

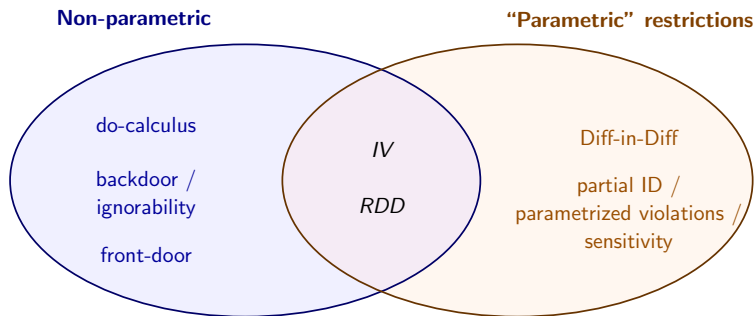
Stat 256: Regression Discontinuity

Chad Hazlett

UCLA

`chazlett@ucla.edu`

Spring 2026



RD in the ID landscape

- So far: adjustment strategies (backdoor/CI, front-door, do-calc), and IV.
- RD is almost obvious in retrospect, but hard to see from a DAG alone.
- Multiple justifications coexist: local randomization, continuity of (CEFs of) POs, IV (for fuzzy RD).
- Unusually, RD is also *always estimation-dependent* — much of the literature is about estimation choices.
- Two main flavors: **sharp** and **fuzzy** RD.

A motivating case: Brazilian electronic voting

Brazil introduced electronic voting machines in 1998, replacing paper ballots.

Roll-out was **not** based on need, ideology, or party — it was based on a **population rule**:

- Municipalities with $> 40,500$ registered voters $\rightarrow$ e-voting in 1998
- Municipalities at or below the threshold $\rightarrow$ paper ballots

Hidalgo (2012) asks: did e-voting change *who* got represented?

- Paper ballots required reading/writing candidate numbers; many voters spoiled their ballots.
- E-voting effectively enfranchised millions of less-educated voters.

What's the causal effect of e-voting on the share of valid votes (and downstream, on left-leaning vote share, social spending, etc.)?

Why we can't just compare

Naive: compare e-voting municipalities to paper-ballot ones.

The threshold cut the country into two very unequal halves:

- Above threshold: bigger, richer, more urban, more literate, more developed.
- Below threshold: small, rural, poorer.

An SOO strategy would require trying to catch all such confounding differences. And we know there is not common support (on population).

The RD trick: give up on the ATE; just compare *near the cutoff*, where a small variation in population would have flipped treatment status.

Result

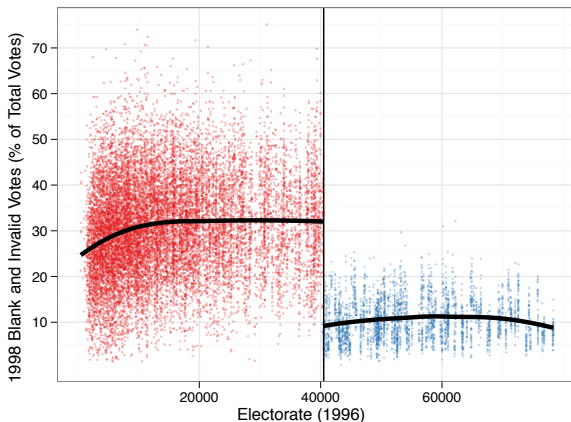


Figure 6: The effect of electronic voting on the percent of null and blank votes. Each dot is a polling station. Polling stations to the left of the vertical black line used paper ballots and polling stations to the right used electronic voting. The black horizontal line is the conditional mean of the outcome estimated with a loess regression.

Generalizing the idea

A binary treatment D_i **determined** by whether a continuous predictor X_i crosses a fixed threshold c :

$$D_i = \mathbb{1}\{X_i > c\} \iff D_i = \begin{cases} 1 & \text{if } X_i > c \\ 0 & \text{if } X_i \leq c \end{cases}$$

Common in rule-based settings: administrative programs, eligibility cutoffs, electoral rules, age cutoffs...

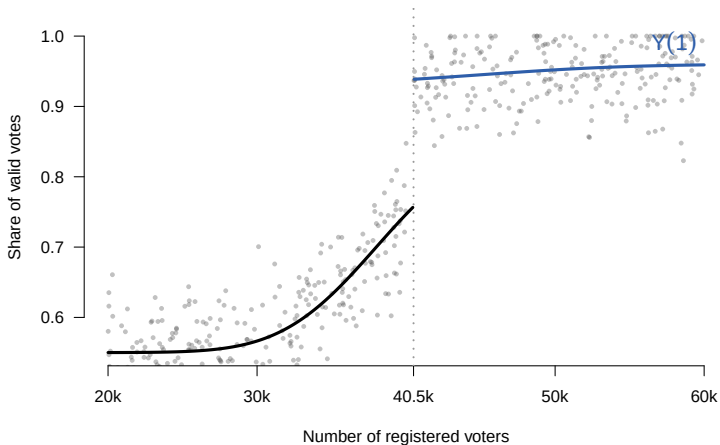
X_i : the **forcing** or **running** variable; c : the **cutoff** or **threshold**.

