MY457/MY557: Causal Inference for Observational and Experimental Studies

Week 11: Regression Discontinuity

Daniel de Kadt

Department of Methodology
LSE

Winter Term 2024

Course Outline

- **Week 1:** The potential outcomes framework
- **Week 2:** Randomized experiments
- **Week 3:** Selection on observables I
- **Week 4:** Selection on observables II
- **Week 5:** Selection on observables III
- Week 6: Reading week
- **Week 7:** Difference-in-differences I
- **Week 8:** Difference-in-differences II
- **Week 9:** Instrumental variables I
- **Week 10:** Instrumental variables II
- **Week 11:** Regression discontinuity

Today

1. Sharp Regression Discontinuity Designs
2. Fuzzy Regression Discontinuity Designs
3. Regression Kink Designs

Table of Contents

A Motivating Example

Do long (short) election-day lines decrease (increase) turnout?

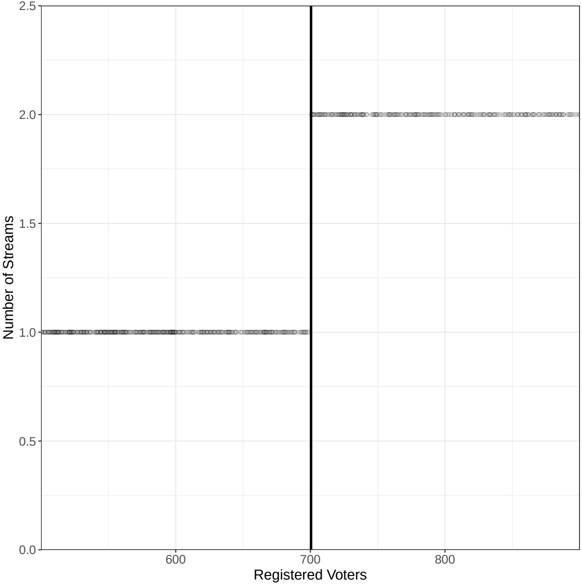
Unsurprisingly, a naïve study of this seems problematic:

- Higher turnout $\rightarrow$ longer lines (reverse causality)
- Longer lines occur where political interest is higher (confounding)
- Shorter lines occur where resourcing is better (confounding)

Harris (2020) studies the case of Kenya's 2017 election, where

- At any polling place, if there were up to 700 registered voters there would be just one stream (line/table).
- If the polling place had 701 registered voters, there would be two.

A Motivating Example: Design in Practice



Sharp RDD: Setup

Let's formalise this research setting:

- $D_i \in \{0,1\}$: Treatment
- X_i: Forcing variable (aka running variable or score) that perfectly determines D_i at cutpoint c:

$$D_i = \mathbf{1}\{X_i > c\} \quad \text{or equivalently} \quad D_i = \begin{cases} 1 & \text{if } X_i > c \\ 0 & \text{if } X_i \leq c \end{cases}$$

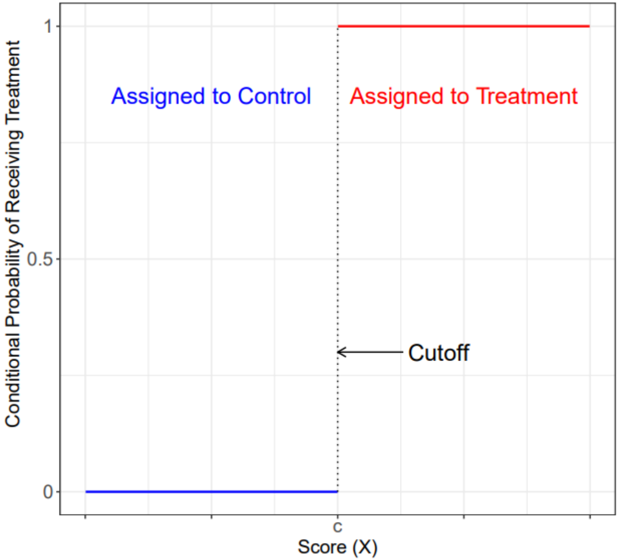
  Note: X_i may be correlated with Y_{0i} and Y_{1i}
- Potential outcomes: $\mathbb{E}[Y_{0i}|X_i]$ and $\mathbb{E}[Y_{1i}|X_i]$, defined for every value of X_i.

This looks kind of like selection on observables... If potential outcomes are a deterministic function of X_i, why not just adjust or control for X_i?

Lack of common support $\rightsquigarrow$ across all i, only one of Y_{0i} and Y_{1i} can be observed for each level of X_i.

Basic RDD intuition: At the cutpoint, we have 'as-if' random variation.

Sharp RDD: Illustrative Treatment Assignment



Source: Cattaneo et al. (2019)

Sharp RDD: Two Schools of Thought

Two frameworks for RDDs: continuity and local randomization.

