Causal Inference I

Mixtape Session

Roadmap

- Introducing Regression Discontinuity Design
  - Basic background
  - Identification Basics
  - Sharp Design
  - Smoothness and Identification
- Estimation
  - Local Regressions
  - Nonparametric estimation
  - Testing for violations
  - Data Visualization
- Examples
  - DWI and recidivism
  - Close election design
  - Fuzzy RDD

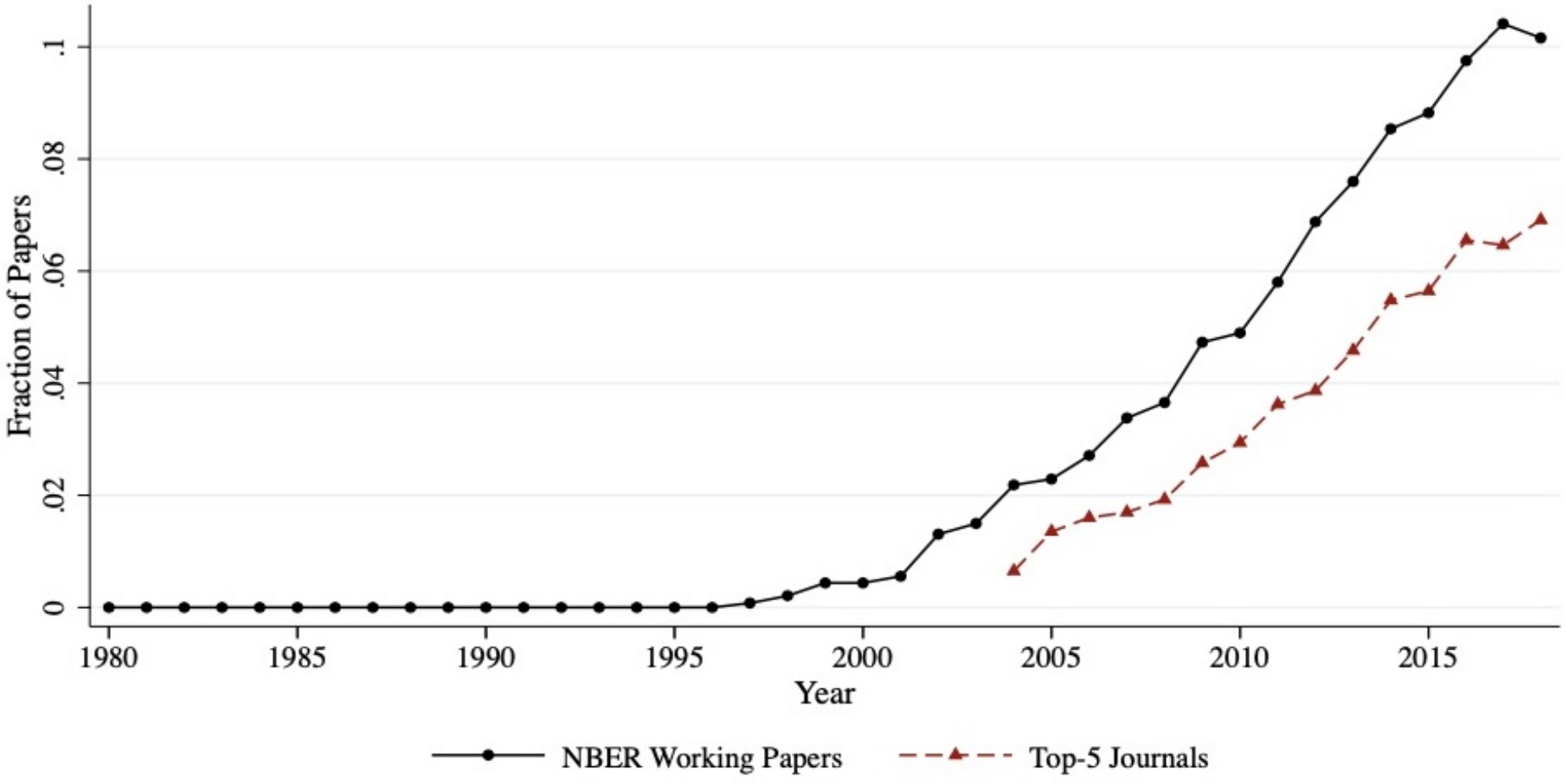
What is regression discontinuity?