X_i may be (and usually is) correlated with the potential outcomes. This isn't an experiment.

Later: *fuzzy* RD, where $X_i > c$ **shifts the probability** of D rather than determining it.

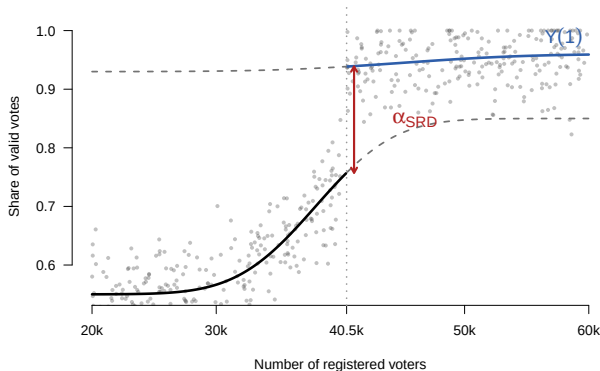
Continuity of potential outcomes

Each municipality has two potential outcomes: $Y(0)$ if paper, $Y(1)$ if e-voting, but we see only one.



Continuity of potential outcomes

Really, think of this in terms of two objects: $\mathbb{E}[Y(0) | X]$ and $\mathbb{E}[Y(1) | X]$:



The effect α_{SRD} is the vertical gap between the tracks *at the cutoff*.

The core assumption is **continuity**: both tracks vary *smoothly* with X at c .

Sharp RD: identification

Identification Assumption (Continuity of conditional expectations)

$\mathbb{E}[Y_{1i} | X_i]$ and $\mathbb{E}[Y_{0i} | X_i]$ are continuous in X_i at $X_i = c$.

Plain language: both potential outcome CEFs move *smoothly* across the running variable, at least at the cutoff.

Identification Result

The treatment effect at the threshold is identified as:

$$\begin{aligned}\alpha_{SRD} &= \mathbb{E}[Y_{1i} - Y_{0i} | X_i = c] \\ &= \lim_{x \downarrow c} \mathbb{E}[Y_i | X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i | X_i = x]\end{aligned}$$

*Without further assumptions, α_{SRD} is **only valid at the threshold**.*

Where the two limits come from

Continuity is exactly what lets us replace each PO at $X = c$ with a one-sided observable limit:

$$\mathbb{E}[Y_{1i} \mid X_i = c] = \lim_{\epsilon \downarrow 0} \mathbb{E}[Y_i \mid X_i = c + \epsilon, D_i = 1]$$

$$\mathbb{E}[Y_{0i} \mid X_i = c] = \lim_{\epsilon \uparrow 0} \mathbb{E}[Y_i \mid X_i = c + \epsilon, D_i = 0]$$

Above the cutoff everyone is treated, below it everyone is untreated, so each one-sided limit is observable.

Estimation is then just:

$$\hat{\alpha}_{SRD} = \hat{\mathbb{E}}[Y_{1i} \mid X_i = c] - \hat{\mathbb{E}}[Y_{0i} \mid X_i = c]$$

Note: not non-parametric—ID invokes a (smoothness) assumption on $\mathbb{E}[Y \mid X]$ at the boundary.

Comments on identification

- We are only saying that $Y(0)$ and $Y(1)$ don't suddenly *jump* at $X = c$.
 - We are *not* saying $Y(0)$ and $Y(1)$ are unrelated to X ,
 - We are *not* saying $Y(0)$ and $Y(1)$ are flat or have the same shape.
- The main thing that would violate this: another treatment kicking in at the same cutoff.
 - If something else changes at $X = c$ that affects Y , then $Y(0)$ (and $Y(1)$) themselves jump.
 - RD would mis-attribute the effect of the other treatment to the one you are asking about.

This is *the main* threat to keep in mind. We'll return to it.

Local randomization

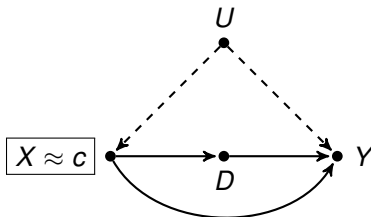
Rather than continuity of POs, some authors speak of “local randomization” (Lee 2008; Skovron & Titiunik 2015), at least as a heuristic.

- cut the data close to $X = c$ and treat D as if randomly assigned
- analyze like an experiment.
- lets some bias creep in whenever $\mathbb{E}[Y(0) | X]$ is not flat over the window

Continuity is the more defensible primitive; local randomization is a useful approximation, no need to rely on it.