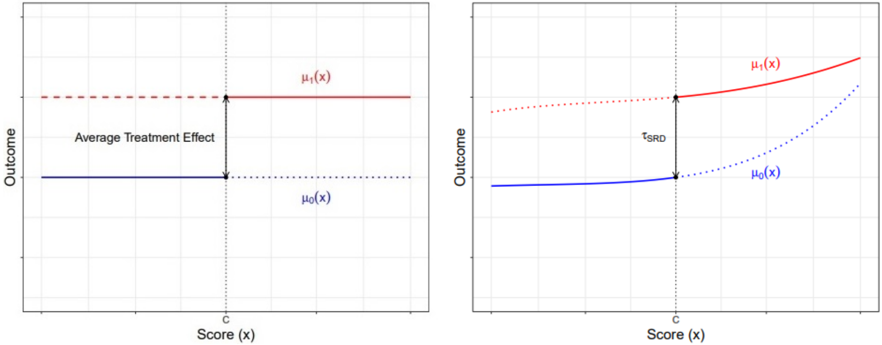
Local randomization is perhaps most intuitive; in fact, this was Thistlethwaite & Campbell's (1960) original view of the RDD.

Intuition: Within some small window around c, all units are as-if randomly assigned a value of X_i, and thus D_i.

Downside is that this requires a strong assumption: Within some known window around c, $(Y_{0i}, Y_{1i}) \perp\!\!\!\perp X_i$.

If satisfied, you can (roughly) use the experiment toolkit for analysis. See Cattaneo et al (2024) for more.

Sharp RDD: Two Schools of Thought



Source: Cattaneo et al. (2019)

Problem: How often is something like the left-hand plot really plausible?

Sharp RDD: Continuity for Identification

We will focus instead on the continuity framework.

Intuition: suppose there is no discontinuity in potential outcomes $\mathbb{E}[Y_{0i}|X_i = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$ at the threshold c.

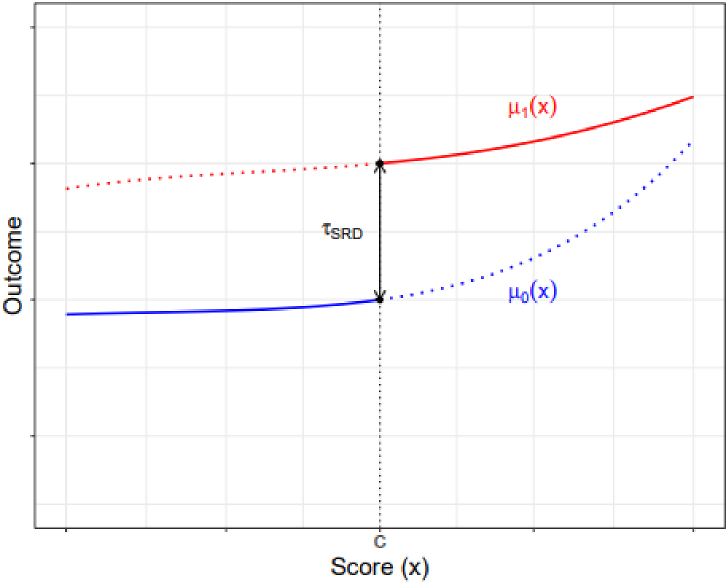
If $\mathbb{E}[Y_{0i}|X_i = x]$ and $\mathbb{E}[Y_{1i}|X_i = x]$ can be approximated by some function of X_i, can estimate missing potential outcomes by extrapolating to $X_i = c$.

Any difference in Y_i at $X_i = c$ is a causal effect!

Estimand: Local Average Treatment Effect (LATE) at the threshold

$$\tau_{SRD} \; = \; \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

Sharp RDD: Continuity in Potential Outcomes



Source: Cattaneo et al. (2019)

Sharp RDD: Continuity for Identification

Assume continuity of average potential outcomes. This implies:

$$\lim_{\varepsilon \uparrow 0} \mathbb{E}[Y_{0i}|X_i = c + \varepsilon] = \mathbb{E}[Y_{0i}|X_i = c]$$

$$\lim_{\varepsilon \downarrow 0} \mathbb{E}[Y_{1i}|X_i = c + \varepsilon] = \mathbb{E}[Y_{1i}|X_i = c]$$

Read: Potential outcomes arbitrarily close to the cutpoint are approximately the same as potential outcomes exactly at the cutpoint.

A simple proof:

$$\begin{aligned}
&\lim_{\varepsilon \downarrow c} \mathbb{E}[Y_i|X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow c} \mathbb{E}[Y_i|X_i = c + \varepsilon] \\
=&\lim_{\varepsilon \downarrow c} \mathbb{E}[Y_{1i}|X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow c} \mathbb{E}[Y_{0i}|X_i = c + \varepsilon] \\
=&\mathbb{E}[Y_{1i}|X_i = c] - \mathbb{E}[Y_{0i}|X_i = c] \quad \because \text{continuity} \\
=&\mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c] = \tau_{SRD}
\end{aligned}$$

Sharp RDD: Parametric Estimation Under Continuity

Under the continuity framework, estimation is an extrapolation problem.

