7. Regression discontinuity designs

LPO.7870: Research Design and Data Analysis II

Sean P. Corcoran

Last time

Experimental and quasi-experimental methods: research designs for causal inference.

- Causal effects and counterfactuals
- Potential outcomes and treatment effects
- Average treatment effects (ATEs)
- Difficulty in estimating ATEs when there are differences in mean potential outcomes between "treated" and "untreated" cases (aka selection bias or omitted variables bias).

Last time

Randomized controlled trials (RCTs):

- Treatment status is *independent* of potential outcomes so "treated" and "untreated" groups should be balanced, on average (no OVB).
- Randomization implies baseline equivalence (researchers should show this to the best of their ability).
- Estimating ATE in an RCT with regression (with and without covariates)

Last time

Evaluating research designs

- Internal validity and threats to internal validity (e.g., in an RCT: partial compliance, attrition).
- External validity and threats to external validity (e.g., nonrepresentative sample, scale-up effects)

Last time

Quasi-experiments: units not assigned to treatment conditions by the researcher, but through variation in individual conditions that are "as good as random" (e.g., natural experiments). A tour of methods:

- Regression discontinuity
- Difference-in-differences
- Panel methods
- Instrumental variables

Tonight

Continuing from Lecture 6: an in-class exercise using the Tennessee STAR data (a RCT).

Regression discontinuity: when assignment to treatment is based on a precise rule using a continuous characteristic (e.g., test score cutoff).

Tonight's sample datasets

We will refer to two datasets tonight (on Github):

1. `STARexample.dta`: select variables from the Tennessee STAR class size experiment.
2. `AEJfigs.dta`: RD example from Mastering Metrics (based on Carpenter & Dobkin, 2009).

Introduction to RD

RD - introduction

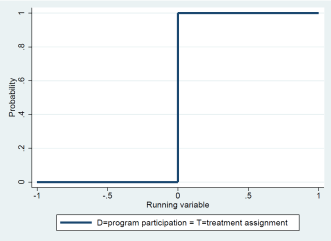
RD can be used when a **precise** rule based on a **continuous** characteristic determines treatment assignment. Examples:

- **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 (PK or KG).
- **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

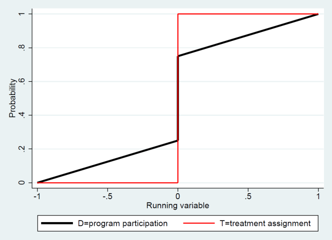
Sharp RD: treatment *assignment* goes from $0 \rightarrow 1$ at a threshold c. Treatment *receipt* goes from 0% to 100% at c (full compliance).



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

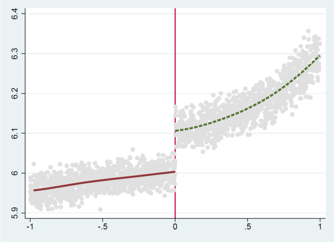
RD - introduction

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

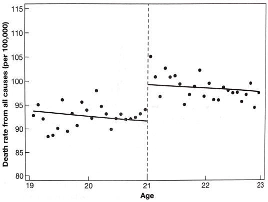
If there is a discrete change in treatment and program participation at c (and the program has a treatment effect) one would expect to see a discrete change in the mean outcome at c.



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

RD - introduction

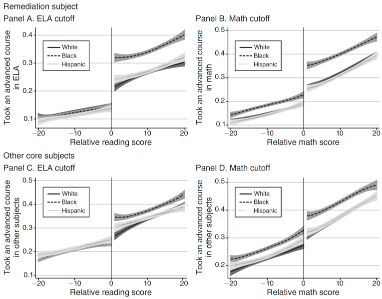
RD typically relies on one of two assumptions, whichever is applicable to the particular study:

- **Continuity assumption**: mean potential outcomes are continuous near c in the absence of any treatment. As long as we have modeled the relationship between Y and the running variable X correctly, the "jump" is the causal effect.

- **Local randomization assumption**: treatment assignment—being above or below c—is random within some neighborhood of c. In other words, mean potential outcomes are balanced on either side of c.

RD - introduction

Figlio & Ozek (2024) looked at a policy in Florida that placed middle school students into remedial classes based on their test score.