A DAG view of sharp RD

One proposal¹:

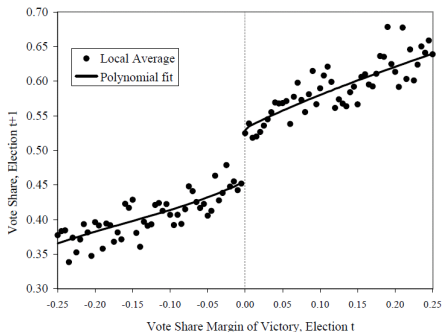


- Choosing units with $X \approx c$ acts like a soft conditioning on X : limits the backdoor $U \rightarrow X \rightarrow D$ and the direct $X \rightarrow Y$.
- We need *some* variation in X to get variation in D — so some bias creeps in.
- Smoothness of $X \leftarrow U \rightarrow Y$ and of $X \rightarrow Y$ near $X = c$ make the bias small.

¹Cf. Steiner et al. (2017) “limiting graphs”; Oldenburg et al. (2016); Huntington-Klein.

Incumbency advantage (Lee, 2008)

Figure IVa: Democrat Party's Vote Share in Election $t+1$, by Margin of Victory in Election t : local averages and parametric fit



- X_i = Dem party vote margin in last election,
- $c = 0$, D_i = won the seat,
- Y_i = Dem party vote share next election.

A few others

- Class size on student achievement (Angrist & Lavy 1999, Maimonides' rule)
- Access to credit (loan offer determined by credit score threshold)
- Mayoral wages on performance (Ferraz & Finan 2011)
- Winning office on politicians' wealth (Eggers & Hainmueller 2009)
- School district boundaries on home values (Black 1999)
- "Close election" RD has become a workhorse:
Lee, Moretti & Butler 2004; Pettersson-Lidbom 2008; Eggers & Hainmueller 2009; many more.

Warning: "close election" designs with treatment other than *winning* (party, gender, education) measure the effect of *electing somebody with that characteristic* — think districts, not individuals.

How to actually estimate

Estimation preferences for RD are still evolving. Three options worth knowing:

- Do something simple: separate linear or quadratic models.
- Local polynomial with automated bandwidth (e.g. `rdrobust`).
- Gaussian process (e.g. `gpss::gp_rdd`).

All three fit $\mathbb{E}[Y | X]$ on each side of c and read off the gap at c . They differ mainly in how they extrapolate to the boundary and how they propagate uncertainty.

Option 1. Simple, separate polynomial models

Suppose $\mathbb{E}[Y_0 | X]$ and $\mathbb{E}[Y_1 | X]$ are separate low-order polynomials of X , and the effect $\mathbb{E}[Y_1 - Y_0 | X]$ can vary with X .

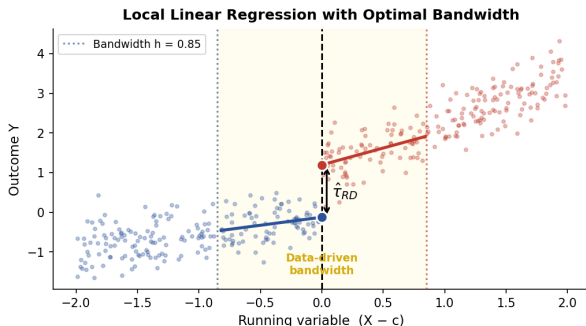
Quadratic with differential slopes:

$$\begin{aligned}\mathbb{E}[Y_i | X_i, D_i] = & \gamma_0 + \gamma_1 \tilde{X}_i + \gamma_2 \tilde{X}_i^2 \\ & + \alpha_0 D_i + \alpha_1 (\tilde{X}_i \cdot D_i) + \alpha_2 (\tilde{X}_i^2 \cdot D_i)\end{aligned}$$

- Typically limit data to a window $[c - h \leq X \leq c + h]$ first.
- $\alpha_0 = \mathbb{E}[Y_1 - Y_0 | X = c]$ is the effect at the cutoff.

IMHO: good for *seeing* your data, not for publication. Quadratic is a nice balance of simple and flexible. Cubic gets risky.

Option 2: rdrobust (Calonico, Cattaneo, Titiunik 2014)



- Local linear regression on each side, triangular kernel
- $\hat{\alpha}$ = gap in fitted values at c
- Automated, principled bandwidth selection
- Bias-corrected, robust CIs

`rddrobust(y, x, c = 0)`

Option 3: Gaussian process (`gpss::gp_rdd`)

Two risks of any local-polynomial fit:

- Polynomials happily extrapolate near the edge — where they have least data and most leverage.
- Uncertainty should *grow* as we approach c because data thins.