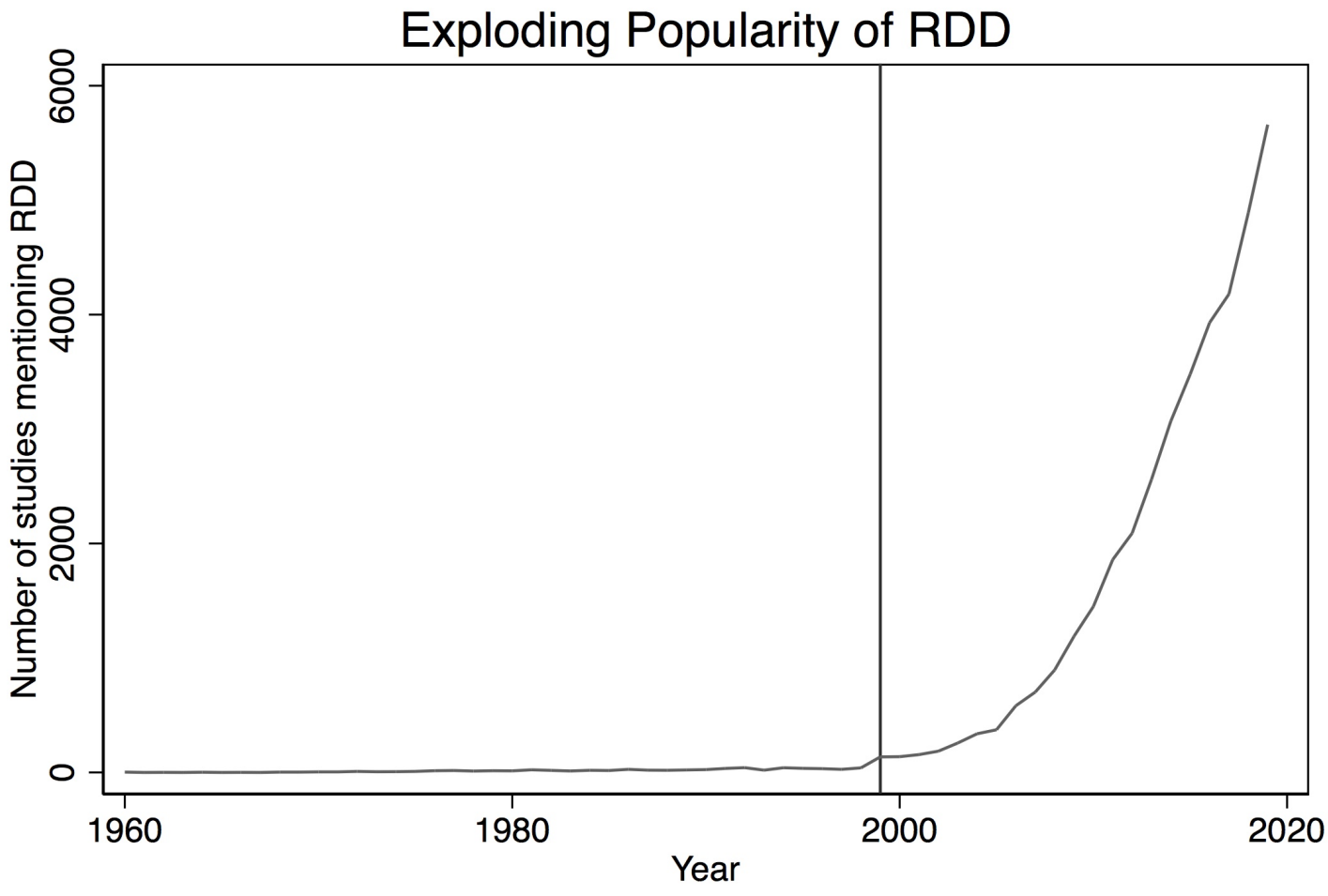
RDD is an extremely popular particular type of research design.

