8. Regression discontinuity

LPO 8852: Regression II

Sean P. Corcoran

RD - introduction

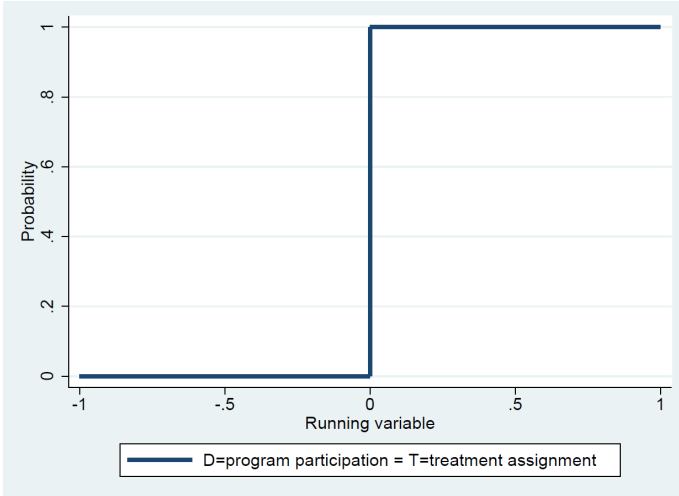
RD can be used when a **precise** rule based on a **continuous** characteristic determines treatment assignment. Examples:

- **Test scores**: can determine school admission, financial aid, summer school, remediation, graduation
- **Income or poverty score**: eligibility for income assistance or benefits, community eligibility for a means-tested anti-poverty program
- **Date**: age cutoff for retirement benefits, health insurance, school enrollment (KG or PK)
- **Elections**: fraction that voted for a particular candidate or ballot measure (e.g., school bond)

The continuous characteristic is typically called a **running variable**, **forcing variable**, or **index**.

RD - introduction

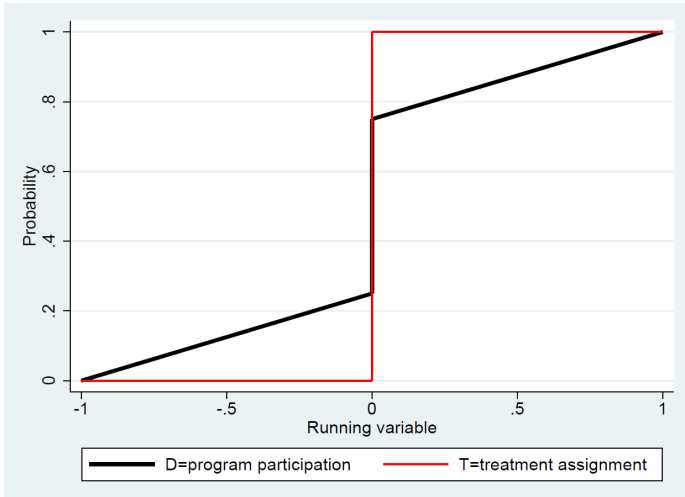
Sharp RD: treatment *assignment* goes from $0 \rightarrow 1$ at a threshold c. Treatment *receipt* goes from 0% to 100% at c (full compliance).



Re-center the running variable so that the threshold value is 0 $(X - c)$.

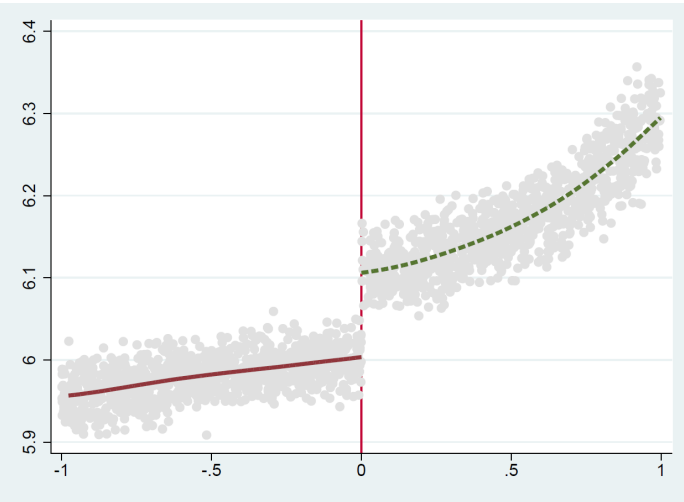
RD - introduction

Fuzzy RD: treatment *assignment* goes from $0 \rightarrow 1$ at a threshold c. Treatment *receipt* increases sharply at c but there is partial compliance.



RD - introduction

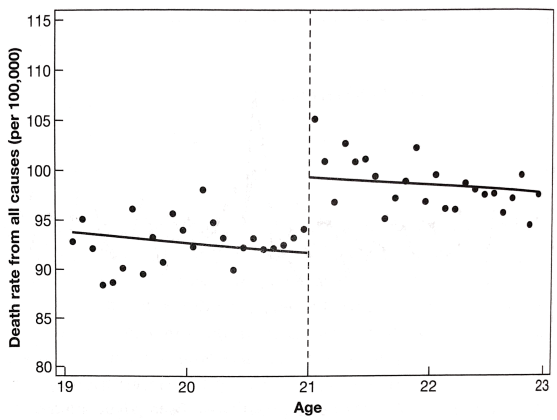
If there is a discrete change in treatment and program participation at c (and the program has a treatment effect) one would expect to see a discrete change in the mean outcome at c.



RD - introduction

Under certain assumptions, this change can be interpreted as the (local) causal effect of the treatment. The challenge is estimating this change. There is often a relationship between the running variable and Y, even in the absence of treatment. We need to carefully estimate this relationship on either side of c, since this is what estimates the treatment effect.

FIGURE 4.2
A sharp RD estimate of MLDA mortality effects



Notes: This figure plots death rates from all causes against age in months. The lines in the figure show fitted values from a regression of death rates on an over-21 dummy and age in months (the vertical dashed line indicates the

Potential outcomes and sharp RD

Each unit has potential outcomes $Y_i(0)$ and $Y_i(1)$, which vary with the running variable X_i. We only observe $Y_i(1)$ for those above c and $Y_i(0)$ for those below c. The observed Y_i:

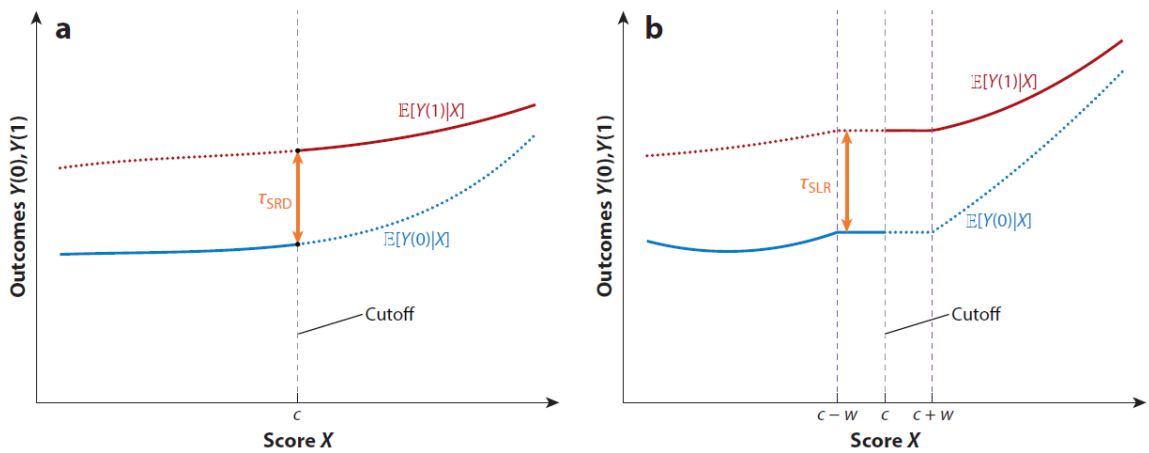
$$Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1) = \begin{cases} Y_i(0) & \text{if } X_i < c \\ Y_i(1) & \text{if } X_i \geq c \end{cases}$$

Here assume assignment to treatment and actual treatment are the same (i.e., sharp RD).

We never observe treated and untreated units with the same X—there is a complete lack of common support. RD estimation requires extrapolation!

Potential outcomes and sharp RD

There are two main frameworks for RD: the continuity framework (a) and local randomization framework (b).



The RD treatment effect at c is τ_{SRD}. Figures: Cattaneo & Titiunik (2022)

Potential outcomes and sharp RD

Under the **continuity assumption**, mean potential outcomes $E[Y(0)|X]$ and $E[Y(1)|X]$ are continuous near c. We need to model the relationship between Y and X below and above c to estimate τ_{SRD}.

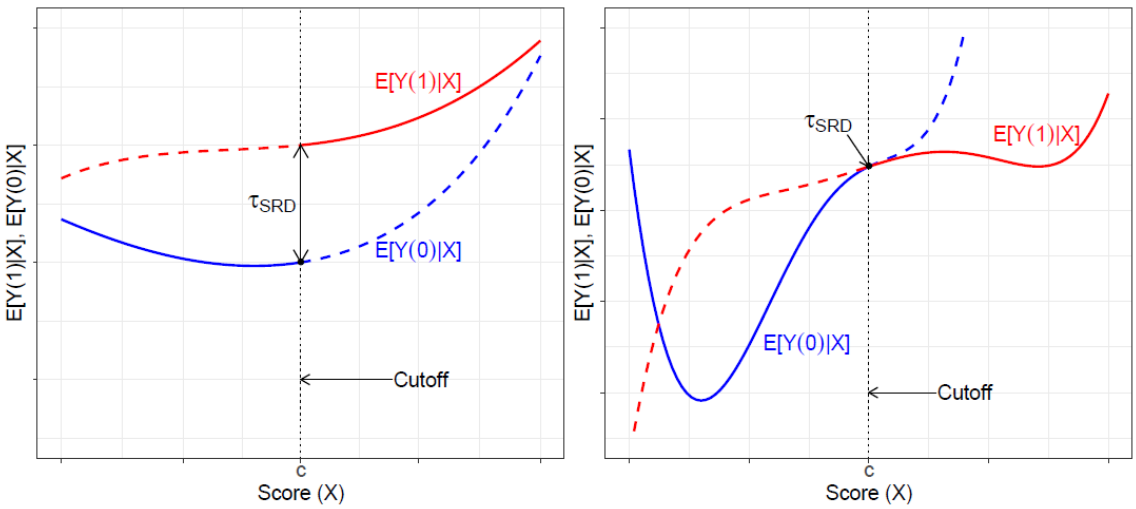
Under the **local randomization assumption**, treatment assignment—being above or below c—is random within a neighborhood $(\pm w)$ of c. Mean potential outcomes do not vary within this neighborhood.

We will focus first and primarily on the continuity framework since it is more common. But there are cases where the randomization assumption arguably holds.