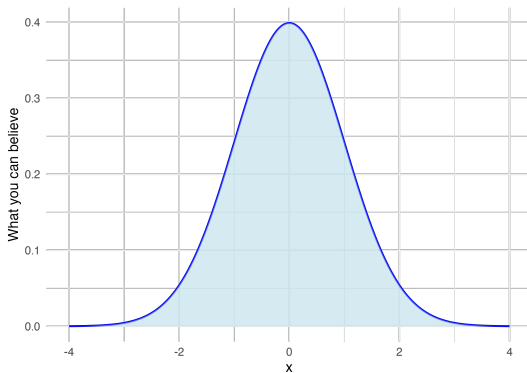
The Gaussian process:

- Fits in a flexible space of smooth functions.
- Carries an honest range of plausible curves through the data → uncertainty grows toward / past the edge.
- Regularizes on each side, avoiding overreaction to a few noisy points.
- Fully automated (Cho, Kim & Hazlett 2025): `gpss::gp_rdd(x, y, cut)`.

A short detour into the GP...

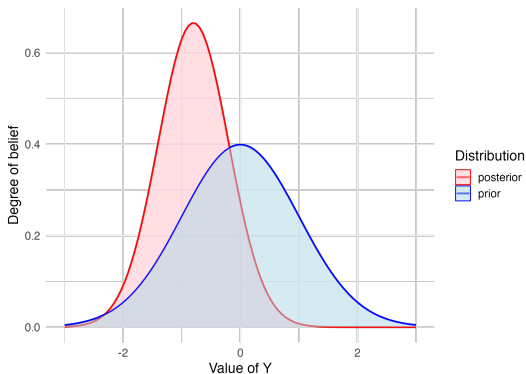
GP intuition: total ignorance

If I tell you Y^* has mean 0 and variance σ^2 , all you can believe is:



GP intuition: informed ignorance

Now I tell you we see $Y_{\text{obs}} = -1$, and $\text{cor}(Y^*, Y_{\text{obs}}) = 0.8$. Then you believe:



GP intuition: neighbors

Now suppose you have $\{X_i, Y_i\}_{i=1}^N$.

Big idea: observations nearer each other in X should have similar Y :

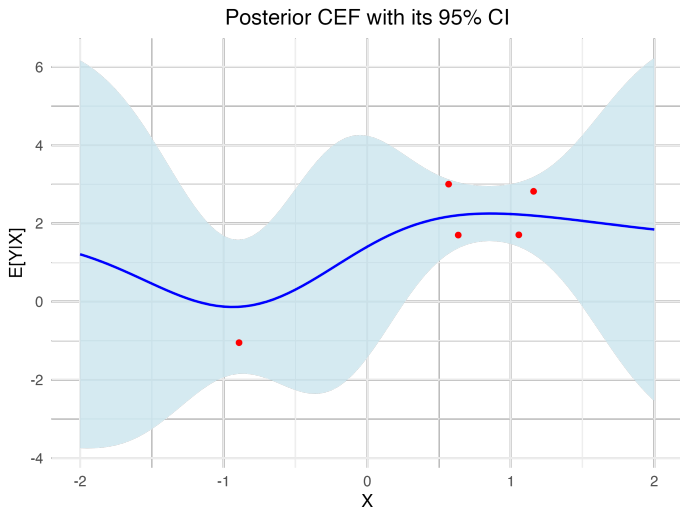
- $\text{cov}(Y_i, Y_j)$ larger when X_i nearer X_j .
- E.g. $\text{cov}(Y_i, Y_j) = k(X_i, X_j) = \exp(-\|X_i - X_j\|^2/b)$.

Now we want to guess about some Y_i :

- If its X_i is *near* an observed unit's X_j , the Y_j (observed) will have high covariance with unknown Y_i ; sharpens our guess about Y_i .
- At X_i *far* from observed points: low covariance with observed Y ...ignorance.

Inferences given a few observations

Given 5 observations of $\{X_i, Y_i\}$, ask for predictions at any X :



Written out

- 1 $Y \sim \mathcal{N}(\mu, \Sigma)$ — defer how μ, Σ are set.
- 2 Pick a kernel $k(\cdot, \cdot)$ with $\text{cov}(Y_i, Y_j) = \sigma_y k(X_i, X_j)$. E.g.
 $k(X_i, X_j) = \exp(-\|X_i - X_j\|^2/b)$.
- 3 Kernel matrix $\mathbf{K}$ stacks the pairwise evaluations: $\mathbf{K}_{i,j} = k(X_i, X_j)$.
- 4 Then

$$Y | X \sim \mathcal{N}(\mu, \sigma_y \mathbf{K}).$$

- 5 That model interpolates every point. Allow noise:

$$Y | X \sim \mathcal{N}(\mu, \sigma_y(\mathbf{K} + \sigma^2 I)).$$

- 6 After centering / scaling Y , take $\mu = 0$, $\sigma_y = 1$:

$$Y | X \sim \mathcal{N}(0, \mathbf{K} + \sigma^2 I).$$

- 7 Conditioning on observations, beliefs about test points Y^* are:

$$Y^* \sim \mathcal{N}\left(\underbrace{\mathbf{K}_* (\mathbf{K} + \sigma^2 I)^{-1} Y}_{\text{coefs.}}, \underbrace{\mathbf{K}_{*,*} + \sigma^2 I}_{\text{prior var.}} - \underbrace{\mathbf{K}_*^\top (\mathbf{K} + \sigma^2 I)^{-1} \mathbf{K}_*}_{\text{reduction from data}}\right)$$

