# Causal Inference Workshop

## Week 4 - Instrumental Variables and Regression Discontinuity
*Application and Implementation*

Causal Inference Workshop

February 12, 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

Regression discontinuity

Instrumental variables

Regression Discontinuity

# (Sharp) Regression discontinuity, DGP

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

- Treatment $D_i$ is not randomly assigned, it is deterministic, but *discontinuous* along a continuous pretreatment running variable $X_i$ a cutoff $c$ (e.g., $D_i = \mathbb{1}\{X_i \geqslant c\}$)

- $D_i$ deterministic function of $X_i$ (no value of $X_i$ with both treatment and control).

- We only observe the outcome under control, $Y_i(0)$, for those units whose running variable (also called *score*) is below the cutoff, and we only observe the outcome under treatment, $Y_i(1)$, for those units whose score is above the cutoff.

- Look at data only in a small neighborhood around $c$ (cutoff), the bandwidth

# (Sharp) Regression discontinuity, potential outcomes

- The following statement shows up often, but it is actually not part of the identification assumptions for the typical RD with continuous $X$, it is for another identification strategy called Local Randomization, which could also be categorized as RD, but typically with discrete $X$ (see Cattaneo et al. (2024)).
  - Average outcome of those right below the cutoff (who are denied treatment) are compared to those right above the cutoff (who receive the treatment) (i.e., $\mathbb{E}[Y_i(d)|X_i < c] = \mathbb{E}[Y_i(d)|X_i \geqslant c]$ for $d = 0, 1$)

- Real assumption needed: $\mathbb{E}[Y_i(1)|X_i]$ and $\mathbb{E}[Y_i(0)|X_i]$ continuous in $X_i$ at $c$.

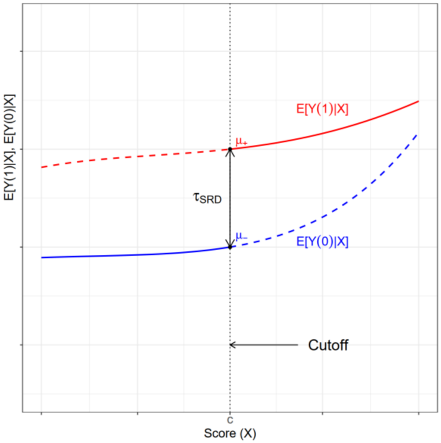



Figure: RD Treatment Effect in Sharp RD Design (Source: Cattaneo et al. (2020))

# (Sharp) Regression discontinuity, identifying assumptions

- Identifying assumptions

| | | |
|---|---|---|
| A1. *local* continuity | $\mathbb{E}[Y_i^1 \vert X_i]$ and $\mathbb{E}[Y_i^0 \vert X_i]$ continuous in $X_i$ at $c$ | other determinants of $Y$ don't jump at $c$ |
| A2. relevance | $D_i = \mathbb{1}[X_i \geqslant c]$ | discontinuity in the dependence of $D_i$ on $X_i$ |

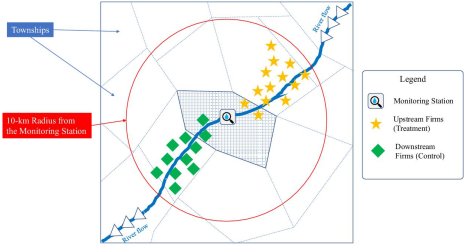
$\rightarrow$ We can attribute a jump in $Y_i$ at $c$ to the causal effect of $D_i$

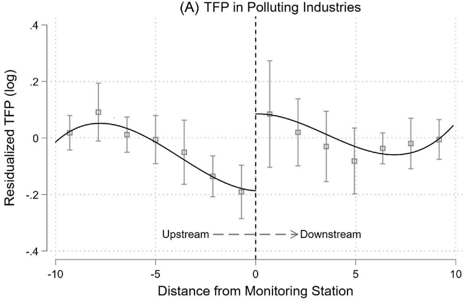
# Regression discontinuity, canonical examples

- Explicit cutoffs in programs (e.g., income in means-tested programs, test scores in gifted-and-talented programs)
- Geographic cutoffs (e.g., school-zone boundaries, such as Black (1999), time zone borders, etc.)
  → e.g., Black (1999) uses house values near elementary school zone boundaries and finds parents are willing to pay 2.5% more for 5% increase in school test scores
- Election cutoffs (e.g., need 50% for win)

# Regression discontinuity, He et al. (2020)

- What is the effect of environmental regulation on firms' productivity?
- A1. The conditional expectation of potential outcomes (productivity) is continuous in $X_i$, the directional distance from the monitoring station
- A2. There is a discontinuity in environmental regulation $D_i$ over the running variable, the directional distance from the monitoring station, $X_i$.



