Causal Inference

Mixtape Session

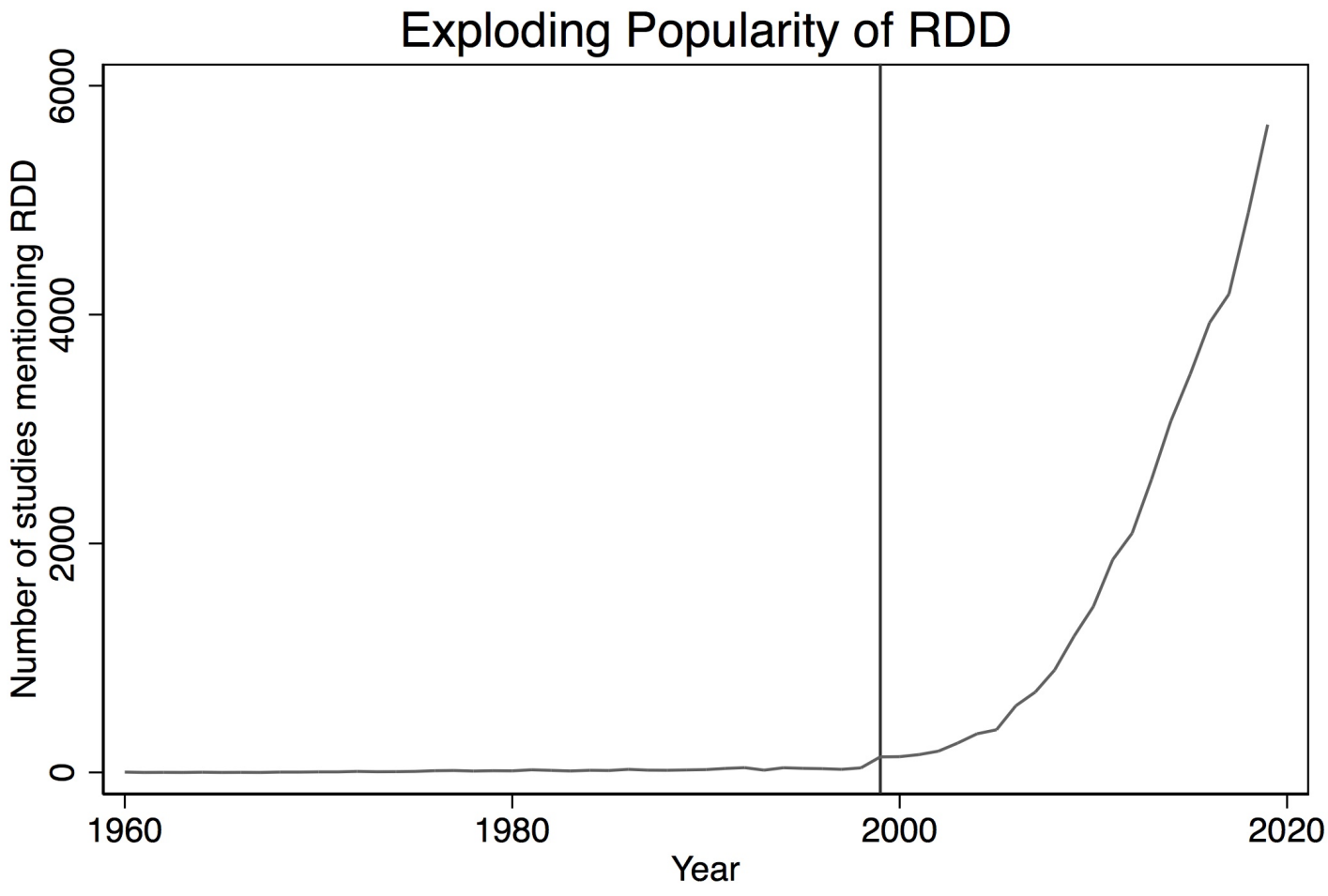
Roadmap

- Regression discontinuity designs
  - Introduction
  - Sharp Design
  - Smoothness, Extrapolation and Estimators
    - Smoothness assumption
    - Nonlinearities
  - Testing for violations
  - Visualization
  - Inference, kernels, bandwidths
  - Sub-RDD: Close election designs

What is regression discontinuity?

RDD is an extremely popular particular type of research design. Sometimes considered the “queen” of the quasi-experimental designs.

- Cook (2008) has a fascinating history of thought on how and why – says it doesn’t show up because we weren’t ready for it
- Donald Campbell, educational psychologist, invented regression discontinuity design (Thistlethwaite and Campbell, 1960), but then it went dormant for decades (Cook 2008).
- Angrist and Lavy (1999) and Black (1999) independently rediscover it. It’s become incredibly popular in economics



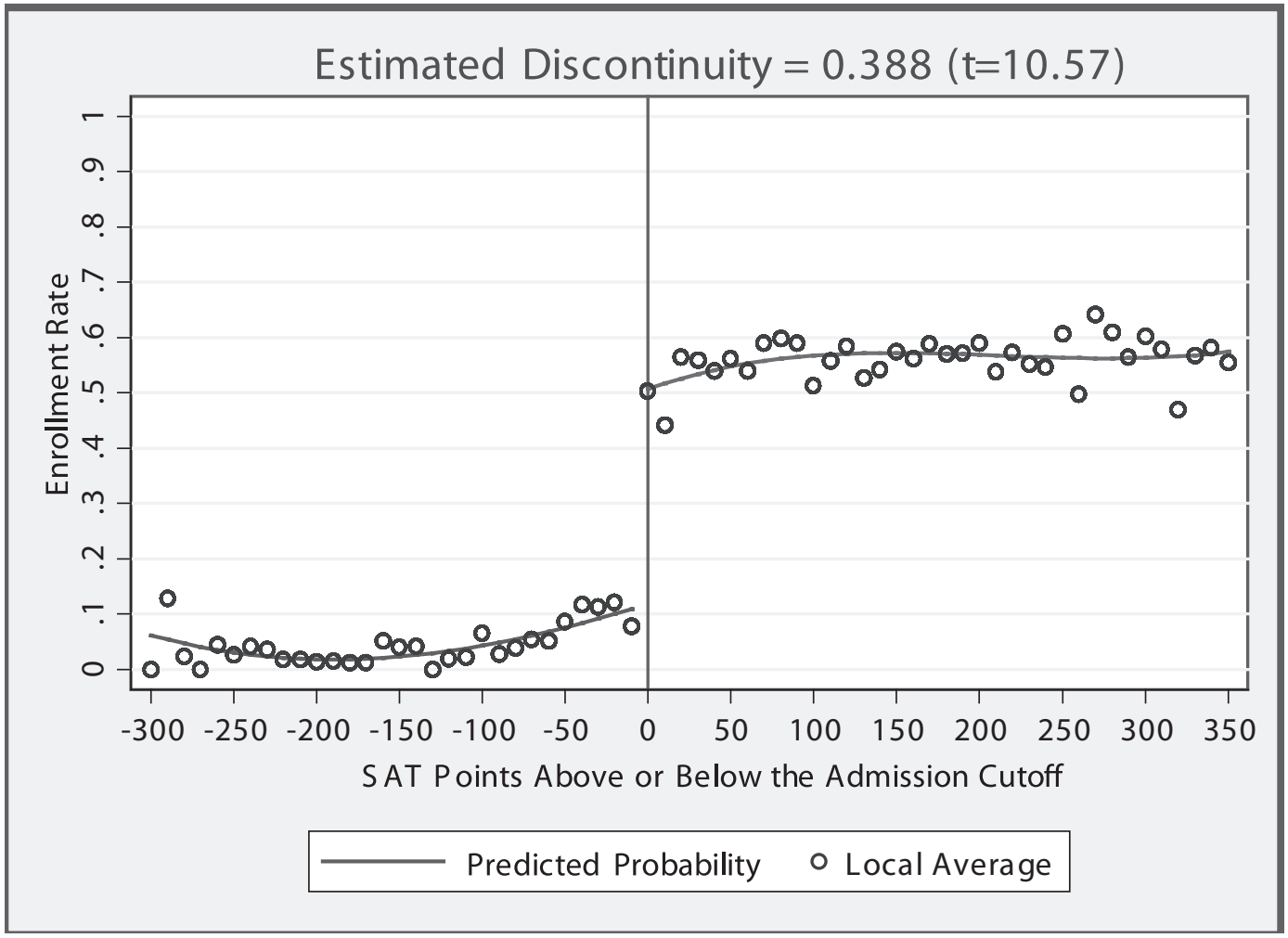
Vertical bar is Angrist and Lavy (1999) and Black (1999)

What is a regression discontinuity design?

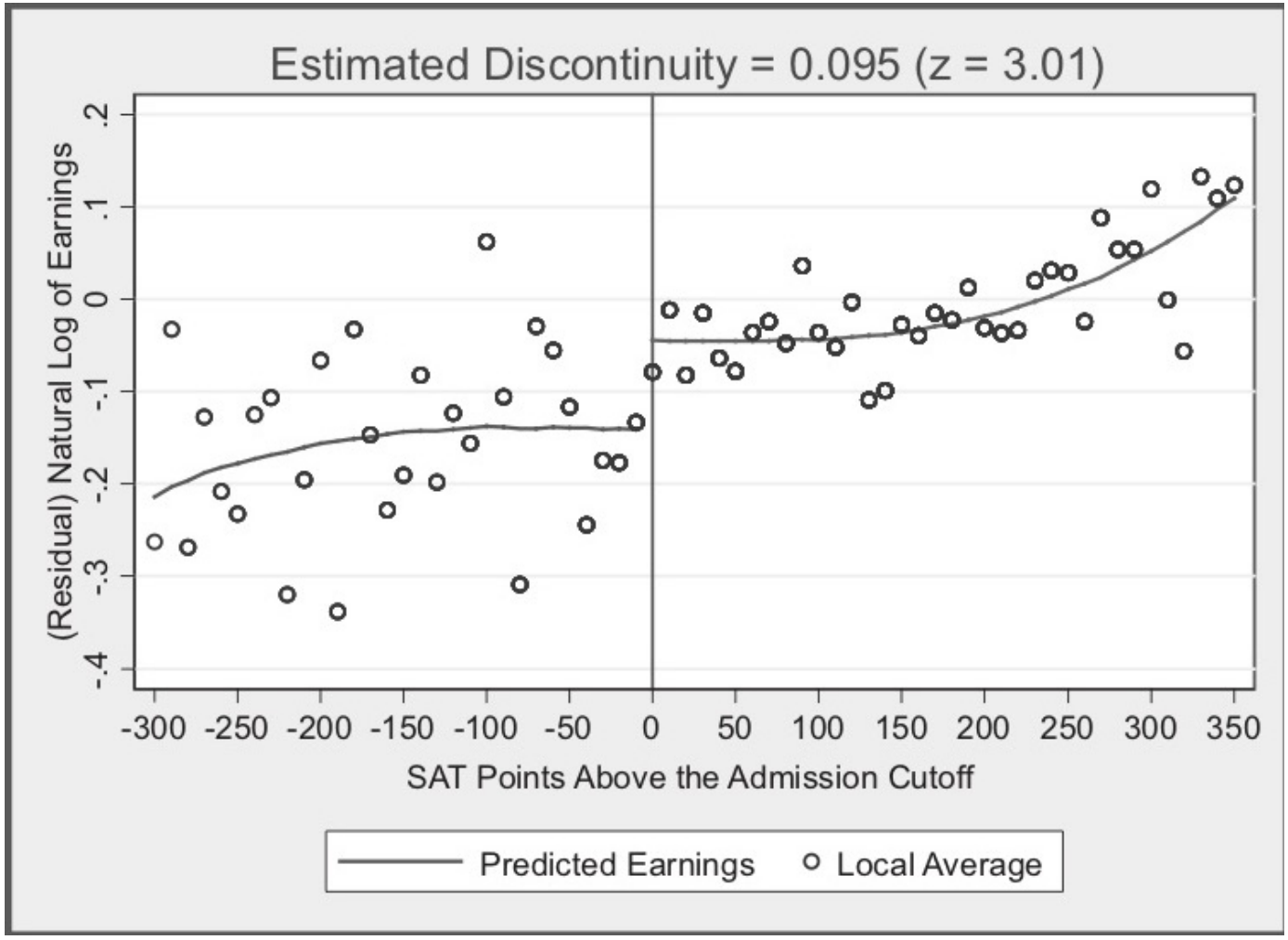
- We want to estimate some causal effect of a treatment on some outcome
- Recall the selection bias in simply comparing treatment and control groups when people choose their own treatments based on gains (i.e., independence is violated)
- But what if treatment assignment occurs abruptly when some underlying variable X called the "running variable" passes a cutoff c_0?
- RDD formalizes the effort to estimate causal effects using just such an event.