Back to RD with a GP

Estimate to the edge of the data, or just past:

- $\widehat{Y(0)}_{X=c}$: train $\mathbb{E}[Y(0) | X]$ on $X < c$, predict at $X = c$.
- $\widehat{Y(1)}_{X=c}$: train $\mathbb{E}[Y(1) | X]$ on $X > c$, predict at $X = c$.

$$\hat{\tau} = \widehat{Y(1)}_{X=c} - \widehat{Y(0)}_{X=c}$$

$$\widehat{\mathbb{V}}(\hat{\tau}) = \widehat{\mathbb{V}}(\widehat{Y(1)}_{X=c}) + \widehat{\mathbb{V}}(\widehat{Y(0)}_{X=c})$$

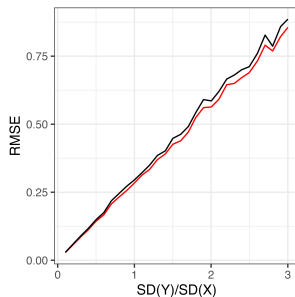
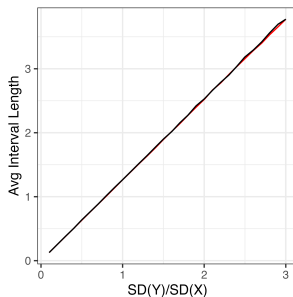
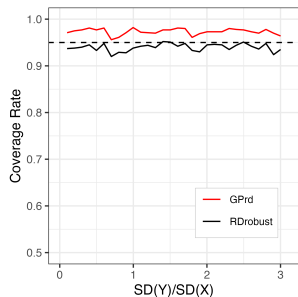
Implemented in `gpss::gp_rdd`; uncertainty grows we approach c .

In practice

- Large samples near the cutoff: `rdrobust` and GP agree.
- Small samples: GP gives better estimates, better coverage, fewer extreme results.
- Very small samples: `rdrobust` sometimes chooses very wide bandwidth, generates extreme estimates, or NA.

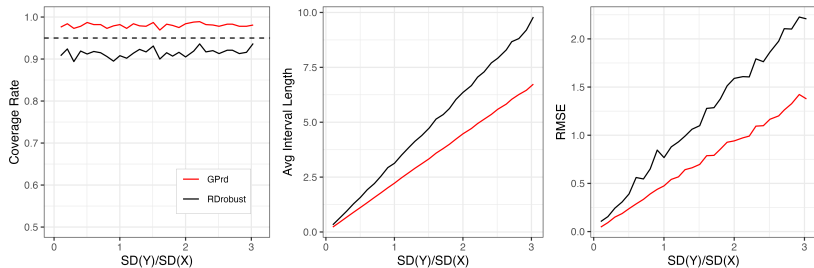
Random Z and Y , large sample

$\tau \in \{0, 3\}$, $N = 500$, rdrobust at defaults vs gp_rdd at defaults.



Random Z and Y , small sample

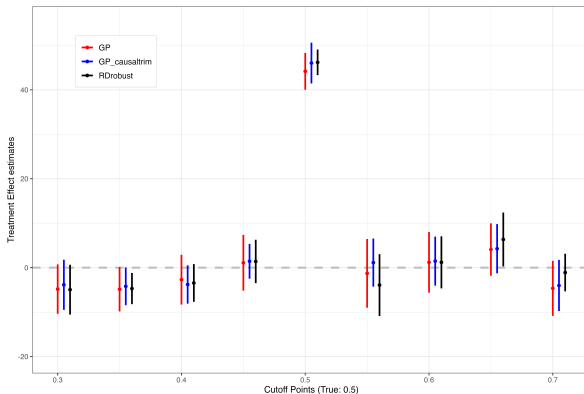
Same DGP, $N = 100$.



Improved RMSE is the big story — GP regularizes on each side, where polynomials can go a little crazy.

