8. Regression discontinuity

LPO 8852: Regression II

Sean P. Corcoran

RD - introduction

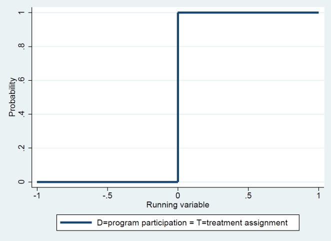
RD can be used when a **precise** rule based on a **continuous** characteristic determines treatment assignment. Examples:

- **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

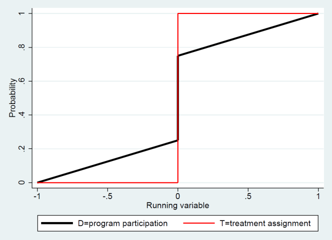
Sharp RD: treatment *assignment* goes from 0 → 1 at a threshold c. Treatment *receipt* goes from 0% to 100% at c (full compliance).



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

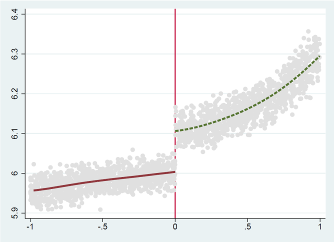
RD - introduction

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



RD - introduction

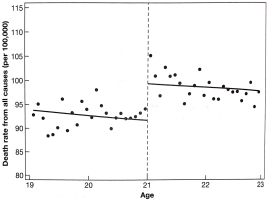
If there is a discrete change in treatment and program participation at c (and the program has a treatment effect) one would expect to see a discrete change in the mean outcome at c.



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:

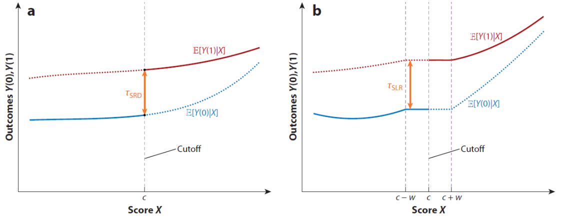
$$Y_i = (1 - D_i)Y_i(0) + D_i Y_i(1) = \begin{cases} Y_i(0) & \text{if } X_i < c \\ Y_i(1) & \text{if } X_i \geq c \end{cases}$$

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 total 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}.

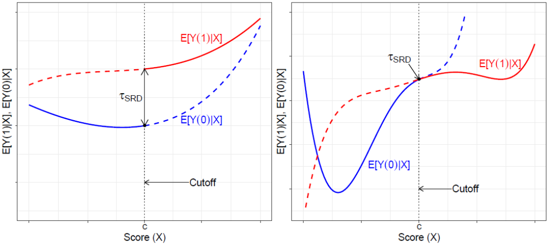
Under the **local randomization assumption**, treatment assignment—being above or below c—is random within a neighborhood ($\pm w$) of c. Mean potential outcomes do not vary within this neighborhood.

We will focus first 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

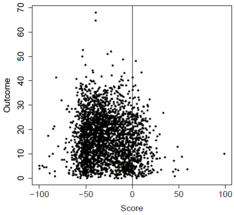


Figure 3.1: Scatter Plot—Meyersson Data

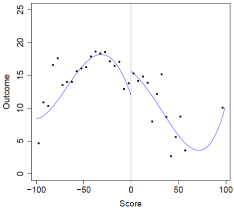


Figure 3.2: RD Plot for Meyersson Data Using 40 Bins of Equal Length

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)

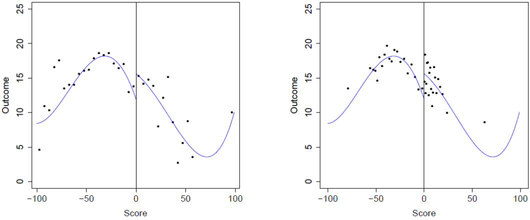
- All bins are equal width
- Number of observations per bin varies (thus, precision varies across bins)

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

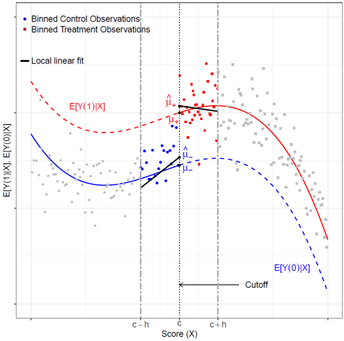


Figure: Cattaneo et al. (2020) *Foundations*

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

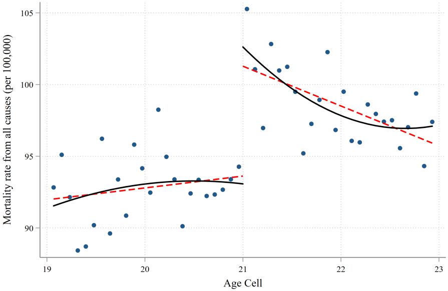
As an example, let $p = 2$ (quadratic):