Tell me what you think is happening

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Tell me what you think is happening



Running and jumping

- Firms, schools and govt agencies have running variables that are used to assign treatments in their rules
- Most effective RDD studies involve programs where running variables assign treatments based on a “hair trigger”
- Happens for a variety of reasons: Good reasons; inexplicable reasons; arbitrary rules; a choice made by necessity and resource constraints
- But whatever the reason, it provides an opportunity for us to estimate causal effects using observational data which is a huge gift

Examples from the literature

- Yelp rounded a continuous score of ratings to generate stars which Anderson and Magruder 2011 used to study firm revenue
- US targeted air strikes in Vietnam using rounded risk scores which Dell and Querubin 2018 used to study the military and political activities of the communist state
- Card, Dobkin, and Maeskas 2008 studied the effect of universal healthcare on mortality and healthcare usage exploiting jumps at age 65
- Almond, et al. 2010 studied the effect of intensive medical attention on health outcomes when a newborn's birthweight fell just below 1,500 grams

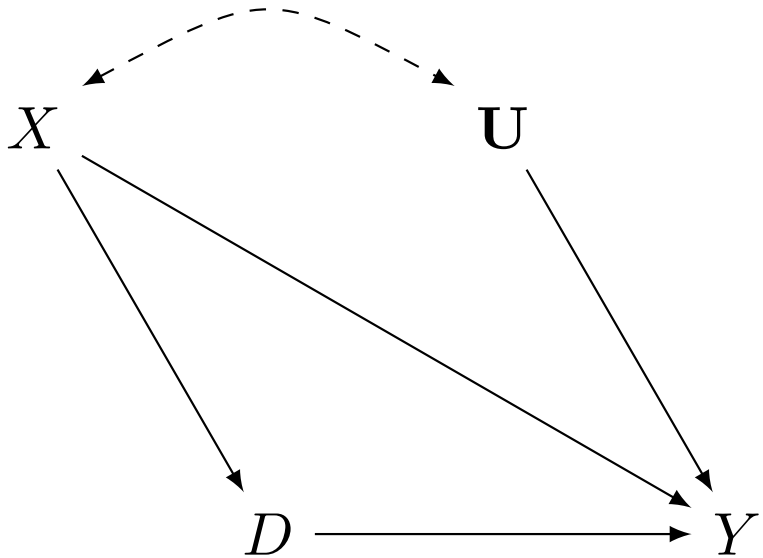
Data requirements

Large sample sizes are characteristic features of the RDD

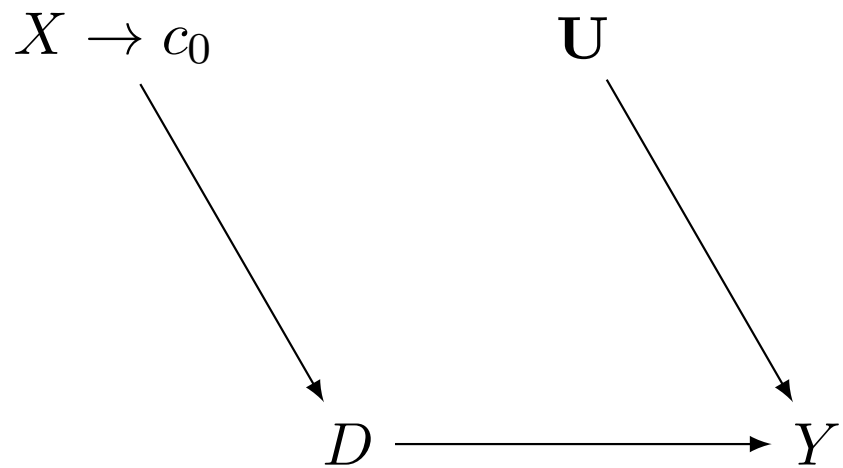
- Usually we think of “trends” as time trends, but in an RDD, trends refer to that “running variable” – but meaning is the same
- If there are strong trends in the running variable, one typically needs a lot more data than if there weren’t
- We ned a lot of data bc we need to fill out the running variable so there is large mass at the cutoff
- Researchers are typically using administrative data or settings such as birth records where there are **many** observations

Might explain why the method never caught on until the 00’s

(A) Data generating graph



(B) Limiting graph



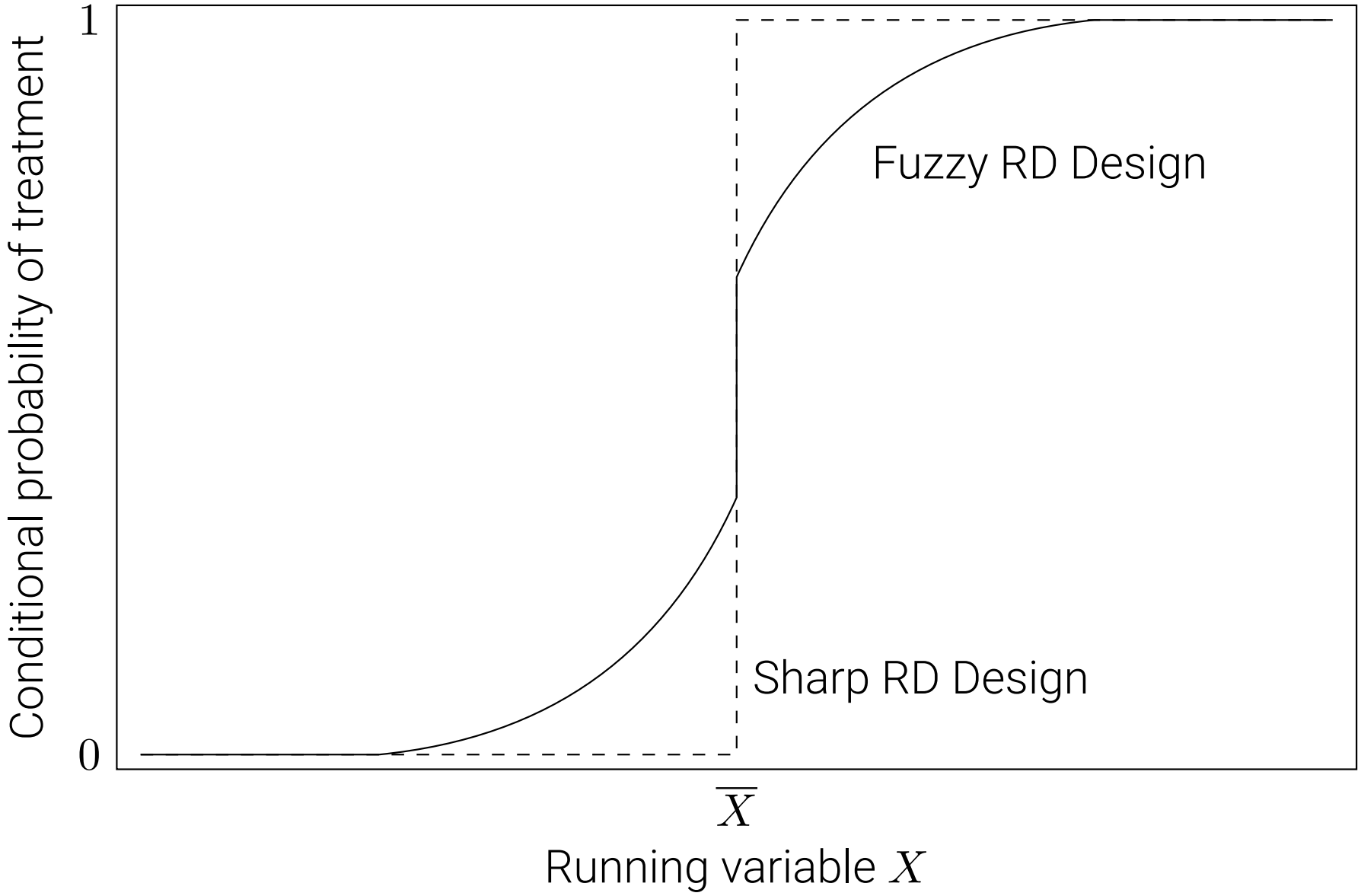


Figure: Sharp vs. Fuzzy RDD

Sharp vs. Fuzzy RDD

- There's traditionally thought to be two kinds of RD designs:
  1. Sharp RDD: Treatment is a deterministic function of running variable, X. Example: Medicare benefits.
  2. Fuzzy RDD: Discontinuous "jump" in the *probability* of treatment when $X > c_0$. Cutoff is used as an instrumental variable for treatment. Example: attending state flagship
- Fuzzy is a type of IV strategy and requires explicit IV estimators like 2SLS; sharp is reduced form IV and doesn't require IV-like estimators – we study it later with IV therefore

Overlap or Common Support

- Notice that in the sharp design, we have people in the treated or untreated along the running variable but not both – no “overlap”
- This means that for a given value of X, we observe *either* the treated group *or* the control group, but not both
- No overlap will rule out all designs that require common support
- (Note: This is not true for RCTs because there units in treatment and control each have the same mean value of X)

Apples to Oranges

- Because we don't have overlap, we can't compare apples to apples (i.e., units with same value of X)
- We have to compare apples ($X < c_0$) to oranges ($X > c_0$)
- Comparing apples to oranges requires something called "extrapolation" which is about prediction across the cutoff of the running variable
- Since it's a type of prediction, it's sensitive to modeling choices like functional form
- This is why a lot of the robustness specifications/tests in an RDD study may use a variety of estimators and a variety of functional forms

Treatment assignment in the sharp RDD

Deterministic treatment assignment ("sharp RDD")

In Sharp RDD, treatment status is a deterministic and discontinuous function of a covariate, X_i:

$$D_i = \begin{cases} 1 \text{ if } & X_i \geq c_0 \\ 0 \text{ if } & X_i < c_0 \end{cases}$$

where c_0 is a known threshold or cutoff. In other words, if you know the value of X_i for a unit i, you know treatment assignment for unit i with certainty.

Example: Medicare: Americans aged 64 are *not* eligible for Medicare, but Americans aged 65 are eligible for Medicare (ignoring disability exemptions). Notice no 64 year olds are in Medicare, and no 65 year olds are in the control group (no overlap)

Treatment effect definition and estimation

Definition of treatment effect

The treatment effect parameter, δ, is the discontinuity in the conditional expectation function:

$$\begin{aligned} \delta &= lim_{X_i \to c_0} E[Y_i^1 | X_i = c_0] - lim_{c_0 \leftarrow X_i} E[Y_i^0 | X_i = c_0] \\ &= lim_{X_i \to c_0} E[Y_i | X_i = c_0] - lim_{c_0 \leftarrow X_i} E[Y_i | X_i = c_0] \end{aligned}$$

The sharp RDD estimation is interpreted as an average causal effect of the treatment at the discontinuity

$$\delta_{SRD} = E[Y_i^1 - Y_i^0 | X_i = c_0]$$

Extrapolation

- In RDD, the counterfactuals are conditional on X.
- We use *extrapolation* in estimating treatment effects with the sharp RDD bc we do not have overlap
  - $\rightarrow$ Left of cutoff, only non-treated observations, $D_i = 0$ for $X < c_0$
  - $\rightarrow$ Right of cutoff, only treated observations, $D_i = 1$ for $X \geq c_0$
- The extrapolation is to a counterfactual

Extrapolation

Estimation methods attempt to approximate the limiting parameter using units left and right of the cutoff

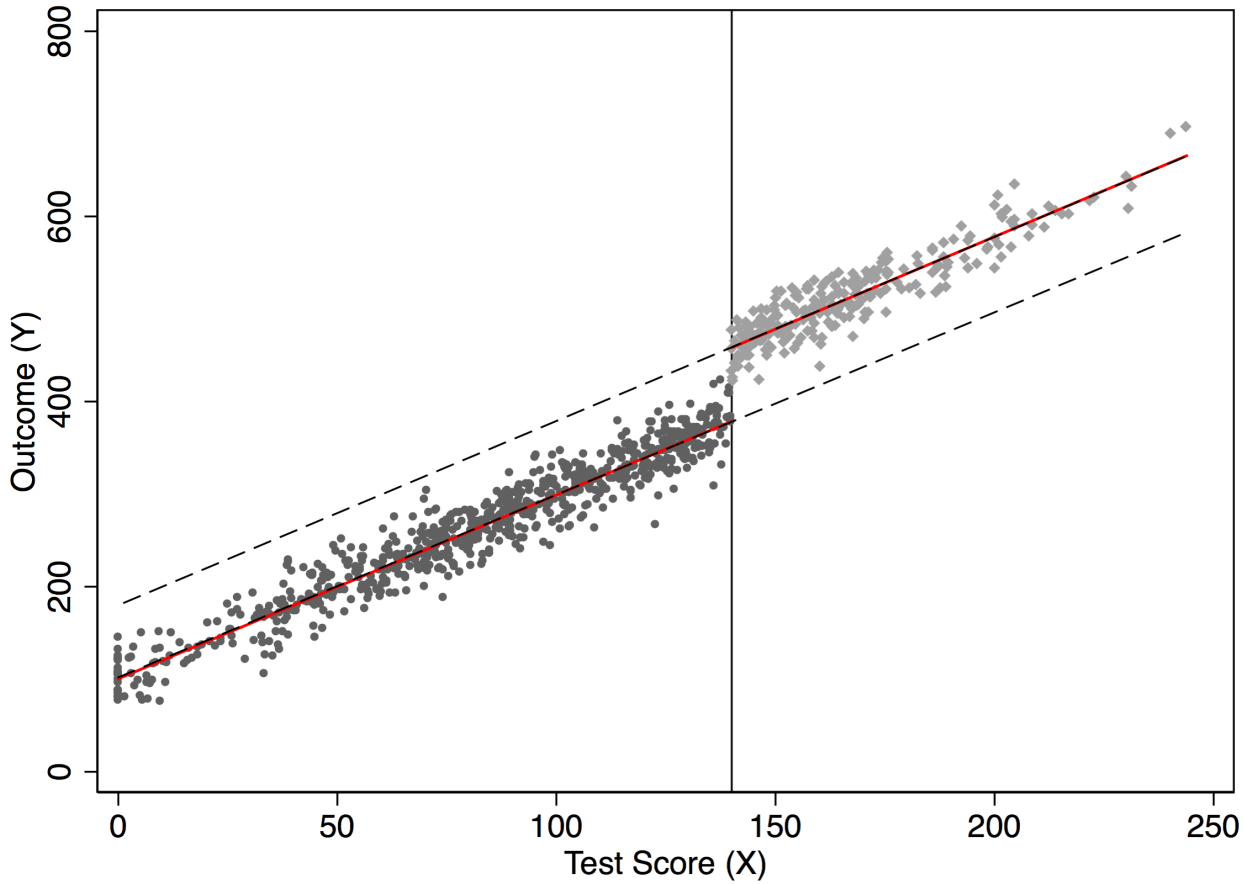


Figure: Dashed lines are extrapolations (Marcelo Perraillon simulated random variables)

Two identification traditions

1. **Smoothness tradition** – Goes back to Hahn, Todd and Van der Klaauw (2001). Emphasizes counterfactual smoothness. More "design-model" approach
2. **Randomization tradition** – But early researchers tended to talk like this: "it is as thought units were randomly assigned around the cutoff". In other words "design-design" approach

Most of my lecture tends to draw upon the smoothness tradition, but I'll briefly mention a randomization inference method based on the randomization tradition too

Smoothness as the identifying assumption

Smoothness (or continuity) of conditional expectation functions (Hahn, Todd and Van der Klaauw 2001)

$E[Y_i^0|X = c_0]$ and $E[Y_i^1|X = c_0]$ are continuous (smooth) in X at c_0.