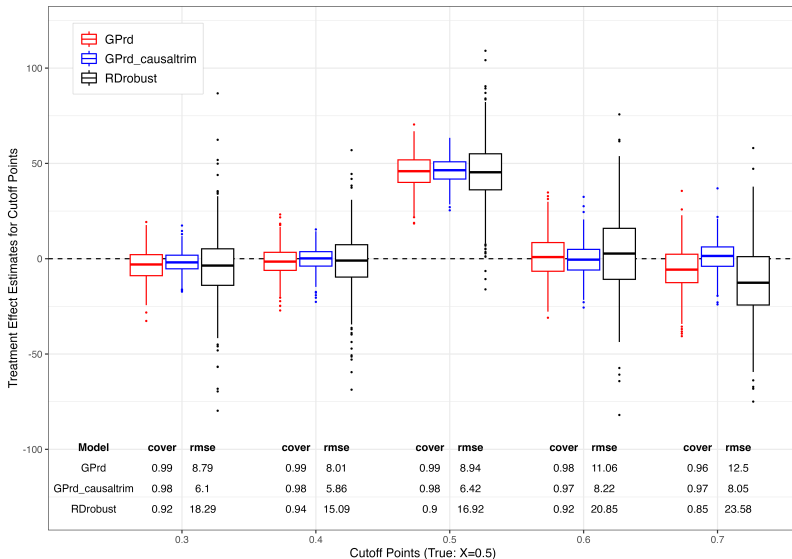
Lee close elections: full sample ($N = 13,577$)

Placebo cutoffs across Lee's close-election data. With this much data, methods barely diverge.

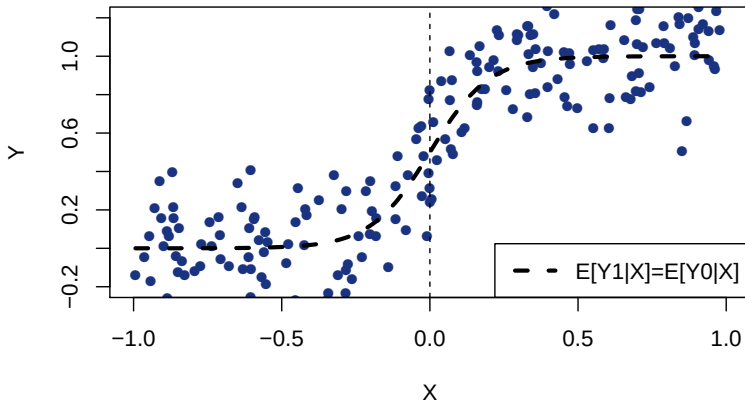


Lee close elections: small subsamples ($N = 200$)

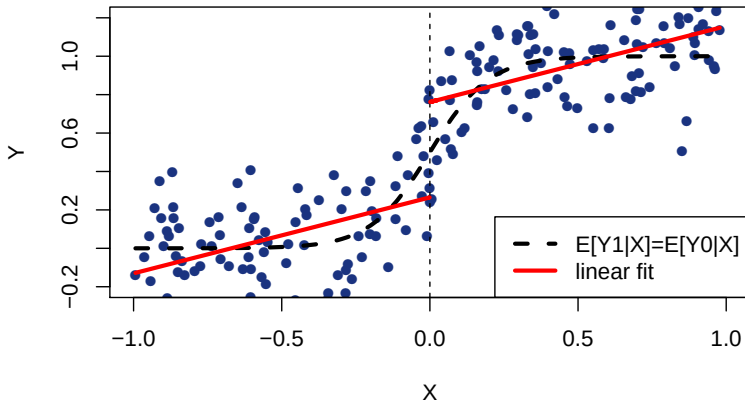
Repeat sampling with $N = 200$ per segment. Now method choice matters.



The risk of misspecification

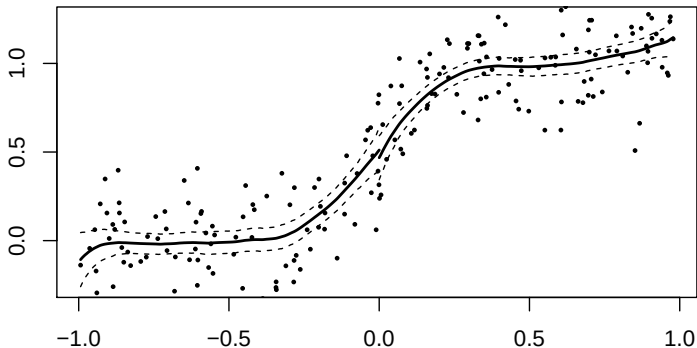


The risk of misspecification



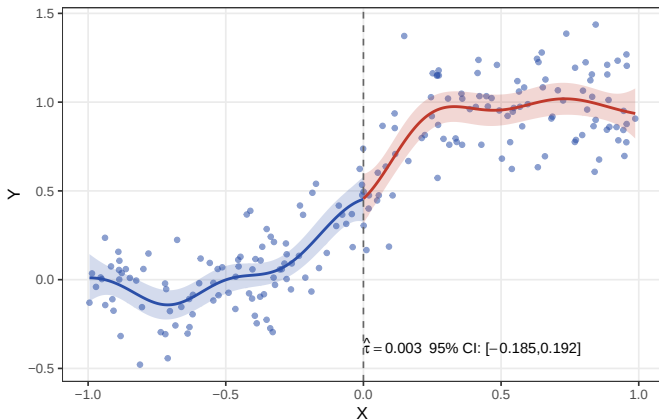
Same example with Imbens–Kalyanaraman

Using IK bandwidth in the RDD package:



Same example, GP estimator

Should be no problem at all with a suitably flexible estimator.

