Regression Discontinuity Designs

Md Moshi Ul Alam

November 14, 2026

What these slides cover

Main slides

1. The idea: treatment assigned by a cutoff
2. Sharp RD: a local RCT at the cutoff
3. Estimation: local linear regression and bandwidth
4. Fuzzy RD: IV at the cutoff (LATE)
5. Is the RD credible? Validity checks
6. Wrap-up: RD and the course toolbox

Backup slides (more detail)

- What determines the "optimal" bandwidth?
- Weights (kernels): who counts how much?
- Fuzzy RD: 2SLS = ratio of two sharp RDs
- Fuzzy RD identifies LATE at the cutoff: derivation

Click a topic to jump to it. In the main slides, ▸ buttons lead to the backup slides, and ◂ Back brings you back.

Where we are

- The fundamental problem: we never see both $Y_i(1)$ and $Y_i(0)$ for the same unit

  $\Rightarrow$ a naive comparison of treated and untreated $=$ ATT $+$ **selection bias**

- Each design we studied is a way to get rid of selection bias:
  - **RCT**: randomization $\Rightarrow$ no selection bias $\Rightarrow$ ATT
  - **D-i-D**: selection bias is allowed but must be stable over time (parallel trends) $\Rightarrow$ ATT
  - **IV**: use only the variation in D_i driven by an exogenous Z_i $\Rightarrow$ LATE (compliers)

- **Today: Regression Discontinuity (RD).** Treatment is assigned by a **rule**: you are treated if a score crosses a **cutoff**

- We will see that RD is a **local RCT** (sharp RD) or a **local IV** (fuzzy RD)

Motivating question

- Does attending an **elite college** raise **earnings**?

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad D_i = 1 \text{ if } i \text{ attends the elite college}$$

Motivating question

- Does attending an **elite college** raise **earnings**?

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad D_i = 1 \text{ if } i \text{ attends the elite college}$$

- Why is OLS biased here? Who attends the elite college?

Motivating question

- Does attending an **elite college** raise **earnings**?

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad D_i = 1 \text{ if } i \text{ attends the elite college}$$

- Why is OLS biased here? Who attends the elite college?
  - Students with higher SAT scores, more motivation, richer parents, ...
  - All of these also raise earnings $\Rightarrow cov(D_i, u_i) \neq 0$: selection bias

Motivating question

- Does attending an **elite college** raise **earnings**?

$$Y_i = \beta_0 + \beta_1 D_i + u_i, \qquad D_i = 1 \text{ if } i \text{ attends the elite college}$$

- Why is OLS biased here? Who attends the elite college?
  - Students with higher SAT scores, more motivation, richer parents, ...
  - All of these also raise earnings $\Rightarrow cov(D_i, u_i) \neq 0$: selection bias
- But suppose the college has a **rule**: admit everyone with an SAT score ≥ 1200
- Think of two students: one scored 1195, the other 1205. How different are they?

Regression Discontinuity: the idea

Setting: treatment is assigned by whether a score crosses a known cutoff.
Examples:

- University admission if SAT score is above a mark (Hoekstra 2009)
- Class size limits: a class is split when enrollment passes 40 (Angrist & Lavy 1999)
- Financial aid eligibility at an income threshold
- Health insurance (Medicare) eligibility at age 65 (Card, Dobkin & Maestas 2008)
- Legal drinking age 21 (Carpenter & Dobkin 2009)
- A party wins the seat if its vote share is above 50% (Lee 2008)

Regression Discontinuity: the idea

Setting: treatment is assigned by whether a score crosses a known cutoff.
Examples:

- University admission if SAT score is above a mark (Hoekstra 2009)
- Class size limits: a class is split when enrollment passes 40 (Angrist & Lavy 1999)
- Financial aid eligibility at an income threshold
- Health insurance (Medicare) eligibility at age 65 (Card, Dobkin & Maestas 2008)
- Legal drinking age 21 (Carpenter & Dobkin 2009)
- A party wins the seat if its vote share is above 50% (Lee 2008)

Units **just above** and **just below** the cutoff are nearly identical: whether a student scores 1195 or 1205 is largely luck on test day. So close to the cutoff, treatment is **as good as randomly assigned**.

Ingredients and notation

- Y_i: outcome (earnings)
- R_i: the **running variable** (or forcing variable) that determines treatment (SAT score)
- c: the known **cutoff** (1200)
- $Z_i = \mathbb{1}(R_i \geq c)$: indicator for being above the cutoff
- $D_i \in \{0, 1\}$: treatment actually received (attending the elite college)

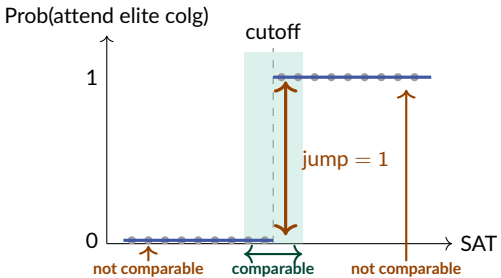
Ingredients and notation

- Y_i: outcome (earnings)
- R_i: the **running variable** (or forcing variable) that determines treatment (SAT score)
- c: the known **cutoff** (1200)
- $Z_i = \mathbb{1}(R_i \geq c)$: indicator for being above the cutoff
- $D_i \in \{0, 1\}$: treatment actually received (attending the elite college)
- Two designs, depending on how $Pr(D_i = 1 \mid R_i)$ changes at c:
  - **Sharp RD**: $D_i = Z_i$. The probability of treatment jumps from 0 to 1 at c
  - **Fuzzy RD**: $D_i \neq Z_i$ for some units. The probability of treatment jumps at c, but by **less than 1**

Sharp RD: SAT cutoff and elite college (perfect compliance)

Example: students who score **above an SAT cutoff** attend an elite college. Outcome: **earnings**. In a **sharp** RD, take-up is perfect: **everyone** above attends, **no one** below does.

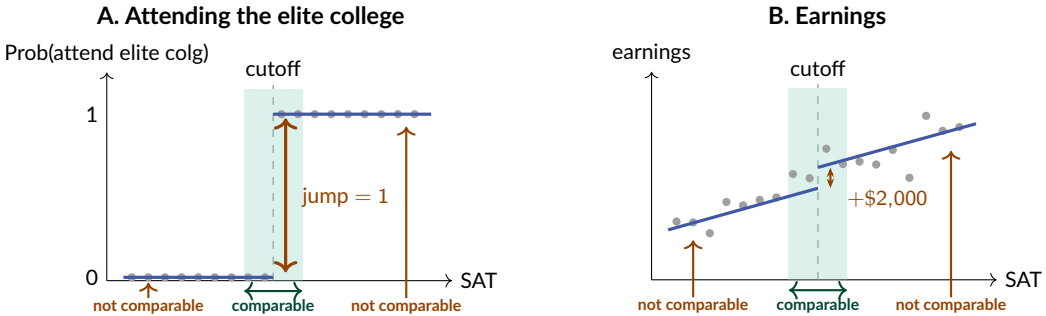
A. Attending the elite college



Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

Sharp RD: SAT cutoff and elite college (perfect compliance)

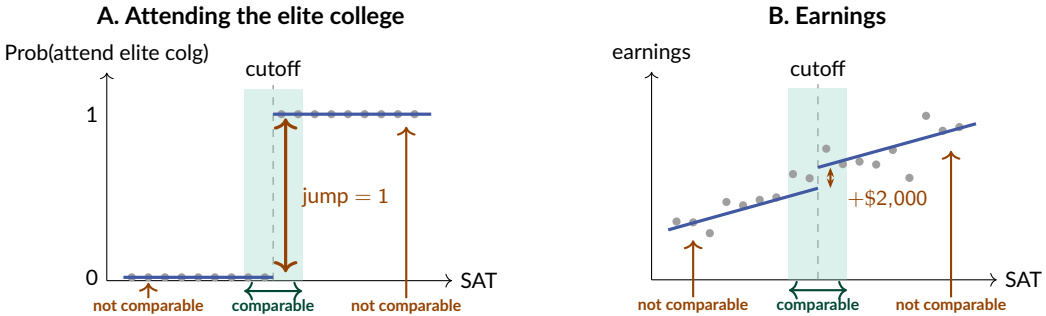
Example: students who score **above an SAT cutoff** attend an elite college. Outcome: **earnings**. In a **sharp** RD, take-up is perfect: **everyone** above attends, **no one** below does.



Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

Sharp RD: SAT cutoff and elite college (perfect compliance)

Example: students who score **above an SAT cutoff** attend an elite college. Outcome: **earnings**. In a **sharp** RD, take-up is perfect: **everyone** above attends, **no one** below does.



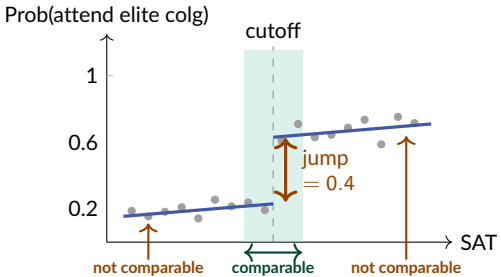
Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

Students just above and just below the cutoff are **nearly identical**; they only differ in whether they attend the elite college or not. So, for students around the cutoff, the **jump in earnings at the cutoff** is the causal effect of attending: **$2,000**.

Fuzzy RD: same cutoff, imperfect compliance

Same example, more realistic: crossing the cutoff makes attending **more likely**, but some students above it do not attend, and some below it still get in. Take-up jumps by **less than 1**.

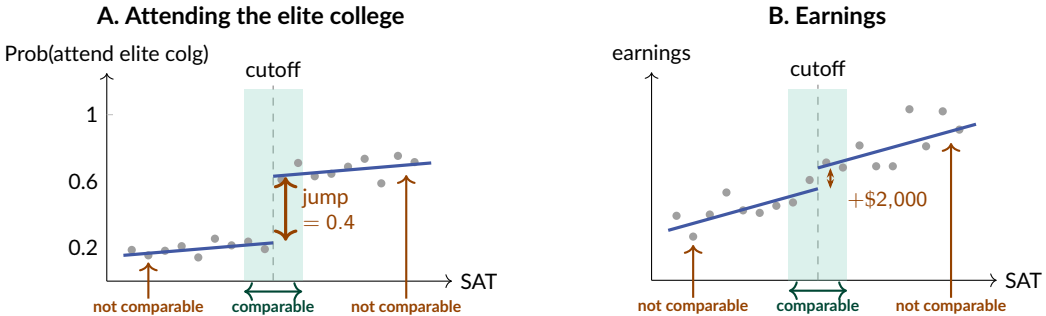
A. Attending the elite college



Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

Fuzzy RD: same cutoff, imperfect compliance

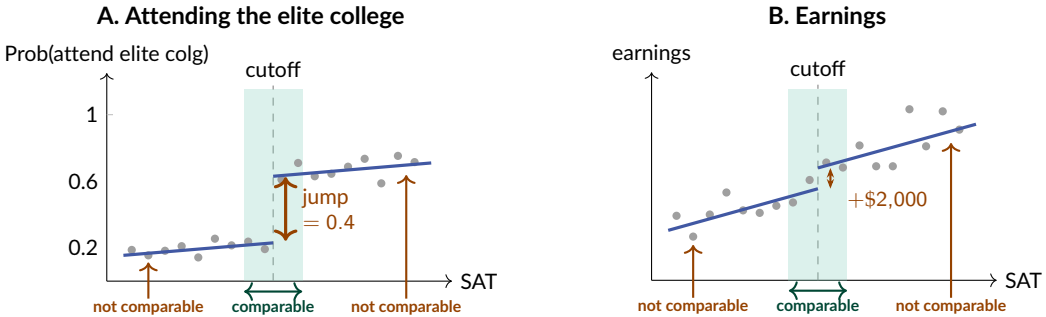
Same example, more realistic: crossing the cutoff makes attending **more likely**, but some students above it do not attend, and some below it still get in. Take-up jumps by **less than 1**.



Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

Fuzzy RD: same cutoff, imperfect compliance

Same example, more realistic: crossing the cutoff makes attending **more likely**, but some students above it do not attend, and some below it still get in. Take-up jumps by **less than 1**.

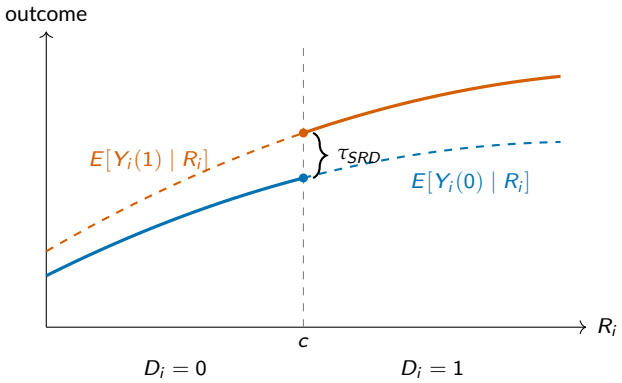


Hypothetical data. Dots in A: share attending in each SAT range; dots in B: individual students. Lines: lines of best fit on each side of the cutoff.

The earnings jump is driven entirely by the extra 40% of students who attend because they crossed the cutoff. So the effect of attending is effect of attending $= \dfrac{\text{jump in earnings}}{\text{jump in attendance}} = \dfrac{\$2{,}000}{0.4} = \mathbf{\$5{,}000}$. Sound familiar? (Wald estimator)

Sharp RD:
a local RCT at the cutoff

Sharp RD in potential outcomes

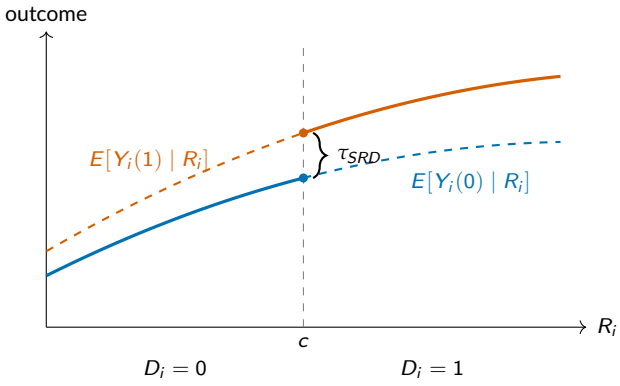


Solid: what we observe. Dashed: counterfactual we never observe.

- Left of c: everyone untreated, we see $Y_i(0)$ only
- Right of c: everyone treated, we see $Y_i(1)$ only
- Both curves can slope: units with higher R_i may earn more even without treatment

  $\Rightarrow$ far from c, treated and untreated are **not** comparable

Sharp RD in potential outcomes



Solid: what we observe. Dashed: counterfactual we never observe.