- This tends to be a place where people confuse potential outcomes with observable outcomes – recall the switching equation!
- If population average *potential outcomes*, Y^1 and Y^0, are smooth functions of X through the cutoff, c_0, then potential average outcomes *won't* jump at c_0.
- Implies the cutoff is exogenous – i.e., nothing else changes related to potential outcomes at c_0
- Unobservables are evolving smoothly, too, through the cutoff

Smoothness is the identifying assumption and untestable

- Smoothness justifies extrapolation because if smoothness is credible, then we can use the data left of the cutoff to "fill in" for the counterfactual right of the cutoff and vice versa
- Causal effect of the treatment will be based on **extrapolation** from the trend, $E[Y_i^0|X < c_0]$, to those values of $X > c_0$ for the $E[Y_i^0|X > c_0]$.
- But like with all design-model approaches, since you can't test it, you have to find evidence for it which is actually a major part of this class and I consider the "art of RDD"

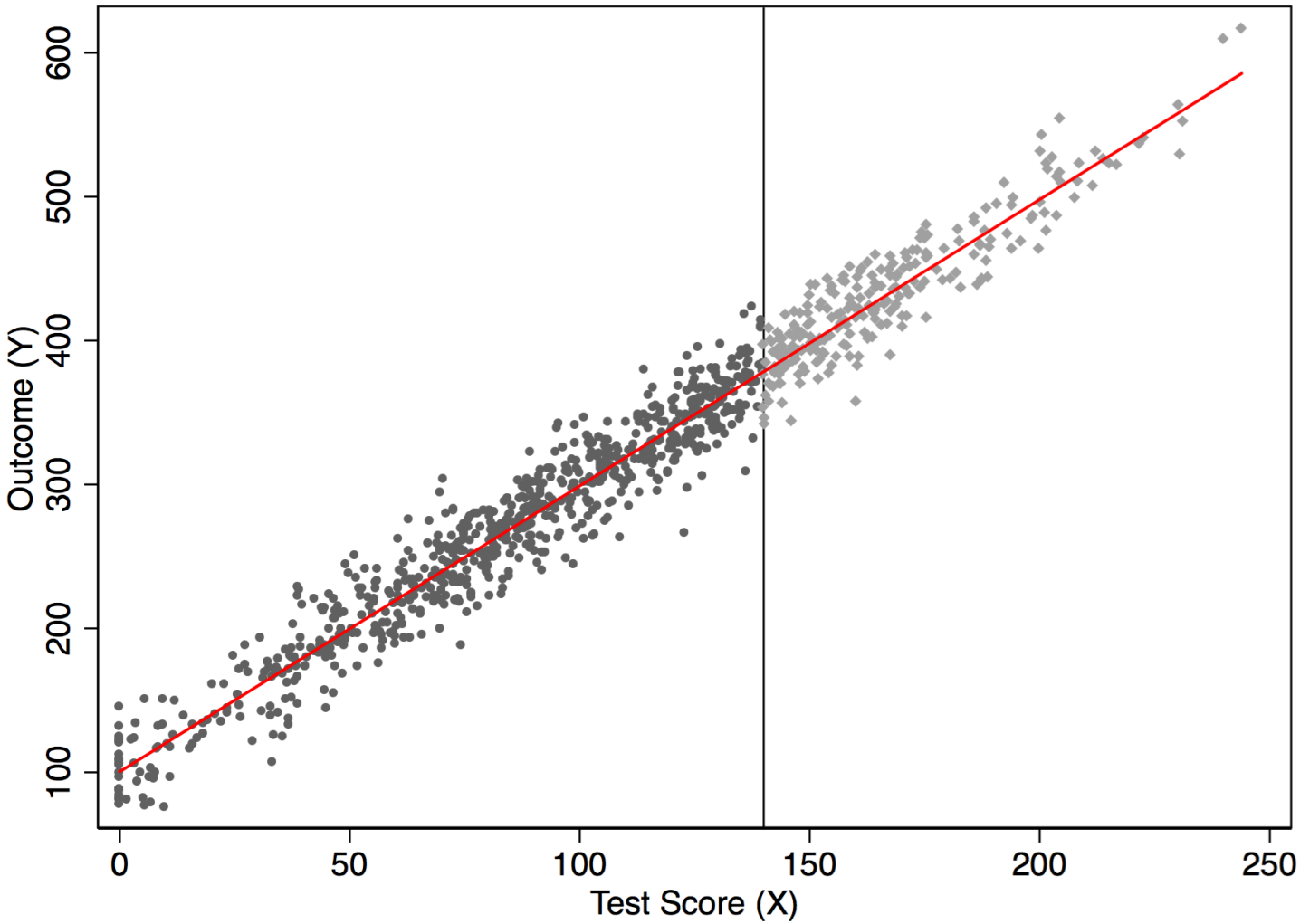
Potential and observable outcomes

- But that's weird Scott because why can't I just compare Y left and right of the cutoff and see if it jumps? Doesn't Y jumping mean smoothness is violated
- No – remember the switching equation?

$$Y = DY^1 + (1 - D)Y^0$$

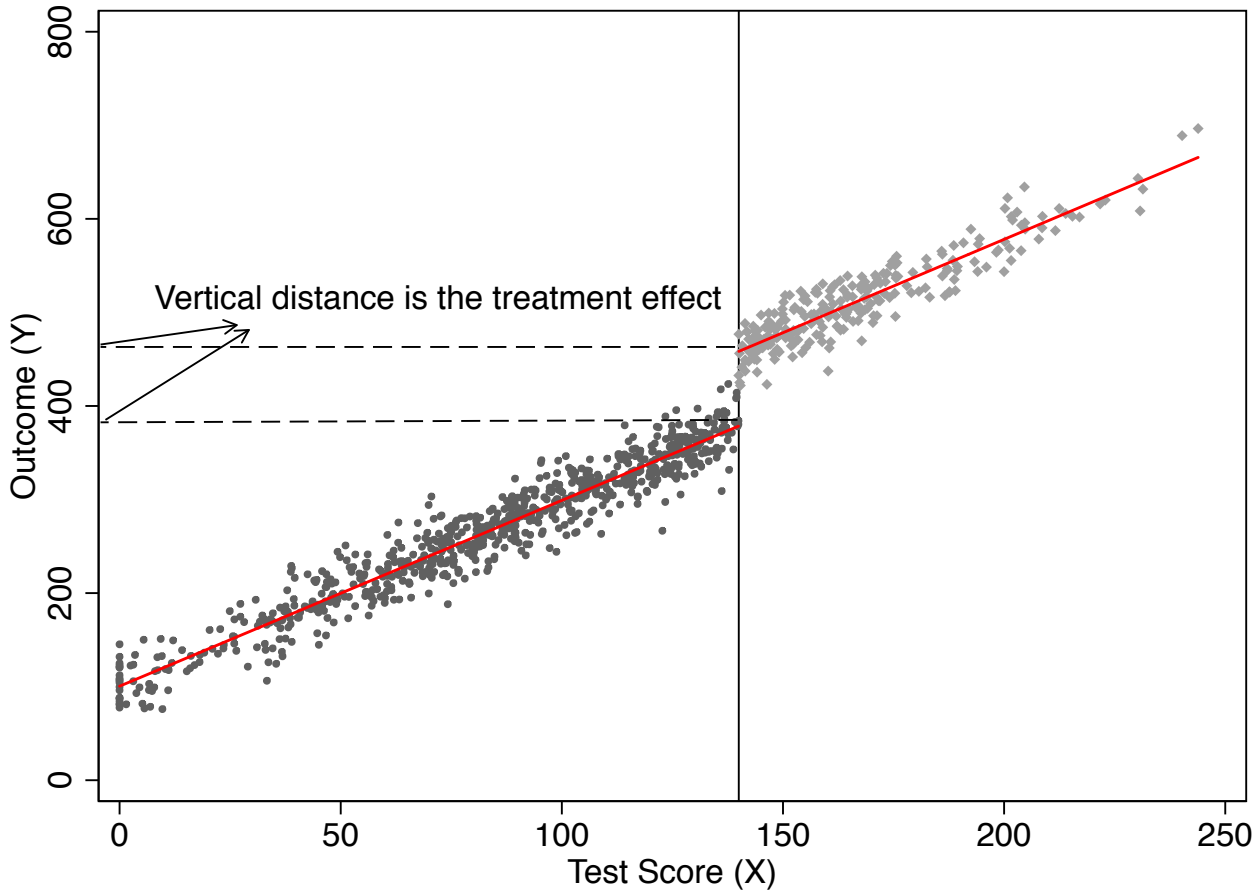
- We don't observe Y^0 or Y^1 unless the switching equation picked it and the switching equation flips in a sharp design when we pass the cutoff
- Smoothness is all about what would have happened to Y^0 had the cutoff not assigned the treatment – subtle idea!

Graphical example of the smoothness assumption



Note these are *potential* not *observable* outcomes (Credit: Marcelo Perraillon simulated random variables)

Graphical example of the treatment effect, not the smoothness assumption



Note that these are *observable*, not *potential* outcomes. (Credit: Marcelo Perraillon simulated random variables)

Re-centering the data

- It is common for authors to transform X by "centering" at c_0:

$$Y_i = \alpha + \beta(X_i - c_0) + \delta D_i + \varepsilon_i$$

- This doesn't change the interpretation of the treatment effect – only the interpretation of the intercept.

Working with the data

- Estimation commonly uses OLS which means we have to think about how we will measure key variables and the model itself
- Couple things that tend to go with estimation
  - → Re-centering the data into relative “event time” (we see this also with diff-in-diff)
  - → Exploring nonlinearities in the data
- Let’s look at these in stages

Re-centering the data

- Example: Medicare and age 65. Center the running variable (age) by subtracting 65:

$$\begin{aligned} Y &= \beta_0 + \beta_1(Age) + \beta_2 Edu + \varepsilon \\ &= \beta_0 + \beta_1(Age - 65) + \beta_2 Edu + \varepsilon \\ &= \beta_0 + \beta_1 Age - \beta_1 65 + \beta_2 Edu + \varepsilon \\ &= \alpha + \beta_1 Age + \beta_2 Edu + \varepsilon \end{aligned}$$

  where $\alpha = \beta_0 - \beta_1 65$.
- All other coefficients, notice, have the same interpretation, except for the intercept.

Regression without re-centering

```
reg y D x
```

Source	SS	df	MS
Model	15842893.9	2	7921446.97
Residual	394715.557	996	396.30076
Total	16237609.5	998	16270.1498

Number of obs	=		999
F(2, 996)	=		19988.47
Prob > F	=		0.0000
R-squared	=		0.9757
Adj R-squared	=		0.9756
Root MSE	=		19.907

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	80.01418	2.144779	37.31	0.000	75.80537	84.22298
x	1.986975	.0186779	106.38	0.000	1.950322	2.023627
_cons	100.3885	1.70944	58.73	0.000	97.03397	103.743

Regression with centering

```
gen x_c = x - 140

reg y D x_c
```

Source	SS	df	MS
Model	15842893.9	2	7921446.97
Residual	394715.554	996	396.300757
Total	16237609.5	998	16270.1498

Number of obs	=	999	
F(2, 996)	=	19988.47	
Prob > F	=	0.0000	
R-squared	=	0.9757	
Adj R-squared	=	0.9756	
Root MSE	=	19.907	

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	80.01418	2.144779	37.31	0.000	75.80537	84.22298
x_c	1.986975	.0186779	106.38	0.000	1.950322	2.023627
_cons	378.565	1.290755	293.29	0.000	376.032	381.0979

Nonlinearity bias

- Smoothness is an assumption about the behavior of the conditional expected potential outcomes as we move across the running variable and through the cutoff
- But just because it's smooth doesn't mean potential outcomes moved in a straight line – it could've been a polynomial (i.e., quadratic)
- What if the trend relation $E[Y_i^0|X_i]$ does not jump at c_0 but rather is simply nonlinear? You could get spurious results
- Ouch: your linear model might identify a treatment effect when there isn't because the functional form had poor predictive properties beyond the cutoff
- Let's look at a simulation

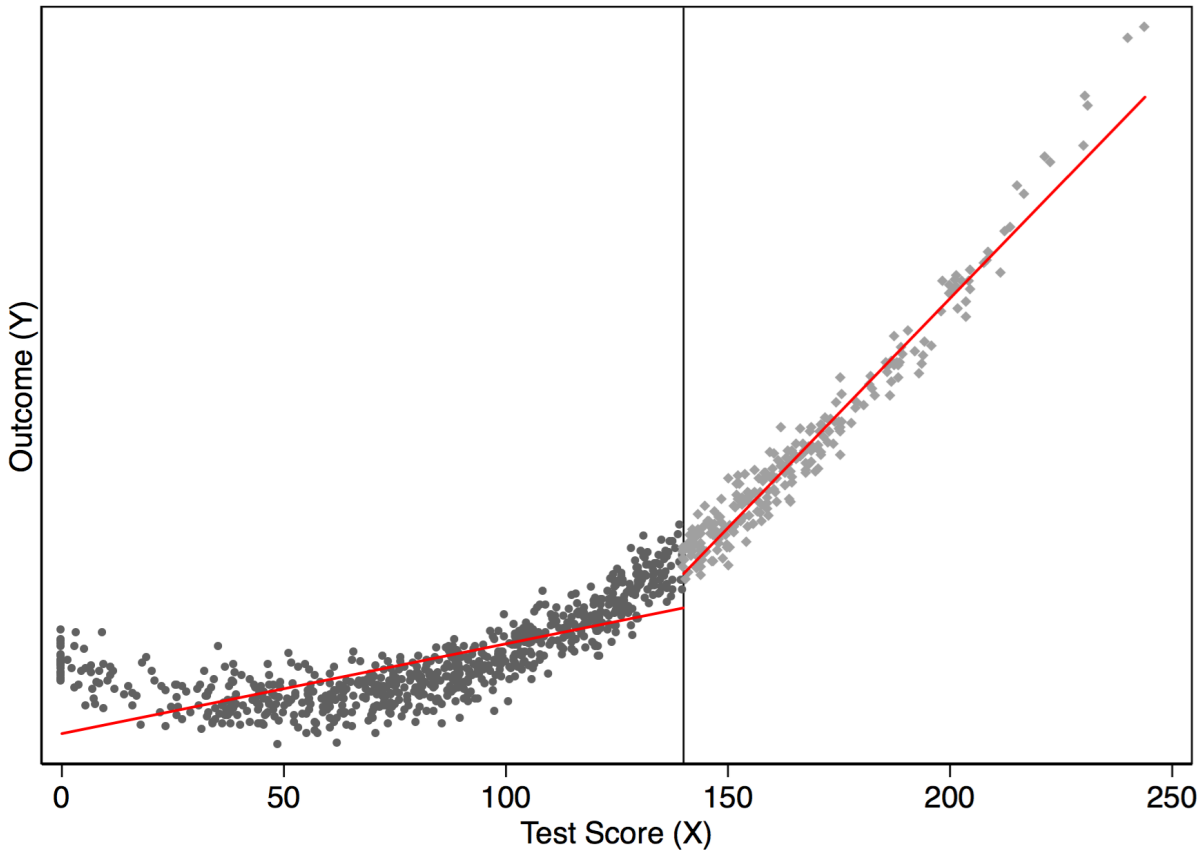
Credit: Marcelo Perraillon simulation

```
gen x2 = x*x

gen x3 = x*x*x

gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)

scatter y x if D==0, msize(vsmall) || scatter y x ///
  if D==1, msize(vsmall) legend(off) xline(140, ///
  lstyle(foreground)) ylabel(none) || lfit y x ///
  if D ==0, color(red) || lfit y x if D ==1, ///
  color(red) xtitle("Test Score (X)") ///
  ytitle("Outcome (Y)")
```



See how the two lines don't touch at c_0 but empirically should? That's bc the linear fit is the wrong functional form and it's going to give us a positive treatment effect when it's really just a modeling problem

Sharp RDD: Nonlinear Case

- Suppose the nonlinear relationship is $E[Y_i^0|X_i] = f(X_i)$ for some reasonably smooth function $f(X_i)$ (e.g., quadratic in X)
- In that case we'd fit the regression model:

$$Y_i = f(X_i) + \delta D_i + \eta_i$$

- But what is this $f(X_i)$? It models the counterfactual values of Y^0 and since we are extrapolating, we need an estimator that we think is extrapolating correctly
- There are 2 common ways of approximating $f(X_i)$

Comment about higher order polynomials

- Before Gelman and Imbens (2018), authors would use “higher order polynomials”
- Higher order polynomials can have overfitting problems leading to poor prediction beyond the cutoff
- Gelman and Imbens (2018) recommend at best a quadratic and you should take their advice seriously
- But as way of exposition, I’m going to be agnostic about that because most papers have done it (including Lee, Moretti and Butler 2004)

Approximate the functional form

1. Use global and local regressions with $f(X_i)$ equalling a p^{th} order polynomial

$$Y_i = \alpha + \delta D_i + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + \eta_i$$

2. Or use some nonparametric kernel method which I'll cover later

Different polynomials on the 2 sides of the discontinuity

- We can generalize the function, $f(x_i)$, by allowing it to differ on both sides of the cutoff by including them both individually and interacting them with D_i.
- In that case we have:

$$\begin{aligned} E[Y_i^0|X_i] &= \alpha + \beta_{01}\tilde{X}_i + \beta_{02}\tilde{X}_i^2 + \cdots + \beta_{0p}\tilde{X}_i^p \\ E[Y_i^1|X_i] &= \alpha + \delta + \beta_{11}\tilde{X}_i + \beta_{12}\tilde{X}_i^2 + \cdots + \beta_{1p}\tilde{X}_i^p \end{aligned}$$

where $\tilde{X}_i$ is the centered running variable (i.e., $X_i - c_0$).

