Econometrics II

Lecture 12: Regression Discontinuity Designs

David Schönholzer

Stockholm University

May 10, 2024

Table of Contents

Basic Setup

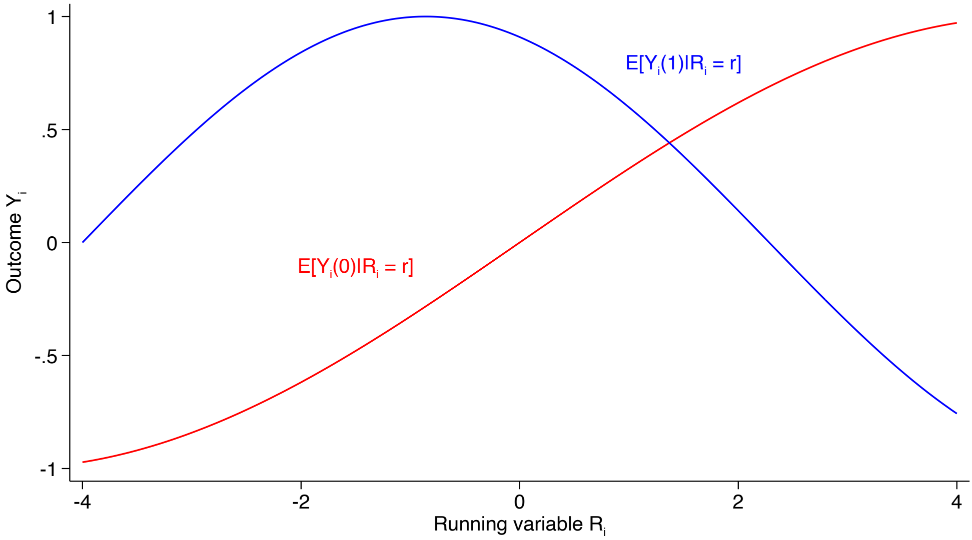
- Consider setting with treatment $D_i \in \{0, 1\}$ and POs $Y_i(1)$ and $Y_i(0)$
- Suppose that D_i is *deterministic* function of observed covariate R_i:

$$D_i = 1[R_i > c]$$

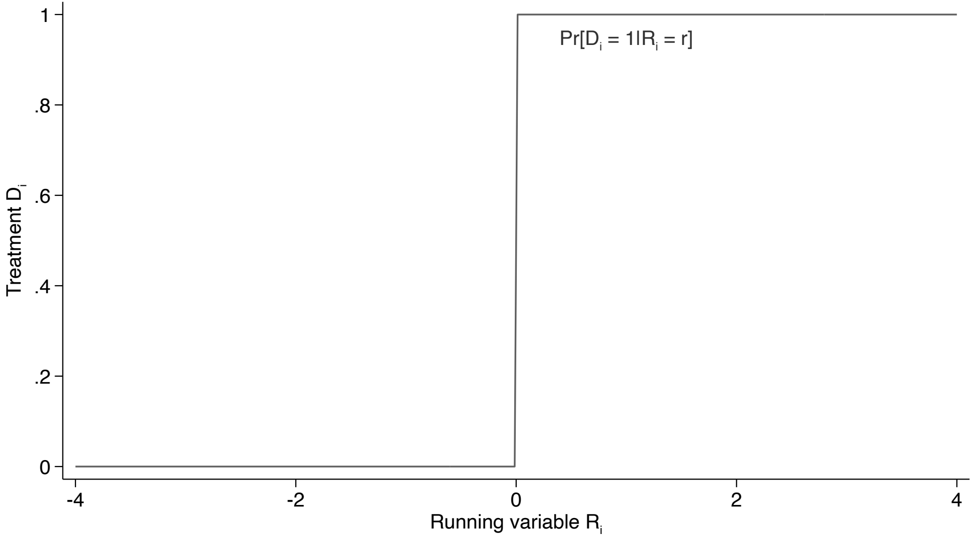
- R_i is called the running variable
- This is a sharp RD because treatment switches from 0 to 1 at threshold
- We observe $Y_i(1)$ when $R_i > c$ and $Y_i(0)$ when $R_i \leq c$
- Example: Scholarship awarded to students above test score threshold

- Basic idea:
  - Compare observations just above and below c
  - Intuitively, treatment may be as good as randomly assigned near $R_i = c$

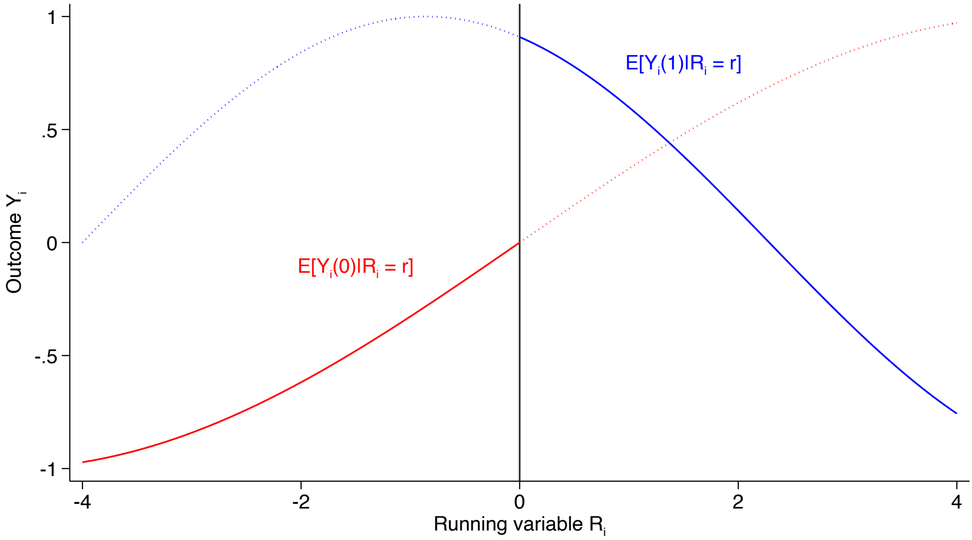
Potential Outcomes



Assignment Probabilities



Observed Outcomes



Sharp RD Identification

- Key identifying assumption: POs are smooth at the threshold c:

$$\lim_{r \uparrow c} \mathbb{E}[Y_i(d) | R_i = r] = \lim_{r \downarrow c} \mathbb{E}[Y_i(d) | R_i = r], \quad d \in \{0, 1\}$$

  - Potential outcome CEFs need to be **continuous at the threshold**
  - The population just below must not be discretely different than above