One very simple approach would be to assume a parameteric model, where τ_{SRD} is constant, and potential outcomes are linear in X_i:

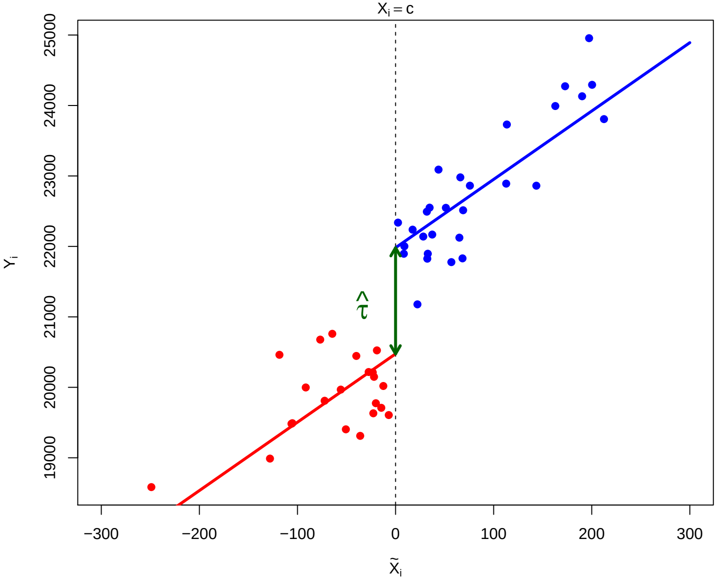
$$Y_{di} = \alpha + \tau_{SRD} d + \beta X_i$$

To estimate τ_{SRD}:

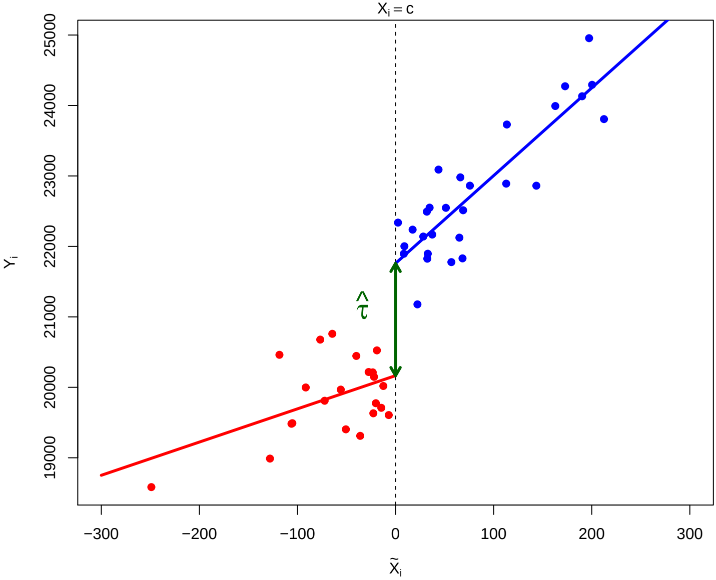
1. Recenter forcing variable: $\tilde{X}_i = X_i - c$
2. Regress $Y_i = \hat{\alpha} + \hat{\tau_{SRD}} D_i + \hat{\beta} \tilde{X}_i$
3. $\hat{\tau_{SRD}}$ gives the LATE.

We could assume a more flexible (realistic?) functional form, e.g. varying slopes in X_i, or polynomial functions of X_i, and fit that regression.

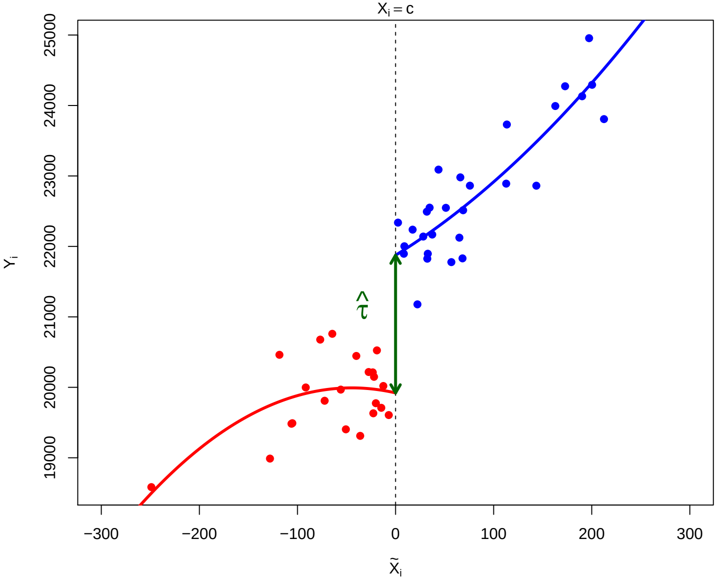
Sharp RDD: Common Slopes Linear Parametric Estimation



Sharp RDD: Varying Slopes Linear Parametric Estimation



Sharp RDD: Varying Polynomial Parametric Estimation



Sharp RDD: Local Polynomial Approximation

Whatever function we choose, we make strong parametric assumptions.