RD effects are local

Under certain assumptions our estimate of τ_{SRD} can be considered causal, but at a single point in the distribution of X. It may have limited external validity. How representative τ is depends on the context:

Figure 2.4: Local Nature of RD Effect



(a) Mild Heterogeneity

(b) Severe Heterogeneity

RD plots

How should we model the relationship between Y and X?

- Linear function?
- Quadratic?
- Higher order polynomial?

A good place to start is a plot of the data, with X centered at c.

It can be difficult, however, to see systematic relationships in a scatterplot of the raw data. The user-written `rdplot` command can help (See also `binscatter`).

RD plots

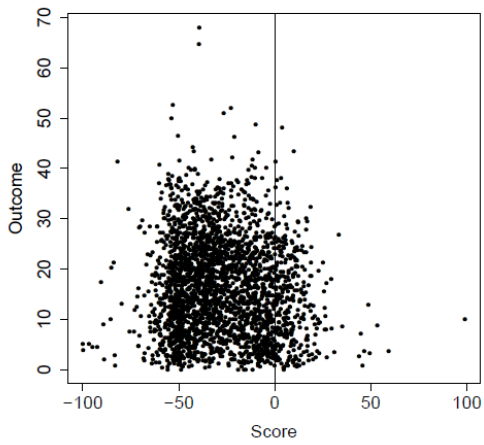


Figure 3.1: Scatter Plot—Meyersson Data

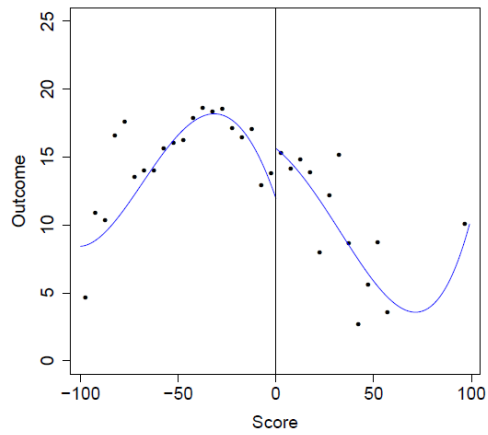


Figure 3.2: RD Plot for Meyersson Data Using 40 Bins of Equal Length

Figures: Cattaneo et al. (2020) *Foundations*

RD plots

A typical RD plot includes:

- **Global polynomial fit** (solid line) based on 4th or 5th order polynomial, fitted separately below and above c
- **Local sample means** (dots) based on binning X and plotting the mean for each bin at the midpoint of the bin

The plot can help you to see:

- Nonlinear relationship between Y and X
- Local variability around the fitted line
- Any evidence of a discontinuity at c
- Any other discontinuities away from c

rdplot

`rdplot` *y x* [, `c(#) nbins(# #) p(#) binselect(`*binmethod*`) kernel(`*kernelfn*`)` *otheroptions*]

- `c(#)` specifies the cutpoint c (default is 0)
- `nbins(# #)` allows you to specify the number of bins on the L and R
- `binselect(`*binmethod*`)` specifies a procedure to determine the number of bins (if not set manually)
- `p(#)` is the order of the polynomial fit (default is 4)
- `kernel(`*kernelfn*`)` allows you to select a kernel: weights that depend on distance from c (default is uniform)
- Other options include confidence intervals, shading, covariates, etc.

RD plots: deciding how bin width is determined

Evenly-spaced bins (ES)

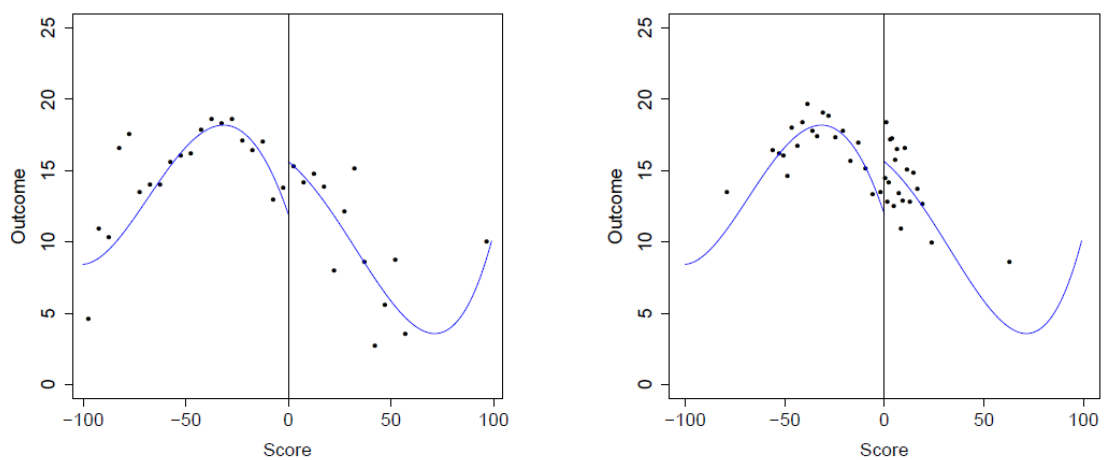
- All bins are equal width
- Number of observations per bin varies (thus, precision varies across bins)

Quantile-spaced bins (QS)

- All bins contain roughly the same number of observations
- Bin width varies
- Advantage: easier to see the density of the running variable

RD plots: deciding how bin width is determined

Figure 3.3: RD Plots—Meyersson Data



(a) 40 Evenly-Spaced Bins

(b) 40 Quantile-Spaced Bins

Figures: Cattaneo et al. (2020) *Foundations*

RD plots: choosing number of bins

Manual

- Can choose your own values `nbins(# #)`
- This is an *ad hoc* decision, however.

Integrated mean squared error (IMSE) method

- Chooses a number of bins that balances bias and variance when estimating local means
- A larger number of bins reduces bias, but there are fewer observations per bin, which leads to more sampling variance.
- Results in means that roughly trace out the global polynomial fit
- Good for seeing general shape or other discontinuities away from c

RD plots: choosing number of bins

Mimicking variance (MV) method

- Chooses a number of bins so that local means have variability that approximates that of the data
- Usually leads to more bins than IMSE
- Provides a better picture of variability—it is less smoothed.

Note IMSE and MV methods will generally select a different number of bins on the L and R of c. If manually choosing the # of bins, you can also choose different values on the L and R.

Cattaneo et al. (2020) recommend starting with MV method, ideally comparing ES and QS for bin widths to show density of scores. Selecting IMSE method is preferable for global features of the regression function.

`rdplot` bin options

Use these options with `rdplot` to select the number and type of bins:

`nbins(# #)` `binselect(es)`: manual # of bins, equally spaced

`nbins(# #)` `binselect(qs)`: manual # of bins, quantile spaced

`binselect(es)`: IMSE method, equally spaced

`binselect(qs)`: IMSE method, quantile spaced

`binselect(esmv)`: MV method, equally spaced

`binselect(qsmv)`: MV method, quantile spaced

Sharp RD estimation

The global polynomial fit in the RD plot can provide a good approximation overall, but a poor approximation at c, which is critical to RD estimation. Additionally, outlying observations far from c may be overly influential.

- When the full range of data are used to fit the relationship between Y and X, this is called a **global** or **parametric** approach.
- When a limited range of data around c are used, this is called a **local**, **flexible**, or **non-parametric** approach.

Current state of the art recommends the non-parametric approach: choosing a bandwidth around c and fitting a local polynomial of low order.

Sharp RD estimation

Figure 4.1: RD Estimation with local polynomial

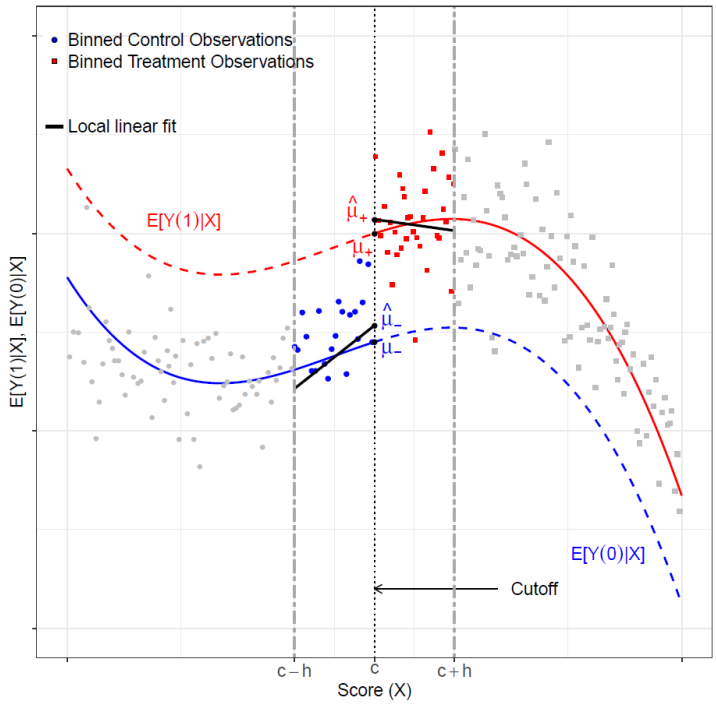


Figure: Cattaneo et al. (2020) *Foundations*

Sharp RD estimation: basic steps

Basic steps for sharp RD estimation under the continuity assumption:

1. Choose a polynomial order p and a kernel function $K()$
2. Choose a bandwidth h
3. Fit a weighted least squares regression of Y on $(X_i - c)$, $(X_i - c)^2$, ..., $(X_i - c)^p$ up to the order p using $K()$ function as weights. Do this for observations below and above the cutoff c. The difference in the two intercepts is $\hat{\tau}_{SRD}$. Note these can be done in the same regression (see next slides).

Sharp RD estimation

Set aside the kernel and bandwidth for now and let $\tilde{X}_i = X_i - c$. The two regression models fit for observations below and above c are:

$$Y_{0i} = \alpha_0 + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + ... + \beta_{0p}\tilde{X}_i^p + u_i \text{ for } X_i < c$$
$$Y_{1i} = \alpha_1 + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + ... + \beta_{1p}\tilde{X}_i^p + u_i \text{ for } X_i \geq c$$

Notice the slope coefficients differ on either side of c. Let $D_i = 1$ if $X_i \geq c$. Then we can pool the data and estimate one regression:

$$\begin{aligned} Y_i = \alpha_0 + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + ... + \beta_{0p}\tilde{X}_i^p \\ +\rho D_i + \beta_1^* D_i\tilde{X}_i + \beta_2^* D_i\tilde{X}_i^2 + \beta_p^* D_i\tilde{X}_i^p + u_i \end{aligned}$$

where $\rho = \alpha_1 - \alpha_0$ (the difference in intercepts at c—this is our $\hat{\tau}_{SRD}$) and $\beta_j^* = (\beta_{1j} - \beta_{0j})$ (the difference in slope coefficients above and below c).