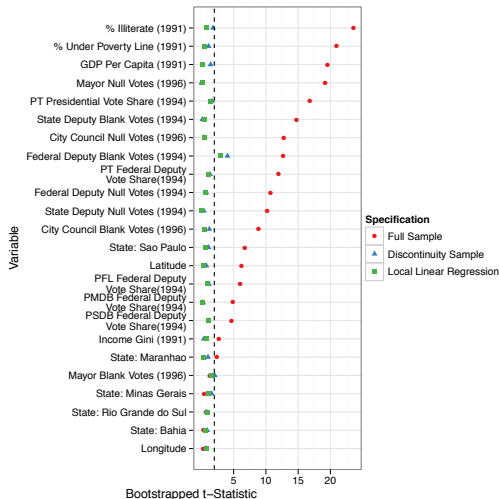


```
gpss::gp_rdd(x, y, cut=0).
```

Falsification checks: the menu

- 1 **Specification sensitivity:** are results stable across estimators / bandwidths? If not, explain.
- 2 **Balance:** do covariates Z jump at the threshold?
- 3 **Placebo outcomes:** does an outcome that *shouldn't* respond to treatment also jump?
- 4 **Placebo thresholds:** do you find jumps at $c^* \neq c$?
- 5 **Sorting:** do units cluster on one side of the cutoff (manipulation)?

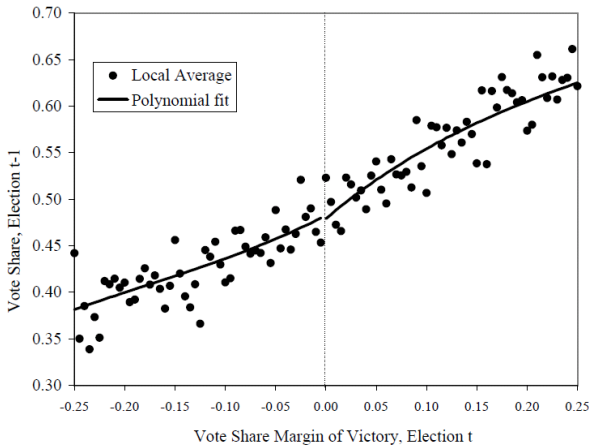
Balance test: back to Hidalgo



Run the same RD on a covariate Z as the “outcome” — hope for no jump.
 Non-falsification $\neq$ verification.

Placebo outcome (Lee 2008)

Most convincing falsification: a covariate (or pre-treatment outcome) plausibly related to potential outcomes but *not* affected by treatment.



Placebo thresholds and sorting

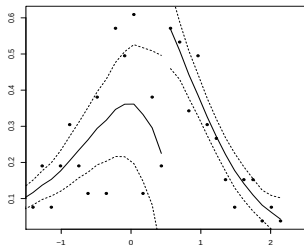
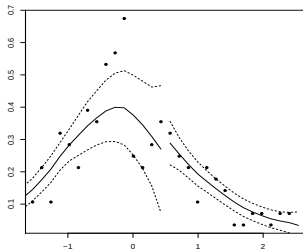
Placebo thresholds: pretend the cutoff is $c^* \neq c$. Run the RD.

- Detecting a jump at c^* demands an explanation.
- Other treatments at c^* ? Non-smoothness in the data?
- Usually fit separate models on each side without crossing the real c , to avoid contamination.

Sorting: can units *self-select* across the cutoff?

- Fine control over X (gaming a test score, withholding income).
- Administrators choosing thresholds strategically.
- Density test (McCrary 2008): bulge in X on one side of c ?
- Not always fatal: sorting would have to be precise *and* widespread to substantially bias the result.

Sorting around the threshold



Look at a histogram.

Formal test: McCrary (2008).

The RD killer: bundled treatments

- Sometimes *multiple things* change at the same threshold.
- Why it's bad:
 - Think of this as a U that jumps at c and influences Y_0 and Y_1 .
 - $\mathbb{E}[Y_d | X]$ jumps at c regardless of our treatment D — continuity is violated.
 - RD mis-attributes the effect of the other treatment to ours.
- RD does not tell you *which* treatment effect you are seeing.
- **Always ask:** *what else changes at c ?* This is where investigatory work and domain expertise pay off.

Fuzzy RD

- The threshold doesn't *determine* treatment, but **shifts the probability** of treatment.
- E.g. scoring above a threshold raises the chance of getting a scholarship from 20% to 70%, but doesn't guarantee it.
- This should remind you of IV:
 - Crossing c is the instrument $Z = \mathbb{1}\{X > c\}$.
 - It changes treatment probability and (we hope) only affects Y through treatment.