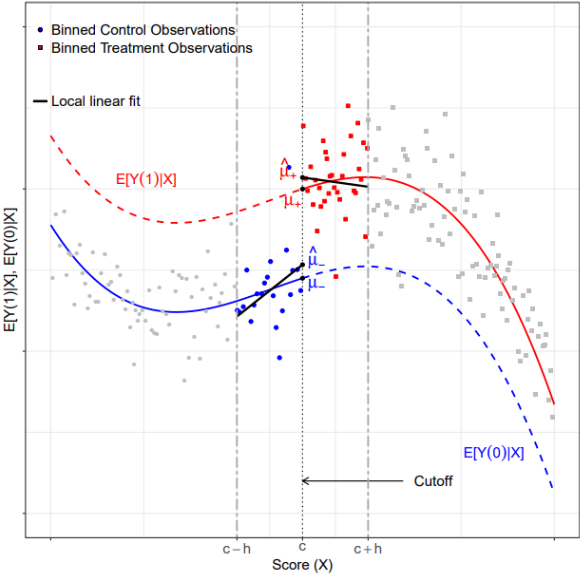
Current state of the art is local polynomial approximation, which offers a non-parametric estimator of τ_{SRD}.

Proceeds as follows:

1. Choose bandwidth or window h
2. Choose polynomial order p and kernel function $K(\cdot)$
3. Fit two weighted regressions (for $X_i > c$ and $X_i \leq c$), as follows:
   a. Treated: Regress Y_i on global constant $\mu_\downarrow$ plus $\sum_{p=1}^{p}(X_i - c)^p$
   b. Control: Regress Y_i on global constant $\mu_\uparrow$ plus $\sum_{p=1}^{p}(X_i - c)^p$
   Weights: For both, separately weight observations by $K(\frac{X_i - c}{h})$
4. Calculate $\hat{\tau_{SRD}} = \hat{\mu_\downarrow} - \hat{\mu_\uparrow}$

Implemented with `rdrobust` in R. See Cattaneo et al. (2019, 2024).

Sharp RDD: Local Polynomial Point Estimation



Source: Cattaneo et al. (2019)

Sharp RDD: Choosing p and $K(\cdot)$

Selecting p:

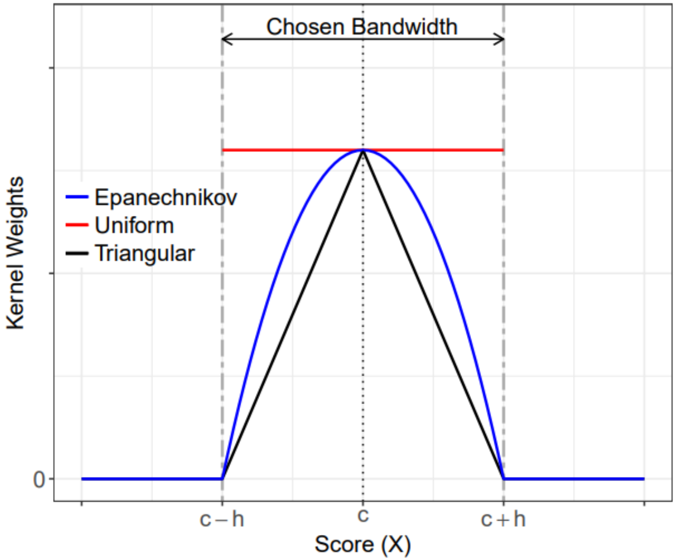
- Lower p will increase bias, but decrease variance
- Higher p will decrease bias, but increase variance
- Default is to set $p = 1$ ('local linear regression') and let h take care of the above.

Selecting $K(\cdot)$:

- Controls the weighting of observations as a function of proximity to c
- Intuitively, we want to up-weight those close to the cutpoint
- Default is a triangular kernel, but uniform or Epanechnikov kernels are sometimes used

Recommendation: Stick with the defaults unless you have a very good justification.

Sharp RDD: Kernel Choices



Source: Cattaneo et al. (2019)

Sharp RDD: Choosing h

Cattaneo et al. (2019) propose the mean-squared-error optimal bandwidth:

$$\text{MSE}(\hat{\tau}_{SRD}) = \text{Bias}^2(\hat{\tau}_{SRD}) + \text{Var}(\hat{\tau}_{SRD}) = (h^{2(p+1)}\mathcal{B})^2 + \frac{1}{nh}\mathcal{V}$$

where:

- $\mathcal{B}$ is bias
- $\mathcal{V}$ is variance

Select h that minimizes this MSE (conditional on p and $K(\cdot)$):