- Left of c: everyone untreated, we see $Y_i(0)$ only
- Right of c: everyone treated, we see $Y_i(1)$ only
- Both curves can slope: units with higher R_i may earn more even without treatment

  $\Rightarrow$ far from c, treated and untreated are **not** comparable
- At c, the observed outcome **jumps** from the blue curve to the red curve
- The size of the jump is the treatment effect at the cutoff:

$$\tau_{SRD} = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

The key identifying assumption: continuity

- Notation: $E[Y_i \mid R_i = c^+]$ is the average outcome for units with R_i **just above** c (the limit as R_i approaches c from above). Similarly c^- for **just below**.
- **Continuity assumption:** $E[Y_i(0) \mid R_i = r]$ and $E[Y_i(1) \mid R_i = r]$ are **smooth** (no jumps) at $r = c$
  - In words: **nothing else** changes discontinuously at the cutoff, **except** treatment
  - The potential outcome curves on the last slide have no breaks; only the observed line does

The key identifying assumption: continuity

- Notation: $E[Y_i \mid R_i = c^+]$ is the average outcome for units with R_i **just above** c (the limit as R_i approaches c from above). Similarly c^- for **just below**.
- **Continuity assumption:** $E[Y_i(0) \mid R_i = r]$ and $E[Y_i(1) \mid R_i = r]$ are **smooth** (no jumps) at $r = c$
  - In words: **nothing else** changes discontinuously at the cutoff, **except** treatment
  - The potential outcome curves on the last slide have no breaks; only the observed line does
- What could break it?
  - **Sorting / manipulation**: if students could retake the SAT until they are just above 1200, those just above differ (more motivated) from those just below
  - **Other policies at the same cutoff**: if a scholarship also kicks in at 1200, the jump mixes two treatments

The key identifying assumption: continuity

- Notation: $E[Y_i \mid R_i = c^+]$ is the average outcome for units with R_i **just above** c (the limit as R_i approaches c from above). Similarly c^- for **just below**.
- **Continuity assumption:** $E[Y_i(0) \mid R_i = r]$ and $E[Y_i(1) \mid R_i = r]$ are **smooth** (no jumps) at $r = c$
  - In words: **nothing else** changes discontinuously at the cutoff, **except** treatment
  - The potential outcome curves on the last slide have no breaks; only the observed line does
- What could break it?
  - **Sorting / manipulation**: if students could retake the SAT until they are just above 1200, those just above differ (more motivated) from those just below
  - **Other policies at the same cutoff**: if a scholarship also kicks in at 1200, the jump mixes two treatments
- Note: we do **not** need D_i to be independent of $(Y_i(1), Y_i(0))$ everywhere. Selection on R_i is allowed! We only need comparability **at** c.

Sharp RD is a local RCT: same decomposition as before

Recall from the RCT lecture: observed difference = ATT + selection bias. Now compare units just above and just below c:

$$\underbrace{E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-]}_{\text{observed jump at the cutoff}} = E[Y_i(1) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^-]$$

$$= \underbrace{E[Y_i(1) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^+]}_{\text{ATT for the treated just above } c}$$

$$+ \underbrace{E[Y_i(0) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^-]}_{\text{selection bias at the cutoff}}$$

Sharp RD is a local RCT: same decomposition as before

Recall from the RCT lecture: observed difference = ATT + selection bias. Now compare units just above and just below c:

$$\underbrace{E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-]}_{\text{observed jump at the cutoff}} = E[Y_i(1) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^-]$$

$$= \underbrace{E[Y_i(1) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^+]}_{\text{ATT for the treated just above } c}$$

$$+ \underbrace{E[Y_i(0) \mid R_i = c^+] - E[Y_i(0) \mid R_i = c^-]}_{\text{selection bias at the cutoff}}$$

- **Continuity** of $E[Y_i(0) \mid R_i]$ at $c \Rightarrow$ selection bias at the cutoff $= 0$
- Continuity of $E[Y_i(1) \mid R_i] \Rightarrow$ "just above" = "at c", so

$$\tau_{SRD} = E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-] = E[Y_i(1) - Y_i(0) \mid R_i = c]$$

- In an RCT, randomization kills selection bias. In a sharp RD, **being close to the cutoff** does.

RCT vs. sharp RD

	RCT	**Sharp RD**
Who is treated?	Decided by a coin flip	Decided by a rule: $D_i = \mathbb{1}(R_i \geq c)$
Why no selection bias?	$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ for everyone	Units at c^+ and c^- are comparable (continuity)
What we estimate	ATT ($=$ ATE) in the whole sample	ATT for units **at the cutoff**
Estimator	Difference in means	Jump in $E[Y_i \mid R_i]$ at c

RCT vs. sharp RD

	RCT	**Sharp RD**
Who is treated?	Decided by a coin flip	Decided by a rule: $D_i = \mathbb{1}(R_i \geq c)$
Why no selection bias?	$D_i \perp\!\!\!\perp (Y_i(1), Y_i(0))$ for everyone	Units at c^+ and c^- are comparable (continuity)
What we estimate	ATT ($=$ ATE) in the whole sample	ATT for units **at the cutoff**
Estimator	Difference in means	Jump in $E[Y_i \mid R_i]$ at c

- The catch: in an RCT we have treated and control units with the same characteristics. In an RD, at any value of R_i, units are **either all treated or all untreated**
- There are (almost) no units exactly at c. We must **extrapolate** from each side to c

  $\Rightarrow$ How we do this extrapolation is the estimation problem

Estimating a sharp RD: local linear regression

Step 0: always start with a picture

- Divide R_i into small bins (e.g., 10 SAT points wide)
- Plot the average Y_i in each bin against the bin's midpoint
- Draw a vertical line at c. Is there a visible jump?
- This is called a **binned scatter plot**. Every RD paper has one (you will see Hoekstra's later)
- In R: `rdplot(y, x, c = cutoff)` from the `rdrobust` package

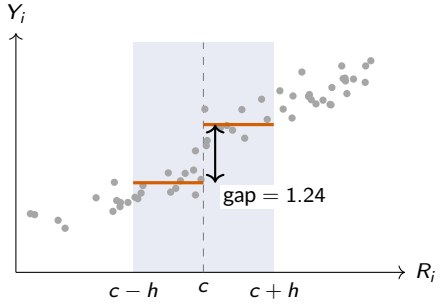
Step 0: always start with a picture

- Divide R_i into small bins (e.g., 10 SAT points wide)
- Plot the average Y_i in each bin against the bin's midpoint
- Draw a vertical line at c. Is there a visible jump?
- This is called a **binned scatter plot**. Every RD paper has one (you will see Hoekstra's later)
- In R: `rdplot(y, x, c = cutoff)` from the `rdrobust` package
- The picture tells you if there is an effect. But to get a number and a standard error, we need to estimate the two limits:

$$E[Y_i \mid R_i = c^+] \quad \text{and} \quad E[Y_i \mid R_i = c^-]$$

Two ways to estimate the jump: a simulated example (true jump $= 0.8$)