Sharp RD estimation

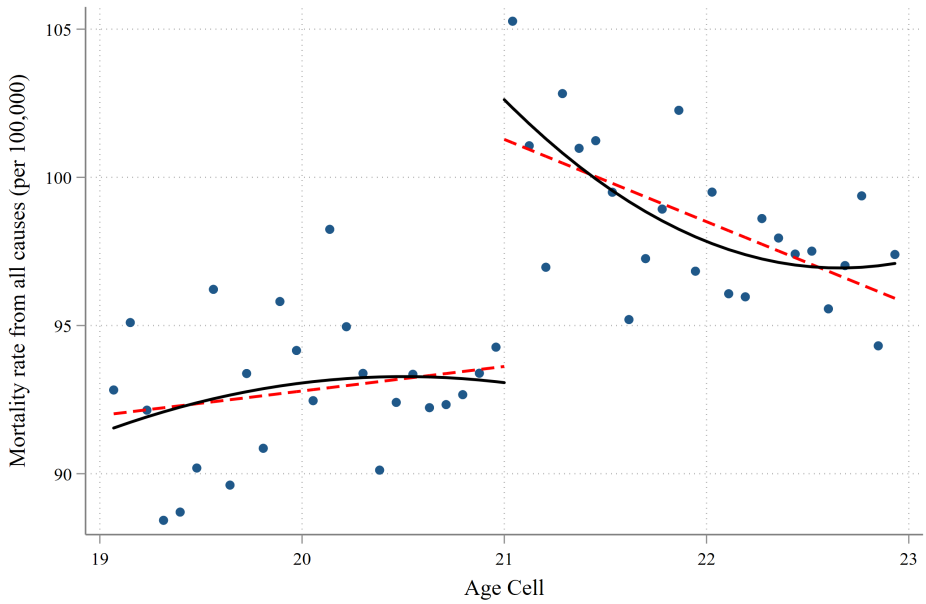
As an example, let $p = 2$ (quadratic):

$$Y_i = \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \rho D_i + \beta_1^* D_i\tilde{X}_i + \beta_2^* D_i\tilde{X}_i^2 + u_i$$

This is very easily estimated using OLS. See in-class example based on *Mastering Metrics* chapter 4 and Carpenter & Dobkin (2009): estimating the mortality effects of legal access to alcohol using the discontinuity in treatment at age 21. Note these data are already binned and limited to ages 19-23.

In-class example: Carpenter & Dobkin 2009

Linear and quadratic fits with different slopes below and above c.



Note: uses *agecell* for x-axis instead of centered version.

Sharp RD estimation

You can improve your estimate of $\hat{\tau}_{SRD}$ by choosing an appropriate bandwidth (h), polynomial order (p) and weighting function, or kernel $K()$.

A **kernel function** assigns weights to the data based on their distance from the cutoff c. It may make sense to give greater weight to observations closer to c. Examples:

- **uniform**: equal weight
- **triangular**: weight declines symmetrically with distance from c (usually recommended)
- **Epanechnikov**: quadratic decline with distance from c

Kernel functions

Figure 4.2: Different Kernel Weights for RD Estimation

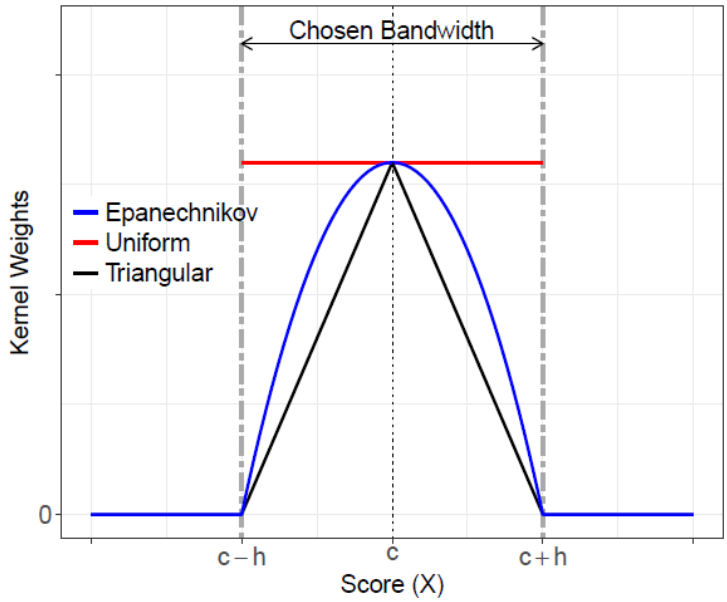


Figure: Cattaneo et al. (2020) *Foundations*

Polynomial order

Choice of polynomial order (p) is usually more consequential than $K()$.

- For a given h, increasing p improves fit but also increases variability of the estimator.
- Higher order polynomials tend to overfit, and have less reliable results near c (where it matters)

The current state of the art recommends a (local) **linear estimator**, but it is common to test sensitivity to higher orders (e.g., quadratic).

Choosing a bandwidth

The most consequential decision in RD is usually the bandwidth h.

Figure 4.3: Bias in Local Approximations

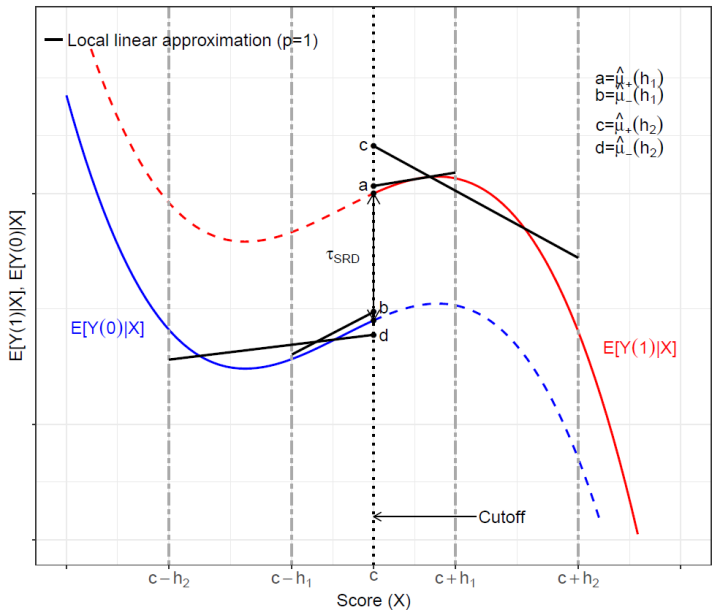


Figure: Cattaneo et al. (2020) *Foundations*

Choosing a bandwidth

A smaller h reduces the risk of misspecification ("smoothing bias") but will tend to increase the estimator's variability due to fewer observations. There is a **bias-variance** tradeoff in the choice of h.

You can choose your own bandwidth, but the decision is *ad hoc* and could lead to specification searching. Need a data-driven way to select h.

The most popular (and recommended) optimal bandwidth is **minimum MSE** (e.g., Imbens & Kalyanaraman, 2012). This minimizes the MSE of a local polynomial RD estimator, given a choice of p and $K()$. See Cattaneo et al. (2020) for a technical discussion.

Sharp RD estimation using `rdrobust`

You could choose an h and $K()$ and implement these using OLS (as in the in-class example above). There are several issues with this, however:

- It will involve multiple steps, and you would also want to use an optimal bandwidth selector first to choose h for you.
- The standard errors will be invalid. The MSE method allows for some bias in the choice of optimal bandwidth. OLS standard errors do not account for the effects of this approximation.

The user-written command `rdrobust` is very flexible and can handle all of these things and more.

`rdrobust`

`rdrobust` *y x* `[, c(#) p(#) h(# #) bwselect(`*bwmethod*`) kernel(`*kernelfn*`) covs(`*covars*`)` *otheroptions*`]`

- `c(#)` specifies the cutpoint c (default is 0)
- `bwselect(`*bwmethod*`)` specifies a procedure to determine the bandwidth (if not set manually)
- `p(#)` is the order of the polynomial fit (default is **1**)
- `kernel(`*kernelfn*`)` (default is **triangular**)
- `covs(`*covars*`)` allows you to include covariates in your model
- Other options

Sharp RD estimation using `rdrobust`

See in the in-class example using data from Carpenter & Dobkin (2009), using `rdrobust` instead of `regress`.

Example using fixed bandwidth of ± 2, uniform kernel, and either linear or quadratic fit:

```
rdrobust all age, c(0) h(2) kernel(uniform) p(1)
rdrobust all age, c(0) h(2) kernel(uniform) p(2)
```


Optimal bandwidth selection using `rdbwselect`

The related command `rdbwselect` will perform optimal bandwidth selection methods. (Note this can also be called by `rdrobust` in one step).

Two common options:

- `mserd`: MSE method, same bandwidth on both sides of c
- `msetwo`: MSE method, different bandwidths on each side

There are other options, see Cattaneo et al. (2020) for details.

rdbwselect

`rdbwselect` *y x* `[, c(#) p(#) bwselect(`*bwmethod*`) kernel(`*kernelfn*`) covs(`*covars*`)` *otheroptions*`]`

- `c(#)` specifies the cutpoint *c* (default is 0)
- `bwselect(`*bwmethod*`)` specifies a procedure to determine the bandwidth (default is **mserd**)
- `p(#)` is the order of the polynomial fit (default is **1**)
- `kernel(`*kernelfn*`)` (default is **triangular**)
- `covs(`*covars*`)` allows you to include covariates in your model
- Other options

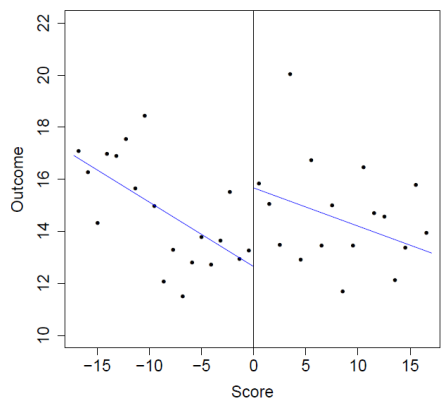
Optimal bandwidth selection

You can pass through the optimal bandwidth (uses `rdbwselect`) to the `rdrobust` and `rdplot` commands.

In `rdrobust`: `rdrobust` *y x*`, bwselect(`*mserd*`)`

See the in-class example code for this and the RD plot code.

Figure 4.4: Local Polynomial RD Effect Illustrated with `rdplot`—Meyersson Data



Standard errors and confidence intervals

As noted earlier, the usual OLS standard errors are invalid. See Cattaneo et al. (2020) for an extended discussion of inference in sharp RD. Their recommendation is to use the **robust bias-corrected** approach to inference (i.e., confidence intervals). Can add `all` option to see inference using different methods.

Including covariates

You can include covariates in your model to improve precision, but if the continuity assumption holds, one would not expect covariates to be systematically different below/above c. Obtaining materially different results with covariates is a red flag.