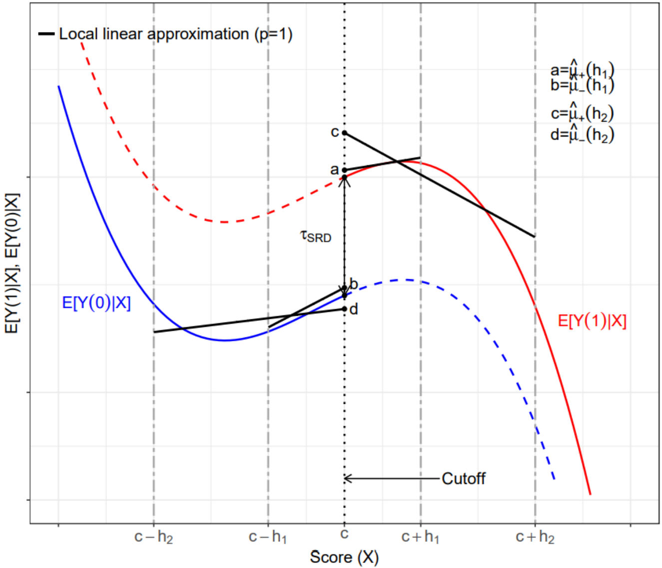
$$h_{MSE} = argmin \left(\frac{\mathcal{V}}{2(p+1)\mathcal{B}^2} \right)^{1/(2p+3)} n^{-1/(2p+3)}$$

e.g. for $p = 1$, $h_{MSE} = n^{-1/5}$

Note: the choice of h can vary on either side of c.

It turns out choice of h is very important...

Sharp RDD: Local Polynomial Sensitivity to h



Source: Cattaneo et al. (2019)

Sharp RDD: Bias from h and Bias-Correction

The bias term is $h^{2(p+1)} \to n^{-4/5}$, a slower convergence rate than n.

Calonico et al. propose an 'undersmoothing' bias-correction:

- Select h_{MSE}, and a smaller $h_* < h_{MSE}$
- Use local polynomial estimator and generate confidence intervals
- Use these CIs instead of those from h_{MSE}

Alternatively, they also propose 'robust bias correction':

- Directly estimate bias term $\mathcal{B}$
- Subtract off $\tau_{\hat{SRD}}$, and generate CIs using this

Can implement both with `rdrobust`, for point estimation, SE estimation, or both.

Sharp RDD: Threats and Falsification

An (incomplete) list of threats and good RDD practices:

1. Smooth instead of discontinuous function of Y_i?
   $\rightsquigarrow$ visualisation of binned points using `rdplot` – jump should be clear
2. Discontinuities in potential confounders?
   $\rightsquigarrow$ balance or continuity tests (using the same specification!)
3. Sorting or manipulation around the threshold?
   $\rightsquigarrow$ McCrary (2008) density test or Cattaneo at al (2020) density test with `rddensity`
4. Sensitivity to researcher choices?
   $\rightsquigarrow$ robustness across choices – computationally cheap
5. Highly localised effects or potential spillovers?
   $\rightsquigarrow$ do(ugh)nut estimation approaches
6. Generally jumpy data creating a 'false discontinuity'?
   $\rightsquigarrow$ placebo cutpoints to benchmark jumpiness

Returning to the Motivating Example: Estimation

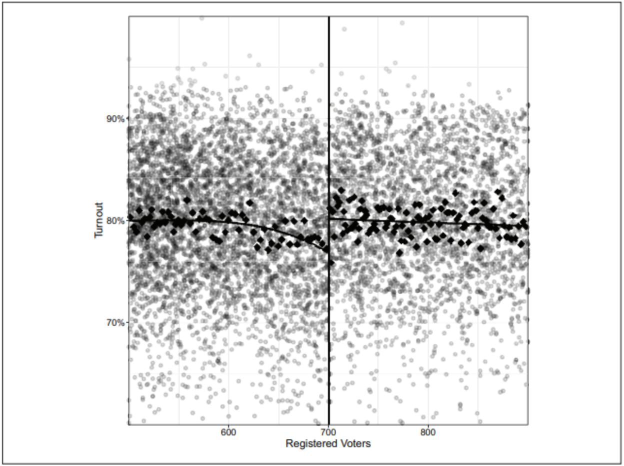


Figure 2. Discontinuity plot—turnout: polling centers above the 700-registered voter cutoff have an additional stream, leading to 2.4% higher voter turnout than polling centers below the cutoff. Dots represent individual polling centers. Squares show bin-specific turnout summaries.