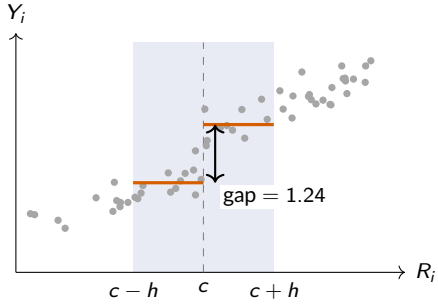
A. Difference in means in a window



Compare average Y_i just below vs. just above

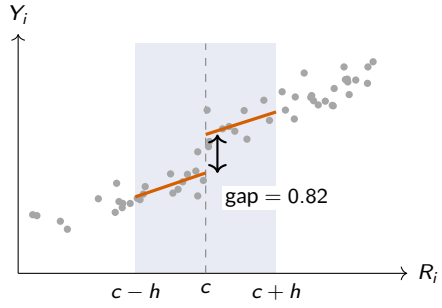
Two ways to estimate the jump: a simulated example (true jump $= 0.8$)

A. Difference in means in a window



Compare average Y_i just below vs. just above

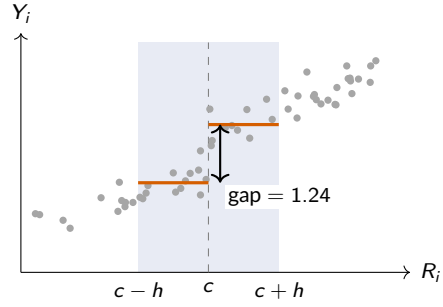
B. Local linear regression



Fit a separate line on each side, compare at c

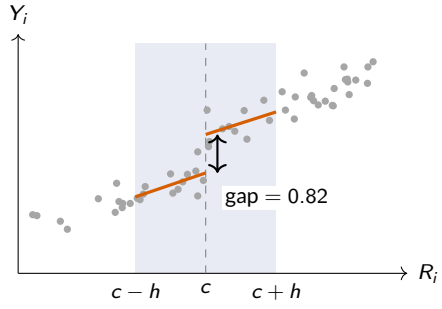
Two ways to estimate the jump: a simulated example (true jump $= 0.8$)

A. Difference in means in a window



Compare average Y_i just below vs. just above

B. Local linear regression



Fit a separate line on each side, compare at c

- The shaded **window** $[c - h,\ c + h]$ contains the data we use. h is the **bandwidth**
- A: the outcome rises with R_i, so the mean above c is pulled up and the mean below c is pulled down $\Rightarrow$ **biased**
- B: a line on each side follows the slope and is evaluated **at** c $\Rightarrow$ much less bias

Local linear regression

- Center the running variable at the cutoff: $\tilde{R}_i = R_i - c$ (so $\tilde{R}_i = 0$ at the cutoff)
- Keep only observations inside the window: $-h \leq \tilde{R}_i \leq h$
- Estimate by OLS:

$$Y_i = \beta_0 + \tau\, D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i, \qquad |\tilde{R}_i| \leq h$$

- $\hat{\tau}$ is the estimated jump at the cutoff: our RD estimate

Local linear regression

- Center the running variable at the cutoff: $\tilde{R}_i = R_i - c$ (so $\tilde{R}_i = 0$ at the cutoff)
- Keep only observations inside the window: $-h \leq \tilde{R}_i \leq h$
- Estimate by OLS:

$$Y_i = \beta_0 + \tau\, D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i, \qquad |\tilde{R}_i| \leq h$$

- $\hat{\tau}$ is the estimated jump at the cutoff: our RD estimate
- This is just a regression with a dummy and an interaction. You know how to do this! Two decisions make it an RD:
  1. **local**: only use data close to c (how close? the **bandwidth** h)
  2. **linear**: a straight line on each side (with possibly different slopes)

Reading the coefficients

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i$$

Below the cutoff $(D_i = 0)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 0] =$

Above the cutoff $(D_i = 1)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 1] =$

Reading the coefficients

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i$$

Below the cutoff $(D_i = 0)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 0] = \beta_0 + \beta_1 \tilde{R}_i$

Above the cutoff $(D_i = 1)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 1] =$

Reading the coefficients

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i$$

Below the cutoff $(D_i = 0)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 0] = \beta_0 + \beta_1 \tilde{R}_i$

Above the cutoff $(D_i = 1)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 1] = (\beta_0 + \tau) + (\beta_1 + \beta_2)\tilde{R}_i$

Reading the coefficients

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i$$

Below the cutoff $(D_i = 0)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 0] = \beta_0 + \beta_1 \tilde{R}_i$

Above the cutoff $(D_i = 1)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 1] = (\beta_0 + \tau) + (\beta_1 + \beta_2)\tilde{R}_i$

- Two lines: intercepts β_0 and $\beta_0 + \tau$; slopes β_1 and $\beta_1 + \beta_2$
- At the cutoff, $\tilde{R}_i = 0$: the gap between the two lines is exactly τ

Reading the coefficients

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \left(D_i \times \tilde{R}_i\right) + u_i$$

Below the cutoff $(D_i = 0)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 0] = \beta_0 + \beta_1 \tilde{R}_i$

Above the cutoff $(D_i = 1)$: $\quad E[Y_i \mid \tilde{R}_i, D_i = 1] = (\beta_0 + \tau) + (\beta_1 + \beta_2)\tilde{R}_i$

- Two lines: intercepts β_0 and $\beta_0 + \tau$; slopes β_1 and $\beta_1 + \beta_2$
- At the cutoff, $\tilde{R}_i = 0$: the gap between the two lines is exactly τ
- **Why center?** Without centering, τ would be the gap between the two lines at $R_i = 0$ (an SAT score of 0!), not at c
- **Why the interaction?** The slope may differ on the two sides. Forcing one common slope can push part of the slope difference into $\hat{\tau}$

Why not use all the data with a polynomial?

- An older approach: use **all** observations and a flexible curve in $\tilde{R}_i$

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \tilde{R}_i^2 + \beta_3 \tilde{R}_i^3 + \beta_4 \tilde{R}_i^4 + \cdots + u_i$$

- You still see it in papers (Hoekstra's figures use polynomials)

Why not use all the data with a polynomial?

- An older approach: use **all** observations and a flexible curve in $\tilde{R}_i$

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \tilde{R}_i^2 + \beta_3 \tilde{R}_i^3 + \beta_4 \tilde{R}_i^4 + \cdots + u_i$$

- You still see it in papers (Hoekstra's figures use polynomials)

- Problems:
  - Observations **far from** c (e.g., SAT 600 or 1600) influence the estimate **at** c. But they are not comparable to students at the cutoff
  - High-order polynomials wiggle a lot near the edges of the data, which is exactly where the cutoff is $\Rightarrow$ spurious jumps

Why not use all the data with a polynomial?

- An older approach: use **all** observations and a flexible curve in $\tilde{R}_i$

$$Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 \tilde{R}_i^2 + \beta_3 \tilde{R}_i^3 + \beta_4 \tilde{R}_i^4 + \cdots + u_i$$

- You still see it in papers (Hoekstra's figures use polynomials)

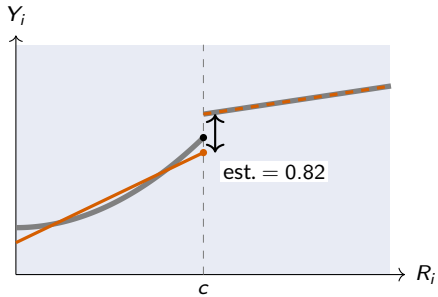
- Problems:
  - Observations **far from** c (e.g., SAT 600 or 1600) influence the estimate **at** c. But they are not comparable to students at the cutoff
  - High-order polynomials wiggle a lot near the edges of the data, which is exactly where the cutoff is $\Rightarrow$ spurious jumps

- Current practice (Gelman & Imbens 2019): **local linear** (sometimes local quadratic) regression close to the cutoff

- A straight line is a good approximation of any smooth curve, **if you zoom in enough**. How much to zoom in is the bandwidth choice

Choosing the bandwidth h: what can go wrong?