In `rdrobust`: `rdrobust` *y x*, `covs(`*covars*`)`

Note the optimal bandwidth will likely change with the new model specification (w/covariates).

Another way to use covariates: splitting the sample by subgroups. This is commonly done, and requires no modification of the steps above (just subset on the group when using `rdrobust`).

RD assumptions

To assess the validity of the sharp RD, consider the assumptions:

- Treatment assignment (and receipt) occurs at a known threshold c.
- The relationship between potential outcomes $Y(1)$, $Y(0)$ and X is **continuous** in the neighborhood of c. There is no reason to expect a sharp break in Y in the absence of treatment.

These imply:

- X has **not been manipulated** to affect who receives treatment.
- There are no other programs or services with the **same eligibility rule** (to avoid confounding with some other treatment).

Some common RD validity tests

The following are commonly performed as validity tests with RD:

- Test for effects at c on **pre-treatment covariates** or **placebo outcomes**. Would expect a null effect.
- Test for **continuity in the density** of the running variable around c ("manipulation test").
- Test for discontinuities elsewhere in the distribution of X (i.e., artificial cutoffs). A "smoothness" test.
- Exclusion of observations near c (**"donut hole"** approach).
- Sensitivity tests for bandwidth choice.

Test for effects on pre-treatment covariates

Use the same `rdrobust` code but swapping in pre-treatment covariates (or placebo outcomes) for y. Note the optimal bandwidth will change (this is OK). Can also show RD plots:

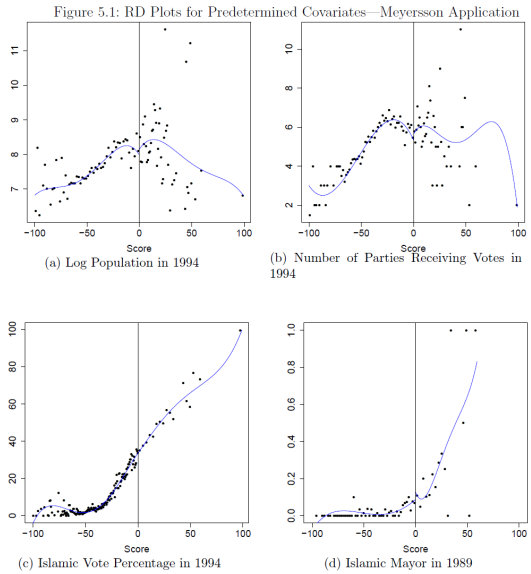


Figure: Cattaneo et al. (2020) *Foundations*

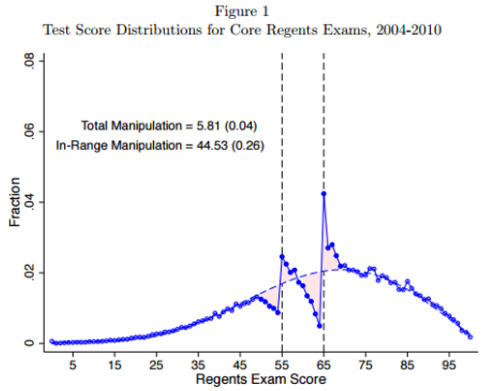
Manipulation test

Manipulation occurs whenever units have their value of X altered in order to affect their treatment status. For example, a teacher might adjust a test score in order to help a student pass or become eligible for a program.

- This may be visible in a histogram, or not if the manipulation goes both ways.
- If manipulation is random or uninformed, such that potential outcomes in the absence of treatment are no different on average for those whose X has been manipulated, then manipulation will not pose a problem. Manipulation is not usually random, however.
- Manipulation may lead one to find an "effect" where there is none.
- Note manipulation is not the same thing as non-compliance (the "fuzzy RD" case).

Manipulation test

Sometimes manipulation is clear from inspecting densities or histograms:



Notes: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in each subject for each student in our sample. Each point shows the fraction of test takers in a score bin with solid points indicating a manipulable score. The dotted line beneath the empirical distribution is a subject specific sixth-degree polynomial fitted to the empirical distribution excluding the manipulable scores near each cutoff. Total manipulation is the fraction of test takers with manipulated scores. In-range manipulation is the fraction of test takers with manipulated scores normalized by the average height of the counterfactual distribution to the left of each cutoff. Standard errors are calculated using the parametric bootstrap procedure described in the text. See the data appendix for additional details on the sample and variable definitions.

Figure: Dee et al. (2011)

Manipulation test

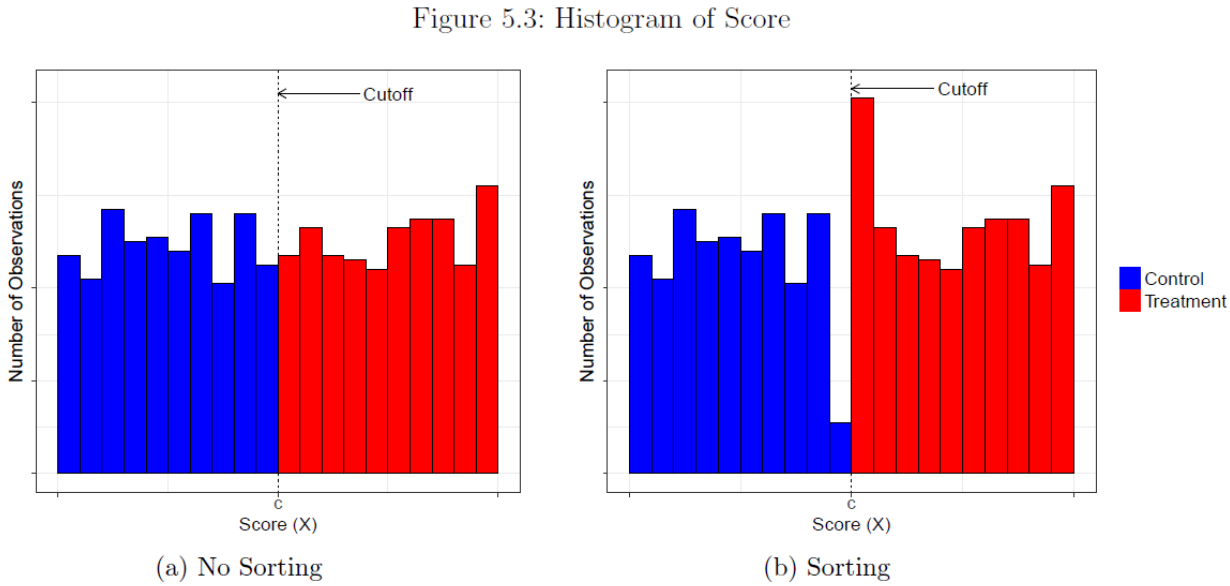


Figure: Cattaneo et al. (2020) *Foundations*

Manipulation test using `rddensity`

You should present both a figure and a formal statistical test. The user-written `rddensity` command can help with this.

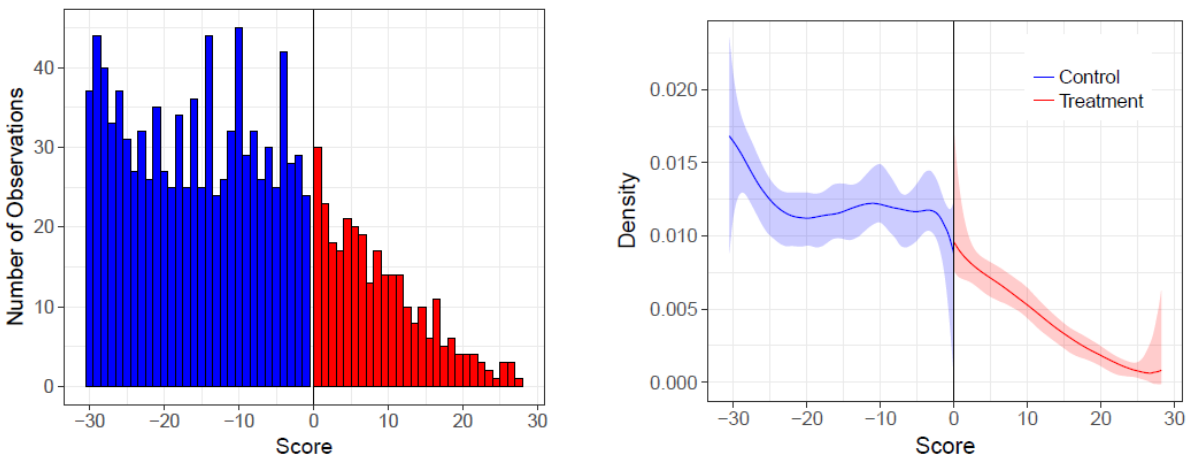
`rddensity` *x, plot otheroptions*

The idea of the statistical test is to fit a local polynomial to the density of X on the L and R of c and testing for a discontinuity at c. This general procedure is commonly referred to as a **McCrary test** (McCrary, 2008) although details of the test vary.

H_0 is "no manipulation," or no discontinuity at c.

Manipulation test using `rddensity`

Figure 5.4: Histogram and Estimated Density of the Score

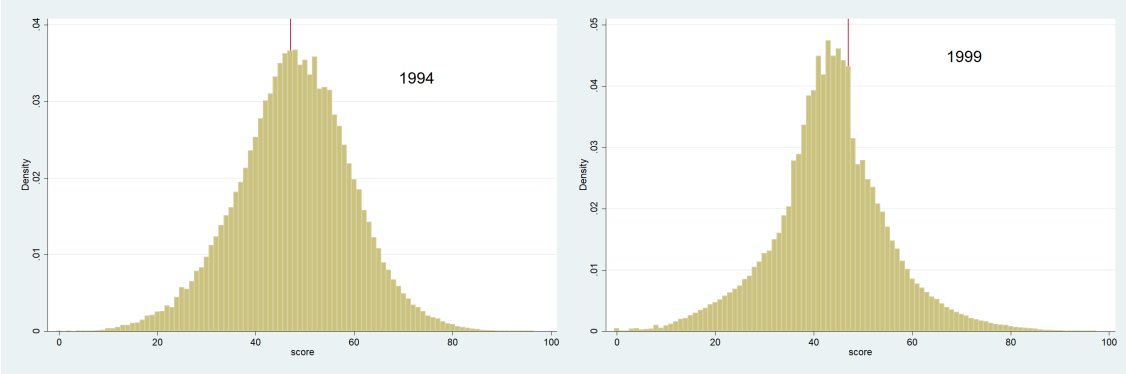


(a) Histogram (b) Estimated Density

Figure: Cattaneo et al. (2020) *Foundations*

Example: Camacho and Conover (2011)

In 1998, Colombia set eligibility threshold for social welfare benefits at a poverty index of 47.



There is evident manipulation in 1999 but the manipulation test also fails in 1994. This test can over-reject with a **discrete** running variable (here, an integer poverty index), even more so when N is large.