Returning to the Motivating Example: Balance Test

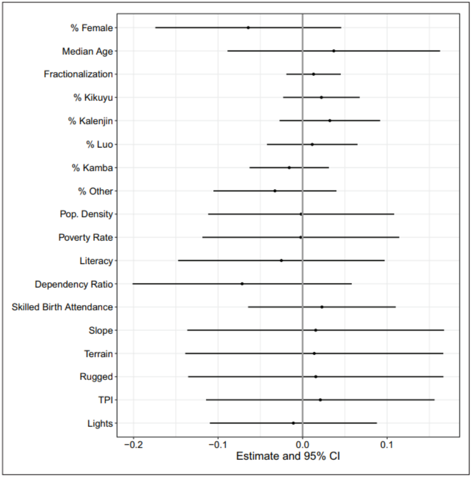
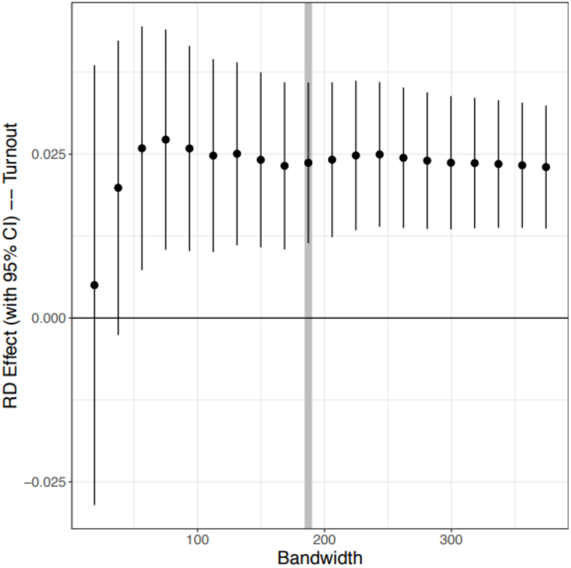
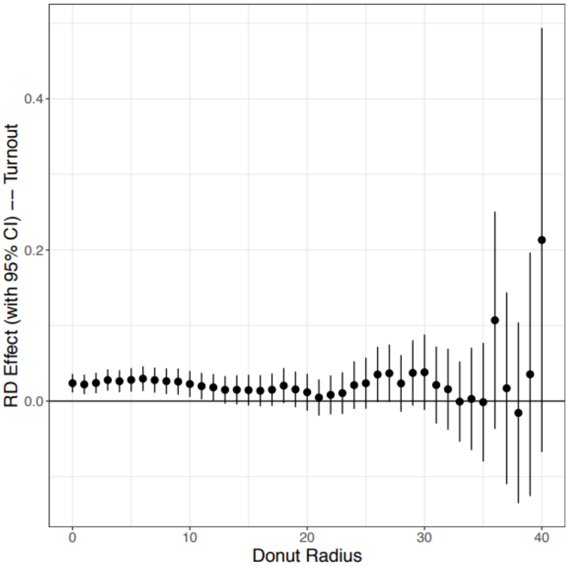


Figure 1. Balance tests: polling centers just above the 700-registered voter cutoff, at which an additional stream is added to a polling center, are similar to those just below the cutoff. Balance tests follow Cattaneo et al. (2020a) and estimate the effect at the discontinuity on predetermined characteristics. Line plots display the 95% confidence interval of the estimated difference at the cutoff.

Returning to the Motivating Example: Sensitivity



Returning to the Motivating Example: Donut RDD



Returning to the Motivating Example: Placebo Cutpoints

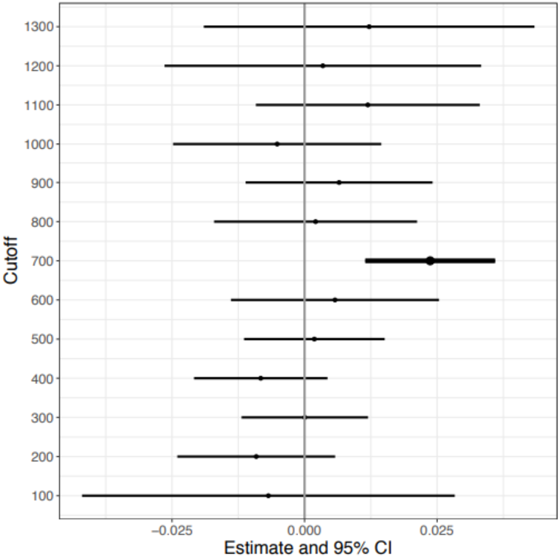


Table of Contents

A New Motivating Example

How does additional schooling affect political beliefs, like Euroscepticism?

You know the drill – want to avoid naïve comparison. (Why?)