- If the assumption holds, then

$$\begin{aligned}
\lim_{r \downarrow c} \mathbb{E}\left[Y_i | R_i = r\right] - \lim_{r \uparrow c} \mathbb{E}\left[Y_i | R_i = r\right] &= \lim_{r \downarrow c} \mathbb{E}\left[Y_i(1) | R_i = r\right] - \lim_{r \uparrow c} \mathbb{E}\left[Y_i(0) | R_i = r\right] \\
&= \mathbb{E}\left[Y_i(1) | R_i = c\right] - \mathbb{E}\left[Y_i(0) | R_i = c\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0) | R_i = c\right] \equiv \tau(c)
\end{aligned}$$

Interpretation

- Note that $\tau(c)$ is the treatment effect only for individuals with $R_i = c \rightarrow$ LATE
- But identification is nonparametric:
  - No assumptions about distribution of $Y_i(d)$
  - Other than continuity of CEFs
- We have a CIA condition:

$$(Y_i(1), Y_i(0)) \perp D_i | R_i$$

  - But no common support: conditional on R_i, always $D_i = 0$ or $D_i = 1$
  - RD is local extrapolation outside support to predict $\mathbb{E}[Y_i(d)|R_i = c]$

Table of Contents

Basic Estimation Problem

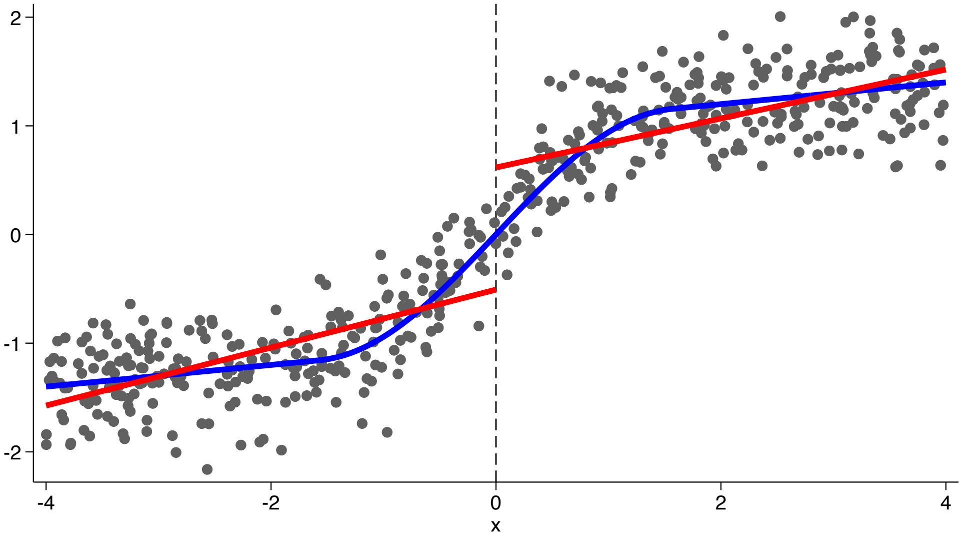
- Implementing RD requires estimating right- and left-side limits:

$$\lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r] \qquad\qquad \lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r]$$

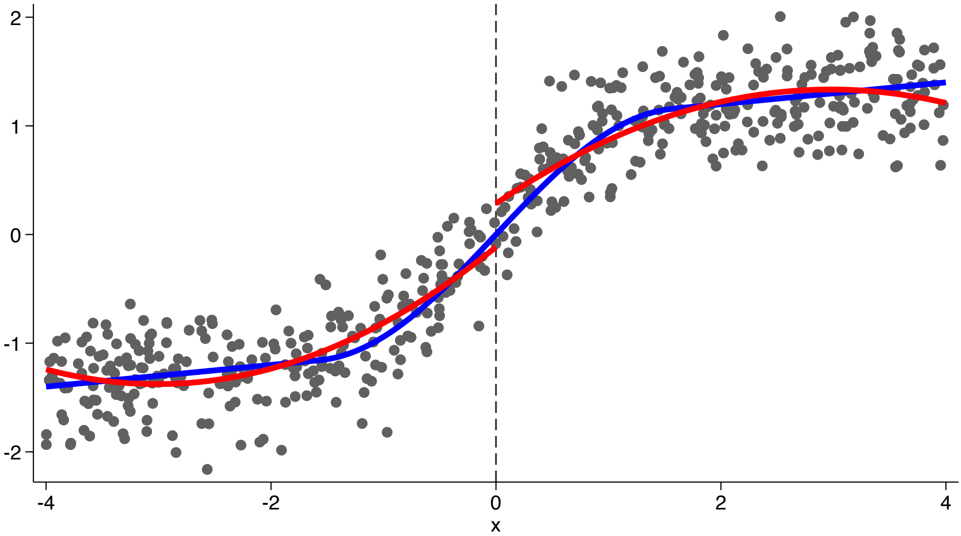
- Since we extrapolate, functional form used to approximate CEF important
  - Not enough to rely on OLS approximation theorems
  - With insufficiently flexible specification, might mistake nonlinearity for effect
  - But too flexible specification reduces precision and may overfit

- How do we balance this tradeoff?

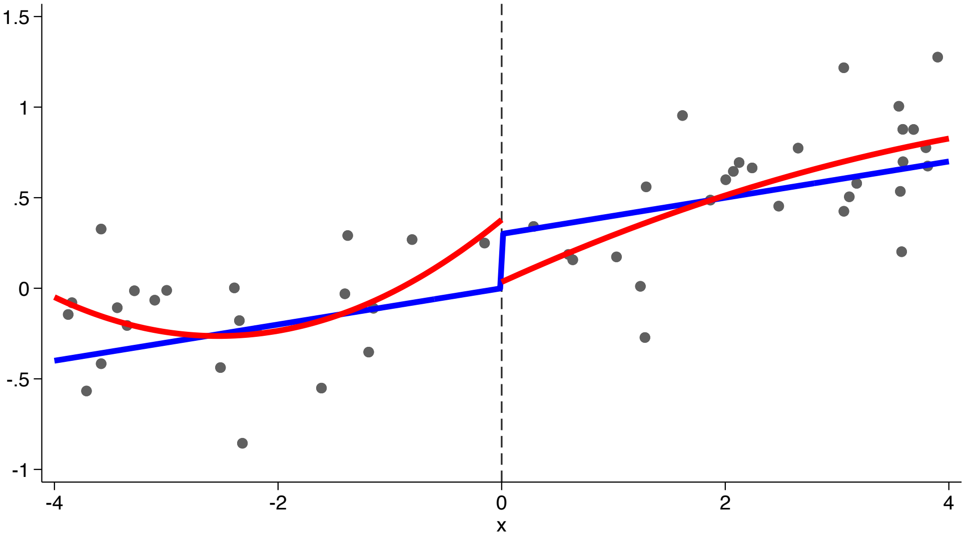
False Positive: Mistaking Nonlinearity as Discontinuity