Tests at alternative cutoffs

Can use `rdrobust` to test for discontinuities at other points. To avoid "contamination" from real cutpoints, include *only* the treated ($X \geq c$) or untreated ($X < c$) observations, depending on where alternative cutoff is located. Example:

```
rdrobust y x if x>=0, c(1)
```

Cattaneo et al. (2020): "evidence of continuity away from the cutoff is neither necessary nor sufficient for continuity at the cutoff, but the presence of discontinuities away from the cutoff can be interpreted as potentially casting doubt on the RD design, at the very least in cases where such discontinuities can not be explained by substantive knowledge of the specific application."

Tests at alternative cutoffs

A graphical summary of tests at alternative cutoffs, including the real cutoff at $X = c$ as a reference:

Figure 5.5: RD Estimation for True and Placebo Cutoffs

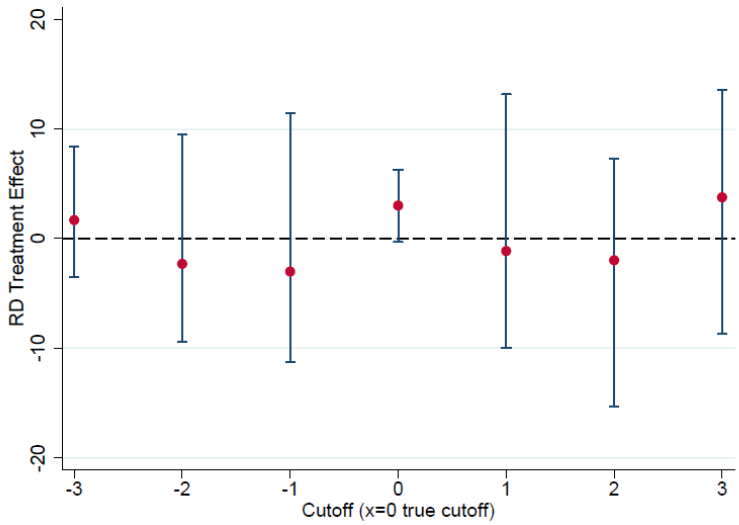


Figure: Cattaneo et al. (2020) *Foundations*

Donut hole robustness test

Units closest to c get the most weight and are also those most susceptible to manipulation (in some applications). Can exclude observations within a certain radius of c to see if/how results change. Example:

`rdrobust` *y x* `if abs(`*x*`)>=0.3`

Figure 5.6: RD Estimation for the Donut-Hole Approach

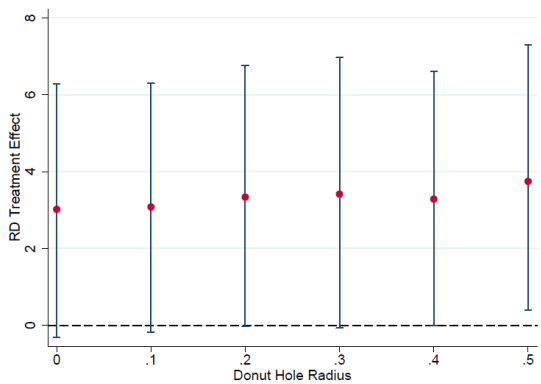


Figure: Cattaneo et al. (2020) *Foundations*

Sensitivity to bandwidth

It is common to see estimates using different bandwidth choices as a robustness check. Could use multiples of the original bandwidth (E.g., $\times 0.5$, $\times 1.5$, $\times 2$)

Results should be interpreted with caution: wider bandwidths can introduce bias, while narrower bandwidths introduce more sampling variability.

Sharp RD under local randomization

Thus far we have relied on the **continuity** framework, where mean potential outcomes are continuous near c. We modeled the relationship between Y and X below and above c in order to estimate τ_{SRD}.

Ideally we would not have to rely on functional form assumptions at all. Under **local randomization**, treatment assignment—being above or below c—is random within a neighborhood ($\pm w$) of c. Mean potential outcomes do not vary within this neighborhood.

RD is often described as "as good as random assignment" around c. But this is not strictly true if we are fitting a polynomial to estimate τ_{SRD}. If local randomization held, we could just compare means on either side.

Sharp RD under local randomization

With local randomization, potential outcomes are not related to the running variable x in a window ($\pm w$) around c:

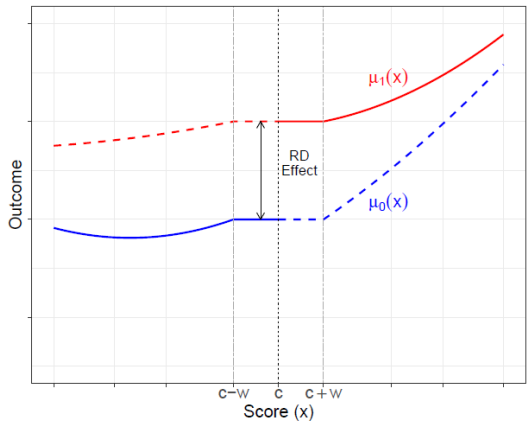


Figure 2.2: Local Randomization RD

Figure: Cattaneo et al. (2023) *Extensions*

Sharp RD under local randomization

The assumption of local randomization is stronger than the assumption of continuity. So when might you want to use this approach?

- If you have a large sample and a continuous running variable where you believe random factors play an important role in whether a unit is just above or below c (e.g., a test, voter turnout).
- If you have a **small sample** near c, making it difficult to reliably approximate the function on either side of c.
- If you have a **discrete running variable**, such that the continuity assumption is not conceptually appropriate.

The local randomization approach is often used as a robustness check to a continuity-based analysis when the running variable is continuous.

Local randomization: estimation and inference

Cattaneo et al. (2023) provide a full description of sharp RD estimation and inference under the local randomization assumption. See their paper for more details. It boils down to comparing means on either side of c.

The most consequential decision: within what window ($\pm w$) can we assume units are "as good as randomly assigned?"

They recommend testing for balance in pre-determined covariates, starting with the smallest possible window around c and getting progressively wider until the null of no difference is rejected. The chosen window is the largest one such that H_0 fails to be rejected inside that window (and all windows within it).

Window selection under local randomization

W_1 is the first window tested, then W_2, etc. W_0 is the widest window where H_0 is not rejected. Can use `rdwinselect` for this.

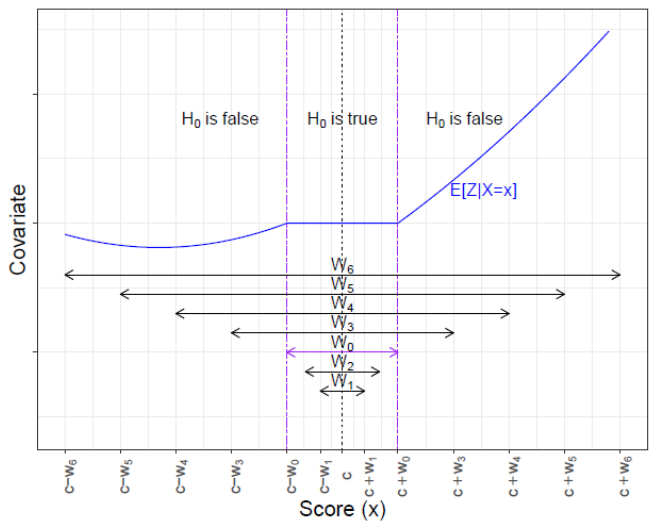


Figure 2.4: Window Selector Based on Covariate Balance

Figure: Cattaneo et al. (2023) *Extensions*

rdwinselect

`rdwinselect` *runvar covs* [, `c(#) wobs(#) wstep(#) seed(#)` *otheroptions*]

- `c(#)` specifies the cutpoint c (default is 0)
- `wobs(#)` tells Stata to increase window width by a fixed number of observations each time.
- `wstep(#)` instead tells Stata to increase window width by a fixed amount each time.
- `seed(#)` sets a random seed, for small-sample randomization inference.
- Other options

By default the smallest window is the one with ≥ 10 obs on each side.

Local randomization: estimation and inference

Once the window is selected, can use `rdrandinf` to estimate the difference in means above and below c, construct confidence intervals, etc.

Inference is based on either small sample (exact) randomization inference, or large sample approximations. See Cattaneo et al. (2023) for details.

Validity tests are important here as well, and are carried out with some modification. See Cattaneo et al. (2023) for details.

- Test for effects on pre-treatment covariates and placebo outcomes
- Manipulation test
- Test for effects at artificial cutoffs
- Sensitivity to window width

rdrandinf

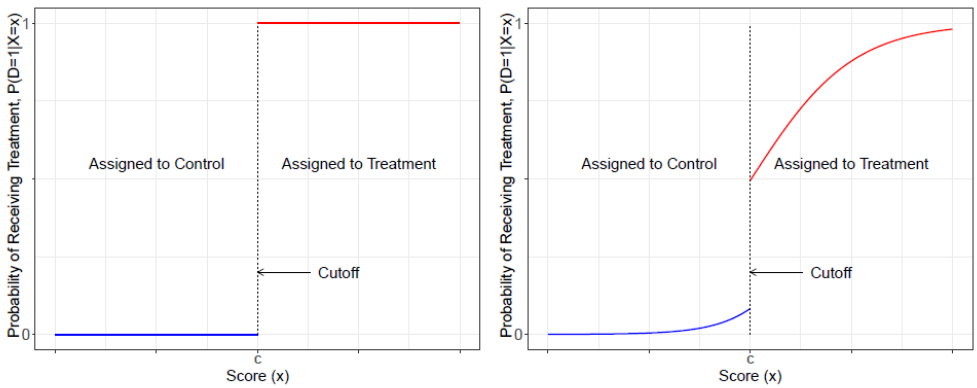
`rdrandinf` *y x* [, `c(#) wl(#) wr(#) seed(#)` *otheroptions*]

- `c(#)` specifies the cutpoint c (default is 0)
- `wl(#)` and `wr(#)` are the lower and upper bounds of the window. You can choose these yourself or use `rdwinselect`.
- `seed(#)` sets a random seed, for small-sample randomization inference.
- Other options

You can select the window and perform estimation all in one step, but it is advisable to make decisions about the window before seeing the results.

Fuzzy RD

Fuzzy RD: treatment *assignment* goes from $0 \rightarrow 1$ at c. Treatment *receipt* (e.g., program participation) increases sharply at c but there is some non-compliance.



(a) Sharp RD (b) Fuzzy RD

Figure 3.1: Conditional Probability of Receiving Treatment in Sharp vs. Fuzzy RD Designs

Figure: Cattaneo et al. (2023) *Extensions*

Fuzzy RD: intent-to-treat effects

Treatment *assignment* changes sharply at c. If the continuity (and/or local randomization) assumption holds, we still can estimate the causal effect of treatment assignment at c using the sharp RD methods described thus far.

This is not the treatment effect at c, but rather the **intent-to-treat**, τ_{ITT}.

In a fuzzy RD context, one should also estimate the effect of treatment assignment on the probability of receiving the treatment, τ_D, using sharp RD methods. This is the **first stage**, for reasons that will become clear.