# (Sharp) Regression discontinuity, estimand and estimator

- Estimand

$$\beta_{RD} = \lim_{x \to c^+} \mathbb{E}[Y_i | X_i = x] - \lim_{x \to c^-} \mathbb{E}[Y_i | X_i = x] = \ldots = \mathbb{E}[Y_i^1 - Y_i^0 | X_i = c]$$

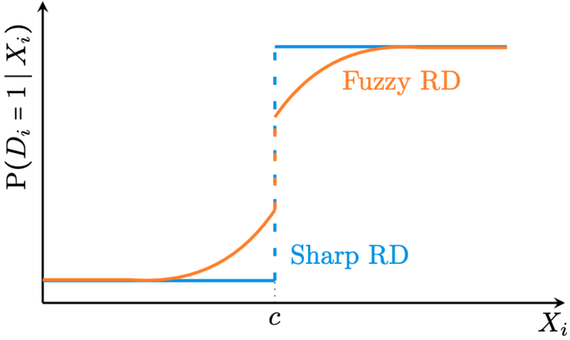
- Estimator

$$Y_i = \alpha + \beta D_i + f(X_i) + e_i$$

  - Use flexible functional forms for $f(X_i)$, such as:
    - local linear regression model: $Y_i = \alpha + \beta D_i + \gamma_1 (X - c) + \gamma_2 (X - c) D + e_i$, with $c - h \leqslant X \leqslant c + h$
    - polynomial regression model with low-degree polynomial (e.g., quadratic, as higher order polynomials can lead to overfitting and introduce bias, see Gelman and Imbens 2019)

# (Fuzzy) Regression discontinuity, estimand and estimator

- In a fuzzy RD, there is imperfect compliance, and at $X_i \geqslant c$, there is a jump but not in treatment assignment but in the *probability* of treatment assignment ($P(D_i = 1|X)$) $\rightarrow$ Discontinuity becomes an instrumental variable for the treatment status $D_i$




(a) RD treatment assignment (sharp & fuzzy)

# (Fuzzy) Regression discontinuity, estimand and estimator

- In a fuzzy RD, there is imperfect compliance, and at $X_i \geqslant c$, there is a jump but not in treatment assignment but in the *probability* of treatment assignment ($P(D_i = 1|X)$)
  $\rightarrow$ Discontinuity becomes an instrumental variable for the treatment status $D_i$

- Estimand

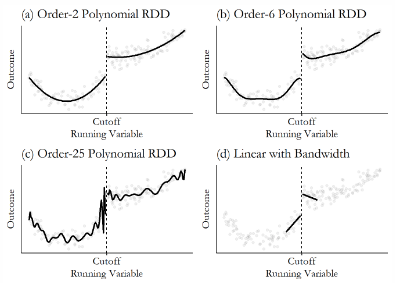
$$\beta_{RD} = \lim_{x \to c^+} \mathbb{E}[Y_i|X_i = x] - \lim_{x \to c^-} \mathbb{E}[Y_i|X_i = x] = \ldots = \mathbb{E}[Y_i^1 - Y_i^0|X_i = c]$$

- Estimator (estimate using 2SLS)

$$\begin{aligned} \text{1st stage: } & D_i = \delta + \gamma Z_i + f(X_i) + u_i \rightarrow \hat{D}_i = \hat{\mathbb{E}}[D_i|X_i] \\ \text{2nd stage: } & Y_i = \tilde{\alpha} + \tilde{\beta}\hat{D}_i + f(X_i) + e_i \end{aligned}$$

# Regression discontinuity, best practices, strengths and weaknesses

- Best practices
  - Choice of $f()$: $f()$ is unknown, so misspecification of the functional form of the DGP may bias the estimator, do robustness checks

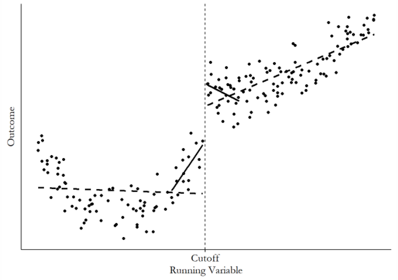



Source: https://theeffectbook.net

  - Bandwith choice can also influence estimate, do robustness checks
  - As in any observational study, adjust for all relevant pre-treatment variables