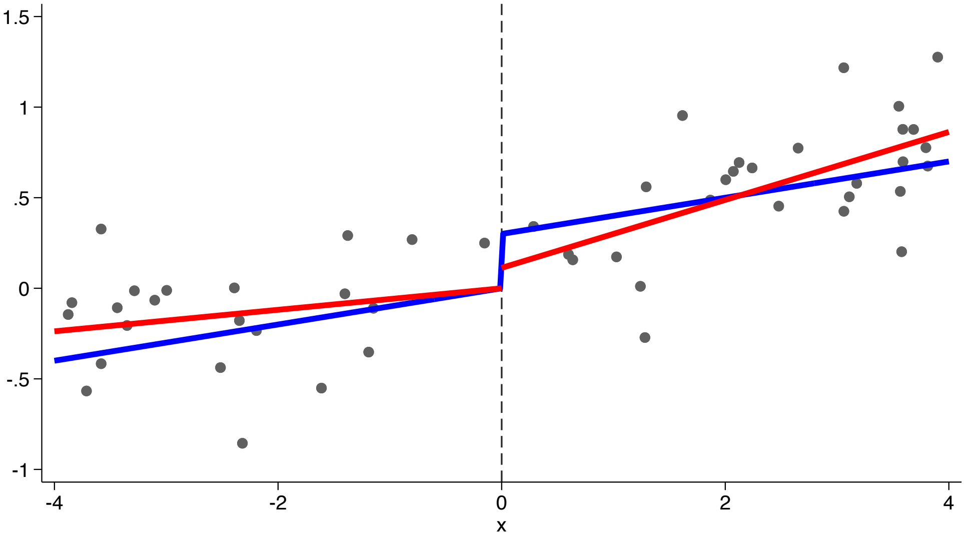
Increased Flexibility May Avoid Detecting False Positive



False Negative: Overfitting



Less Flexibility May Avoid Overfitting



Estimation: Global Parametric versus Local Nonparametric

1. Global parametric: approximate $\mathbb{E}[Y_i|R_i]$ using polynomials
   - Traditional approach
   - Can work with moderately large datasets ($N \approx 1000$)
   - Usually uses all data, even when far from cutoff
2. Local nonparametric: approximate $\mathbb{E}[Y_i|R_i]$ using nonparametric techniques
   - More modern approach
   - Typically requires large datasets ($N > 1000$)
   - Uses only small subset of data close to the cutoff

- Often useful to compare across methods – stability suggests robust finding

Global Parametric Specification

- Estimating equation:

$$Y_i = \alpha + \underbrace{\beta 1\left[R_i > c\right]}_{\text{Discontinuity}} + \underbrace{\sum_{k=1}^{K} \gamma_{0k}\left(R_i - c\right)^k}_{\text{Polynomial for all data}} + \underbrace{\sum_{k=1}^{K} \gamma_{1k} 1\left[R_i > c\right]\left(R_i - c\right)^k}_{\text{Polynomial only on right side}} + \varepsilon_i$$

- K-th order polynomials to capture CEFs on either side
- Using separate coefficients $\gamma_{0k} + \gamma_{1k}$ and γ_{1k} on either side

- Key parameter: β, the size of the jump at the threshold
- Treatment effect estimate: $\widehat{\beta}$

Local (i.e. Nonparametric) Linear Specification

- Estimating equation for **left-hand side CEF** using **local linear**:

$$\left(\widehat{\alpha}_0, \widehat{\delta}_0\right) = \arg\min_{\alpha_0, \delta_0} \sum_{i=1}^{N} 1\left[R_i \leq c\right] K\left(\frac{R_i - c}{h}\right) \left(Y_i - \alpha_0 - \delta_0 \left(R_i - c\right)\right)^2$$

- Similarly, for the **right-hand side CEF** and **local linear**, we run:

$$\left(\widehat{\alpha}_1, \widehat{\delta}_1\right) = \arg\min_{\alpha_1, \delta_1} \sum_{i=1}^{N} 1\left[R_i > c\right] K\left(\frac{R_i - c}{h}\right) \left(Y_i - \alpha_1 - \delta_1 \left(R_i - c\right)\right)^2$$

- Here, $\widehat{\alpha}_0$ is an estimate of $\lim_{r \uparrow c} \mathbb{E}[Y_i | R_i = r]$ and $\widehat{\alpha}_1$ of $\lim_{r \downarrow c} \mathbb{E}[Y_i | R_i = r]$
- Hence the treatment effect estimate is $\widehat{\alpha}_1 - \widehat{\alpha}_0$
- Note that $K(\cdot)$ is a Kernel with bandwidth h
- Linear term eliminates "boundary bias" (Fan and Gijbels 1992)

Visualization

- RD is appealing because the design is relatively simple and transparent
  - Key relationships: outcome & running variable; treatment & running variable
  - These relationships can be plotted $\rightarrow$ visual evidence of discontinuity
  - In practice, people won't believe an RD if the result isn't visible in a plot
- Plotting tips:
  - Automated plotting procedure `rdplot` (Calonico et al, 2014b)
  - May use own plotter using e.g. `scatter` of bin means and `lpoly`
  - `binscatter` also has an `, rd` option

Table of Contents

Choosing Global Versus Local

- Remaining issues:
  1. Global versus local?
  2. How to choose order of polynomial?
  3. How to choose bandwidth?

- No real conceptual distinction: need to choose bandwidth & polynomial order
  1. Bandwidth: how much data to include on either side of cutoff
  2. Polynomial order: how much flexibility to allow
- Typical choices:
  1. Global parametric: infinite or large bandwidth, higher-order polynomial
  2. Local nonparametric: small bandwidth, first-order polynomial (linear)
- But under uniform kernel, given bandwidth, and poly order they are the same!