Fuzzy RD: nonrandom take-up

The challenge for estimating the treatment effect at c is that units may strategically decide to take up (or not) the treatment based on their expected gains from it. There is no reason to believe potential outcomes are unrelated to the decision to take up treatment.

Potential treatment take-up

Potential treatment take-up, analogous to potential outcomes:

$D_i(0)$ is i's treatment take-up/receipt when $T_i = 0$
$D_i(1)$ is i's treatment take-up/receipt when $T_i = 1$

$D_i(T_i)$		Types:
$T_i = 0$	$T_i = 1$	
0	1	Compliers
1	1	Always-takers
0	0	Never-takers
1	0	Defiers

There are now four potential outcomes that depend on both treatment assignment and take-up: $Y_i(T_i, D_i)$.

Fuzzy RD as IV

Treatment assignment T_i at c is potentially a good instrumental variable for treatment receipt D_i. If so, we know from IV that the (local) average treatment effect of D_i on the outcome Y would be:

$$\tau_{FRD} = \frac{\tau_{ITT}}{\tau_D}$$

The numerator and denominator are both sharp RD parameters that we know how to estimate.

Call τ_{FRD} the fuzzy RD parameter. Under what conditions does it provide the local average treatment effect (LATE) at c?

Fuzzy RD as IV

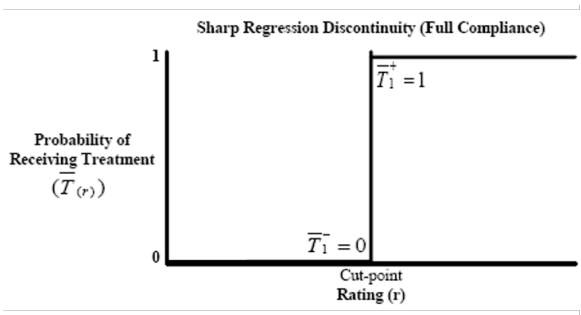
Conditions under which τ_{FRD} provides the LATE:

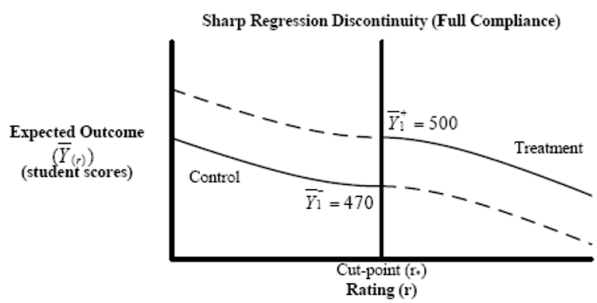
- Exclusion restriction: treatment assignment (being above/below c) cannot affect potential outcomes (or potential treatments) except through treatment received.
- Monotonicity: we assume there are no defiers at or near c (or inside the window W if using the local randomization approach).

Note a implication of this is that compliance decisions (i.e., whether one will comply with their treatment assignment or not) do not change abruptly at c.

Example from Bloom (2009, 2012)

Sharp RD:

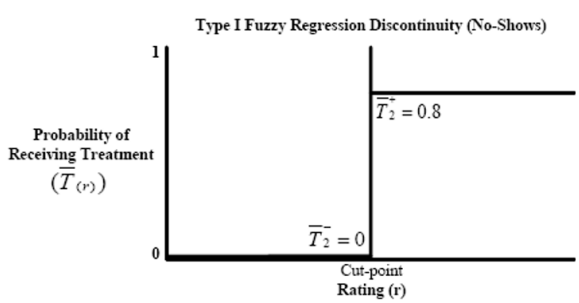


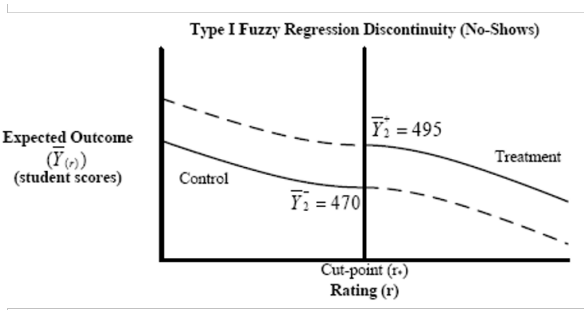


$\tau_{RD} = 500 - 470 = 30$

Example from Bloom (2009, 2012)

Fuzzy RD (one-sided non-compliance):





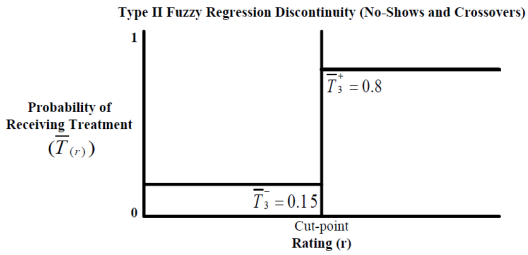
$\tau_{ITT} = 495 - 470 = 25$
$\tau_D = 0.80 - 0 = 0.80$
$\tau_{FRD} = \tau_{ITT}/\tau_D = 25/0.80 = 31.25$

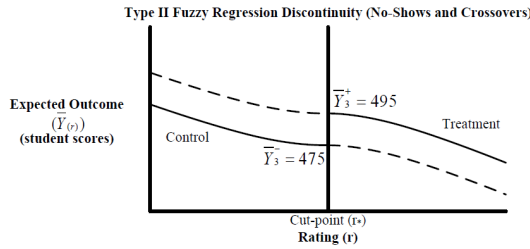
Example from Bloom (2009, 2012)

Above: there are no "always-takers" (else we would see them below c); there are only compliers and never-takers. Above c, it is clear that 80% are compliers and 20% are never-takers. The ITT is a weighted average of τ_{FRD} (for compliers) and 0 (for never-takers). The fuzzy RD parameter "scales up" the ITT to reflect the effect on the compliers.

Example from Bloom (2009, 2012)

Fuzzy RD (two-sided non-compliance):





$\tau_{ITT} = 495 - 475 = 20$
$\tau_D = 0.80 - 0.15 = 0.65$
$\tau_{FRD} = \tau_{ITT}/\tau_D = 20/0.65 = 38.46$

Example from Bloom (2009, 2012)

Above: there are both "always-takers" and "never-takers" (in addition to compliers). *Assuming the proportion of always-takers and never-takers is constant in the neighborhood of c*, we can estimate the never-taker rate to be 20% and always-taker rate to be 15%. The rest (65%) are compliers.

Again, the ITT is a weighted average of τ_{FRD} (for compliers) and 0 (for never-takers). The fuzzy RD parameter "scales up" the ITT to reflect the effect on the compliers.

Note: in practice the proportion of non-compliant cases can vary with the running variable x, but it must be continuous through c.

Fuzzy RD estimation

Since τ_{FRD} is constructed as the ratio of two sharp RD parameters, one could just use `rdrobust` or `rdrandinf` to estimate those parameters. However:

- In `rdrobust`, the two bandwidths may (and probably will) differ. (If interested in the ITTs themselves, it's fine to use differing bandwidths). In `rdrandinf`, this is not an issue since the window is based on pre-determined covariates.
- You will lack a standard error, confidence intervals, p-value for τ_{FRD}

These two commands include a `fuzzy` option that (in the case of `rdrobust`) will choose one optimal bandwidth and report standard errors, etc., for the ratio.

Fuzzy RD: example

Access to financial aid program based on poverty index (SISBEN) in Colombia. First stage:

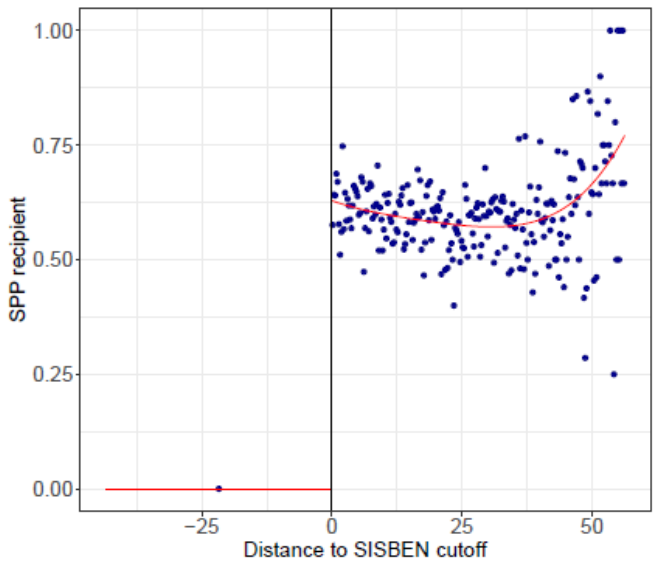


Figure 3.2: RD Plot: First Stage—SPP data

Fuzzy RD: example

Reduced form (intent-to-treat) effect on enrollment in a high-quality college:

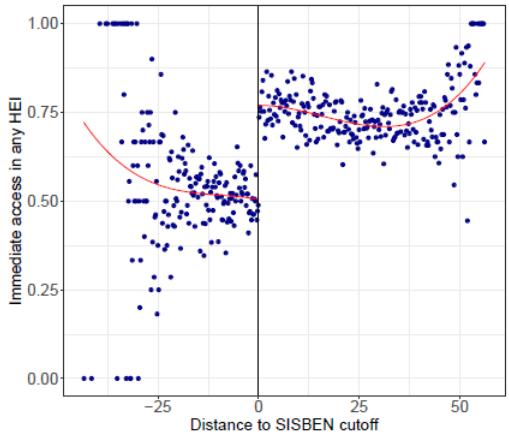


Figure 3.3: RD Plot: Intention-to-treat ($p = 3$)—SPP data

Figure: Cattaneo et al. (2023) *Extensions*

Fuzzy RD: example

`rdrobust` with `fuzzy` option:

R Snippet 3.3

```
> out <- rdrobust(Y, X1, fuzzy = D)
> summary(out)
First-stage estimates.
================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     0.619     0.017    35.857     0.000     [0.585 , 0.653]
        Robust         -         -    29.885     0.000     [0.575 , 0.656]
================================================================================
Treatment effect estimates.
================================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
================================================================================
  Conventional     0.434     0.034    12.773     0.000     [0.368 , 0.501]
        Robust         -         -    11.026     0.000     [0.366 , 0.524]
================================================================================
```

Fuzzy RD validity tests

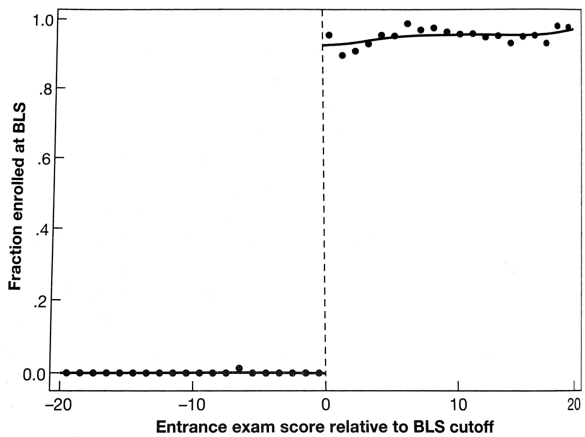
Validity tests should be carried out in the same manner as sharp RD, following the continuity or local randomization approach. Focus on *assignment* to treatment, not actual treatment (which is nonrandom).