Invented by Donald Campbell, educational psychologist, it went dormant for decades until around 1999 when I got “rediscovered” by Josh Angrist, Victor Lavy and Sandra Black

Often thought to be the most “credible” of the observational designs, even though it does not depend on randomization for identification

B: Regression Discontinuity





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

Regression discontinuity design

- We want to estimate some causal effect of a treatment on some outcome
- But we can’t compare two groups (treated and not treated) because of the self selection which creates selection bias
- But what if treatment assignment was forced on units because the firm or agency uses a multi valued variable and splits the sample when units are above or below some threshold?
- RDD formalizes this setup and under some assumptions will identify causal effects

Breakout: RDD pictures

Each of the following 3 (but 4 slides) pictures is telling a story with distinct statistical objects

Help me understand the questions at the bottom of each picture by discussing what you think is happening

RDD picture # 1

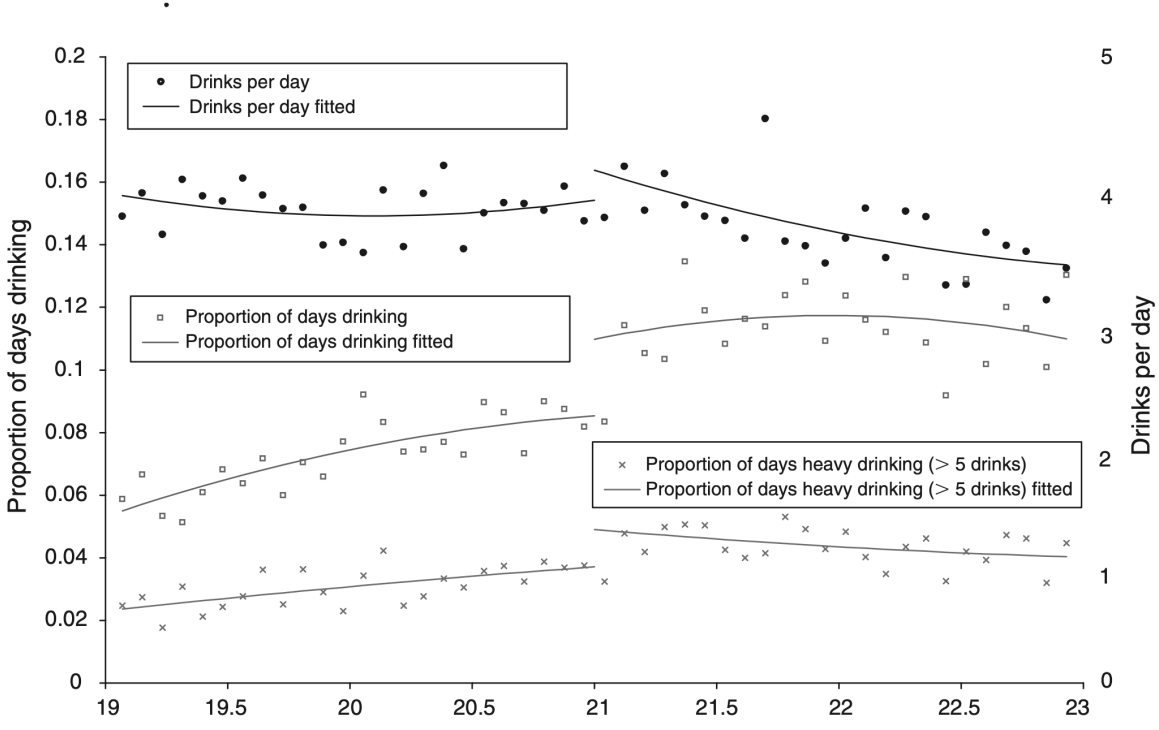


FIGURE 2. AGE PROFILE OF DRINKING INTENSITY

Notes: People can report their drinking for the last week, month, or year; 71 percent of respondents used a reference period of one week or one month. Average number of drinks per day is for people who reported some drinking.

Question: What is age doing here? What's the story being told?