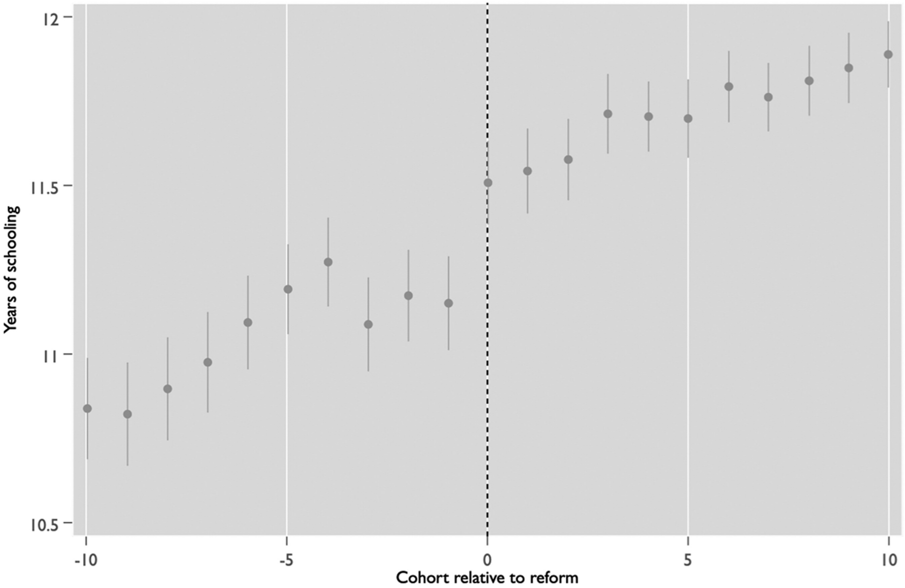
Kunst, Kuhn, and van der Werfhorst (2019) study survey respondents (i) in 12 European countries (k):

- Compulsory schooling reforms were passed between 1947 and 1983, affecting only those younger than a certain age
- Construct a forcing variable $X_i = [\text{Y.o.B}_i]–[\text{Y.o.B First Affected}_k]$

Notes:

- The authors observe year-of-birth, so forcing variable is discretised.
- This is really an example of a regression discontinuity in time (RDiT), where the forcing variable is a function of time. See Hausman & Rapson (2018) for review and best practices

A Motivating Example: Design in Practice



Fuzzy RDD: Setup

Formalising this research setting:

- $Z_i \in \{0,1\}$: Encouragement
- $D_i \in \{0,1\}$: Treatment, a probabilistic function of Z_i
- X_i: Forcing variable perfectly determines Z_i with cutpoint c

$$Z_i = \mathbf{1}\{X_i > c\} \quad \text{or equivalently} \quad Z_i = \begin{cases} 1 & \text{if } X_i > c \\ 0 & \text{if } X_i \leq c \end{cases}$$

Note: The reduced form (effect of Z_i on Y_i) is just a sharp RDD.

Assumptions:

1. Both $\mathbb{E}[D_{zi} \mid X_i = x]$ (p.o. for treatment) and $\mathbb{E}[Y_{zi} \mid X_i = x]$ (p.o. for dependent variable) are continuous in x around $X_i = c$ for $z = 0, 1$
2. IV assumptions: Monotonicity, exclusion restriction, relevance of Z_i

Fuzzy RDD: Identification

Estimands:

1. Local ITT of encouragement at the threshold

$$\tau_{LITT} \;=\; \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c]$$

2. LATE for compliers at the threshold

$$\tau_{FRD} \;=\; \mathbb{E}[Y_{1i} - Y_{0i} \mid \text{unit } i \text{ is a complier and } X_i = c]$$

Identification results:

1. Under just continuity in $\mathbb{E}[Y_{zi} \mid X_i = x]$:

$$\tau_{LITT} = \lim_{\varepsilon \downarrow 0} \mathbb{E}[Y_i | X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbb{E}[Y_i | X_i = c + \varepsilon]$$

2. Under both continuity assumptions + IV assumptions:

$$\tau_{FRD} \;=\; \frac{\lim_{\varepsilon \downarrow 0} \mathbb{E}[Y_i | X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbb{E}[Y_i | X_i = c + \varepsilon]}{\lim_{\varepsilon \downarrow 0} \mathbb{E}[D_i | X_i = c + \varepsilon] - \lim_{\varepsilon \uparrow 0} \mathbb{E}[D_i | X_i = c + \varepsilon]}$$

Fuzzy RDD: Estimation

Parametric estimation for τ_{FRD}:

1. Code instrument: $Z = \mathbf{1}\{X > c\}$
2. Fit 2SLS:

$$\text{First Stage: } D_i = f(X_i) + \beta Z_i + \varepsilon_i$$

$$\text{Second Stage: } Y_i = f(X_i) + \alpha \hat{D}_i + \nu_i$$

Note: Specification of f is flexible but must be same in both stages

Non-parametric estimation:

1. τ_{LITT} can be estimated using local polynomial approximation, as the LATE was for a sharp RDD. Why?
2. Proportion of compliers can likewise be estimated with D_i as the outcome
3. τ_{FRD} (for compliers at the threshold) is just $\frac{\tau_{LITT}}{Pr(Compliers|X_i=c)}$

Whatever you do, it is critical that you test and visualise the first stage. A weak (or non-existent) first stage generates severe bias, and misleads.

Sharp and Fuzzy RDD: Internal and External Validity

Note that, like IV, both sharp and fuzzy RDDs focus on specific sub-populations (so 'local' has different meanings):

- IV estimates the LATE for compliers.
- Sharp RDD estimates the LATE on the subpopulation with X_i close to c
- Fuzzy RDD does both – the LATE for compliers with X_i close to c

Only with strong assumptions (e.g., continuity and homogeneous treatment effects across all values of X) can we move from LATE to a more general estimand!