The fuzzy RD estimator is an IV estimator, so a **weak instrument** is problematic. If the first stage—how assignment to treatment is related to actual treatment—is small, this can lead to unreliable estimates. Cattaneo et al. (2023) recommend an F-statistic of 20 or higher.

Abdulkadiroğlu, Pathak, & Roth (2014)

This paper used RD to estimate the effect of attending an elite selective high school (e.g., Boston Latin). First stage:

FIGURE 4.6
Enrollment at BLS

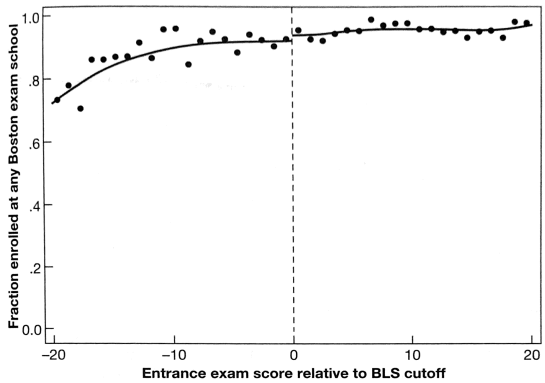


Notes: This figure plots enrollment rates at Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression estimated separately on either side of the cutoff (indicated by the vertical dashed line).

Abdulkadiroğlu, Pathak, & Roth (2014)

Defining "treatment" was difficult in this context since students who failed to qualify for Boston Latin often enrolled in *another* selective school.

FIGURE 4.7
Enrollment at any Boston exam school

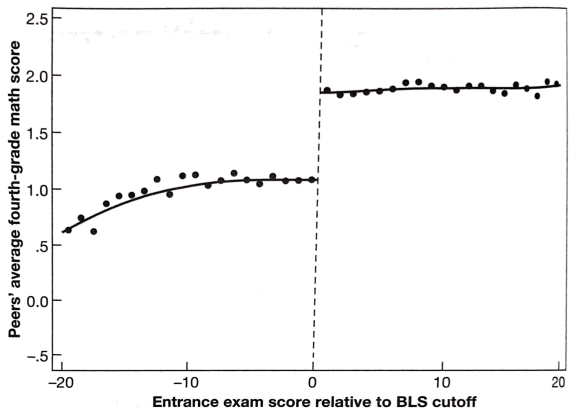


Notes: This figure plots enrollment rates at any Boston exam school, conditional on admissions test scores, for Boston Latin School (BLS) applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

Abdulkadiroğlu, Pathak, & Roth (2014)

Still, admission to Boston Latin was associated with a sharp jump in the average test scores of one's peers:

FIGURE 4.8
Peer quality around the BLS cutoff

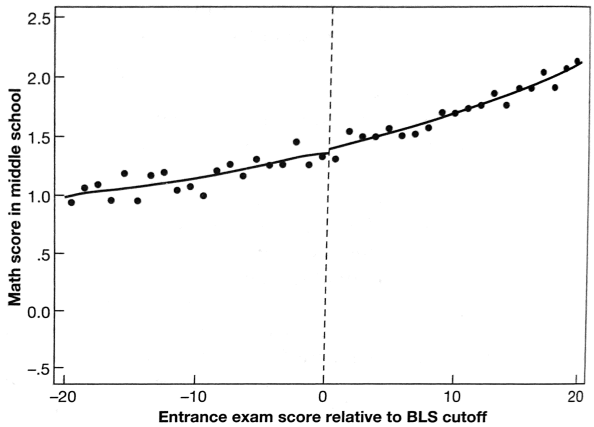


Notes: This figure plots average seventh-grade peer quality for applicants to Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the admissions cutoff. Peer quality is measured by seventh-grade schoolmates' fourth-grade math scores. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

Abdulkadiroğlu, Pathak, & Roth (2014)

The intent-to-treat found no impact of admission to Boston Latin on a wide variety of outcomes:

FIGURE 4.9
Math scores around the BLS cutoff



Notes: This figure plots seventh- and eighth-grade math scores for applicants to the Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the admissions cutoff. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff (indicated by the vertical dashed line).

RD extensions

The regression discontinuity design has received significant attention and is a very active and rapidly developing literature. Some important extensions:

- **Discrete running variables**: when the running variable takes on a finite number of values (perhaps a small number).
- **Multi-dimensional RD designs**:
  - **Multi-cutoff designs**: when treatment is associated with different cutoff values
  - **Multi-score designs**: when treatment depends on 2 or more running variables
- **Geographic RD designs**: a type of multi-score design where treatment depends on being on one side of a geographic boundary

I will briefly introduce the issues that arise with these; see Cattaneo et al. (2023) for details.

Discrete running variables

Some running variables are discrete in that they take on a finite number of values. This results in **mass points** (values of x where the density is concentrated). Examples: some test scores, age, date of birth.

The first step in this case is to determine how many unique mass points you have, and how many observations per mass point.

Discrete running variables

Issues with discrete running variables:

- The continuity framework—where we fit a local polynomial to estimate the treatment effect at c—may not be appropriate.
- Local polynomial estimation will behave as if each mass point is a single observation. So, the effective number of observations is smaller. Collapsing the data to one observation per mass point will often yield similar results.
- Density tests for manipulation may be more likely to fail.

If the number of mass points is large—particularly near c—then the local polynomial approach may provide a good approximation. If not, Cattaneo et al. recommend the local randomization framework.

Discrete running variables

Why the continuity framework may not be appropriate:

- Consider 5 mass points: $[-2, -1, 0, 1, 2]$ with $c = 0$
- The set of control observations will never get "close enough" to 0 to infer an effect at $x = 0$
- Since there is no density just to the left of c, the estimate is not a treatment effect at c

Whether -1 is "close enough" to 0 in the above example depends on its scale. The running variable could be in days or years, for example.

Discrete running variables

Local randomization can help, since the inference is now to units in the neighborhood of c (i.e., within the window w).

- Note this changes one's interpretation of the treatment effect (it's no longer the treatment effect at c, but in the window).
- The smallest possible window is often obvious: from the first discrete point to the left of c to c. If more observations are needed, can use covariate-based window selection to increase the window width.

Discrete running variables

Earlier RD literature recommended adjusting standard errors for clustering on individual values of the running variable when the running variable is discrete (Lee & Card, 2008; Lee & Lemieux, 2010).

However, more recent papers strongly advise against this. Kolesár & Rothe (2018) proposed "honest" confidence intervals when the running variable is discrete. See Github page for a link to the Stata package `rdhonest`.

Discrete running variables and heaping

One issue that may arise with discrete running variables is *heaping*. This often occurs with self-reported data (e.g., rounding) and discrete variables with limited precision. Non-random heaping can result in bias.

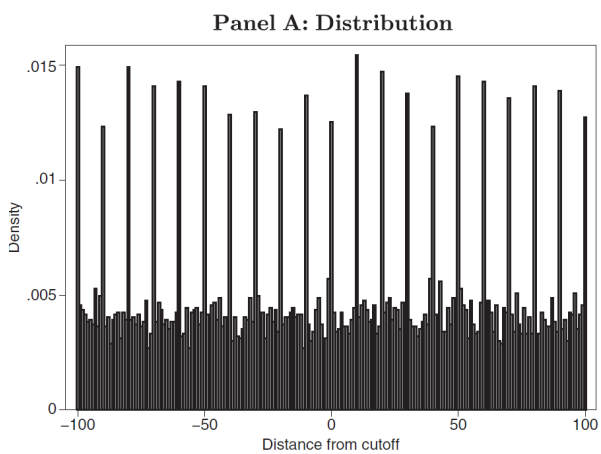


Figure: Barreca, Lindo, & Waddell, 2016

Discrete running variables and heaping

Example: birthweight. Not all hospitals have the most modern scales. Infants at the "heaps" (rounded) are more likely to be non-white.

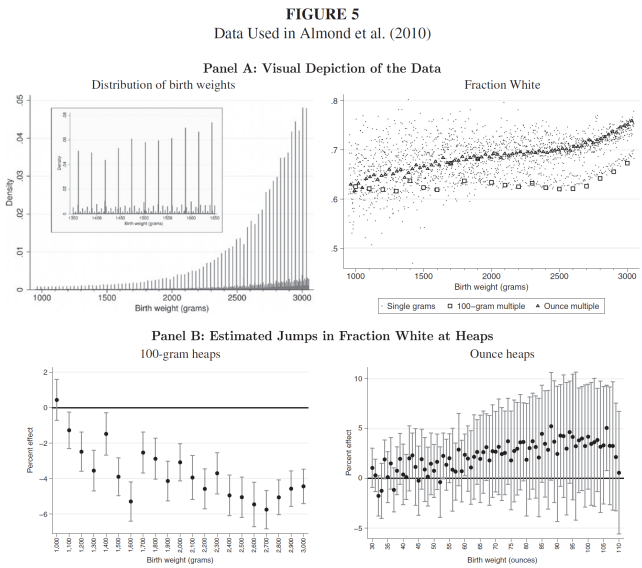


Figure: Barreca, Lindo, & Waddell, 2016

Discrete running variables and heaping

Barreca, Lindo, & Waddell (2016) offer some diagnostic tests for non-random heaping and some recommendations of solutions when heaping is non-random.