Sharp RD estimation

Sharp RD estimation

Suppose the RD is *sharp*—that is, everyone above c is treated. Also, suppose the *continuity assumption* applies. The goal is to model the relationship between the outcome Y and the running variable X, both below and above c. Decisions to make:

1. Should the model be linear? Quadratic? Some other polynomial?
2. Should the slope(s) be the same on the left and right of c?
3. What bandwidth should we use, if any?
4. Should other covariates be included in the regression?

An RD **bandwidth** is a limitation on the data used in the analysis. For example, we may limit the observations to those within a certain distance from c.

Sharp RD estimation

Linear regression: same slope on either side of c. Assume X has been re-centered, so that $X = 0$ at c. Define $D = 1$ if $X > 0$ (treated).

$$Y = \beta_0 + \beta_1 X + \beta_2 D + u$$

This model has an intercept shift at c ($X = 0$).

Sharp RD estimation

Linear regression: *different* slopes on either side of c.

$$Y = \beta_0 + \beta_1 X + \beta_2 D + \beta_3 (D \times X) + u$$

This model has an intercept shift at c, and the slope is allowed to change when $X > 0$. (There is an interaction between X and D).

Sharp RD estimation

Quadratic regression: same slopes on either side of c.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 D + u$$

This model has an intercept shift at c ($X = 0$).

Sharp RD estimation

Quadratic regression: *different* slopes on either side of c.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 D + \beta_4 (D \times X) + \beta_5 (D \times X^2) + u$$

This model has an intercept shift at c, and the slopes are allowed to change when $X > 0$.

...and so on.

Sharp RD estimation

There are lots of options. Which model specification should you choose?

- The goal of this analysis is to estimate the "jump" at c as accurately as possible.
- A higher-order polynomial provides a better fit to the data, but may come at a cost of poorly estimating the endpoints.
- Allowing slopes on either side of c is more flexible, putting fewer constraints on the model.

The current state of the art recommends using a linear or quadratic model (usually no higher terms), with different slopes on either side of c. It also recommends limiting the analysis to a narrow bandwidth. (There are procedures available for finding the "optimal" bandwidth).

Example RD application

Example: Carpenter & Dobkin 2009

In the United States, young adults can legally consume alcohol when they turn 21. What is the effect of legal access to alcohol on traffic fatalities? Example based on *Mastering Metrics* chapter 4.

- What is the outcome variable of interest?
- What is the *running variable*?
- Where would we expect to see a discontinuity, if there were a treatment effect?

Example: Carpenter & Dobkin 2009

The dataset *AEJfigs.dta* can be found on Github.

- Each observation is an *age cell*. Persons age 19-23 have been divided into 50 equal-width cells.
- For each age cell, we have mortality rates (deaths per 100,000) for all causes, and for different reasons (e.g., motor vehicle accidents).
- For purposes of the RD, create a re-centered *age* at the legal drinking age (let *age* = *agecell* - 21).
- Reaching legal drinking age is the "treatment." The discontinuity in treatment is sharp. Create a treatment variable *over21* = *age* $\geq$ 0)
- We might predict a discontinuity in mortality rates at 21 (*age* = 0).

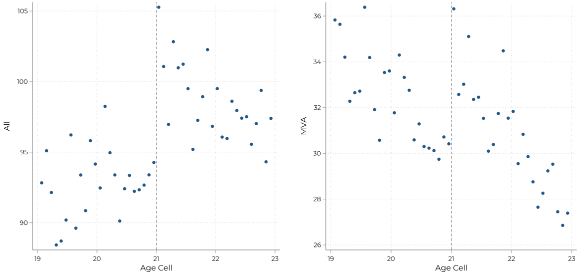
Example: Carpenter & Dobkin 2009

Do the following:

1. Create a scatterplot of all-causes fatality vs. *age*. Do the same for motor vehicle fatalities. Do you see evidence of a discontinuity? What does the shape of the relationship between Y and X look like? Does it differ on either side of c?
2. Estimate the RD for all-causes and motor vehicle fatalities, assuming a *linear* model and the same slope on both sides of c. Interpret the coefficients.
3. Repeat, but allow the slopes to vary on either side of c.
4. Estimate the RD for all-causes and motor vehicle fatalities, assuming a *quadratic* model and the same slope on both sides of c.
5. Repeat, but allow the slopes to vary on either side of c.