Large h: use all the data

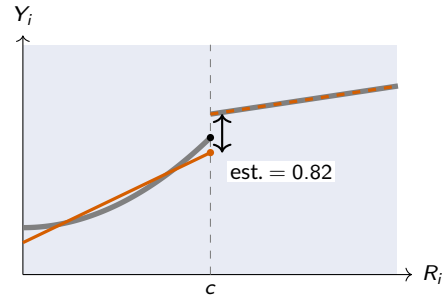


The line misses the curvature near $c \Rightarrow$ **bias**

Gray: true $E[Y_i \mid R_i]$ (true jump $= 0.5$). Red: fitted lines inside the shaded window.

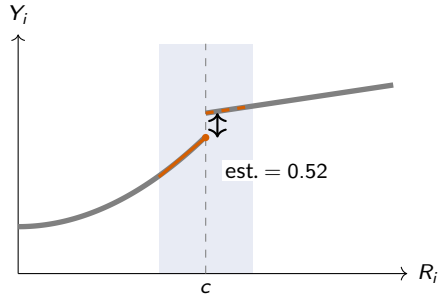
Choosing the bandwidth h: what can go wrong?

Large h: use all the data



The line misses the curvature near $c \Rightarrow$ **bias**

Small h: zoom in

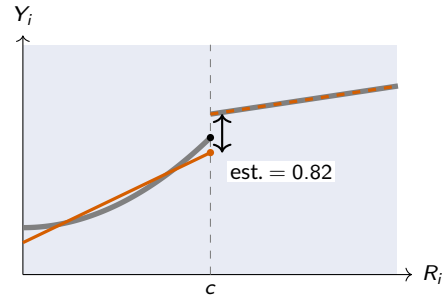


Near c the curve is almost a line $\Rightarrow$ little bias

Gray: true $E[Y_i \mid R_i]$ (true jump $= 0.5$). Red: fitted lines inside the shaded window.

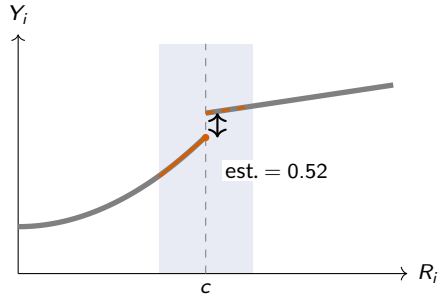
Choosing the bandwidth h: what can go wrong?

Large h: use all the data



The line misses the curvature near $c \Rightarrow$ **bias**

Small h: zoom in



Near c the curve is almost a line $\Rightarrow$ little bias

Gray: true $E[Y_i \mid R_i]$ (true jump $= 0.5$). Red: fitted lines inside the shaded window.

- The catch: with a small h there are **few observations** in the window $\Rightarrow$ the estimated lines are **noisy** (large standard error)

The bandwidth is a bias–variance trade-off

	Small h	**Large** h
Units in the window are comparable to those at c?	Yes	Less so
Straight line a good approximation?	Yes	Maybe not
$\Rightarrow$ **Bias**	Low	High
Number of observations used	Few	Many
$\Rightarrow$ **Variance** (standard error)	High	Low

The bandwidth is a bias–variance trade-off

	Small h	**Large** h
Units in the window are comparable to those at c?	Yes	Less so
Straight line a good approximation?	Yes	Maybe not
$\Rightarrow$ **Bias**	Low	High
Number of observations used	Few	Many
$\Rightarrow$ **Variance** (standard error)	High	Low

- In practice, h is chosen by the data: the "optimal" bandwidth minimizes the mean squared error $=$ bias^2 $+$ variance (Imbens & Kalyanaraman 2012; Calonico, Cattaneo & Titiunik 2014). `rdrobust` does this for you
- Good practice: report the estimate with the optimal h, then show it is similar with $h/2$ and $2h$
- The optimal h shrinks as the sample grows: more data $\Rightarrow$ we can afford to zoom in
  - What determines the optimal h?

Implementation in R

- Standard errors: use heteroskedasticity-robust SEs, as usual
- `rdrobust` chooses *h*, uses the triangular kernel ▸ Weights, and reports a **robust** confidence interval that also accounts for the remaining bias of the straight-line approximation

```
df$Rt <- df$R - cutoff                  # center the running variable
df$D  <- as.numeric(df$Rt >= 0)         # sharp RD: treated if above
h <- 50                                 # a bandwidth

# 1. By hand, uniform kernel: OLS inside the window
m1 <- lm(Y ~ D * Rt, data = subset(df, abs(Rt) <= h))
# 2. By hand, triangular kernel: weighted least squares
df$w <- pmax(0, 1 - abs(df$Rt) / h)
m2 <- lm(Y ~ D * Rt, data = df, weights = w)
coeftest(m2, vcov = vcovHC(m2, type = "HC1"))   # robust SEs

# 3. The standard package: optimal h, triangular kernel, robust CI
library(rdrobust)
summary(rdrobust(df$Y, df$R, c = cutoff))
rdplot(df$Y, df$R, c = cutoff)                  # binned scatter plot
```

Fuzzy RD:
IV at the cutoff

Hoekstra (2009): does attending the flagship state university pay off?

- Applications to a state flagship university, linked to later earnings (ages 28–33) from UI tax records
- Running variable: SAT score, re-centered at the admission cutoff
- Crossing the cutoff raises enrollment by **0.388**
- Not 0 to 1: some below enroll anyway, many above go elsewhere

  $\Rightarrow$ **fuzzy RD**
- This is the **first stage**

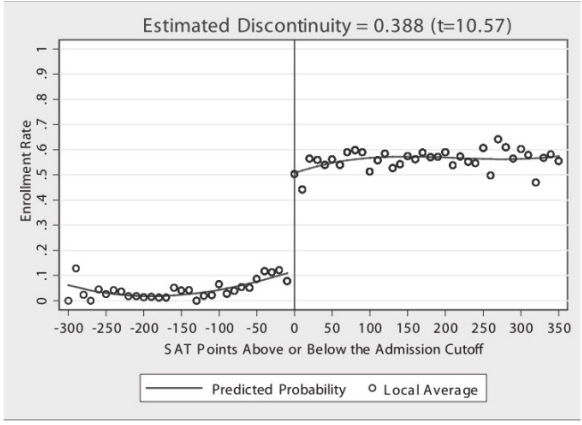


Figure 6.2: Attending the state flagship university as a function of re-centered standardized test scores. Reprinted from Hoekstra (2009).