Global versus Local: Polynomial Order

- Global estimators can rely heavily on data far from cutoff
    - This might lead to very sensitive estimates (Gelman and Imbens 2014)
    - But if CEF is well-approximated by polynomial, legitimate and efficient
- Choosing optimal polynomial order:
    - In contrast to optimal bandwidth, not much known about optimal poly order
    - Generally, higher orders may be suboptimal (Pei et al. 2021)
- In practice: typically want to show robustness to tuning choices
- If you have the power, local linear generally preferred

Table of Contents

Diagnostics Issues

- RD identifying assumption: PO distributions smooth around threshold
  - $1[R_i > c]$ must be as good as randomly assigned near $R_i = c$
  - May be violated if individuals can exactly control R_i
  - Example: self-reported income, people misreport for program eligibility
  - Just like in DID, this is **untestable**, but diagnostics can assess plausibility

- Two diagnostics:
  1. Covariate balance: check that no sorting on **characteristics**
  2. Bunching: check that no sorting by **frequency**

RD Diagnostic I: Covariate Balance

- If people sort in the neighborhood of $R_i = c$, expect imbalance in covariates
- Motivates a check for whether there is a discontinuity in $\mathbb{E}[X_i | R_i = c]$ at $R_i = c$
- Analogous to balance check in baseline variables in an RCT
- Implement by using X_i as the RD dependent variable
    - Often done visually for primary variables of concern
    - Balance table for comparison of larger number of characteristics

RD Diagnostic II: Bunching

- Instead of the mean **value** of X_i, the **frequency** could exhibit a jump
- McCrary (2008) proposes a test for whether the RD is compromised:
  - Logic: Imagine individuals strategically locate above/below threshold
  - Would then expect "bunching" on the side of threshold that is more preferable

- More generally, sorting may generate anomalies in the distribution of R_i
- McCrary suggests looking for a discontinuity in the density of R_i near c
- Typically implemented visually

- Sometimes, a "donut" strategy can ameliorate sorting concerns
  - Basic idea: leave out observations very close to threshold (Barreca et al 2011)
- Bunching is bad for credibility of RD, but may in itself be interesting:

  $\rightarrow$ Bunching estimators in public economics

Table of Contents

Kernels and Bandwidths

- What kernel to use? Many plausible choices
- Popular choice: “edge” (or triangle) kernel:

$$K(u) = 1[|u| \leq 1] \times (1 - |u|)$$

  - Has optimality properties in boundary estimation problems (Cheng et al 1997)
  - It is also intuitively appealing: generates weighting function:

$$K\left(\frac{R_i - c}{h}\right) = 1\left[|R_i - c| \leq h\right] \times \left(1 - \frac{|R_i - c|}{h}\right)$$

  - h can be interpreted as cutoff distance for data inclusion
  - Weights fall linearly from 0 to 1 in included sample

- Should not make big difference, but variation across kernels rarely shown

Optimal Bandwidth

- Recent literature on optimal choice of bandwidth h
- Bias/variance tradeoff: smaller bandwidth reduces bias but reduces precision
- Intuitively, if little curvature in CEF, bias from large h is small

- Imbens and Kalyanaraman (IK, 2012):
  - Asymptotic approximation of MSE of RD estimator
  - Derive MSE-minimizing bandwidth
  - Optimal bandwidth depends on curvature of CEF near discontinuity
  - Use plug-in estimators of parameters governing curvature

Robust Confidence Intervals

- The IK bandwidth is well-suited for estimation
- But what about inference?

- Calonico, Cattaneo and Titiunik (CCT, 2014) show it works poorly for inference
  - Non-negligible bias term in the estimate
  - Naive inference can lead to misleading confidence intervals

- Alternative by CCT:
  - CCT advocate using a second, smaller bandwidth to eliminate this bias
  - "Undersmooth", i.e. prioritize small bias
  - Report robust standard error using this smaller bandwidth

Optimal Bandwidth in Practice

- Unfortunately, CCT bandwidth often much smaller than IK bandwidth
- CCT generates large confidence intervals
- May limit feasibility of many potential RD projects:
  - Lots of data required to detect an effect with CCT robust CIs
  - Noisy effects are very hard to write up and sell

- In practice, there is some slack with using alternatives to CCT
  - Primacy of visualization: compelling figures guide robustness requirements
  - Plotting effect and CIs for different bandwidths can be useful

- Implementation of IK and CCT: see `rdrobust`

Table of Contents

Fuzzy RD Setup

- In many cases, assignment probability does not jump from zero to one at c
  - Instead of treatment being deterministic with R_i, its probability jumps
  - Other factors besides running variable may affect treatment assignment

- Suppose that

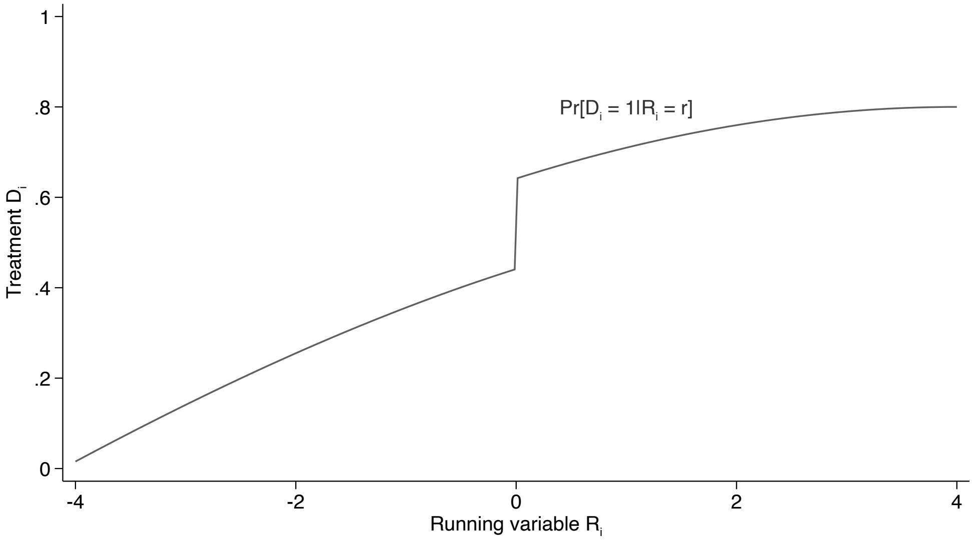
$$\lim_{r \uparrow c} \Pr\left(D_i = 1 | R_i = r\right) \neq \lim_{r \downarrow c} \Pr\left(D_i = 1 | R_i = r\right)$$

  i.e. treatment probability jumps at $R_i = c$, but not necessarily from zero to one