# Regression discontinuity, best practices, strengths and weaknesses

- Best practices
  - Choice of $f()$: $f()$ is unknown, so misspecification of the functional form of the DGP may bias the estimator, do robustness checks
  - Bandwith choice can also influence estimate, do robustness checks




Source: https://theeffectbook.net

  - As in any observational study, adjust for all relevant pre-treatment variables

# Regression discontinuity, best practices, strengths and weaknesses

- Best practices
  - Choice of $f()$: $f()$ is unknown, so misspecification of the functional form of the DGP may bias the estimator, do robustness checks
  - Bandwith choice can also influence estimate, do robustness checks
  - As in any observational study, adjust for all relevant pre-treatment variables

- Strengths & weaknesses
  + All about finding “jumps” in the probability of treatment as we move along some $X$; much potential in economic applications as geographic boundaries and administrative or organizational rules often create usable discontinuities
  - Risk being underpowered
  - Parameter estimates are very “local”, so their external validity may be low

# Outline

Workshop outline

Regression discontinuity

Instrumental variables

Regression Discontinuity

# Outline

Workshop outline

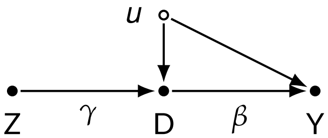
Regression discontinuity

Instrumental variables
- IV examples
- IV coding

Regression Discontinuity
- 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
  $\rightarrow$ We use the instrument to isolate variation in $D$ that is unrelated to $u$ and recover $\beta$

- Identifying assumptions:

| | |
|---|---|
| A1. Exclusion Restriction | $cov[Z_i, u_i] = 0$ |
| A2. Relevance | $cov[Z_i, D_i] \neq 0$ |

# An SDev-y IV example: Deryugina et al. (2019)

- 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

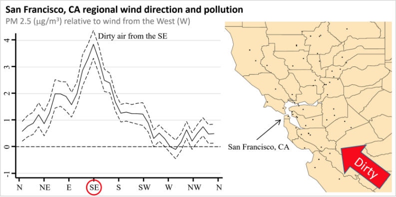



**Figure 2. Relationship between daily average wind direction and PM 2.5 concentrations for counties in and around the Bay Area, CA.** The left panel shows regression estimates of equation (A1) from the Online Appendix, where the dependent variable is the county average daily PM 2.5 concentration and 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.

# An SDev-y IV example: Deryugina et al. (2019)

$$Y_{cdmy} = \beta \text{PM2.5}_{cdmy} + X'_{cdmy}\gamma + \alpha_c + \alpha_{sm} + \alpha_{my} + \epsilon_{cdmy}. \quad (1)$$

Index: county $c$ on day $d$ in month $m$ and year $y$. Wind instrument

$$\text{PM2.5}_{cdmy} = \sum_{g \in \mathcal{G}} \sum_{b=0}^{2} \beta_b^g \mathbf{1}[G_c = g] \times WINDDIR_{cdmy}^{90b} + X'_{cdmy}\sigma + \alpha_c + \alpha_{sm} + \alpha_{my} + \epsilon_{cdmy}. \quad (2)$$

- Each variable in the set $WINDDIR_{cdmy}^{90b}$ is equal to 1 if the daily average wind direction in county $c$ falls in the 90-degree interval $[90b, 90b + 90)$ and 0 otherwise. The omitted category is the interval [270,360).
- They use the k-mean cluster algorithm to classify all the pollution monitors in the United States into 100 spatial groups based on their locations. The variable $\mathbf{1}[G_c = g]$ is an indicator for county $c$ being classified into monitor group $g$.

# Outline

Workshop outline

Regression discontinuity

Instrumental variables
- IV examples
- IV coding

Regression Discontinuity
- 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

Regression discontinuity

Instrumental variables

Regression Discontinuity

# Outline

Workshop outline

Regression discontinuity

Instrumental variables
- IV examples
- IV coding

Regression Discontinuity
- RDD examples
- RDD coding

# RDD recap

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

- Treatment $D_i$ is not randomly assigned, it is deterministic, but *discontinuous* along a continuous pretreatment running variable $X_i$ around the cutoff $c$ (e.g., $D_i = \mathbb{1}\{X_i \geqslant c\}$)