Table of Contents

A Motivating Example (from my PhD thesis ☺)

Does electoral pivotality affect political participation?

One determinant of pivotality is the number of voters per race. The higher (lower) the number of voters, the lower (higher) each voter's pivotality.

As usual, be wary of a naïve study!

Consider South Africa, where the size of local races is set by a kinked formula:

- Within local governments, the number of councillors is determined by formula
- Councillors are added as a function of registered voters in the area
- At certain thresholds (in terms of registered voters), the rate at which councillors are added changes

Kinked Formula for Seat Allocation

Formula for determination of number of councillors for category B municipalities

2.(a) The formula for determining the number of councillors of a category B municipality is –

(i) in respect of such a municipality that has less than 7 501 registered voters on its segment of the national common voters roll:

$y = 5;$

(ii) in respect of such a municipality that has between 7 500 and 100 001 registered voters on its segment of the national common voters roll:

$y = (x \div 2\,055) + 2;$ and

(iii) in respect of such a municipality that has more than 100 000 registered voters on its segment of the national common voters roll:

$y = (x \div 8\,333) + 48.$

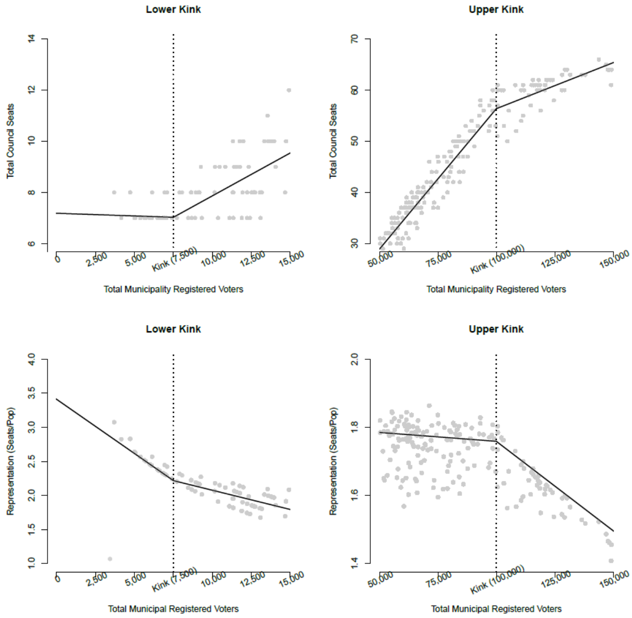
(b) In applying the formulae referred to in paragraph (a)-

(i) y represents the number of councillors;

(ii) x represents the number of registered voters on the municipality's segment of the national common voters roll on 5 March 2014; and

(iii) fractions are to be disregarded.

Kinked Formula for Seat Allocation



RKD: Setup

Setup is the similar to the SRD case (or the FRD case, if there is non-compliance):

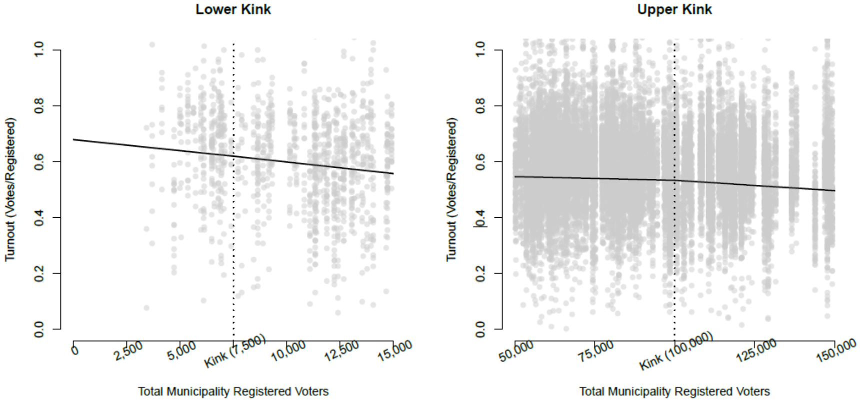
- Y_i: outcome of interest
- X_i: the forcing variable, with cutpoint c
- $W_i = w(X_i)$: a continuous variable, which is a function of X_i, and that function changes at $X_i = c$

Difference: treatment effect is not (exclusively) in terms of a level shift in Y_i, but a slope shift in the relationship between Y_i and X_i, driven by a change in the slope of the relationship between W_i and X_i.

We call this estimand the Local Average Response (LAR):

$$\tau_{LAR} = \frac{\lim_{x \downarrow c} \frac{d\mathbb{E}[Y_i | X_i = x]}{dx} - \lim_{x \uparrow c} \frac{d\mathbb{E}[Y_i | X_i = x]}{dx}}{\lim_{x \downarrow c} \frac{dw(x)}{dx} - \lim_{x \uparrow c} \frac{dw(x)}{dx}}$$

Kinked Formula for Seat Allocation



"Be very very quiet, we're hunting exogenous variation in D"