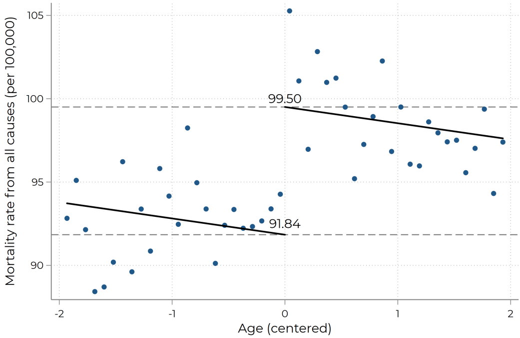
See the do-file on Github for an example of how I graphed the results.

Example: Carpenter & Dobkin 2009



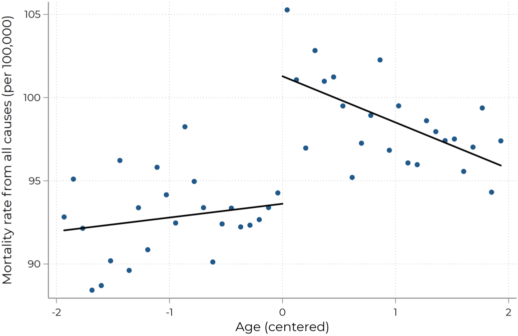
Note: used *agecell* for x-axis rather than *age* centered here

Example: Carpenter & Dobkin 2009



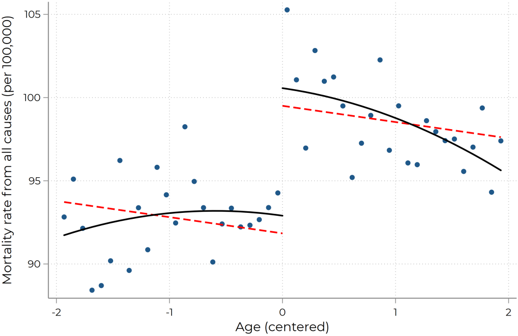
Intercept from the left: 91.84. From the right: 91.84 + **7.66** = 99.50

Example: Carpenter & Dobkin 2009



Intercept from the left: 93.618. From the right: 93.618+7.66 = 101.278.

Example: Carpenter & Dobkin 2009



RD validity

RD assumptions

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

- Treatment assignment (and receipt) occurs at a known threshold c.
- The relationship between potential outcomes 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:

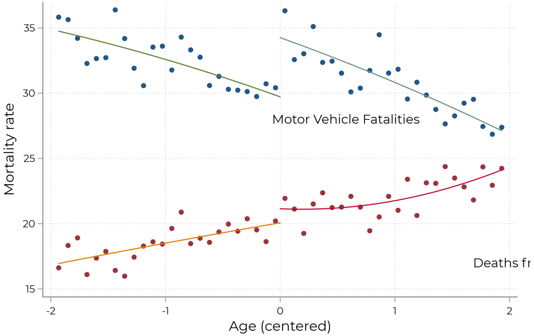
- Test for effects at c on **pre-treatment covariates** or **placebo outcomes**. Would not expect any discontinuities here.
- Test for **continuity in the distribution** of the running variable around c ("manipulation test").
- Test for discontinuities elsewhere in the distribution of X (i.e., artificial cutoffs). A "smoothness" test.
- Exclusion of observations near c (**"donut hole"** approach).
- Sensitivity tests for bandwidth choice.

Example: Carpenter & Dobkin 2009

What kinds of validity checks might we do with Carpenter & Dobkin?

- Given the nature of the running variable in this analysis (age), we don't expect manipulation.
- If we had population covariates by age, we could test for discontinuities at those in age 21. But again, would not expect to see anything here.
- Placebo tests: estimate the RD for an outcome that we *don't* expect to see an "effect" on. Try mortality by internal causes.