Lines to the left, lines to the right of the cutoff

- Re-centering at c_0 ensures that the treatment effect at $X_i = c_0$ is the coefficient on D_i in a regression model with interaction terms
- As Lee and Lemieux (2010) note, allowing different functions on both sides of the discontinuity should be the main results in an RDD paper

Different polynomials on the 2 sides of the discontinuity

To derive a regression model, substitute the observed outcome for the potential outcome using a slight modification of the switching equation (Krueger 1999)

$$E[Y|X] = E[Y^0|X] + \left(E[Y^1|X] - E[Y^0|X]\right) D$$

Regression equation

- Regression model you estimate is:

$$\begin{aligned} Y_i &= \alpha + \beta_{01}\tilde{x}_i + \beta_{02}\tilde{x}_i^2 + \cdots + \beta_{0p}\tilde{x}_i^p \\ &\quad + \delta D_i + \beta_1^* D_i\tilde{x}_i + \beta_2^* D_i\tilde{x}_i^2 + \cdots + \beta_p^* D_i\tilde{x}_i^p + \varepsilon_i \end{aligned}$$

  where $\beta_1^* = \beta_{11} - \beta_{01}$, $\beta_2^* = \beta_{21} - \beta_{21}$ and $\beta_p^* = \beta_{1p} - \beta_{0p}$

- Notice the interactions of D with the re-centered running variables – they model the dynamics in the running variable above and below the cutoff
- But the parameter of interest, the treatment effect, is the coefficient at c_0 or $\widehat{\delta}$

Polynomial simulation example (Credit: Marcelo Perraillon simulation)

```
capture drop y x2 x3

gen x2 = x*x
gen x3 = x*x*x
gen y = 10000 + 0*D - 100*x +x2 + rnormal(0, 1000)

reg y D x x2 x3
predict yhat

scatter y x if D==0, msize(vsmall) || scatter y x
  if D==1, msize(vsmall) legend(off) xline(140,
  lstyle(foreground)) ylabel(none) || line yhat x
  if D ==0, color(red) sort || line yhat x if D==1,
  sort color(red) xtitle("Test Score (X)")
  ytitle("Outcome (Y)")
```

Polynomial simulation example (Credit: Marcelo Perraillon simulation)

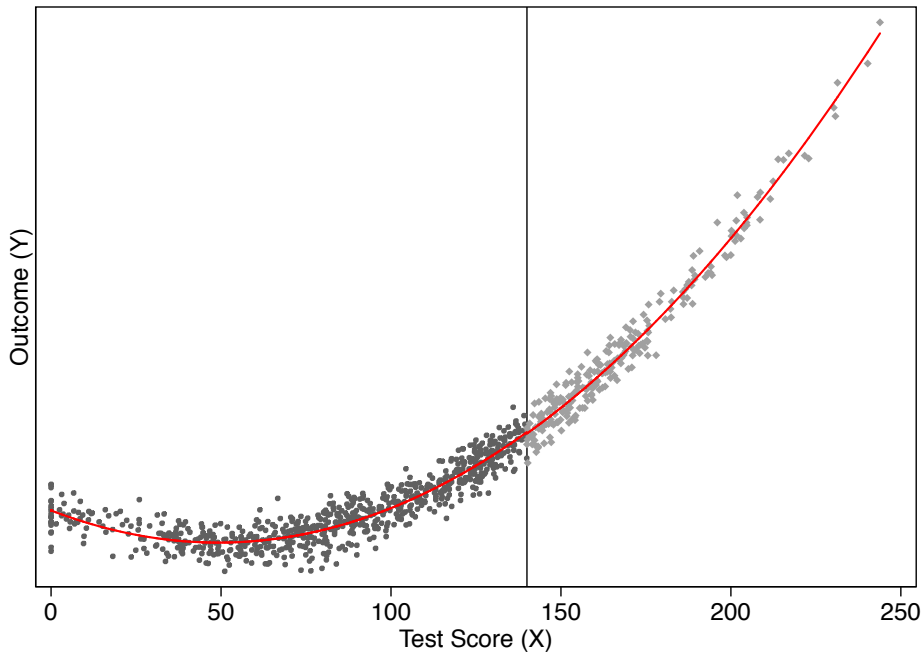


Figure: Third degree polynomial. Actual model second degree polynomial. (Credit: Marcelo Perraillon simulation).

Notice: no more gap at c_0 once we model the function $f(x)$

Stata simulation

```
gen x2_c = x2  _  140

gen x3_c = x3  _  140

reg y D x x2

reg y D x_c x2_c
```

Polynomial simulation example

Source	SS	df	MS
Model	3.7863e+10	3	1.2621e+10
Residual	957507024	995	962318.617
Total	3.8821e+10	998	38898361.8

Number of obs	= 999
F(3, 995)	=13115.22
Prob > F	= 0.0000
R-squared	= 0.9753
Adj R-squared	= 0.9753
Root MSE	= 980.98

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	-115.5381	127.4967	-0.91	0.365	-365.7314	134.6552
x	-98.57582	2.285769	-43.13	0.000	-103.0613	-94.09034
x2	1.000001	.0122767	81.45	0.000	.9759098	1.024092
_cons	9864.218	111.1206	88.77	0.000	9646.16	10082.28

Notice: no more gap at c_0 once we model the function $f(x)$ (e.g., D is insignificant once we include polynomials)

Polynomial simulation example

Source	SS	df	MS
Model	3.7863e+10	3	1.2621e+10
Residual	957507020	995	962318.613
Total	3.8821e+10	998	38898361.8

Number of obs	= 999
F(3, 995)	=13115.22
Prob > F	= 0.0000
R-squared	= 0.9753
Adj R-squared	= 0.9753
Root MSE	= 980.98

y	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
D	-115.5381	127.4967	-0.91	0.365	-365.7315	134.6552
x_c	-98.57582	2.285769	-43.13	0.000	-103.0613	-94.09034
x2_c	1.000001	.0122767	81.45	0.000	.9759098	1.024092
_cons	-3796.397	227.7894	-16.67	0.000	-4243.4	-3349.394

And centering did nothing to the interpretation of the main results (D), only to the intercept.

Can we evaluate credibility of smoothness assumption?

- Design-model based approaches cannot appeal to independence (Rubin, “we know how the science works” with randomization)
- Design-model based approaches appeal to *restrictions* on potential outcomes and there is no such “science” for that
- You main results are only causal insofar as smoothness is a credible belief, and since counterfactuals are missing data problems, you have to build your case indirectly
- Lots of sensitivity checks, placebos and alternative approaches

Main Challenges

Classify your concern regarding smoothness violations into two categories:

- Manipulation on the running variable
- Endogeneity of the cutoff

Most robustness is aimed at building credibility around these,

Manipulation of your running variable score

- Treatment is not as good as randomly assigned around the cutoff, c_0, when agents are able to manipulate their running variable scores. This happens when:
  1. the assignment rule is known in advance
  2. agents are interested in adjusting
  3. agents have time to adjust
  4. administrative quirks like nonrandom heaping along the running variable

  Examples include re-taking an exam, self-reported income, certain types of non-random rounding.
- Since necessarily treatment assignment is no longer independent of potential outcomes, it's likely this implies smoothness has been violated

Test 1: Manipulation of the running variable

Manipulation of the running variable

Assume a desirable treatment, D, and an assignment rule $X \geq c_0$. If individuals sort into D by choosing X such that $X \geq c_0$, then we say individuals are manipulating the running variable.

Also can be called “sorting on the running variable” – same thing

Imagine this RCT

- Imagine a treatment (statin) that people widely believe will save lives
- Now imagine a team randomly assigns patients to treatment or control to study the effect of the statin on heart attacks within 10 years
- 200 patients are placed in two different waiting rooms – 100 in A and 100 in B
- If people know which room has the statin, people want to move between rooms, and people have time to move into the rooms, how many people will be in A? B?

Choosing your running variable value

- Probably everyone in room B will move to room A if they
  1. Want to be in room A (i.e., something to gain)
  2. Have time to go there
  3. Know about it ahead of time
- If all three, then we probably will see 200 people in A and 0 people in B

Implications

- We cannot test for smoothness because it involves potential outcomes, but
- we can test for whether units are stacking just above or below the cutoff
- What other reason could explain a bunch of people making just barely enough money to qualify for some treatment except that they *manipulated* their position on the running variable (e.g., income)
- This was the idea that Justin McCrary had when he wrote his paper on using density tests to test for “sorting on the running variable” (McCrary 2008)
- Highly cited paper in a symposium in the Journal of Econometrics on RDD (lots of classics)

McCrary Density Test

- Assumes a null where the *density* is continuous at the cutoff point
- Under the alternative hypothesis, the density increases at the cutoff as people sort onto the desirable side of the cutoff
- This is oftentimes visualized with confidence intervals illustrating the effect of the discontinuity on density - you need no jump to pass this test
- Not perfect, but pretty ingenious and is based on rational choice when you think about it

Steps for a density test in RDD

1. Count observations for a chosen bin [*ed: how big should you make the bin*]
2. Estimate your nonlinear OLS model with quadratics in the running variable on the *counts*
3. Do you reject the null at the cutoff?

There are updates to McCrary (2008) using other density tests but this is the basic idea

McCrary Density Ttest

- A discontinuity in the density is “suspicious” because it *suggests* manipulation which may be evidence that smoothness doesn’t hold
- Density tests are *mandatory* for every analysis using RDD with cross-sectional data
- How do you know if you can do a density test? If you were able to calculate conditional expectations in the first place
- Contemporary method in R and Stata is `rddensity` which is based on several papers by Cattaneo like Cattaneo, Jansson and Ma (2020)
- Density tests are not helpful for RD in time (Hausman and Rapson 2018), which is its own beast and not all that straightforward

Simulations of density tests

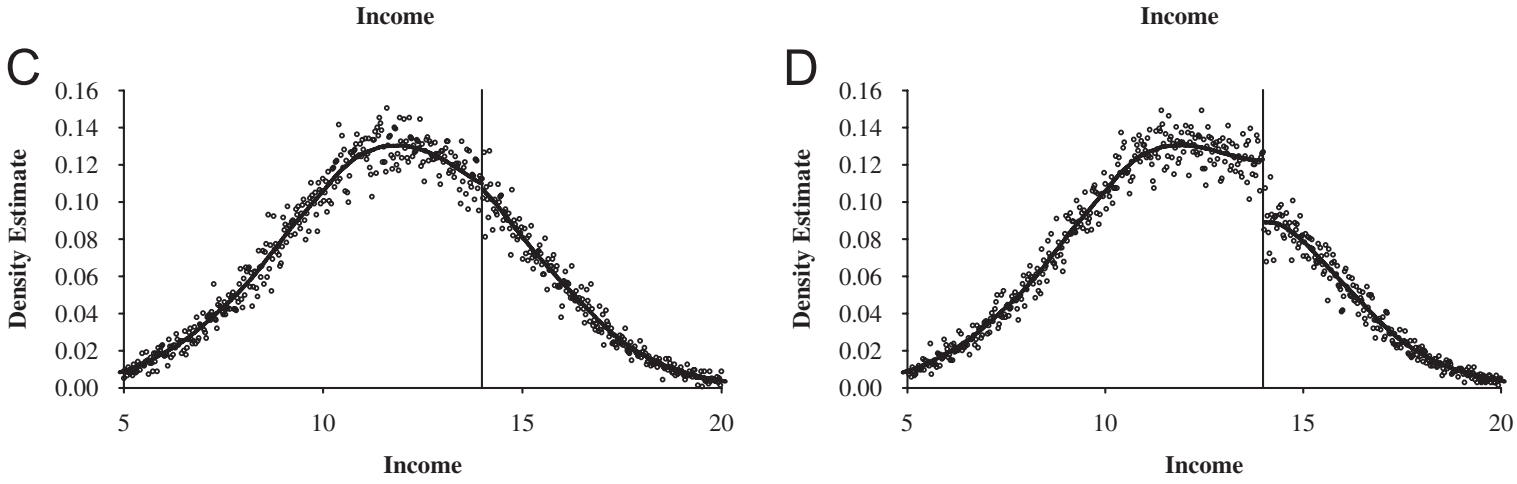


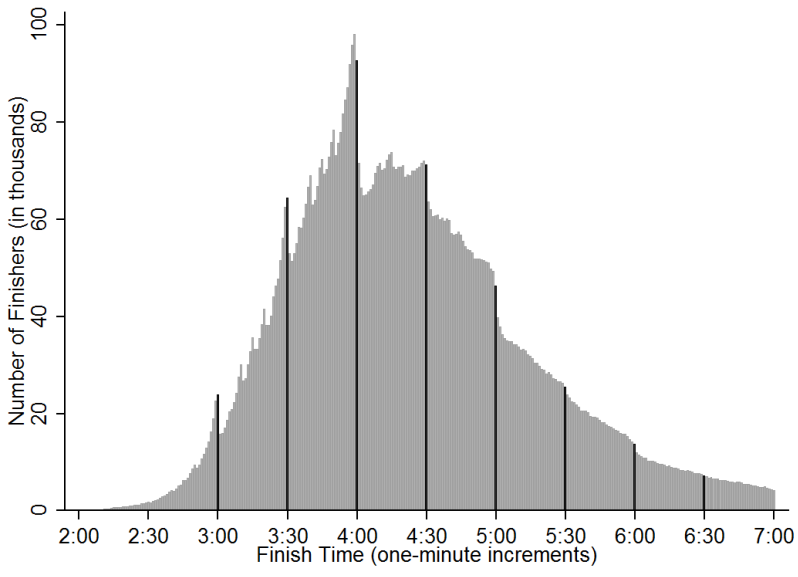
Figure: From McCrary (2008). Left shows failing to reject. Right shows rejection of the null.