Hoekstra (2009): earnings

- Outcome: log earnings, after removing year, experience and cohort effects (residuals)
- Earnings jump by **0.095** log points at the cutoff
- This is the **reduced form**: the effect of being **above the cutoff**, not of enrolling

Hoekstra (2009): earnings

- Outcome: log earnings, after removing year, experience and cohort effects (residuals)
- Earnings jump by **0.095** log points at the cutoff
- This is the **reduced form**: the effect of being **above the cutoff**, not of enrolling
- Only 38.8% more students enroll when crossing the cutoff. So the effect of **enrolling** must be larger than 0.095

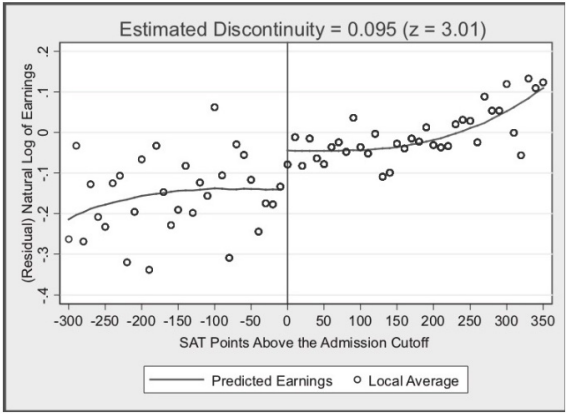


Figure 6.3: Future earnings as a function of re-centered standardized test scores. Reprinted from Hoekstra (2009).

The fuzzy RD estimand: a Wald estimator at the cutoff

- Scale the jump in the outcome by the jump in treatment:

$$\tau_{FRD} = \frac{E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-]}{Pr(D_i = 1 \mid R_i = c^+) - Pr(D_i = 1 \mid R_i = c^-)} = \frac{\text{jump in } Y_i}{\text{jump in } D_i}$$

The fuzzy RD estimand: a Wald estimator at the cutoff

- Scale the jump in the outcome by the jump in treatment:

$$\tau_{FRD} = \frac{E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-]}{Pr(D_i = 1 \mid R_i = c^+) - Pr(D_i = 1 \mid R_i = c^-)} = \frac{\text{jump in } Y_i}{\text{jump in } D_i}$$

- Compare with the Wald estimator from the IV lecture:

$$\beta_{IV} = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{Pr(D_i = 1 \mid Z_i = 1) - Pr(D_i = 1 \mid Z_i = 0)}$$

- Same formula, with $Z_i = \mathbb{1}(R_i \geq c)$, comparing only units **close to the cutoff**

The fuzzy RD estimand: a Wald estimator at the cutoff

- Scale the jump in the outcome by the jump in treatment:

$$\tau_{FRD} = \frac{E[Y_i \mid R_i = c^+] - E[Y_i \mid R_i = c^-]}{Pr(D_i = 1 \mid R_i = c^+) - Pr(D_i = 1 \mid R_i = c^-)} = \frac{\text{jump in } Y_i}{\text{jump in } D_i}$$

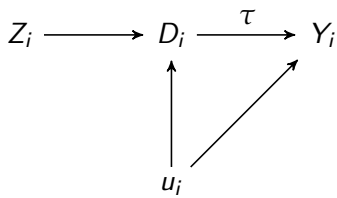
- Compare with the Wald estimator from the IV lecture:

$$\beta_{IV} = \frac{E(Y_i \mid Z_i = 1) - E(Y_i \mid Z_i = 0)}{Pr(D_i = 1 \mid Z_i = 1) - Pr(D_i = 1 \mid Z_i = 0)}$$

- Same formula, with $Z_i = \mathbb{1}(R_i \geq c)$, comparing only units **close to the cutoff**
- Hoekstra: $\hat{\tau}_{FRD} = 0.095 / 0.388 \approx 0.24 \Rightarrow$ enrolling at the flagship raises earnings by roughly 24 log points, for students induced to enroll by crossing the cutoff
- Sharp RD is the special case where the denominator $= 1$

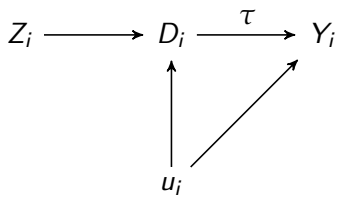
Fuzzy RD is IV: the instrument is “being above the cutoff”

- Instrument: $Z_i = \mathbb{1}(R_i \geq c)$
- **First stage**: jump in $Pr(D_i = 1)$ at c
- **Reduced form**: jump in $E[Y_i]$ at c
- $\tau_{FRD} = \dfrac{\text{reduced form}}{\text{first stage}}$, exactly as in IV



Fuzzy RD is IV: the instrument is "being above the cutoff"

- Instrument: $Z_i = \mathbb{1}(R_i \geq c)$
- **First stage**: jump in $Pr(D_i = 1)$ at c
- **Reduced form**: jump in $E[Y_i]$ at c
- $\tau_{FRD} = \dfrac{\text{reduced form}}{\text{first stage}}$, exactly as in IV



The IV assumptions, applied at the cutoff:

- **Relevance**: the probability of treatment jumps at c (first stage $\neq 0$). *Testable*: look at the first-stage picture
- **Exogeneity / exclusion**: crossing c affects Y_i **only** through D_i, and units just above and just below are comparable (continuity). *Not testable*
- **Monotonicity**: crossing c never makes anyone **less** likely to be treated (no defiers)

Who are the compliers in a fuzzy RD?

Groups	Below c ($Z_i = 0$)	Above c ($Z_i = 1$)	In Hoekstra
Never takers (NT)	$D_i = 0$	$D_i = 0$	go elsewhere anyway
Defiers (DF)	$D_i = 1$	$D_i = 0$	ruled out (monotonicity)
Compliers (C)	$D_i = 0$	$D_i = 1$	enroll **because** admitted
Always takers (AT)	$D_i = 1$	$D_i = 1$	enroll anyway (e.g., athletes)

Who are the compliers in a fuzzy RD?

Groups	Below c ($Z_i = 0$)	Above c ($Z_i = 1$)	In Hoekstra
Never takers (NT)	$D_i = 0$	$D_i = 0$	go elsewhere anyway
Defiers (DF)	$D_i = 1$	$D_i = 0$	ruled out (monotonicity)
Compliers (C)	$D_i = 0$	$D_i = 1$	enroll **because** admitted
Always takers (AT)	$D_i = 1$	$D_i = 1$	enroll anyway (e.g., athletes)

- Just as in the IV lecture, always takers and never takers cancel out of the Wald ratio
- Under the assumptions above, fuzzy RD identifies a **LATE at the cutoff**:

$$\tau_{FRD} = E[Y_i(1) - Y_i(0) \mid \text{complier},\ R_i = c]$$

- It is **doubly local**: only compliers, and only those at the cutoff
- First stage $= Pr(\text{complier at } c)$: in Hoekstra, 38.8% of students at the cutoff ▸ Derivation

Estimating a fuzzy RD: local linear 2SLS

Inside the window $|\tilde{R}_i| \leq h$, with $Z_i = \mathbb{1}(\tilde{R}_i \geq 0)$ as the instrument for D_i:

$$\begin{aligned}
\text{First stage:} \quad & D_i = \theta_0 + \theta_1 Z_i + \theta_2 \tilde{R}_i + \theta_3 (Z_i \times \tilde{R}_i) + \varepsilon_i \\
\text{Reduced form:} \quad & Y_i = \Pi_0 + \Pi_1 Z_i + \Pi_2 \tilde{R}_i + \Pi_3 (Z_i \times \tilde{R}_i) + e_i \\
\text{Structural:} \quad & Y_i = \beta_0 + \tau D_i + \beta_1 \tilde{R}_i + \beta_2 (Z_i \times \tilde{R}_i) + u_i
\end{aligned}$$