Fuzzy RD: identification

Identification Assumption (Local-randomization for fuzzy RD)

For sufficiently small ϵ with $c - \epsilon < X_i < c + \epsilon$, the usual IV assumptions hold:

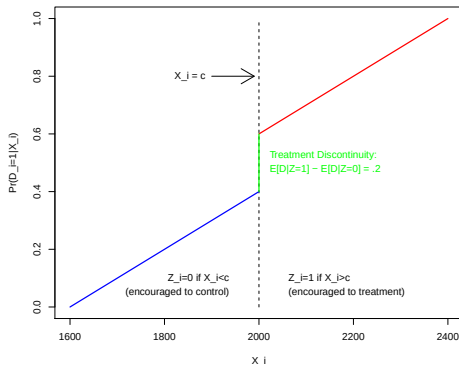
- 1 Ignorability (and exclusion): $\{Y_{1i}, Y_{0i}, D_{1i}, D_{0i}\} \perp\!\!\!\perp Z_i$
- 2 Relevance: $\mathbb{E}[D_{1i}] \neq \mathbb{E}[D_{0i}]$
- 3 Monotonicity: $D_{1i} \geq D_{0i} \forall i$

Identification Result

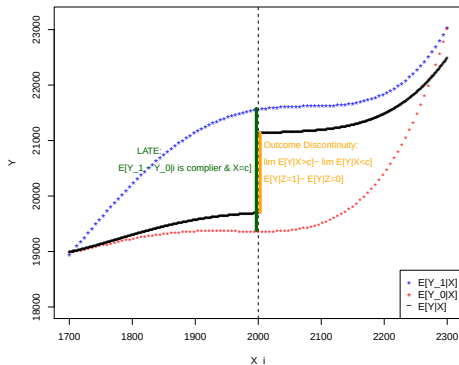
$$\begin{aligned}
 \alpha_{FRD} &= \mathbb{E}[Y_{1i} - Y_{0i} \mid X_i = c, i \text{ is a complier}] \\
 &= \frac{\lim_{x \downarrow c} \mathbb{E}[Y_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[Y_i \mid X_i = x]}{\lim_{x \downarrow c} \mathbb{E}[D_i \mid X_i = x] - \lim_{x \uparrow c} \mathbb{E}[D_i \mid X_i = x]} \\
 &= \frac{\text{outcome discontinuity}}{\text{treatment discontinuity}}
 \end{aligned}$$

Fuzzy RD: jumps in $\mathbb{E}[D | X]$ and $\mathbb{E}[Y | X]$

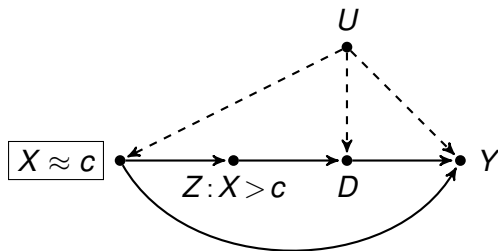
$$\mathbb{E}[D | X]$$



$$\mathbb{E}[Y | X]$$



A DAG view of fuzzy RD



- $Z = \mathbb{1}\{X > c\}$ explicitly marked — it will be the (near-)instrument.
- “Loosely conditioning” on $X \approx c$ closes most of $X \leftarrow U \rightarrow Y$.
- ...making Z close to a good instrument, with LATE interpretation.

Doing fuzzy RD

The recipe (just like IV at the cutoff):

- 1 **Reduced form**: estimate the jump in Y at the cutoff (the sharp RD on Y).
- 2 **First stage**: estimate the jump in D at the cutoff (the sharp RD on D).
- 3 Divide.

Toy numbers: crossing c raises Y by \$5,000 *and* raises scholarship probability from 20% to 70%.

Then $\hat{\alpha}_{\text{LATE}} = 5000/0.5 = \$10,000$.

Interpretation & credibility:

- The estimand is **doubly local**: at the cutoff, and only for compliers.
- Same threat as before: no *other* treatments switching at c .

`rdrobust(y, x, fuzzy = d, c = 0)` does this for you.

Internal vs. external validity

Internal validity:

- Continuity of POs is often defensible $\rightarrow$ RD is one of the happiest observational designs.
- Main threat: **bundled treatments** at the same cutoff.

External validity:

- You learn the average effect for units with X_i near c .
- Fuzzy RD: also restricted to compliers near c .
- No information about effects elsewhere on the X distribution. Don't pretend.
- Often used *in combination* with broader-target designs (SOO) that have less credible assumptions.

Takeaways

- RD turns an administrative rule into something close to an experiment — locally.
- Identification: **continuity**. Plain English: no other (impactful) treatments at the same cutoff.
- Estimation: use an automated tool (`rdrobust`, `gp_rdd`).
- Falsification battery: balance, placebo outcomes, placebo thresholds, sometimes sorting.
- Most important: **investigatory work to look for other treatments at the same cutoff.**