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

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

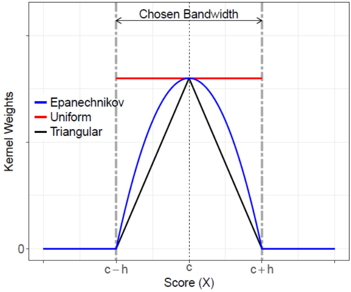


Figure: Cattaneo et al. (2020) *Foundations*

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

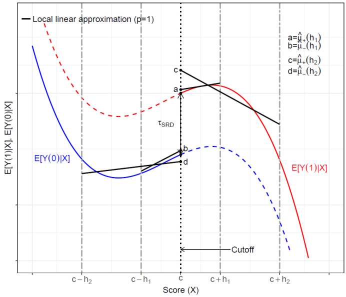


Figure: Cattaneo et al. (2020) *Foundations*

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 acknowledges 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

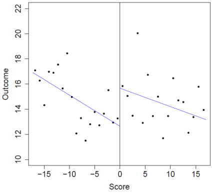
Optimal bandwidth selection

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 much 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.

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

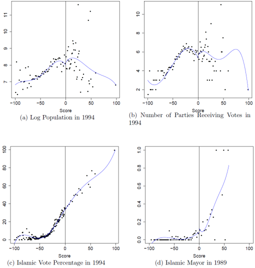
The following are commonly performed as validity tests with RD:

- 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. Can also show RD plots:

Figure 5.1: RD Plots for Predetermined Covariates—Meyersson Application



(a) Log Population in 1994

(b) Number of Parties Receiving Votes in 1994

(c) Islamic Vote Percentage in 1994

(d) Islamic Mayor in 1989

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

Manipulation test

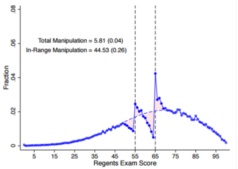
Manipulation occurs whenever units have their value of X altered in order to affect their treatment status. For example, a teacher might adjust a test score in order to help a student pass or become eligible for a program.

- 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 <u>not</u> the same thing as non-compliance (the "fuzzy RD" case).

Manipulation test

Sometimes manipulation is clear from inspecting densities or histograms:

Figure 1
Test Score Distributions for Core Regents Exams, 2004-2010

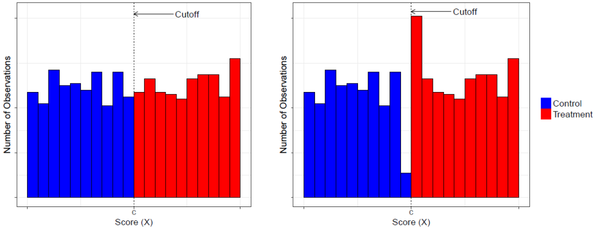


Notes: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in 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 5.3: Histogram of Score



(a) No Sorting

(b) Sorting

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.

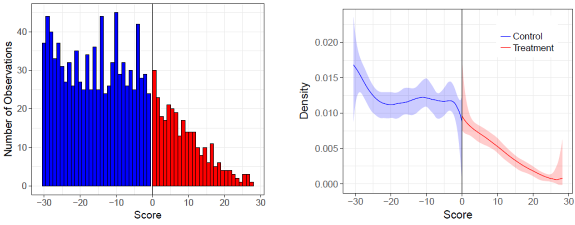
```
rddensity x, plot otheroptions
```

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



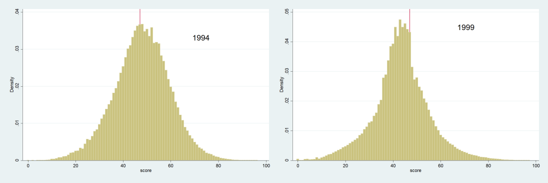
(a) Histogram

(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

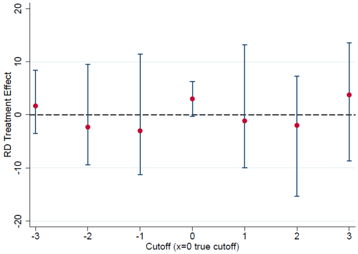


Figure: Cattaneo et al. (2020) *Foundations*

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

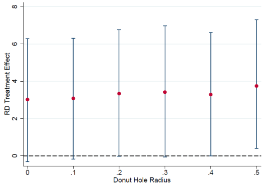


Figure: Cattaneo et al. (2020) *Foundations*

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.