RDD picture # 2

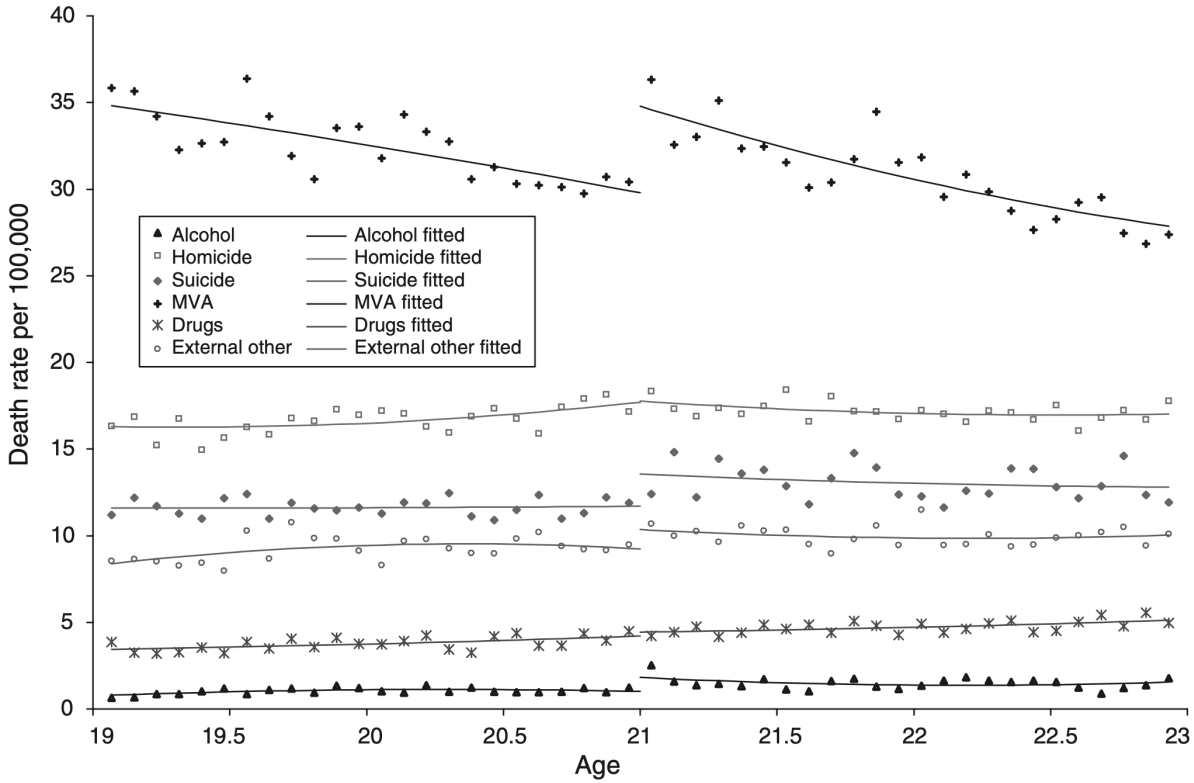


FIGURE 4. AGE PROFILES FOR DEATH RATES BY EXTERNAL CAUSE

Notes: See notes to Figure 3. The categories are mutally exclusive. The order of precedence is homicide, suicide, MVA, deaths with a mention of alcohol, and deaths with a mention of drugs. The ICD-9 and ICD-10 Codes are in Appendix C.

Question:What do you think each dot represents? What do you think each line represents? What's the story being told?