I have consistently found manipulation on the running variable when evaluating SNAP (food stamps) using income cutoffs fwiw

Density tests in marathons

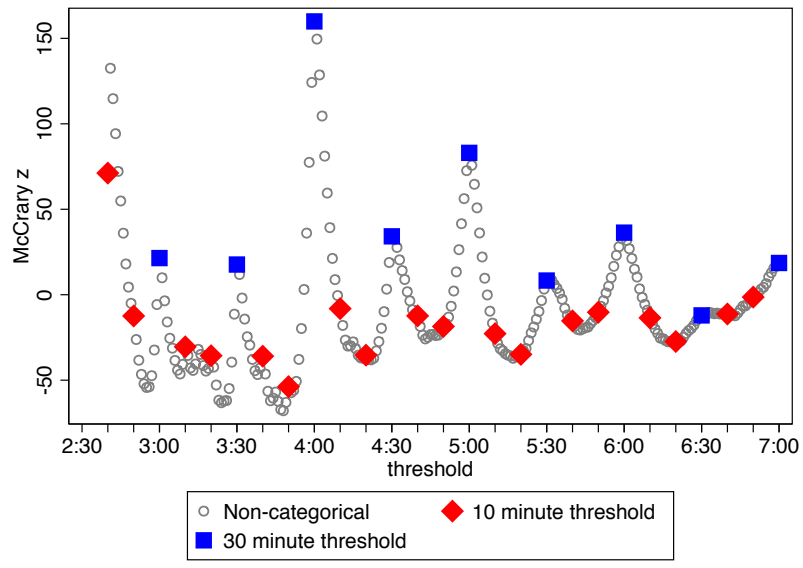
Do people maximize speed in a marathon? Or do they have "reference dependent" times (e.g., making Boston)? Let's look at raw data by running variable and tests for rejection across the running variable

Figure 2: Distribution of marathon finishing times ($n = 9,378,546$)



NOTE: The dark bars highlight the density in the minute bin just prior to each 30 minute threshold.

Figure 3: Running McCrary z-statistic



NOTE: The McCrary test is run at each minute threshold from 2:40 to 7:00 to test whether there is a significant discontinuity in the density function at that threshold.

Figure: From Allen, Dechow, Pope and Wu (2013) "Reference-Dependent Preferences: Evidence from Marathon Runners"

Does NICU save premature babies' lives?

- Premature babies both receive expensive medical spending *and* are more likely to perish.
- What is the causal effect of NICU on infant mortality?
- Think of our aliens example – naive correlations suffer from *severe* selection bias
- One team used RDD to answer it, but how? Valid RDD need at minimum a cutoff and a running variable. What might that be in this context?

Almond et al. (2010) RDD strategy

- Almond, et al. (2010) attempted to estimate the causal effect of medical expenditures on health outcomes using RDD
- In the US, newborns whose birthweight falls below 1500 grams are placed in intensive care bc 1500 is the “very low birth weight” range
- Compare those just above with those below 1500 using a variety of estimators and visualizations
- They used hospital administrative records and found 1-year infant mortality decreased by 1pp just below 1500 grams compared to just above
- Concluded these medical expenditures are cost-effective (compounded value of life over typical lifespan)
- But now let’s look at density tests

Heaping problem

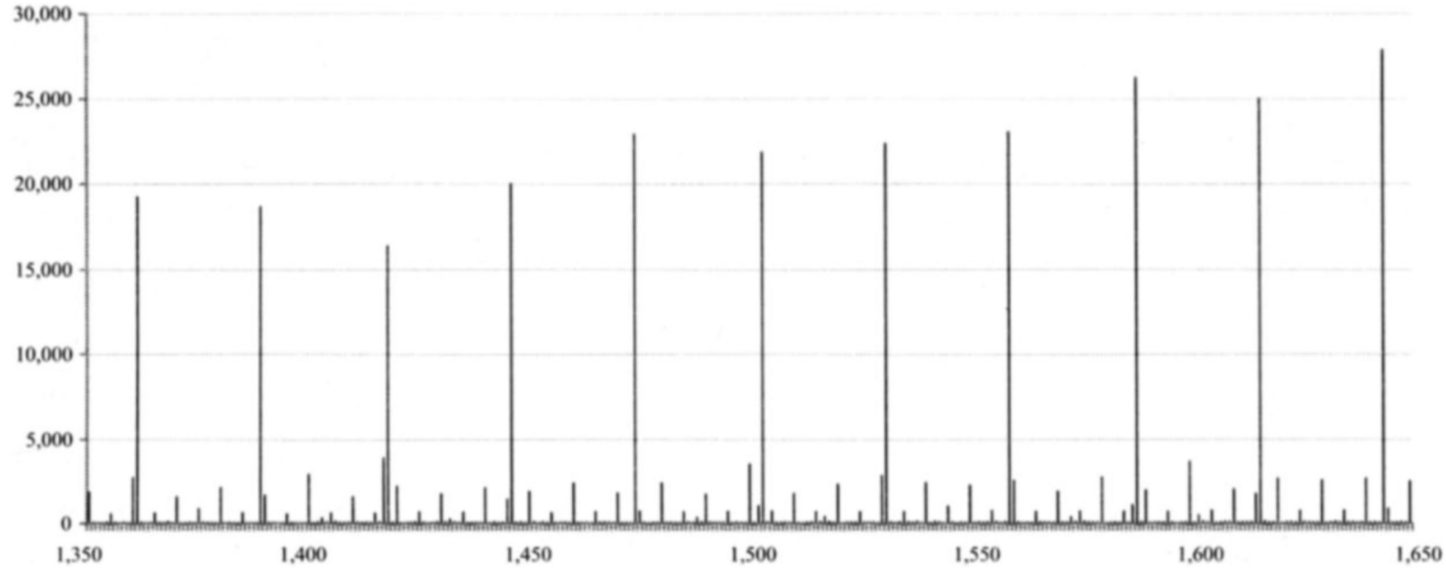


FIGURE I
Frequency of Births by Gram: Population of U.S. Births between 1,350 and 1,650 g
NCHS birth cohort linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text.

Figure: Distribution of births by gram from Almond, et al. 2010

Heaping along the Running variable

- Isn’t this a discontinuity in the number of units at the cutoff? Sure looks like it
- But it’s technically non-random “heaping” at spaded out intervals (and pretty bad too)
- Heaping is excess number of units at certain points along the running variable

Why heaping at birth?

- One hypothesis is that nature just allocates births at grams spread out 100 grams apart but that's unlikely
- More likely someone is rounding – but who? And why?
  - → Maybe some scales in some hospitals are less sophisticated
  - → Maybe rounding practices are more common in some types of hospitals than others
  - → Maybe parents and staff push for rounding to get favorable treatment
- All of these are technically manipulation examples, but they seem qualitatively different don't they?
- Heaping can make it hard to reject

Authors pass the McCrary density test

- Almond, et al. 2010 used the McCrary density test and found no evidence of manipulation
- How could that be? It's right there clear as day
- Problem is density tests are not designed for detecting heaping because there's so few observations just left and right of these cut points compared to the heaps
- In this scenario, the heaping is associated with high mortality children who are outliers compared to newborns both to the left and to the right
- Researchers using RDD are particularly wedded to eyeball tests for reasons like this

Academic war begins

A team (Barreca, Guldi, Lindo and Waddell 2011) challenged the papers' findings

> *"This [heaping at 1500 grams] may be a signal that poor-quality hospitals have relatively high propensities to round birth weights but is also consistent with manipulation of recorded birth weights by doctors, nurses, or parents to obtain favorable treatment for their children. Barreca, et al. 2011 show that this nonrandom heaping leads one to conclude that it is "good" to be strictly less than any 100-g cutoff between 1,000 and 3,000 grams."*

QJE prints their paper as a comment and then a reply by Almond, et al. (2011) where they address some of their concerns

Donut hole RDD

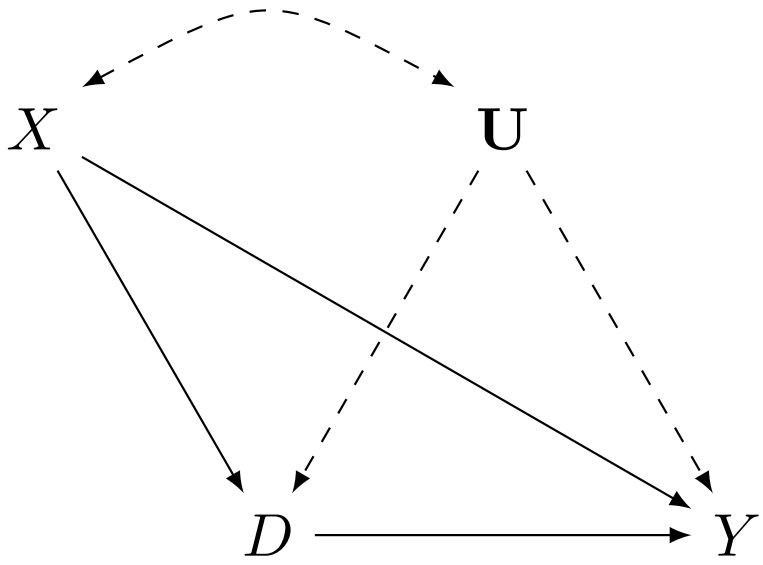
- RDD compares means as we approach c_0 from either direction along X
- Estimates should not logically be sensitive to the observations at the cutoff – if it is, then smoothness may be violated
- Barreca, et al. (2016) suggest dropping units in the vicinity of 1500 grams, and re-estimate the model – if it changes, heaping may be creating a problem
- They call this a “donut” RDD bc you drop the units at the cutoff (the “donut hole”) and estimate your model on the units in the neighborhood instead

Newborn mortality and medical expenditure

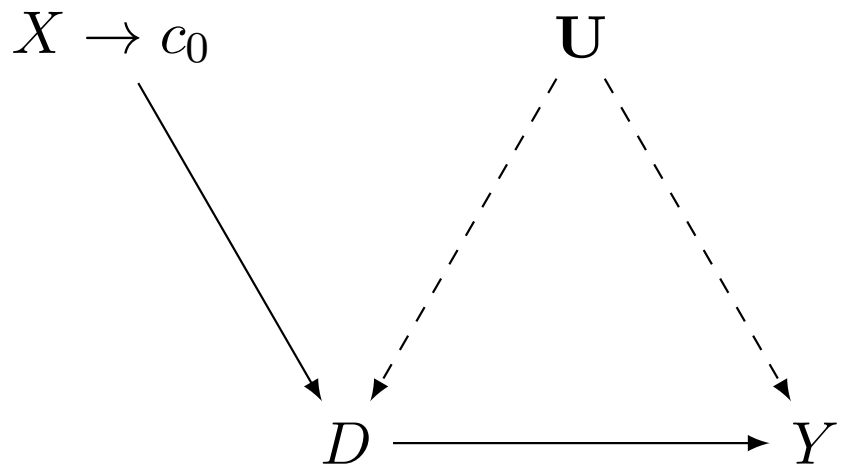
- Dropping units (e.g., trimming) always changes the parameter we're estimating
- In this case, dropping at the threshold reduced sample size by 2%
- But the strength of this practice is that it allows for the possibility that units at the heap differ markedly due to selection bias than those in the surrounding area
- Donut RDD analysis found effect sizes that were 50% smaller than Almond, et al (2010) found
- Be careful with heaping

Endogenous cutoffs

(A) Data generating graph



(B) Limiting graph



Endogeneous cutoffs

- RDD blocks the backdoor path from $D \leftarrow X \leftarrow ? \rightarrow U \rightarrow Y$; but *assumes* that the backdoor path $D \leftarrow U \rightarrow Y$
- But if cutoffs are endogenous, then it is there, which means absent the treatment, smoothness would've been violated *anyway*
- Smoothness isn't guaranteed by an RDD unless $D \leftarrow U \rightarrow Y$ isn't present – which is why it is *the* critical identifying assumption

Endogeneous cutoffs

- Examples of endogenous cutoffs
  - → Age thresholds used for policy (i.e., person turns 18, and faces more severe penalties for crime) is correlated with other variables that affect the outcome (i.e., graduation, voting rights, etc.)
  - → Age 65 is correlated with factors that directly affect healthcare expenditure and mortality such as retirement
- But some of these can be weakly defended with balance tests (observables), or may be directly testable through placebos assuming you have the data

Evaluating smoothness through balance

- Balance tests and placebo tests are related but distinct
- We can't directly test smoothness bc we are missing counterfactuals
- Ask yourself: why should average values of exogenous covariates jump if potential outcomes are smooth through the cutoff?
- If there are exogenous (non collider) covariates strongly associated with potential outcomes but exogenous to them, then they should be the same on either side of the cutoff if smoothness holds
- In this sense, balance tests are indirect searching for evidence supporting smoothness

Balance implementation

Don't make it hard – do what you did to Y, only to Z

- Choose other noncolliders associated with potential outcomes, Z
- Create similar graphical plots as you did for Y
- Could also conduct the parametric and nonparametric estimation on Z
- You do **not** want to see a jump around the cutoff, c_0