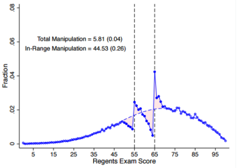
Example: Carpenter & Dobkin 2009



Evidence of manipulation

Sometimes manipulation is clear from inspecting densities or histograms:

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

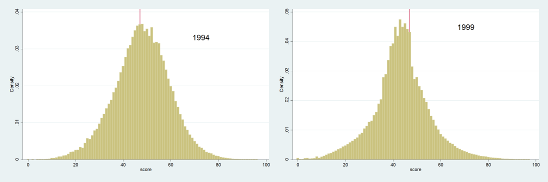


Notes: This figure shows the test score distribution around the 55 and 65 score cutoffs for New York City high school test takers between 2004-2010. Core exams include English Language Arts, Global History, U.S. History, Math A/Integrated Algebra, and Living Environment. We include the first test in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

Evidence of manipulation

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



There is evident manipulation in 1999. Figure: Camacho and Conover (2011).

RD internal vs. external validity

If the main RD assumptions hold, RD has high internal validity. One can make strong claims to causal inference.

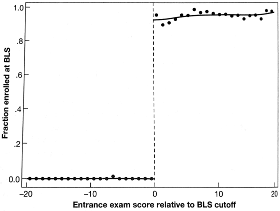
However, RD estimates are *local* to the threshold. They generally have low external validity. It's hard to extrapolate beyond the cut point.

Example papers

Abdulkadiroğlu, Pathak, & Roth (2014)

This paper used RD to estimate the effect of attending an elite selective high school in Boston or NYC (e.g., Boston Latin). This was a "fuzzy" RD since not all students who were admitted enrolled:

Figure 4.6
Enrollment at BLS

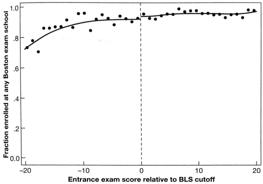


Notes: This figure plots enrollment rates at Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression estimated separately on either side of the cutoff (indicated by the vertical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

Abdulkadiroğlu, Pathak, & Roth (2014)

Also, 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

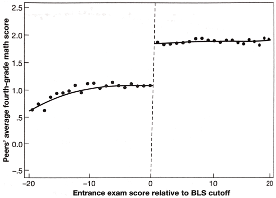


Notes: This figure plots enrollment rates at any Boston exam school, conditional on admissions test scores, for Boston Latin School (BLS) applicants scoring near the BLS admissions cutoff. Solid lines show fitted values from a local linear regression, estimated separately on either side of the cutoff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

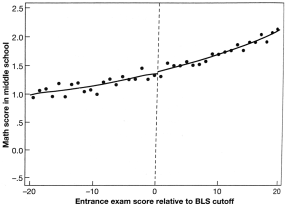


Notes: This figure plots average seventh-grade peer quality for applicants to Boston Latin School (BLS), conditional on admissions test scores, for BLS applicants scoring near the admissions cutoff. Peer quality is measured by seventh-grade schoolmates' fourth-grade math scores. Solid lines show fitted 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. (ITT since there was not perfect compliance):

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).

Example: Dee & Wyckoff 2015

This paper examined whether a new teacher evaluation system in D.C. (IMPACT) affected retention and subsequent performance. IMPACT assigned an overall evaluation score, with several consequential thresholds:

- Dismissal threat: minimally effective vs. effective
- Financial incentives: highly effective vs. effective

Example: Dee & Wyckoff 2015

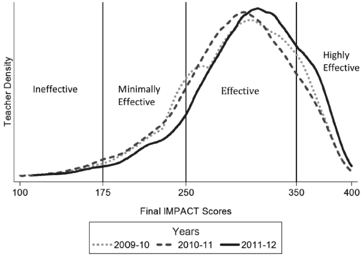
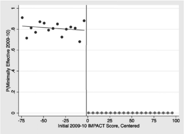


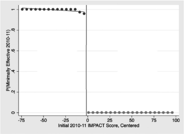
Figure 1. Distribution of IMPACT Scores, AY 2009–10 through AY 2011–12.

Example: Dee & Wyckoff 2015

Treatment: classification as "minimally effective." Note the score has been re-centered.