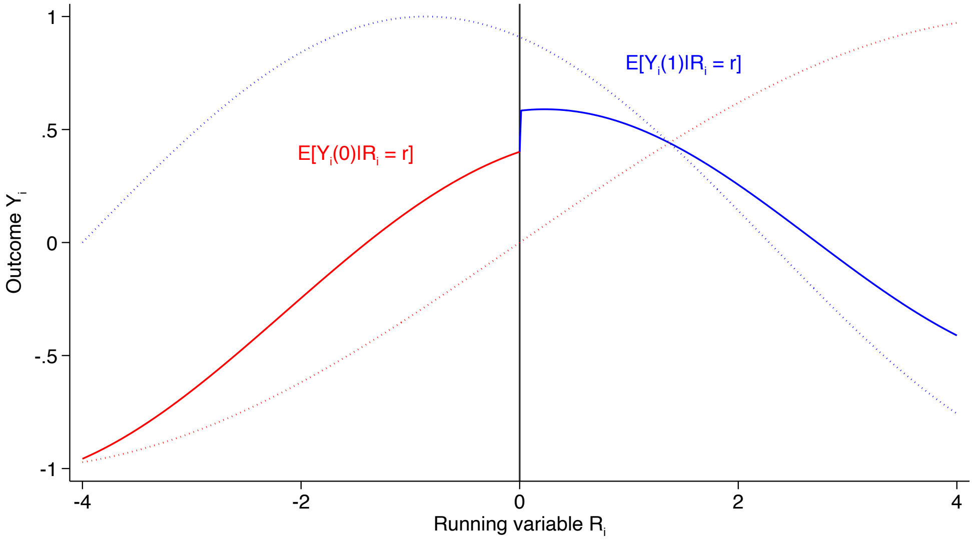
- This is the **fuzzy RD** scenario

- Example:
  - Income threshold determines program eligibility
  - But not all households particpate

Assignment Function in Fuzzy RD



Observed Outcome in Fuzzy RD



Fuzzy RD Assumptions

- As before, assume $Y_i(1)$ and $Y_i(0)$ are smooth around the threshold
- Let $D_i(1)$ and $D_i(0)$ denote treatment status above/below $R_i = c$
- **New assumptions**:
  1. Potential treatment continuity: $D_i(1)$ and $D_i(0)$ are smooth around the threshold
  2. Monotonicity: crossing threshold weakly increases treatment probability:

$$D_i(1) \geq D_i(0) \text{ for all } i$$

- Can again investigate comparison of individuals above and below threshold:

$$\lim_{r \downarrow c} \mathbb{E}\left[Y_i | R_i = r\right] - \lim_{r \uparrow c} \mathbb{E}\left[Y_i | R_i = r\right]$$

with our usual relationship that $Y_i = Y_i(0) + \left[Y_i(1) - Y_i(0)\right] D_i$

Difference in Outcomes and Treatment near Threshold

- We get: $\lim_{r\downarrow c} \mathbb{E}\left[Y_i|R_i = r\right] - \lim_{r\uparrow c} \mathbb{E}\left[Y_i|R_i = r\right] =$

$$
\begin{aligned}
&= \lim_{r\downarrow c} \mathbb{E}\left[Y_i(0) + \left(Y_i(1) - Y_i(0)\right) D_i | R_i = r\right] \\
&\quad - \lim_{r\uparrow c} \mathbb{E}\left[Y_i(0) + \left(Y_i(1) - Y_i(0)\right) D_i | R_i = r\right] \\
&= \lim_{r\downarrow c} \mathbb{E}\left[\left(Y_i(1) - Y_i(0)\right) D_i(1) | R_i = r\right] - \lim_{r\uparrow c} \mathbb{E}\left[\left(Y_i(1) - Y_i(0)\right) D_i(0) | R_i = r\right] \\
&= \mathbb{E}\left[\left(Y_i(1) - Y_i(0)\right)\left(D_i(1) - D_i(0)\right) | R_i = c\right] \\
&= \mathbb{E}\left[Y_i(1) - Y_i(0) | D_i(1) > D_i(0), R_i = c\right] \times \Pr\left(D_i(1) > D_i(0) | R_i = c\right)
\end{aligned}
$$

  where the last step uses monotonicity

- Similarly, comparing compliance probabilities at the threshold;

$$
\lim_{r\downarrow c} \mathbb{E}\left[D_i|R = c\right] - \lim_{r\uparrow c} \mathbb{E}\left[D_i|R = c\right] = \Pr\left(D_i(1) > D_i(0) | R_i = c\right)
$$

Fuzzy RD is IV

- Then we can take the ratio:

$$\frac{\lim_{r\downarrow c}\mathbb{E}\left[Y_i|R_i=r\right]-\lim_{r\uparrow c}\mathbb{E}\left[Y_i|R_i=r\right]}{\lim_{r\downarrow c}\mathbb{E}\left[D_i|R=c\right]-\lim_{r\uparrow c}\mathbb{E}\left[D_i|R=c\right]}=\mathbb{E}\left[Y_i(1)-Y_i(0)|D_i(1)>D_i(0),R_i=c\right]$$

- i.e. the jump in the outcome CEF, over the jump in the treatment probability
- this identifies the LATE of:
  (a) individuals who are at the threshold... (as in strict RD)
  (b) individuals who comply with the treatment (i.e. not always/never takers)
- Look familiar?

$\rightarrow$ **Fuzzy RD is IV**!

- Using a threshold indicator $Z_i = 1[R_i > c]$ as instrument for treatment
- ... in the neighborhood of the threshold

Fuzzy RD Implementation

- Can implement fuzzy RD with global 2SLS approach:

$$D_i = \lambda + \pi 1\left[R_i > c\right] + \sum_{k=1}^{K} \theta_{0k}\left(R_i - c\right)^k + \sum_{k=1}^{K} \theta_{1k} 1\left[R_i > c\right]\left(R_i - c\right)^k + \xi_i$$

$$Y_i = \alpha + \beta \widehat{D}_i + \sum_{k=1}^{K} \gamma_{0k}\left(R_i - c\right)^k + \sum_{k=1}^{K} \gamma_{1k} 1\left[R_i > c\right]\left(R_i - c\right)^k + \varepsilon_i$$

- $Z_i = 1[R_i > c]$ is the excluded instrument
- Alternatively, can estimate each of the four limits using local linear approach
    - The two limits of the outcome above/below threshold
    - The two limits of the treatment probability

Table of Contents

Multiple Discontinuities with Scalar Running Variable

- There may be more than one discontinuity along running variable
- Three approaches:
  1. Nonparametric solution: estimate each RD separately, then average across
     - Only feasible if discontinuities are far enough from one another
     - Requires sufficient precision on each RD separately
  2. Impose stationarity:
     - Concretely, running variable is distance to **nearest** discontinuity
     - Can then again use local linear methods
  3. Fully parametric:
     - Global approach with polynomials
     - Discontinuities may or may not be restricted to be the same

Multiple Discontinuity Example: Fredriksson et al (2013)