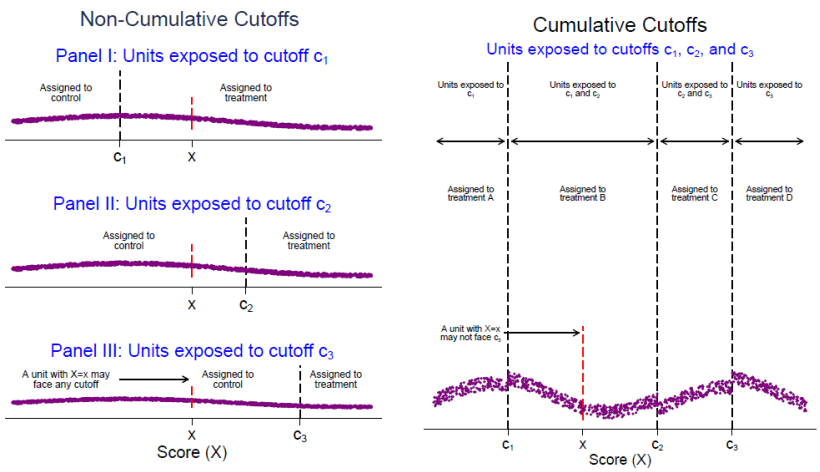
Multi-cutoff RD

A **multi-cutoff RD design** is when treatment is associated with different (multiple) cutoff values. There are two basic types:

- Non-cumulative: typically when different subpopulations are subject to different cutoffs (e.g., different regions have different program eligibility). Often, the same treatment is given at different cutoffs.
- Cumulative: units may be eligible for additional treatments at higher cutoffs (e.g., eligible for a higher "dose" of the treatment at higher c)

Multi-cutoff RD

Non-cumulative vs. cumulative cutoffs:



(a) Non-cumulative Cutoffs (b) Cumulative Cutoffs

Figure 5.1: Cumulative vs. Non-cumulative Cutoffs in Multi-Cutoff RD Design

Figure: Cattaneo et al. (2023) *Extensions*

Multi-cutoff RD: issues

Main implications/considerations:

- Potential outcomes of affected units likely differ at different cutpoints, and treatment effects may differ at different c. Have to think about how you are defining the treatment effect, and what treatment effect is of interest to you.
- With cumulative cutoffs, there is a lack of common support in x for units facing different cutoffs.
- With cumulative cutoffs, it is ambiguous who is "exposed" to each treatment threshold. Some units will be treated for some cutpoints but untreated for others.

Multi-cutoff RD: approaches

There are different approaches to RD settings with multiple cutoffs:

- Estimate **cutoff-specific effects**. Simply estimate RDs separately at different cutoffs. (With cumulative cutoffs, may have to make a decision about which units are exposed to a given cutoff. Also, effects are *cumulative*—i.e., relative to the previous cutoff).
- **Normalizing and pooling**: normalize the running variable such that $x = 0$ at each cutoff. Produces one point estimate that is equal to a weighted average of the cutoff-specific effects.
- **Extrapolation**: estimating treatment effects away from c (requires additional assumptions).

Multi-cutoff RD: cutoff-specific effects

Estimating RD effects at c_1 and c_2 in non-cumulative case:

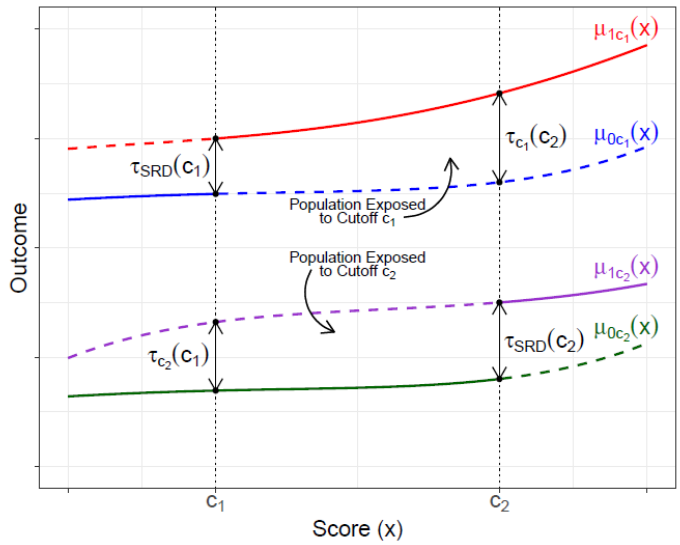


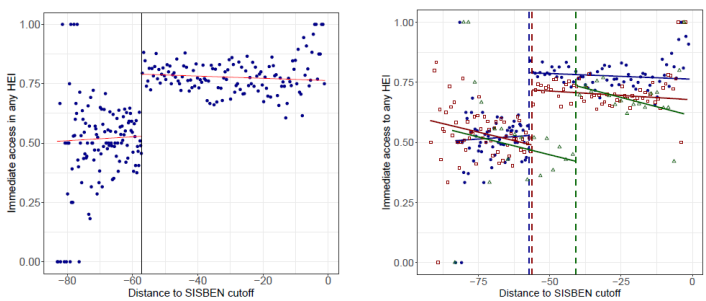
Figure 5.2: Multi-Cutoff RD Design with Two Non-cumulative Cutoffs

Figure: Cattaneo et al. (2023) *Extensions*

`rdmc` and `rdmcplot`

Cattaneo et al. developed the `rdmulti` package for Stata which includes `rdmc` for estimation and `rdmcplot` for plotting.

- `rdmc` will estimate RD effects (using `rdrobust`) at multiple cutoffs, as well as provide a weighted and pooled estimate.
- `rdmcplot` produces nice RD plots when there are multiple cutoffs.



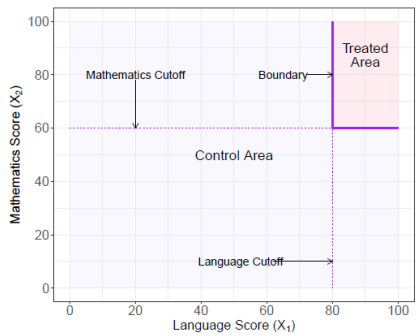
(a) Cutoff $C_i = 57.21$ (J_-^{MV}, J_+^{MV}, $p = 1$) (b) All Cutoffs ($J_-^{MV}/2$, $J_+^{MV}/2$, $p = 1$)

Figure 5.4: RD Plots—SPP data

Multi-score RD

A **multi-score RD design** is when treatment depends on 2 or more running variables. Example: must achieve above a minimum score on two subject-area exams to graduate high school.

Issue: the presence of two or more running variables and thresholds means there are multiple ways in which a unit can be on the margin of treatment. This creates a **boundary** for treatment.



(a) Hard-threshold Assignment

Multi-score RD

"Analogously to the multi-cutoff RD, the parameters of interest in the multi-score RD design change because there is no longer a single cutoff at which the probability of treatment assignment changes discontinuously; instead the probability of treatment assignment changes discontinuously at an often uncountable collection of locations along the boundary" (Cattaneo et al., 2023).

As in the multi-cutoff RD case, potential outcomes and treatment effects may differ at different locations along the boundary. How (if at all) should these be combined?

Multi-score RD: approaches

There are different approaches to RD settings with multiple running variables:

- **Estimate location-specific effects**: choose specific locations along the boundary to focus on.
- **Collapse scores to one dimension**: for example, calculate the shortest distance of each unit to the boundary and then pool all observations. (Like normalizing and pooling).

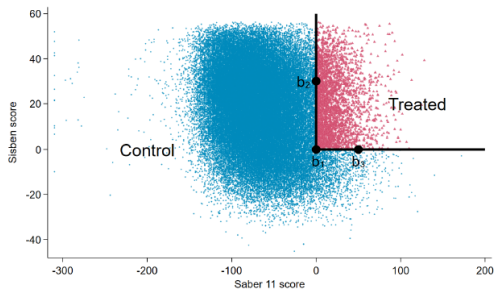


Figure 5.6: Treated and Control Regions—SPP data

`rdms`

The Stata command `rdms` is part of the `rdmulti` package and can be used for estimation of RDs when there are multiple running variables.

- `rdms` will estimate RD effects (using `rdrobust`) and can provide a pooled estimate.

Multi-score RD: approaches

See Reardon & Robinson (2012) and Wong et al. (2013) for a full discussion of multi-score RD in an education context. They contrast five different approaches:

- Response surface RD: fit a polynomial in multiple dimensions in a single regression model.
- Frontier RD: subset the data on all but one of the running variables and the model the discontinuity on that one dimension.
- Fuzzy frontier RD: uses IV to estimate RD effects.
- Distance-based RD: collapse to one dimension, and use the shortest distance to the boundary.
- Binding-score RD: focus on the minimum of the running variables, which perfectly determines assignment.

Papay, Murnane, & Willett (2014) is another often-cited example of multi-score RD in an education context. They examine high school exit exams in Massachusetts (passing two is required).

Geographic RD

A spatial or **geographic RD design** is one in which treatment depends on being on one side of a geographic boundary. It is a type of multi-score design where the scores are the unit's latitude and longtiude.

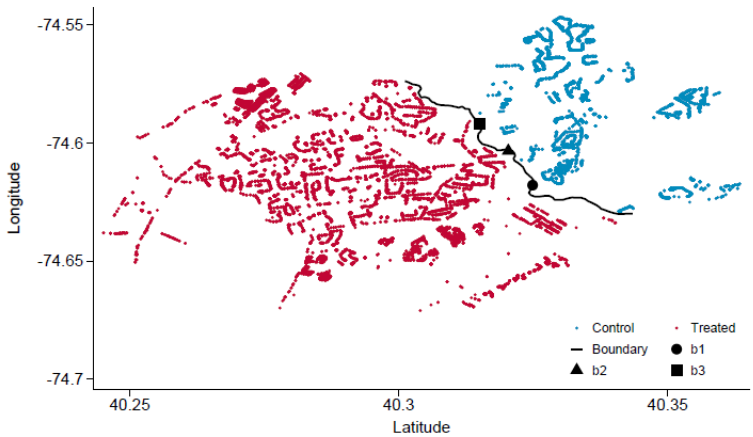


Figure 5.7: Treated and Control Geographic Areas—Media Market data

Figure: Cattaneo et al. (2023) *Extensions*. From Keele & Titunik (2015) on exposure to swing-state political advertisements (PA vs. NJ).

Geographic RD

A canonical paper using a boundary design is Black (1999), who estimated the housing price premium associated with better-performing schools. She compared homes within the same school district but within a narrow band on either side of an attendance zone boundary:

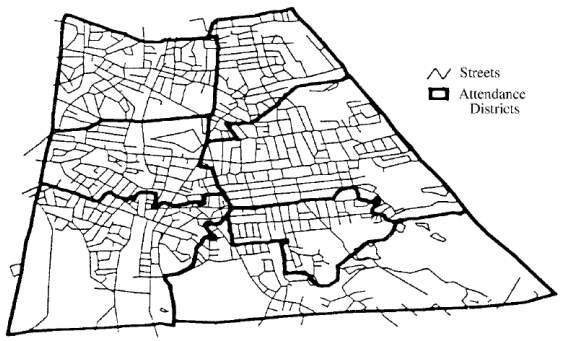


FIGURE I
Example of Data Collection for One City: Melrose Streets, and Attendance District Boundaries

She found housing prices were 2.5% higher for every 5% increase in test scores.

Geographic RD

Issues/considerations: similar to the multi-score design.

- Potential outcomes and treatment effects may differ at different boundary points. Have to think about how you are combining these estimates (if at all).
- Some boundary points may lack density; it makes good sense to focus on boundary points where there is a lot of mass.
- In many geographic applications, there may be a lack of density at the boundary (e.g., industrial areas, rivers, parks).

Additional resources and references

See Github for lots of additional resources and sample studies.

- Jacob et al. (2012) *A Practical Guide to Regression Discontinuity*
- What Works Clearinghouse Handbook (2020) - standards for regression discontinuity
- Power calculations in RD (Schochet, 2009; Deke and Dragoset, 2012)