RDD picture # 3

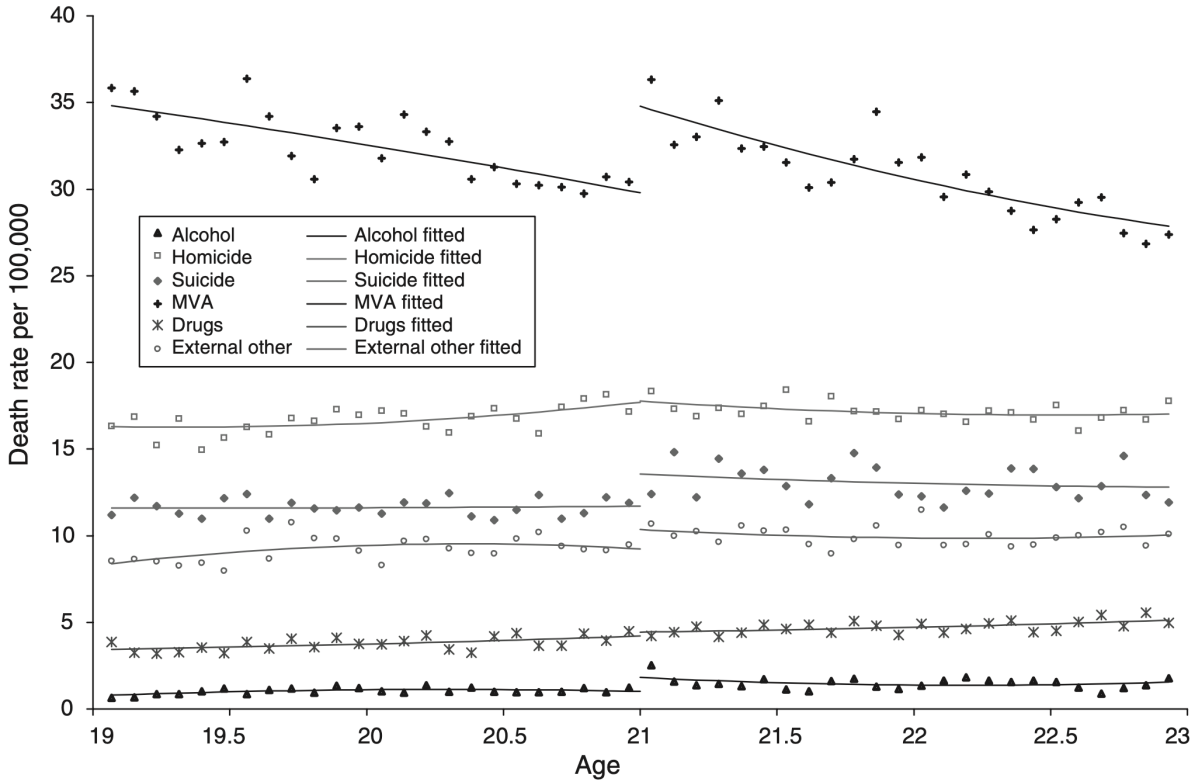


FIGURE 4. AGE PROFILES FOR DEATH RATES BY EXTERNAL CAUSE

Notes: See notes to Figure 3. The categories are mutally exclusive. The order of precedence is homicide, suicide, MVA, deaths with a mention of alcohol, and deaths with a mention of drugs. The ICD-9 and ICD-10 Codes are in Appendix C.

Question:Where is the treatment effect on the graph? Treatment effect *of what*? Give a rough approximation of the treatment effect?

Examples of RDD

- Rather than cover examples at the end, I'd like to cover it at the front
- I'll keep jargon to a minimum, and skip over identification issues
- Hope is that this sparks some ideas for your own work

Rounding to nearest star

- Several interesting papers use a “rounding to nearest star” approach (**tons** of these papers, both in industry and social science applications)
  - → Michael Luca (2011) looks at restaurant revenue
  - → Anderson and Magruder (2011) look at wait times at restaurants,
  - → Dell and Querubin (2018) look at bombing runs in Vietnam war
- Let’s summarize one of them at a high level to get us going

Good Restaurants are Popular

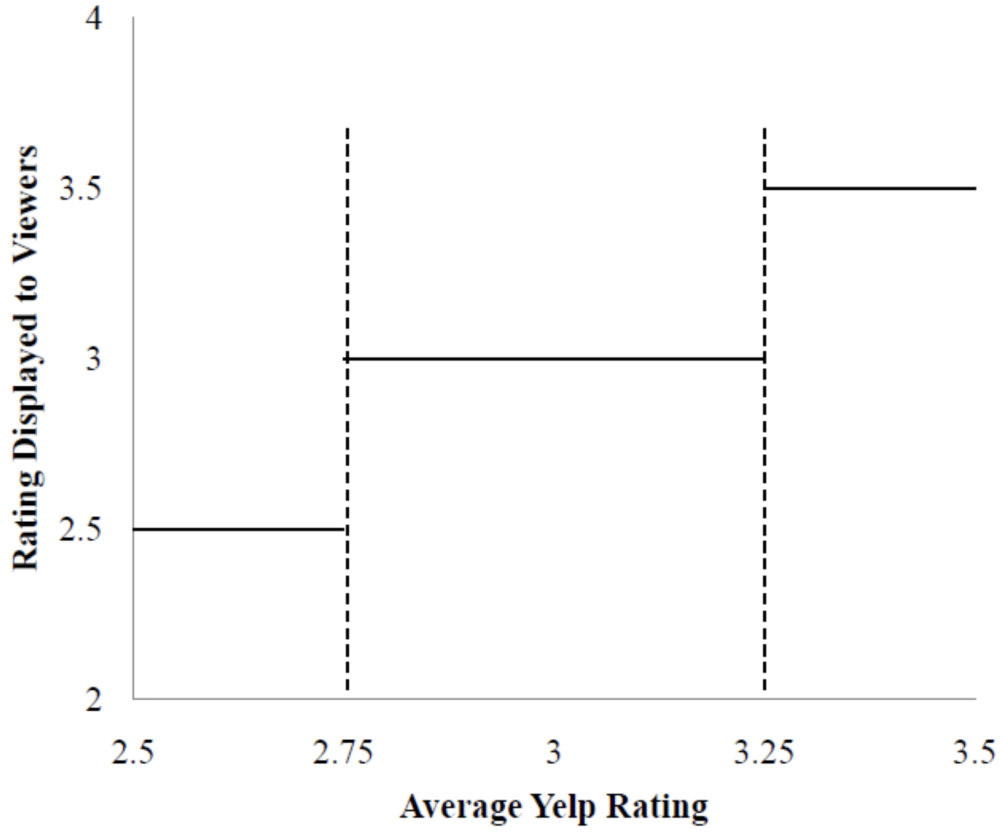
- Anderson and Magruder (2011) wonder if higher rated stars make restaurants popular or if it's the other way around
- They'll look at higher rated stars effect on longer wait times
- But stars are based on unobserved quality
- They will separate selection bias from causality using rounded Yelp average reviews with merged wait times