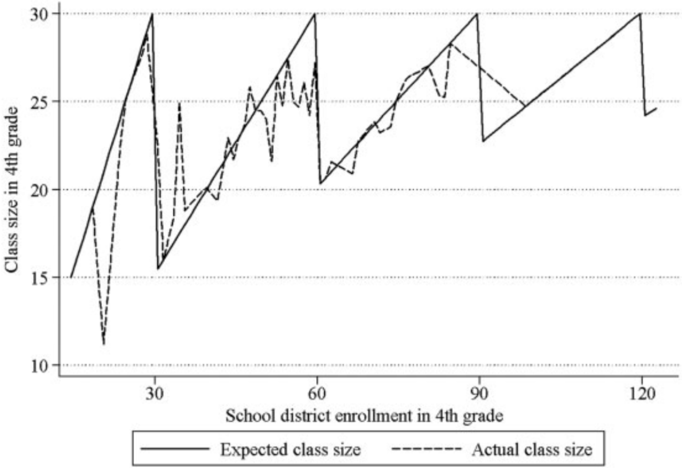


FIGURE II

Expected and Actual Class Size in Grade 4 by Enrollment in Grade 4

Residualized Outcome Around Standardized Running Variable

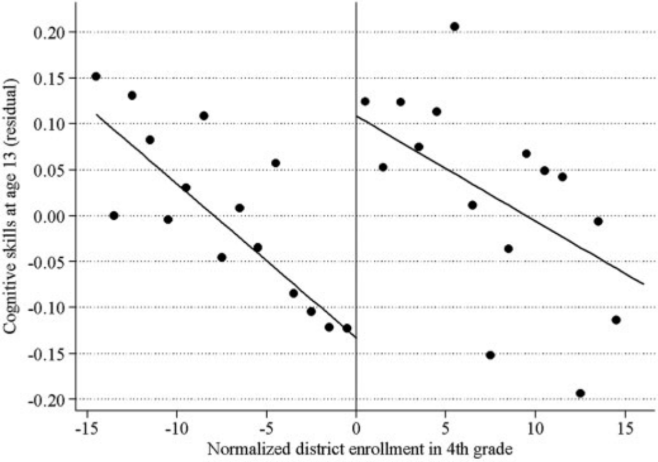


FIGURE VI

Cognitive Ability at Age 13 by Enrollment in Grade 4

Fuzzy Multidimensional RD Example: Londoño-Vélez et al (2020)

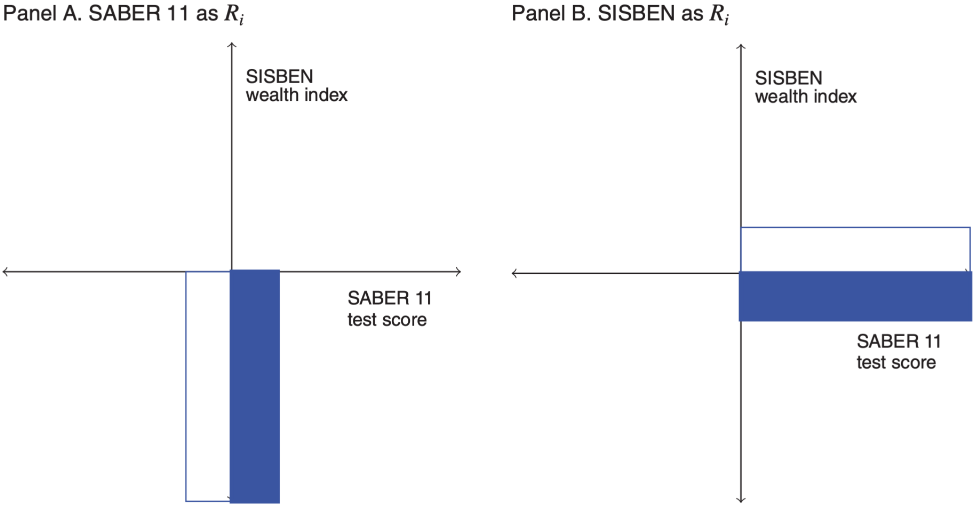


FIGURE 2. ILLUSTRATION OF THE TWO TYPES OF COMPLIERS

Assignment Probabilities Along Each Running Variable

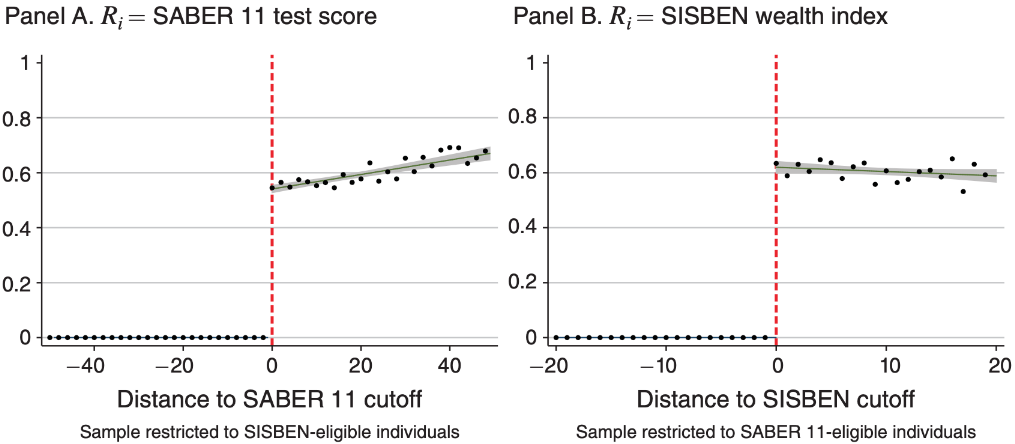


FIGURE 3. DISCONTINUITY IN THE PROBABILITY OF RECEIVING SPP FINANCIAL AID

Actual and Placebo Outcomes

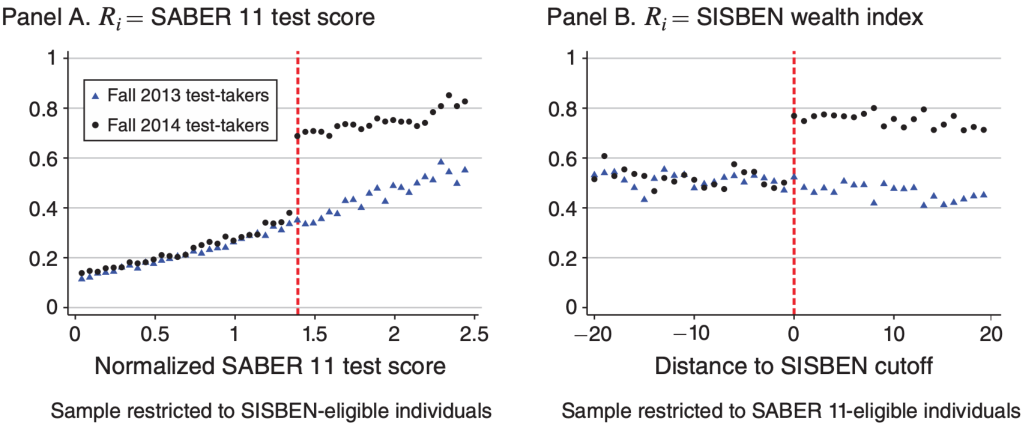


FIGURE 5. PLACEBO TEST USING PRE-TREATMENT PERIOD

Table of Contents

Boundary Discontinuity Design

- Many researchers used institutional boundaries as discontinuities
- Popular in urban, development, and economic history
- Challenges:
  1. Non-trivial two-dimensional boundary: it may arbitrarily move through space
  2. Individuals usually actively sort near boundaries