- Each of the first stage and reduced form is a **sharp RD** regression, with D_i or Y_i as the outcome
- 2SLS gives $\hat{\tau} = \hat{\Pi}_1 / \hat{\theta}_1$ ▸ Why?
- Use the **same** bandwidth and weights in both stages; $\tilde{R}_i$ and $Z_i \times \tilde{R}_i$ are exogenous controls

```
df$Z <- as.numeric(df$Rt >= 0)                          # above the cutoff
library(AER)
m <- ivreg(Y ~ D + Rt + Z:Rt | Z + Rt + Z:Rt,
           data = subset(df, abs(Rt) <= h))             # uniform kernel
summary(rdrobust(df$Y, df$R, c = cutoff, fuzzy = df$D)) # the package way
```

The big picture: RD and the designs we already know

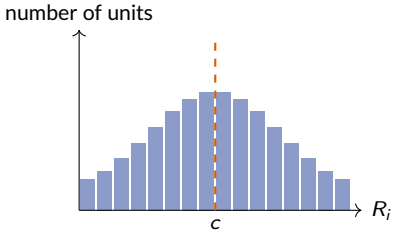
	Everywhere	**Only at the cutoff**
Assignment = treatment	RCT	Sharp RD
(perfect compliance)	ATT	ATT at c
Assignment shifts treatment	IV (Wald)	Fuzzy RD
(imperfect compliance)	LATE	LATE at c

- The "assignment" is a coin flip in an RCT, an instrument Z_i in IV, and $Z_i = \mathbb{1}(R_i \geq c)$ in RD
- RD replaces "randomly assigned" with "as good as randomly assigned **close to the cutoff**"
- Sharp RD = fuzzy RD with a first stage of 1, just like an RCT with perfect compliance is an IV with a first stage of 1

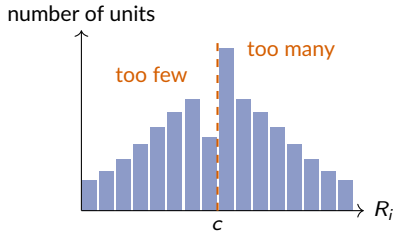
Is the RD credible?
Validity checks

The main threat: manipulation of the running variable

No manipulation



Bunching just above the cutoff



- If people know the cutoff and can **precisely** control R_i, they move to the side they prefer
- Example: families reporting income just below a financial aid threshold; teachers "rounding up" exam scores to a passing grade
- Then those just above are **not** comparable to those just below: continuity fails
- RD works when R_i has some **randomness** people cannot control (luck on test day), or the cutoff is **unknown** to them

Checklist: how to assess an RD

1. **Density test**: is there bunching in the number of units at c?
   McCrary (2008); in R: `rddensity`
2. **Covariate balance**: run the RD with a **pre-determined** covariate (gender, parents' income, high school GPA) as the outcome. There should be **no jump**
   Same logic as the balance table in an RCT
3. **Placebo cutoffs**: estimate "jumps" at fake cutoffs (e.g., $c - 100$, $c + 100$). There should be none
4. **Bandwidth sensitivity**: the estimate should not change much with $h/2$, $2h$, or another kernel
5. **Other things at the same cutoff?** At age 65, Medicare starts **and** many people retire. A jump at 65 mixes the two

Checklist: how to assess an RD

1. **Density test**: is there bunching in the number of units at c?
   McCrary (2008); in R: `rddensity`
2. **Covariate balance**: run the RD with a **pre-determined** covariate (gender, parents' income, high school GPA) as the outcome. There should be **no jump**
   Same logic as the balance table in an RCT
3. **Placebo cutoffs**: estimate "jumps" at fake cutoffs (e.g., $c - 100$, $c + 100$). There should be none
4. **Bandwidth sensitivity**: the estimate should not change much with $h/2$, $2h$, or another kernel
5. **Other things at the same cutoff?** At age 65, Medicare starts **and** many people retire. A jump at 65 mixes the two

- As with parallel trends in D-i-D, continuity itself is **not testable**. These checks make it more (or less) believable

RD effects are local

Important limitation

RD estimates the effect **at the cutoff**. The fuzzy RD estimates it only for **compliers at the cutoff**.

- **What RD tells us:** the effect for students right at the admission cutoff
- **What RD cannot tell us:** the effect for students far from the cutoff (very high or very low SAT scores)
- This affects **external validity**
- But often the cutoff is exactly where policy is made: “What if we lowered the admission cutoff by a few points?” is a question about students **at** the cutoff

Wrap-up

Regression discontinuity: takeaways

- RD uses a rule: treatment changes when a running variable crosses a cutoff
- Key assumption: **continuity**. Nothing but treatment jumps at $c \Rightarrow$ no manipulation of R_i
- **Sharp RD** ($D_i = Z_i$): a local RCT. Jump in Y_i = ATT at the cutoff
- **Fuzzy RD**: a local IV. $\dfrac{\text{jump in } Y_i}{\text{jump in } D_i}$ = LATE for compliers at the cutoff
- **Estimation**: local linear regression on each side of c, inside a bandwidth h
  - h trades off bias (large h) against variance (small h); choose it with the data and show robustness
  - Weights (kernels) let units nearer c count more; they matter less than h
  - Fuzzy RD: the same regressions, estimated by 2SLS
- Always: plot the data, test for manipulation and covariate balance

The course toolbox for causal inference

Design	Where variation comes from	Key assumption	What we get
RCT	Random assignment	Randomization done right	ATT (= ATE)
Panel FE	Changes within a unit over time	No time-varying confounders	Effect of within-unit changes
D-i-D	Policy change for some groups	Parallel trends, no anticipation	ATT
IV	An instrument Z_i	Relevance, exogeneity, monotonicity	LATE (compliers)
RD	A cutoff in a running variable	Continuity at c (no manipulation)	ATT / LATE at c

In every case, the question is the same: **what is the comparison group, and why is it a good counterfactual?**

Backup slides

What determines the “optimal” bandwidth?

- The data-driven bandwidth minimizes the mean squared error of $\hat{\tau}$. Its formula has the form

$$h^* = C \times \left(\frac{\text{variance of } Y_i \text{ near } c}{\text{density of } R_i \text{ at } c \ \times \ (\text{curvature of } E[Y_i \mid R_i] \text{ near } c)^2} \right)^{1/5} \times N^{-1/5}$$

- C is a constant that depends on the kernel

Back

What determines the “optimal” bandwidth?

- The data-driven bandwidth minimizes the mean squared error of $\hat{\tau}$. Its formula has the form

$$h^* = C \times \left(\frac{\text{variance of } Y_i \text{ near } c}{\text{density of } R_i \text{ at } c \;\times\; (\text{curvature of } E[Y_i \mid R_i] \text{ near } c)^2} \right)^{1/5} \times N^{-1/5}$$

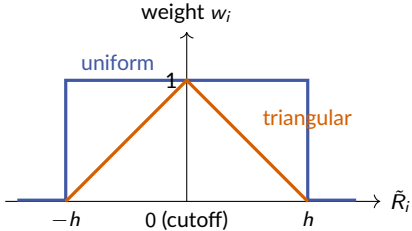
- C is a constant that depends on the kernel
- Reading the formula:
  - **Noisier outcome** $\Rightarrow$ larger h: we need more observations to average out the noise
  - **More curvature** near c $\Rightarrow$ smaller h: a straight line is only good very close to c
  - **More data near** c (higher density, larger N) $\Rightarrow$ smaller h: we can afford to zoom in
- You never compute this by hand: `rdrobust` estimates each piece from the data

Weights (kernels): who counts how much inside the window?