Data

- **Context**: San Francisco (Luca had looked at Seattle which has lower rated restaurants)
- **Scraped Yelp**: “When clicking on an individual business, Yelp.com displays the entire history of reviews for that business. We downloaded the history for each restaurant on Yelp.com and recorded the date of the review, the rating assigned (1-5), and the reviewer’s unique user identifier. We then reconstructed the average rating and total number of reviews for each restaurant at every point in time.”
- **Outcome**: Reservation availability data from “large online reservation website” which lists real-time reservation availability for hundreds of restaurants in SF then merged with Yelp data

Yelp used to show the rounded star only

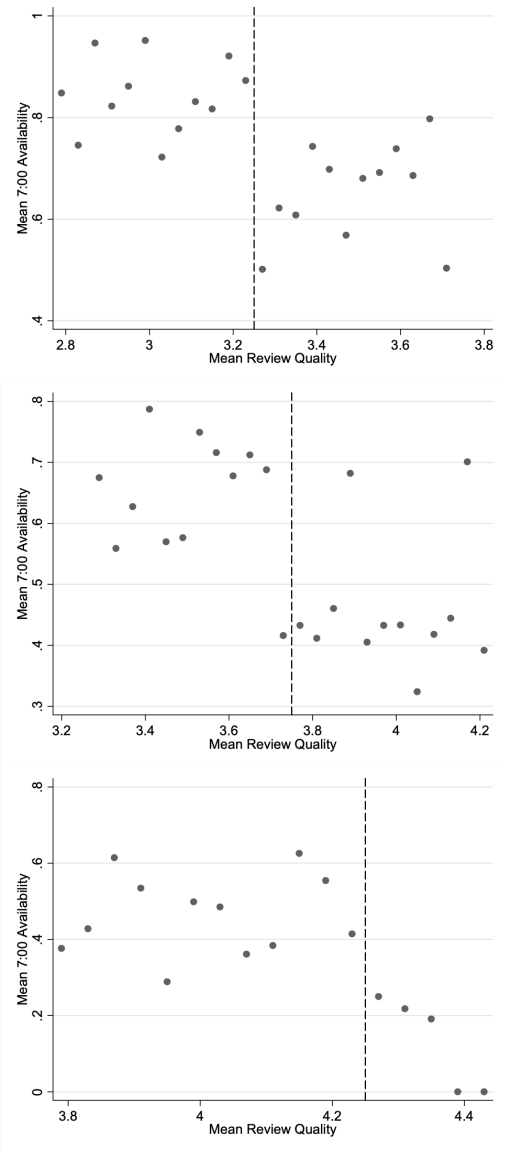
Figure 3: Yelp Displays Each Restaurant's Rounded Average Rating



Notes: Yelp prominently displays a restaurant's rounded average rating. Each time a restaurant's rating crosses a rounding threshold, the restaurant experiences a discontinuous increase in the displayed average rating.