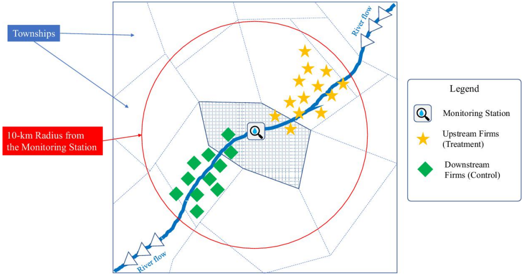
- Identifying assumptions

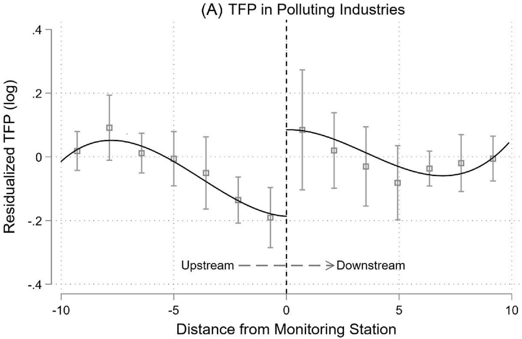
| | |
|---|---|
| A1. *local* continuity | other determinants of $Y$ don't jump at $c$ |
| A2. relevance | discontinuity in the dependence of $D_i$ on $X_i$ |

$\rightarrow$ We can attribute jump in $Y_i$ at $c$ to $D_i$'s causal effect

# SDev-y RDD example: He et al. (2020)

- What is the effect of environmental regulation on firms' productivity?

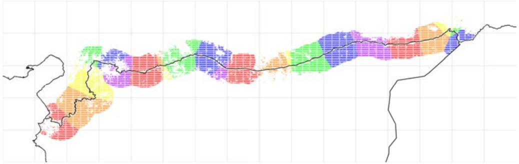


# SDev-y RDD example: Wuepper et al. (2023)

Figure 3: Constructing Border Segments

**Panel A: Turkey - Syria border**



Question: The effect of being on a certain side of the border on crop yield. Estimate the following equation for each border and year separately for all borders

$$y_{ibt} = \alpha_{s[i,b]t} + \beta_{bt} I_{[i \in H]} + \gamma_{s[i,b]t} \mathbf{X}_i + \delta_{bt} \mathbf{Z}_{it} + \epsilon_{ibt} \quad (3)$$

- $y_{ibt}$ are log of the annual maximum EVI) of pixel $i$ in year $t$ along border $b$ of a country-pair. $I_{[i \in H]}$ indicates for the country with the higher country code.
- $\alpha_{s[i,b]t}$ is border-segment by year fixed effects.
- $\gamma_{s[i,b]t}$ are border-segment specific coefficients that are allowed to vary by year and include three time-invariant variables. $\mathbf{X}_i$: a smooth function in longitude, latitude as well as the cross-term.

# Outline


Workshop outline

Regression discontinuity

Instrumental variables
- IV examples
- IV coding

Regression Discontinuity
- 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 I

Heavily based on Claire Palandri's 2022 version and Anna Papp's 2024 version of the Causal Inference Workshop.


Black, Sandra E. 1999. "Do Better Schools Matter? Parental Valuation of Elementary Education." *The Quarterly Journal of Economics* 114 (2): 577–599. ISSN: 00335533, 15314650, accessed January 26, 2024. http://www.jstor.org/stable/2587017.

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.

Cattaneo, Matias D., Nicolas Idrobo, and Rocío Titiunik. 2024. *A Practical Introduction to Regression Discontinuity Designs: Extensions.* Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. 2020. *A Practical Introduction to Regression Discontinuity Designs: Foundations.* Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.

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.

# References II


Gelman, Andrew, and Guido Imbens. 2019. "Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs." *Journal of Business & Economic Statistics* 37 (3): 447–456. https://doi.org/10.1080/07350015.2017.1366909. eprint: https://doi.org/10.1080/07350015.2017.1366909. https://doi.org/10.1080/07350015.2017.1366909.

He, Guojun, Shaoda Wang, and Bing Zhang. 2020. "Watering down environmental regulation in China." *The quarterly journal of economics* 135 (4): 2135–2185.

Wuepper, David, Haoyu Wang, Wolfram Schlenker, Meha Jain, and Robert Finger. 2023. *Institutions and Global Crop Yields.* Working Paper, Working Paper Series 31426. National Bureau of Economic Research. https://doi.org/10.3386/w31426. http://www.nber.org/papers/w31426.