- Two general approaches:
  1. Nonparametric local RD
  2. Parametric RD with spatial polynomial in latitude and longitude

- Caveats:
  - Even passing diagnostics, concerns about sorting in unobservables
    - Exclude areas with geographic discontinuities (e.g. highways)
    - Include boundary segment fixed effects
  - May require model to convince people that RD is interesting
  - Even without sorting, spatial spillovers may be important

He et al (2020): Design

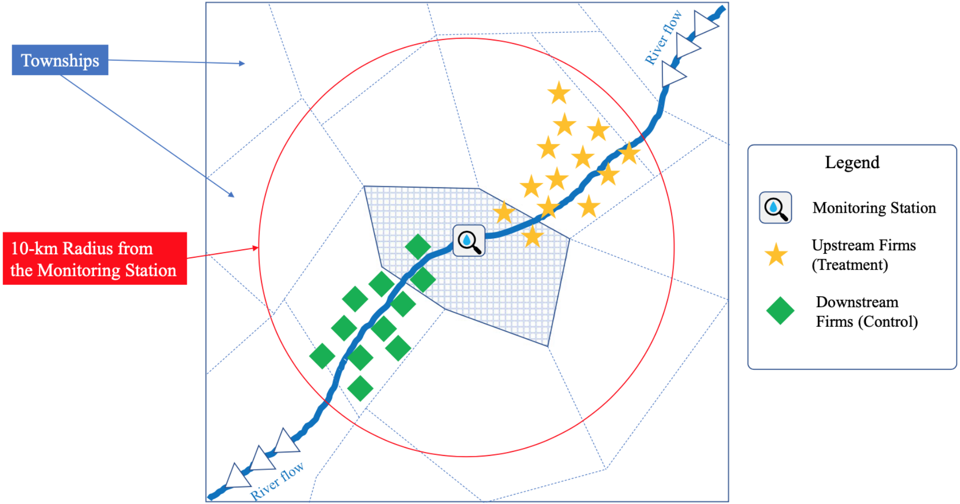
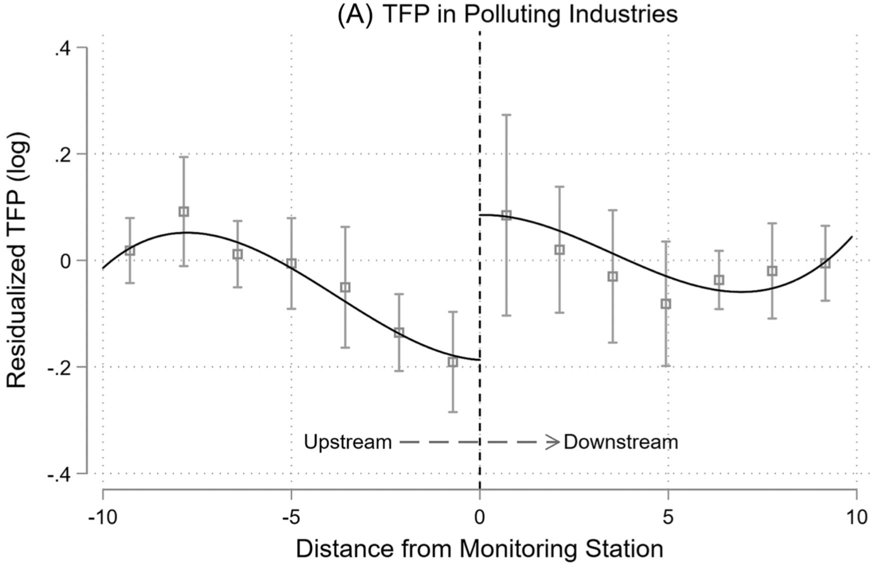
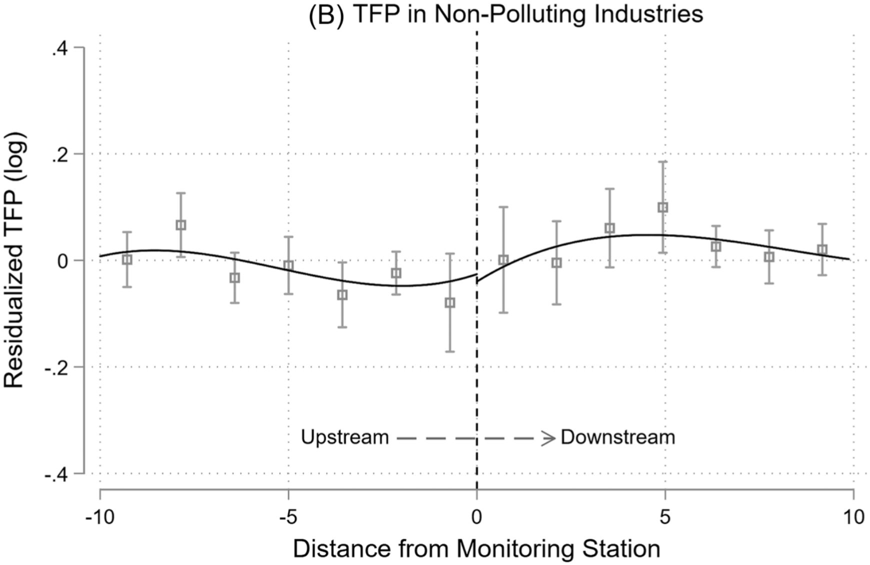


