.org/10.1257/app.1.1.164. https://www.aeaweb.org/articles?id=10.1257/app.1.1.164.

Deryugina, Tatyana, Garth Heutel, Nolan H. Miller, David Molitor, and Julian Reif. 2019. "The Mortality and Medical Costs of Air Pollution: Evidence from Changes in Wind Direction." *American Economic Review* 109 (12): 4178–4219. https://doi.org/10.1257/aer.20180279. https://www.aeaweb.org/articles?id=10.1257/aer.20180279.