Visualizing Balance

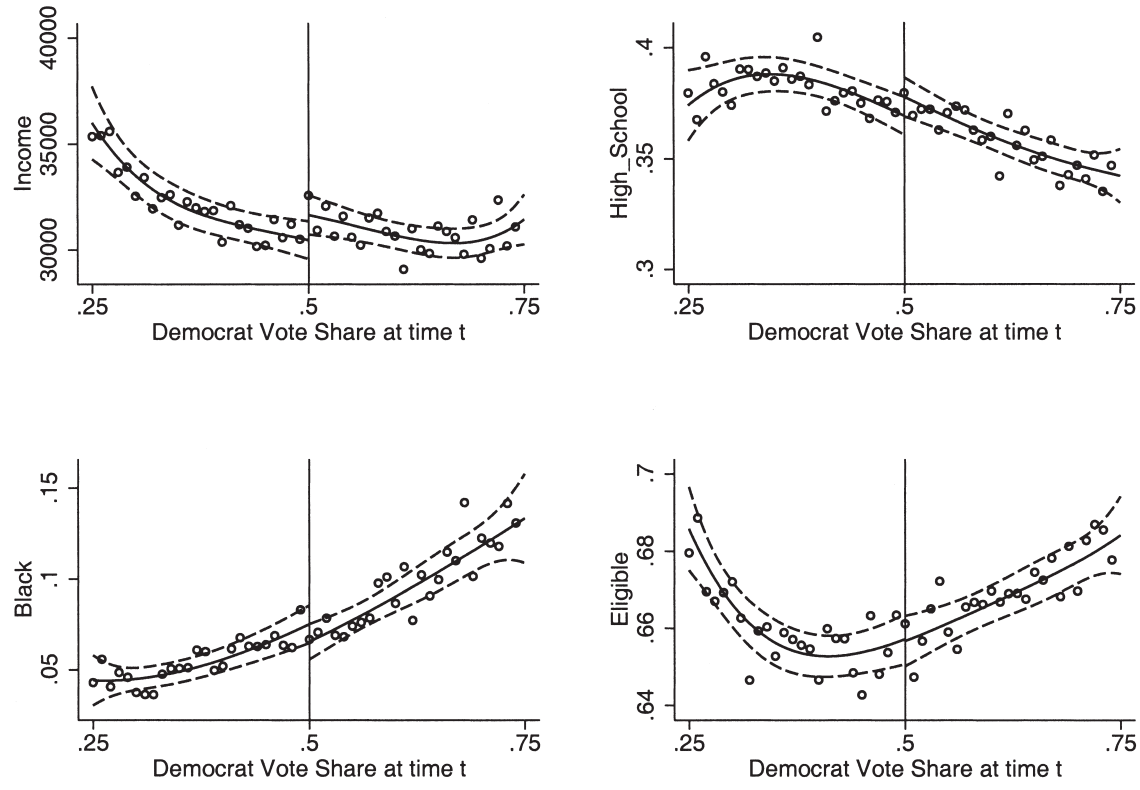


Figure: Figure 3 from Lee, Moretti and Butler (2004), "Do Voters Affect or Elect Policies?" *Quarterly Journal of Economcis*. Panels refer to (top left to bottom right) the following district characteristics: real income, percentage with high-school degree, percentage black, percentage eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share. The continuous line represents the predicted values from a fourth-order polynomial in vote share fitted separately for points above and below the 50 percent threshold. The dotted line represents the 95 percent confidence interval.

Placebos at non-discontinuous points

- Placebos in time are common with panels; placebo in running variables are their equivalent in RDD
- Imbens and Lemieux (2010) suggest we look at one side of the discontinuity (e.g., $X < c_0$), take the median value of the running variable in that section, and pretend it was a discontinuity, c_0'
- Then test whether in reality there is a discontinuity at c_0'. You do **not** want to find anything.
- Remember though: smoothness at placebo points is neither necessary nor sufficient for smoothness in the potential outcomes at the cutoff
- So there are Type I and Type II risks of error with this

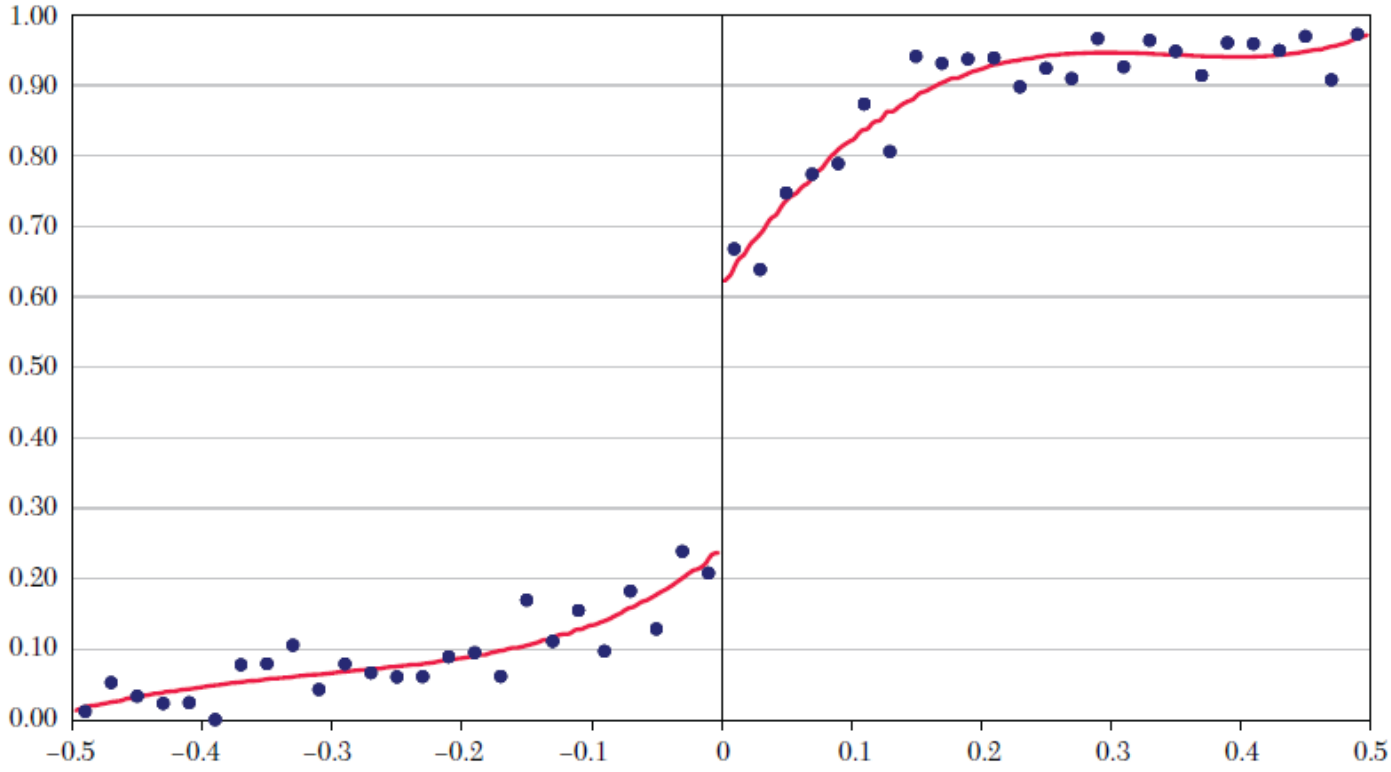
Pictures, pictures and more pictures

- Synthetic control and RDD are visually intense
- Eyeball tests are rampant (and deservedly) in RDD studies
- Even if your main results are all parametric, you'll still want to present at least some nonparametric style pictures according to Imbens and Lemieux (2010)
- Let's review some of the graphs you have to include

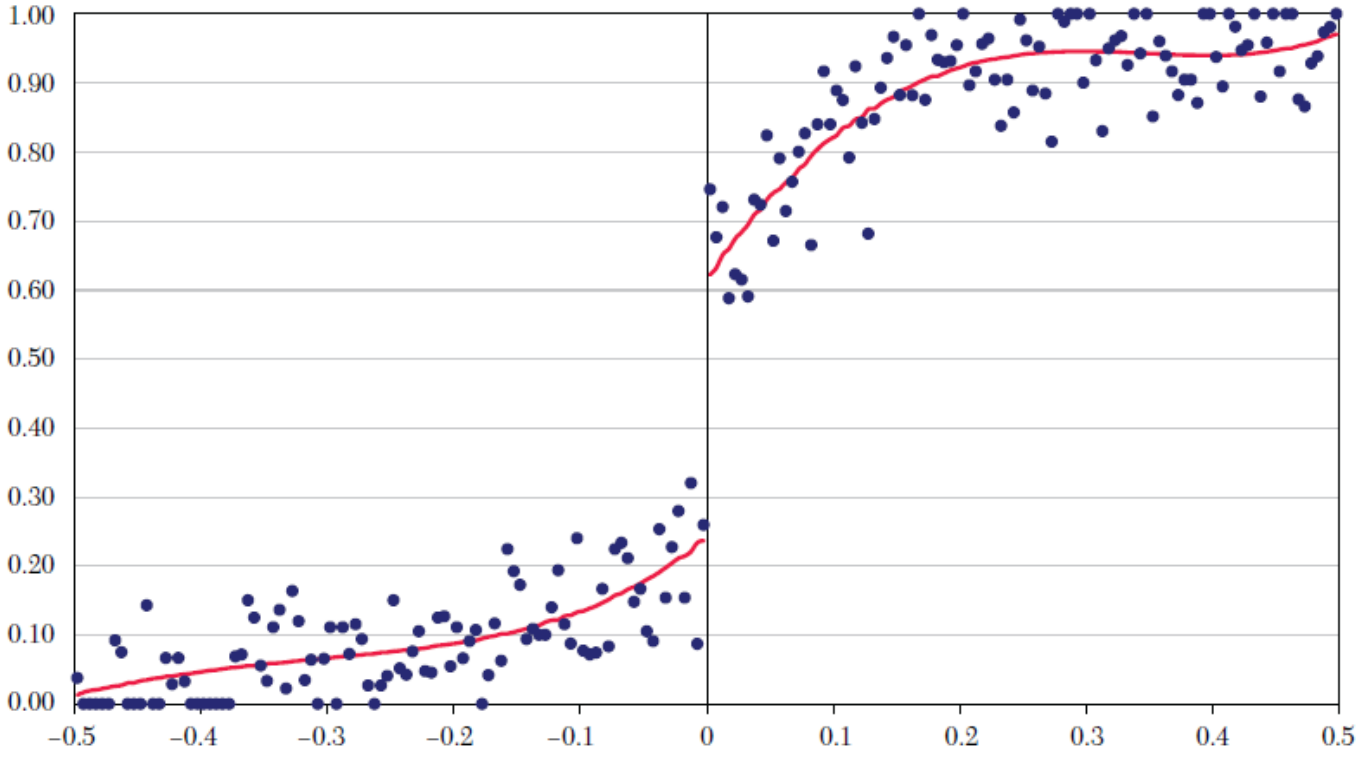
Outcomes

1. **Outcome by running variable, (X_i):**
   - → Construct bins and average the outcome within bins on both sides of the cutoff
   - → Look at different bin sizes when constructing these graphs
   - → Plot the running variables, X_i, on the horizontal axis and the average of Y_i for each bin on the vertical axis
   - → Consider plotting a relatively flexible regression line on top of the bin means, but some readers prefer an eyeball test without the regression line to avoid “priming”

Example: Outcomes by Running Variables



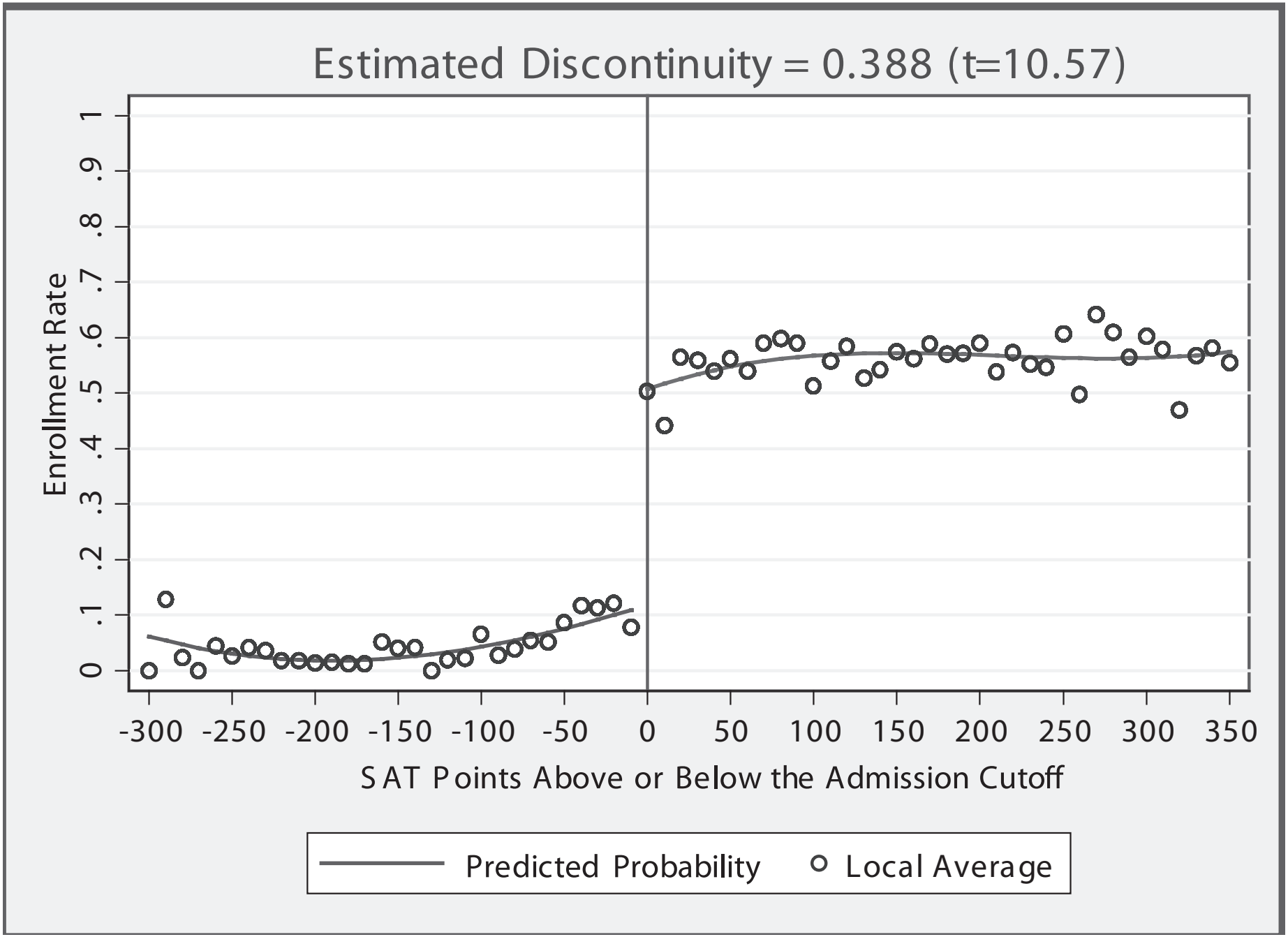
Example: Outcomes by Running Variables with smaller bins