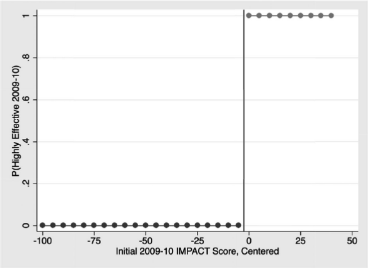
(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010-11

Example: Dee & Wyckoff 2015

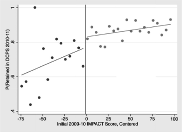
Treatment: classification as "highly effective"



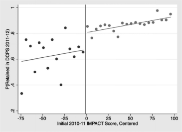
(c) Highly Effective, AY 2009-10

Example: Dee & Wyckoff 2015

Outcome: retention in next year.



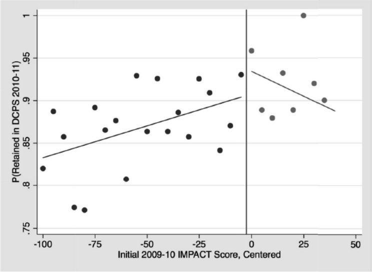
(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010 to 2011

Example: Dee & Wyckoff 2015

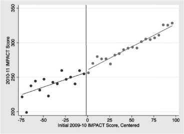
Outcome: retention in next year.



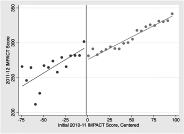
(c) Highly Effective, AY 2009-10

Example: Dee & Wyckoff 2015

Outcome: subsequent IMPACT score.



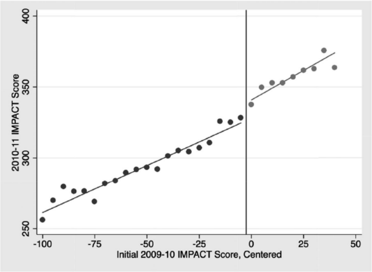
(a) Minimally Effective, AY 2009-10



(b) Minimally Effective, AY 2010-11

Example: Dee & Wyckoff 2015

Outcome: subsequent IMPACT score.



(c) Highly Effective, AY 2009-10

Example: Dee & Wyckoff 2015

RD estimates:

Table 4. Reduced-form RD estimates, minimally and HE ITT.

	Dependent variable					
	Retained in DCPS, year $t+1$			IMPACT score, year $t+1$		
Sample	(1)	(2)	(3)	(4)	(5)	(6)
	Independent variable: ME ITT					
Full sample	-0.0915^{***}	-0.0675^{**}	-0.0730^{**}	5.841	5.793	4.146
	(0.0318)	(0.0291)	(0.0294)	(3.736)	(3.657)	(3.652)
AY 2009–10	−0.0603	−0.0345	−0.0414	−3.233	−2.200	−2.595
	(0.0423)	(0.0390)	(0.0392)	(5.033)	(4.925)	(4.790)
AY 2010–11	-0.132^{***}	-0.112^{***}	-0.112^{***}	18.35^{***}	16.37^{***}	12.60^{**}
	(0.0481)	(0.0432)	(0.0426)	(5.334)	(5.296)	(5.229)
	Independent variable: HE ITT					
AY 2009–10	0.0263	0.0298	0.0264	12.87^{***}	12.87^{***}	10.93^{***}
	(0.0275)	(0.0236)	(0.0245)	(2.914)	(2.882)	(2.760)
Teacher controls	no	yes	Yes	no	yes	yes
School-fixed effects	no	no	Yes	no	no	yes

Notes: $^{***}P < 0.01$; $^{**}P < 0.05$; $^{*}P < 0.1$. Robust standard errors in parentheses. All models condition on a linear spline of the assignment variable.

Example: Dee & Wyckoff 2015

Dee and Wyckoff performed lots of validity checks:

- Test for manipulation: did teachers (or school administrators) manipulate their scores?
- Discontinuities in baseline characteristics: did observable characteristics also "jump" at the cut point(s)?
- Placebo tests: were there discontinuities at *other* thresholds?

Next time

Difference-in-differences designs

- See S&W chapter 6
- *Mastering Metrics* chapter 5 is also useful
- Example study: Gamoran & An (2016) on school segregation