FIGURE II

Main Effects of Spatial Discontinuity



Placebo of Non-Polluting Firms



RD Estimates

TABLE I

THE UPSTREAM–DOWNSTREAM TFP GAP

	Polluting industries			Nonpolluting industries		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: No control						
RD in TFP (log)	0.34	0.37	0.32	− 0.03	0.04	0.01
(downstream − upstream)	(0.57)	(0.59)	(0.56)	(0.15)	(0.18)	(0.18)
Bandwidth (km)	4.203	3.889	3.622	5.887	5.168	4.522
Panel B: Station FE + industry FE absorbed						
RD in TFP (log)	0.36**	0.38**	0.34**	0.03	0.04	− 0.02
(downstream − upstream)	(0.17)	(0.17)	(0.15)	(0.09)	(0.09)	(0.09)
Bandwidth (km)	5.723	5.523	5.144	5.890	5.479	6.091
Panel C: Station by industry FE absorbed						
RD in TFP (log)	0.27*	0.29**	0.29**	0.02	0.04	0.03
(downstream − upstream)	(0.15)	(0.15)	(0.14)	(0.06)	(0.06)	(0.07)
Bandwidth (km)	4.496	4.333	4.689	5.692	5.204	4.430
Obs.	6,224	6,224	6,224	11,502	11,502	11,502
Kernel	Triangle	Epanech.	Uniform	Triangle	Epanech.	Uniform

Notes. Each cell in the table represents a separate RD regression. The running variable is the distance between a firm and a monitoring station, where negative (positive) distance means firms are located to the upstream (downstream) of the monitoring stations. The positive coefficients indicate that downstream firms have higher TFP than upstream firms. TFP is estimated using the Olley and Pakes (1996) method, with "upstream polluting" added as an additional state variable. The discontinuities at monitoring stations are estimated using local linear regressions and MSE-optimal bandwidth proposed by Calonico, Cattaneo, and Titiunik (2014) for different kernel weighting methods. Standard errors clustered at the monitoring station level are reported below the estimates. * significant at 10%, ** significant at 5%, *** significant at 1%.

Difference-in-Discontinuity

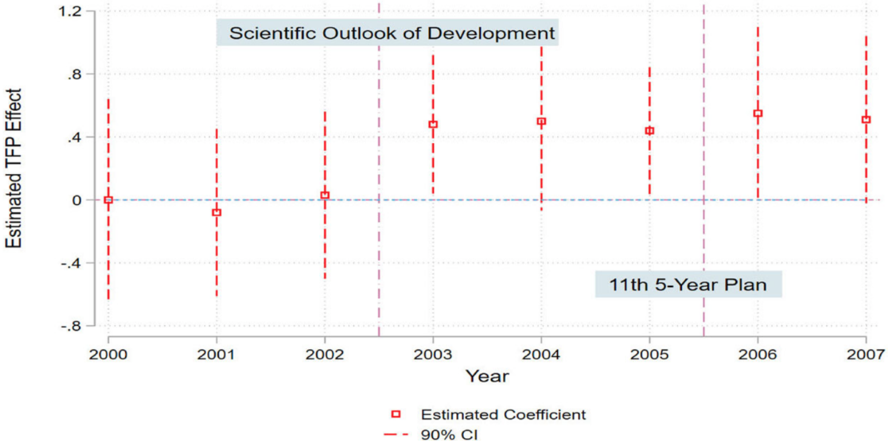


FIGURE V

RD Estimates by Year

Table of Contents

RD with a Discrete Running Variable

- In some cases, the running variable is discrete
- Example: treatment assignment based on test score with few questions
- Estimation techniques generally work, despite discreteness
- However, inference is problematic:
  - Card and Lee (2008) suggest clustering on values of running variable
  - But Kolesar and Rothe (2018) show that these CI have poor coverage
  - Propose alternative with guaranteed coverage under restrictions on CEF

- Conceptually distinct issue:
  - True running variable is continuous
  - But only observe running variable rounded at discrete values
  - See Dong (2015) for this issue

Table of Contents

Regression Kink Design

- Extension of RDD: the **regression kink design** (RKD) by Card et al (2015)
- Recall RDD logic:
  - Binary treatment $D_i(R_i)$, e.g. $D_i = 1[R_i > c]$ in strict RDD
  - Exploit discontinuous level change in treatment (probability)
- Basic RKD logic:
  - Continuous treatment $S_i(R_i)$, such as monthly unemployment payment
  - Exploit discontinuous slope change ("kink") in treatment intensity

- Suppose $S_i = b(R_i)$, e.g. $S_i = \gamma_0 R_i + 1[R_i > c] \times \gamma_1 R_i$ with $\gamma_1 > 0$
- Specifically, $b(\cdot)$ is a continuous function with kink at c
- Let $f_i(s)$ denote i's potential outcome if treated with intensity $s = S_i$

RKD Logic and Estimation

- Then under mild regularity conditions:

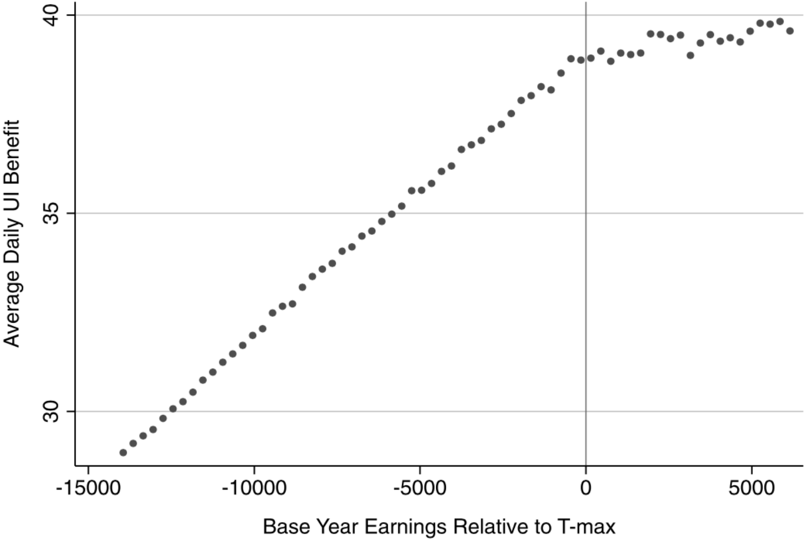
$$\frac{\lim_{r\downarrow c} \frac{d\mathbb{E}[Y_i|R_i=r]}{dr} - \lim_{r\uparrow c} \frac{d\mathbb{E}[Y_i|R_i=r]}{dr}}{\lim_{r\downarrow c} b'(r) - \lim_{r\uparrow c} b'(r)} = \mathbb{E}\left[f_i'(S_i)|R_i = c\right]$$

  which says the ratio of ...

  - ... discontinuity in outcome derivative ...
  - ... over discontinuity in the treatment derivative ...
  - ... identifies the **average marginal effect** of treatment ...
  - ... for individuals at the threshold

- As before, assumption is that POs are smooth around threshold
  - Any kink in outcome CEF must then be due to treatment
  - Diagnostics: check for kinks in covariates or bunching near c
- Estimation: analogous to RDD; see `rdrobust` for RKD implementation
- **Even more data-intensive than RDD**!

RKD Example: Card et al (2015)



RKD Example: Outcome