Probability of treatment

2. **Probability of treatment by running variable if fuzzy RDD**
   - → In a fuzzy RDD, you also want to see that the treatment variable jumps at c_0
   - → This tells you whether you have a first stage ("bite")
   - → Let's look at that again from earlier Hoekstra (2008) and enrollment at the flagship

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



McCrary Density

3. **Density of the running variable**
   - → One should plot the number of observations in each bin.
   - → This plot allows to investigate whether there is a discontinuity or heaping in the distribution of the running variable at the threshold
   - → Heaping or discontinuities in the density suggest that people can manipulate their running variable score
   - → This is an indirect test of the identifying assumption that each individual has imprecise control over the assignment variable, which may violate smoothness

Density of the running variable

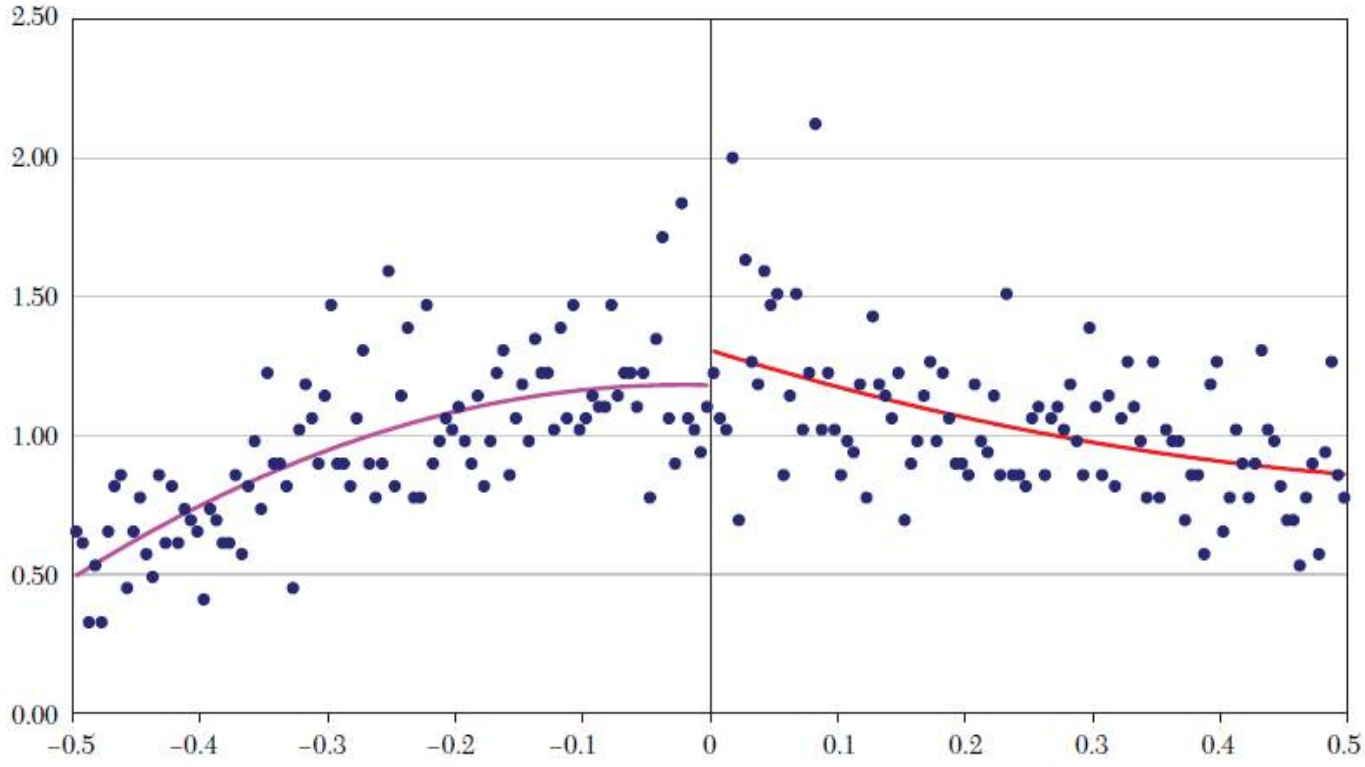


Figure 16. Density of the Forcing Variable (Vote Share in Previous Election)

Balance pictures

4. **Covariates by a running variable**
   - → Construct a similar graph to the outcomes graph but use a noncollider covariate as the “outcome”
   - → Balance implies smoothness through the cutoff, c_0.
   - → If noncollider covariates jump at the cutoff, one is probably justified to reject that potential outcomes aren’t also probably jumping there

Example: Covariates by Running Variable

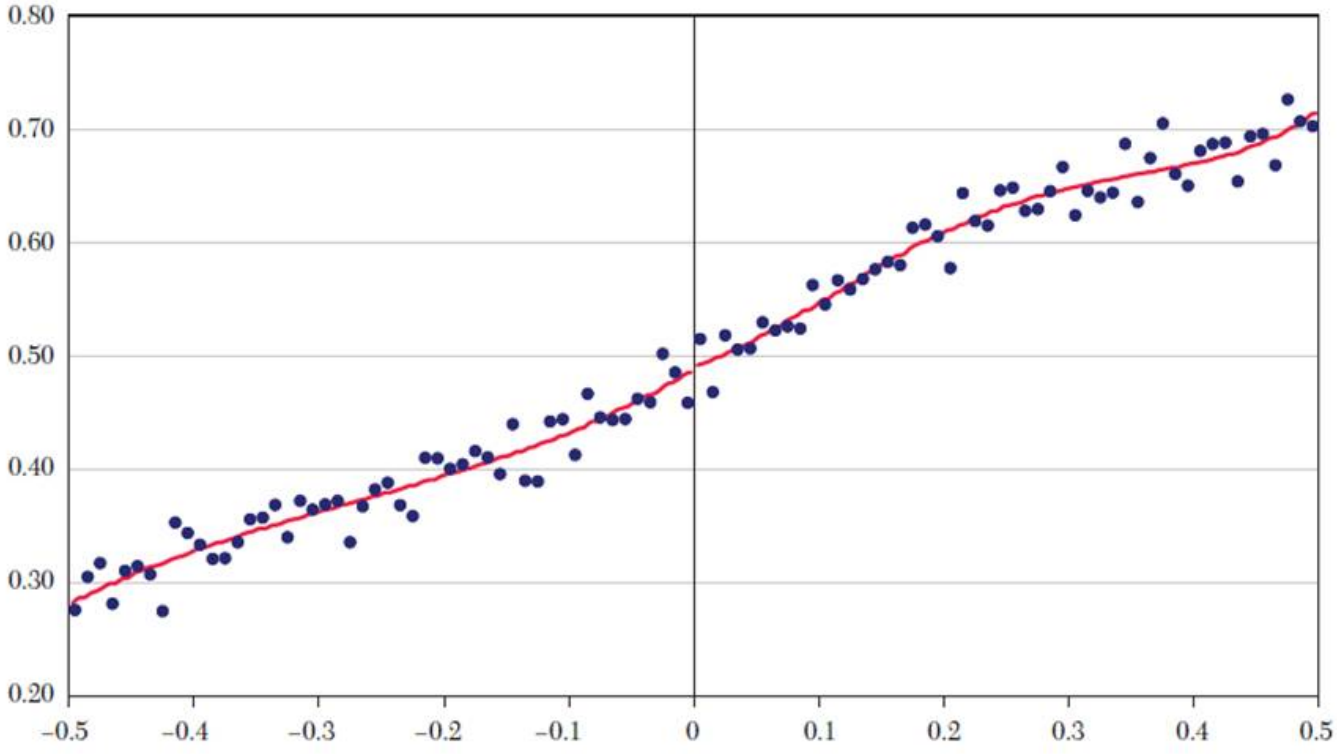


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

Inference – honesty

- Lee and Card (2008) and Lee and Lemieux (2010) recommend clustering standard errors on the running variable
- Kolesár and Rothe (2018) provide extensive theoretical and simulation-based evidence that this is not good; you'd be better off just with heteroskedastic robust
- They propose two alternative confidence intervals that achieve correct coverage in large samples – called "honest" (great intro! Still studying this procedure)
- Unavailable in Stata, but is available in R – RDHonest – at `https://github.com/kolesarm/RDHonest`

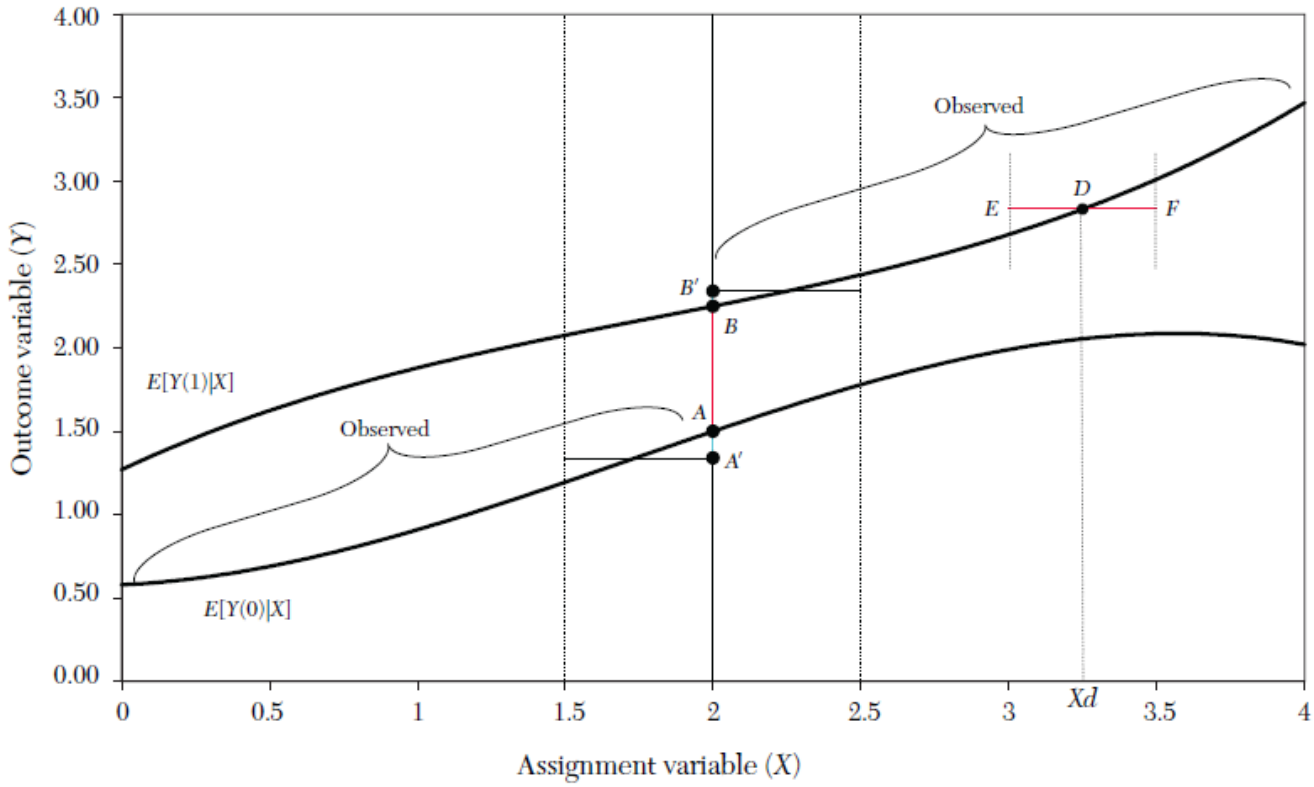
Inference – randomization inference

- Recall the “randomization” tradition – randomization leads directly into an alternative inference we covered earlier
- Cattaeneo, et al. (2015) say to consider that the cutoff is a randomized experiment
- Use randomization inference which is a test of the null of no individual unit level treatment effect at the cutoff
- Requires randomly assigning treatment and estimating test statistics using approximate simulations to calculate approximate p-values

Parametric vs. nonparametric approaches

- Least squares approaches, because it models the counterfactual using functional forms, is parametric
- As a result, it can have poor predictive properties on counterfactuals above/below the cutoff
- Another way of approximating $f(X_i)$ is to use a nonparametric kernel which has its own problems; just not that one.

Kernel regression



- While the "true" effect is AB, with a certain bandwidth a rectangular kernel would estimate the effect as $A'B'$
- There is therefore systematic bias with the kernel method if the $f(X)$ is upwards or downwards sloping

Kernel weighted local polynomial regression

- The nonparametric one-sided kernel estimation problems are called “boundary problems” at the cutoff (Hahn, Todd and Van der Klaauw 2001)
- Kernel estimation (such as lowess) may have poor properties because the point of interest is at a boundary
- They proposed to use “local linear nonparametric regressions” instead

Local linear regression with weights

- Local linear nonparametric regression substantially reduces the bias
- Think of it as a weighted regression restricted to a window – kernel provides the weights to that regression.

$$(\widehat{a}, \widehat{b}) \equiv_{a,b} \sum_{i=1}^{n} (y_i - a - b(x_i - c_0))^2 K\left(\frac{x_i - c_0}{h}\right) 1(x_i > c_0)$$

where x_i is the value of the running variable, c_0 is the cutoff, K is a kernel function and $h > 0$ is a suitable bandwidth

Animation of a local linear regression

https://twitter.com/page_eco/status/958687180104245248

Estimation

- Stata's `poly` estimates kernel-weighted local polynomial regressions.
- A rectangular kernel would give the same result as $E[Y]$ at a given bin on X. The triangular kernel gives more importance to observations close to the center.
- This method will be sensitive to how large the bandwidth (window) you choose

Optimal bandwidths

- A rectangular kernel would give the same result as taking $E[Y]$ at a given bin on X whereas the triangular kernel gives more importance to the observations closer to the center.
- While estimating this in a given window of width h around the cutoff is straightforward, it's more difficult to choose this bandwidth (or window), and the method is sensitive to the choice of bandwidth.

Bandwidths

- Several methods for choosing the optimal bandwidth (window), but it's always a trade off between bias and variance
- In practical applications, you want to check for balance around that window
- Standard error of the treatment effects can be bootstrapped but there are also other alternatives
- You could add other variables to nonparametric methods.