- Local linear regression is **weighted least squares**: choose the coefficients to minimize

$$\sum_i w_i \left(Y_i - \beta_0 - \tau D_i - \beta_1 \tilde{R}_i - \beta_2 D_i \tilde{R}_i\right)^2$$

- Each squared residual is multiplied by a weight $w_i \geq 0$
- $w_i = 0$ outside the window



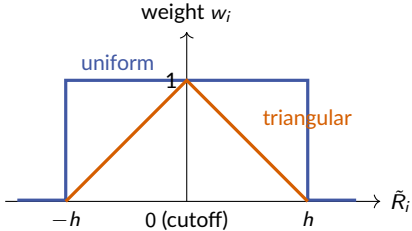
  - **Uniform**: $w_i = 1$ inside the window. This is plain OLS on the observations with $|\tilde{R}_i| \leq h$
  - **Triangular**: $w_i = 1 - |\tilde{R}_i|/h$. Units closer to the cutoff count more; the default in `rdrobust`

Weights (kernels): who counts how much inside the window?

- Local linear regression is **weighted least squares**: choose the coefficients to minimize

$$\sum_i w_i \left(Y_i - \beta_0 - \tau D_i - \beta_1 \tilde{R}_i - \beta_2 D_i \tilde{R}_i\right)^2$$

- Each squared residual is multiplied by a weight $w_i \geq 0$
- $w_i = 0$ outside the window



  - **Uniform**: $w_i = 1$ inside the window. This is plain OLS on the observations with $|\tilde{R}_i| \leq h$
  - **Triangular**: $w_i = 1 - |\tilde{R}_i|/h$. Units closer to the cutoff count more; the default in `rdrobust`
  - Bandwidth decides **who is in**; the kernel decides **how much each one counts**. In practice, results are much more sensitive to h than to the kernel

Back

Fuzzy RD: 2SLS = ratio of two sharp RDs

- Recall from the IV lecture: with one endogenous variable and one instrument, and the **same** exogenous controls in every equation,

$$\hat{\beta}_{2SLS} = \frac{\text{reduced form coefficient on } Z_i}{\text{first stage coefficient on } Z_i}$$

- In the fuzzy RD, the controls are $\tilde{R}_i$ and $Z_i \times \tilde{R}_i$, the same in all equations, so

$$\hat{\tau}_{FRD} = \frac{\hat{\Gamma}_1}{\hat{\theta}_1} = \frac{\text{sharp RD estimate with } Y_i \text{ as the outcome}}{\text{sharp RD estimate with } D_i \text{ as the outcome}}$$

Fuzzy RD: 2SLS = ratio of two sharp RDs

- Recall from the IV lecture: with one endogenous variable and one instrument, and the **same** exogenous controls in every equation,

$$\hat{\beta}_{2SLS} = \frac{\text{reduced form coefficient on } Z_i}{\text{first stage coefficient on } Z_i}$$

- In the fuzzy RD, the controls are $\tilde{R}_i$ and $Z_i \times \tilde{R}_i$, the same in all equations, so

$$\hat{\tau}_{FRD} = \frac{\hat{\Gamma}_1}{\hat{\theta}_1} = \frac{\text{sharp RD estimate with } Y_i \text{ as the outcome}}{\text{sharp RD estimate with } D_i \text{ as the outcome}}$$

- So you can do a fuzzy RD in two steps:
  1. Run the local linear regression of Y_i on Z_i, $\tilde{R}_i$, $Z_i \times \tilde{R}_i$: the jump in Y_i
  2. Run the same regression with D_i as the outcome: the jump in D_i
  3. Divide
- This gives the right point estimate. Use `ivreg` or `rdrobust` for the standard error
- Requirement: the **same** h and weights in both regressions

Back

Fuzzy RD identifies LATE at the cutoff

- Near c, each unit is a never taker, always taker, complier or defier. Continuity $\Rightarrow$ the shares of each group and their average potential outcomes are the same just above and just below c
- Monotonicity $\Rightarrow$ no defiers

Back

Fuzzy RD identifies LATE at the cutoff

- Near c, each unit is a never taker, always taker, complier or defier. Continuity $\Rightarrow$ the shares of each group and their average potential outcomes are the same just above and just below c
- Monotonicity $\Rightarrow$ no defiers
- Denominator (first stage):

$$Pr(D_i = 1 \mid c^+) - Pr(D_i = 1 \mid c^-) = \big[Pr(AT) + Pr(C)\big] - Pr(AT) = Pr(C)$$

Fuzzy RD identifies LATE at the cutoff

- Near c, each unit is a never taker, always taker, complier or defier. Continuity $\Rightarrow$ the shares of each group and their average potential outcomes are the same just above and just below c
- Monotonicity $\Rightarrow$ no defiers
- Denominator (first stage):

$$Pr(D_i = 1 \mid c^+) - Pr(D_i = 1 \mid c^-) = \big[Pr(AT) + Pr(C)\big] - Pr(AT) = Pr(C)$$

- Numerator (reduced form):

$$E[Y_i \mid c^+] = E[Y(1) \mid AT]\, Pr(AT) + E[Y(1) \mid C]\, Pr(C) + E[Y(0) \mid NT]\, Pr(NT)$$
$$E[Y_i \mid c^-] = E[Y(1) \mid AT]\, Pr(AT) + E[Y(0) \mid C]\, Pr(C) + E[Y(0) \mid NT]\, Pr(NT)$$

Always takers and never takers cancel: numerator $= \big(E[Y(1) \mid C] - E[Y(0) \mid C]\big)\, Pr(C)$

Fuzzy RD identifies LATE at the cutoff

- Near c, each unit is a never taker, always taker, complier or defier. Continuity $\Rightarrow$ the shares of each group and their average potential outcomes are the same just above and just below c
- Monotonicity $\Rightarrow$ no defiers
- Denominator (first stage):

$$Pr(D_i = 1 \mid c^+) - Pr(D_i = 1 \mid c^-) = \big[Pr(AT) + Pr(C)\big] - Pr(AT) = Pr(C)$$

- Numerator (reduced form):

$$E[Y_i \mid c^+] = E[Y(1) \mid AT]\, Pr(AT) + E[Y(1) \mid C]\, Pr(C) + E[Y(0) \mid NT]\, Pr(NT)$$
$$E[Y_i \mid c^-] = E[Y(1) \mid AT]\, Pr(AT) + E[Y(0) \mid C]\, Pr(C) + E[Y(0) \mid NT]\, Pr(NT)$$

  Always takers and never takers cancel: numerator $= \big(E[Y(1) \mid C] - E[Y(0) \mid C]\big)\, Pr(C)$

- Ratio:

$$\tau_{FRD} = \frac{\big(E[Y(1) \mid C] - E[Y(0) \mid C]\big)\, Pr(C)}{Pr(C)} = E[Y_i(1) - Y_i(0) \mid \text{complier},\ R_i = c]$$

- (All expectations are for units at the cutoff.) Same steps as the LATE derivation in the IV lecture

◂ Back