Bandwidths

- Imbens and Kalyanaraman (2012), and more recently Calonico, et al. (2017), have proposed methods for estimating “optimal” bandwidths which may differ on either side of the cutoff.
- Calonico, et al (2017) propose local-polynomial regression discontinuity estimators with robust confidence intervals
- Stata ado package and R package are both called `rdrobust`

DO VOTERS AFFECT OR ELECT POLICIES? EVIDENCE FROM THE U. S. HOUSE*

DAVID S. LEE
ENRICO MORETTI
MATTHEW J. BUTLER

There are two fundamentally different views of the role of elections in policy formation. In one view, voters can *affect* candidates' policy choices: competition for votes induces politicians to move toward the center. In this view, elections have the effect of bringing about some degree of policy compromise. In the alternative view, voters merely *elect* policies: politicians cannot make credible promises to moderate their policies, and elections are merely a means to decide which one of two opposing policy views will be implemented. We assess which of these contrasting perspectives is more empirically relevant for the U. S. House. Focusing on elections decided by a narrow margin allows us to generate quasi-experimental estimates of the impact of a "randomized" change in electoral strength on subsequent representatives' roll-call voting records. We find that voters merely *elect* policies: the degree of electoral strength has no effect on a legislator's voting behavior. For example, a large *exogenous* increase in electoral strength for the Democratic party in a district does not result in shifting both parties' nominees to the left. Politicians' inability to credibly commit to a compromise appears to dominate any competition-induced convergence in policy.

Implementation

- The following paper is a seminal paper in public choice both scientifically and methodologically – the close election RDD
- I call the close election RDD a type of sub-RDD in that it's widely used in political science and economics to the point that it's taken on a life of its own
- Let's take everything we've done and apply it by replicating this paper using programs I've provided

Public choice

There are two fundamentally different views of the role of voters in a representative democracy.

1. **Convergence**: Voters force candidates to become relatively moderate depending on their size in the distribution (Downs 1957).

   "Competition for votes can force even the most partisan Republicans and Democrats to moderate their policy choices. In the extreme case, competition may be so strong that it leads to 'full policy convergence': opposing parties are forced to adopt identical policies" – Lee, Moretti, and Butler 2004.

2. **Divergence**: Voters pick the official and after taking office, she pursues her most-preferred policy.

Falsification of either hypothesis had been hard

- Very difficult to test either one of these since you don't observe the counterfactual votes of the loser for the same district/time
- Winners in a district are selected based on their policy's conforming to unobserved voter preferences, too
- Lee, Moretti and Butler (2004) develop the "close election RDD" which has the aim of determining whether convergence, while theoretically appealing, has any explanatory power in Congress
- The metaphor of the RCT is useful here: maybe close elections are being determined by coin flips (e.g., a few votes here, a few votes there)

Outcome is Congress person's liberal voting score

- **Liberal voting score** is a report card from the Americans for Democratic Action (ADA) for the House election results 1946-1995
  - → Authors use the ADA score for all US House Representatives from 1946 to 1995 as their voting record index
  - → For each Congress, ADA chooses about twenty high-profile roll-call votes and creates an index varying 0 and 100 for each Representative of the House measuring liberal voting record

Democratic “voteshare” is the running variable

- **Voteshare** from the same races
  - → The running variable is `voteshare` which is the share of all votes that went to a Democrat.
  - → They use a close Democratic victory to check whether convergence or divergence is correct (what’s smoothness here?)
  - → Discontinuity in the running variable occurs at `voteshare`$= 0.5$. When `voteshare`> 0.5, the Democratic candidate wins.
- I’ll show `lmb1.do` to `lmb10.do` (and R) at times just so we can all see the simple estimation methods ourselves.

Remember these results

TABLE I
RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

	Total effect			Elect component	Affect component
	γ	π_1	$(P^D_{t+1} - P^R_{t+1})$	$\pi_1[(P^D_{t+1} - P^R_{t+1})]$	$\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$
Variable	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2)*(col. (3))	(col. (1)) − (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2	47.6	0.48		
	(1.9)	(1.3)	(0.02)		
				22.84	−1.64
				(2.2)	(2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

Figure: Lee, Moretti, and Butler 2004, Table 1.

Nonparametric estimation

- Hahn, Todd and Van der Klaauw (2001) emphasized using local polynomial regressions
- Estimate $E[Y|X]$ in such a way that doesn't require committing to a functional form
- That model would be something general like

$$Y = f(X) + \varepsilon$$

Nonparametric estimation (cont.)

- We'll do this estimation just rolling $E[ADA]$ across the running variable $voteshare$ visually
- Stata has an option to do this called `cmogram` and it has a lot of useful options, though many people prefer to graph it themselves bc it gives more flexibility.
- We can recreate Figures I, IIA and IIB using it

Future liberal voting score

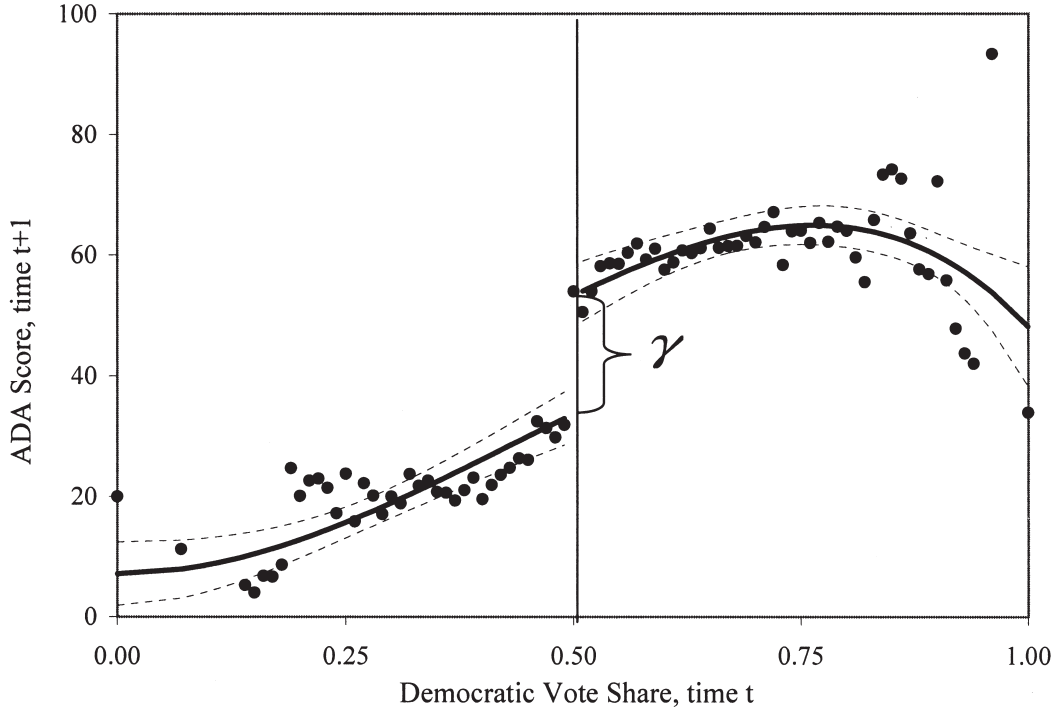


FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t. Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Affect”}} + \underbrace{\pi_1(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Elect”}}.$$

Figure: Lee, Moretti, and Butler 2004, Figure I. $\gamma \approx 20$

Contemporaneous liberal voting score

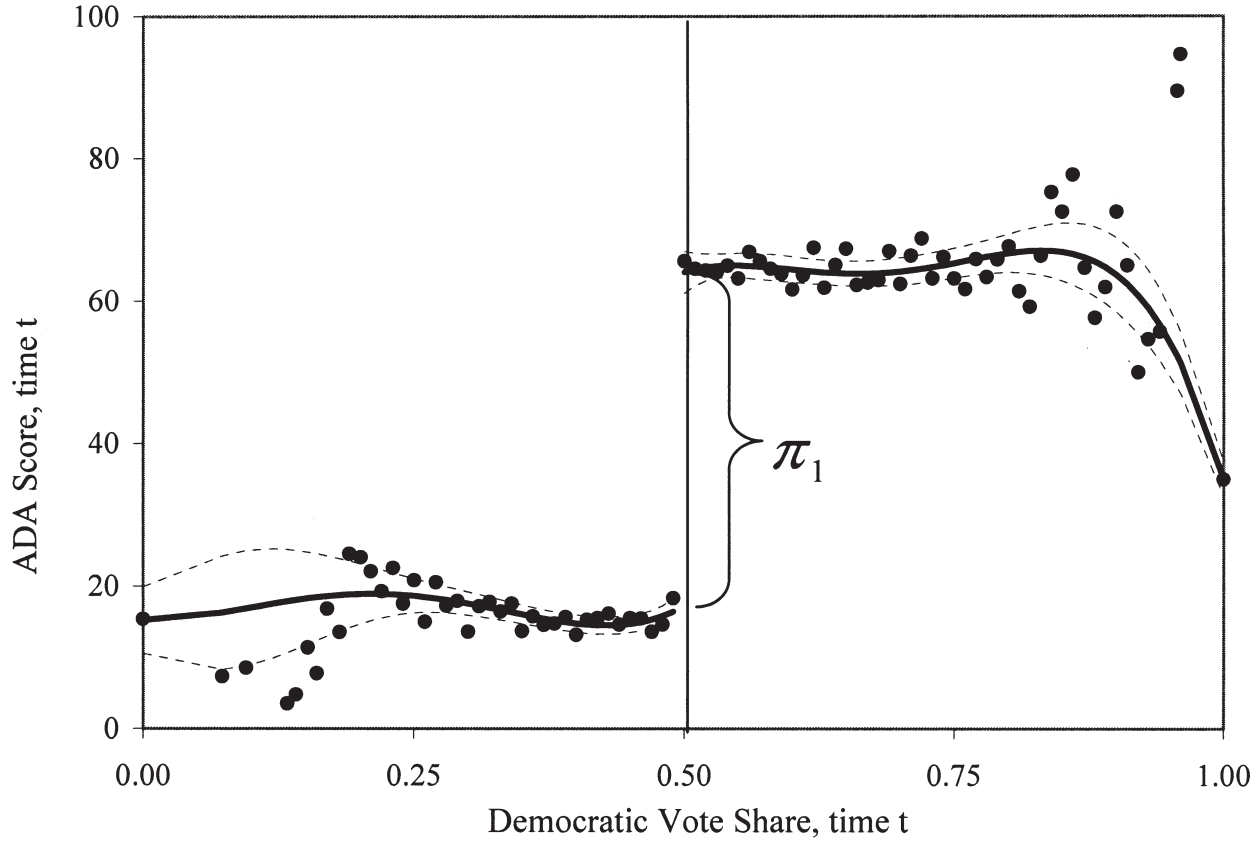


FIGURE IIa
Effect of Party Affiliation: π_1

Figure: Lee, Moretti, and Butler 2004, Figure IIa. $\pi_1 \approx 45$

Incumbency advantage

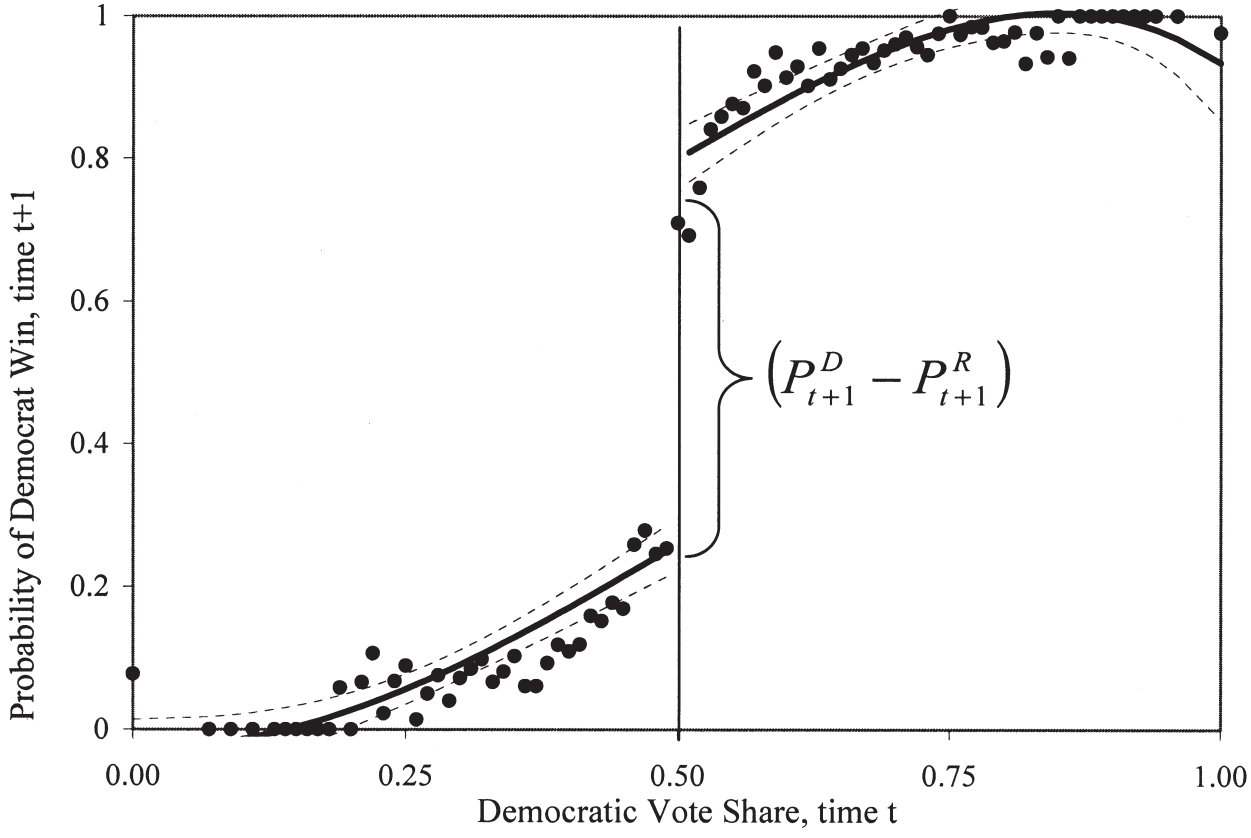


FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P^D_{t+1} - P^R_{t+1})$

Figure: Lee, Moretti, and Butler 2004, Figure IIb. $(P^D_{t+1} - P^R_{t+1}) \approx 0.50$

Replication of LMB

- We can replicate their results – both the table and the figure
- Let's look at it together using our code
- I'll walk us through some extensions I've done in Stata and an exercise I leave to you is to do it in R and python

Concluding remarks

- Caughey and Sekhon (2011) questioned the finding (not the design per se) saying that bare winners and bare losers in the US House elections differed considerably on pretreatment covariates (imbalance), which got worse in the closest elections
- Eggers, et al. (2014) evaluated 40,000 close elections including the House in other time periods, mayor races, and other types of US races including nine other countries
- They couldn't find another instance where Caughey and Sekhon's critique applied
- Assumptions behind close election design therefore probably holds and is one of the best RD designs we have