Wait times and star: pictures

Fig. 2: Reservation Availability at 7:00 pm by Average Yelp Rating



Wait times and star: OLS

Table 2: Regression Discontinuity Results at Individual Thresholds

	6:00 Availability			7:00 Availability			8:00 Availability		
Yelp Display Rating	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
3.5 Yelp Stars	-0.079			-0.213 **			-0.150 *		
	(0.086)			(0.096)			(0.080)		
4 Yelp stars		-0.101			-0.192 **			-0.095	
		(0.075)			(0.093)			(0.086)	
4.5 Yelp stars			0.004			-0.113			-0.119
			(0.185)			(0.127)			(0.149)
Yelp Rating	-0.228	0.145	-0.131	0.082	0.024	-0.022	0.088	0.008	-0.321
	(0.201)	(0.203)	(0.230)	(0.216)	(0.255)	(0.271)	(0.180)	(0.218)	(0.276)
Yelp Rating*Yelp Star	0.372	-0.275	-2.934 **	-0.057	-0.048	-1.817 ***	-0.080	-0.329	-1.324
	(0.287)	(0.309)	(1.342)	(0.335)	(0.375)	(0.674)	(0.282)	(0.352)	(0.869)
Observations	8,705	11,858	5,597	8,705	11,858	5,597	8,705	11,858	5,597

Notes 1. Contains RD estimates of the effects of an additional Yelp half-star on availability

2. Availability measures indicate whether the reservations were available at that time on Thursday, Friday, or Saturday when queried 36 hours in advance

3. Standard errors are clustered at the restaurant level

4. Stars denote significance levels: 10% (*), 5% (**), and 1% (***)

RDD Words and Pictures

- Just keep in mind as do this – RDD is a method of mimicking the experimental design, as opposed to merely a regression model
- There's a lot of new terminology if you're new to RDD
- Tons of pictures, but tons of weird concepts too

RDD language

- **Running variable**: a usually continuous score that some firm or agency uses to assign treatments to units

RDD language

- **Trends**: Not time trends; changing outcomes along the running variable (e.g., food gets better for restaurants with higher star ratings)

RDD language

- **Cutoff** or **threshold**: a particular value at a point on the running variable above which the firm or agency assigns treatments to unit

RDD language

- **Discontinuity** and/or **Jump**: Since we are *estimating* breaks in the outcome right at the cutoff, and when that happens we say that there is a “discontinuity”

RDD language

- **Regression**: Many of the models are simple difference in means, local regressions from OLS or global regressions from OLS

Some common types of RDD

- **Rounding**: Restaurants get “stars” (e.g., Michelin) and more revenue, but the ones that get more stars were probably better restaurants even without the star. Yelp rounds to the nearest half star even though the underlying “score” is continuous. The continuous score assigns stars
- **Close elections**: Majority voting, with a cutoff of 0.5, assigns different political parties into power
- **Algorithms**: Uber’s surge pricing based on underlying demand measures
- **Biology**: Age and drinking; Birthweight and ICU medical care

Data requirements

Large sample sizes are characteristic features of the RDD

- If there are strong trends in the running variable, one typically needs a lot more data than if there weren't
- If the observations tend to be noisy, we need more data than if it was less noisy
- We need a lot of data bc we need significant mass at the running variable to reject the null
- Rewards people with access to firm level data since it can be large

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

Switching equation

Recall our definitions around notation Y vs Y^0 and Y^1

1. Potential outcomes have superscripts, Y_i^1, Y_i^0. They refer to a person i hypotheticals in worlds with or without treatment *regardless if that happened*
2. One of these two is chosen when a treatment, D_i, is chosen according to the switching equation

$$Y_i = D_i Y_i^1 + (1 - D_i) Y_i^0$$

3. Observed outcomes have no superscripts, Y_i. They refer to a person i observed outcome which is selected from one of their two potential outcomes (with binary treatments)

The idea of a counterfactual is based on (1), not (2)

Why did I bring that up?

- Something about RDD makes the subtlety of potential outcomes even more subtle
- The identifying assumptions are expressed purely in terms of $E[Y^1]$ and $[Y^0]$ which we'll see
- But you'll be looking for changes in $E[Y]$, and so it's very easy for this to really be confusing
- So just remember – we move from potential outcomes to realized outcomes based on treatment assignment

Sharp vs. Fuzzy RDD

- There's two classes 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

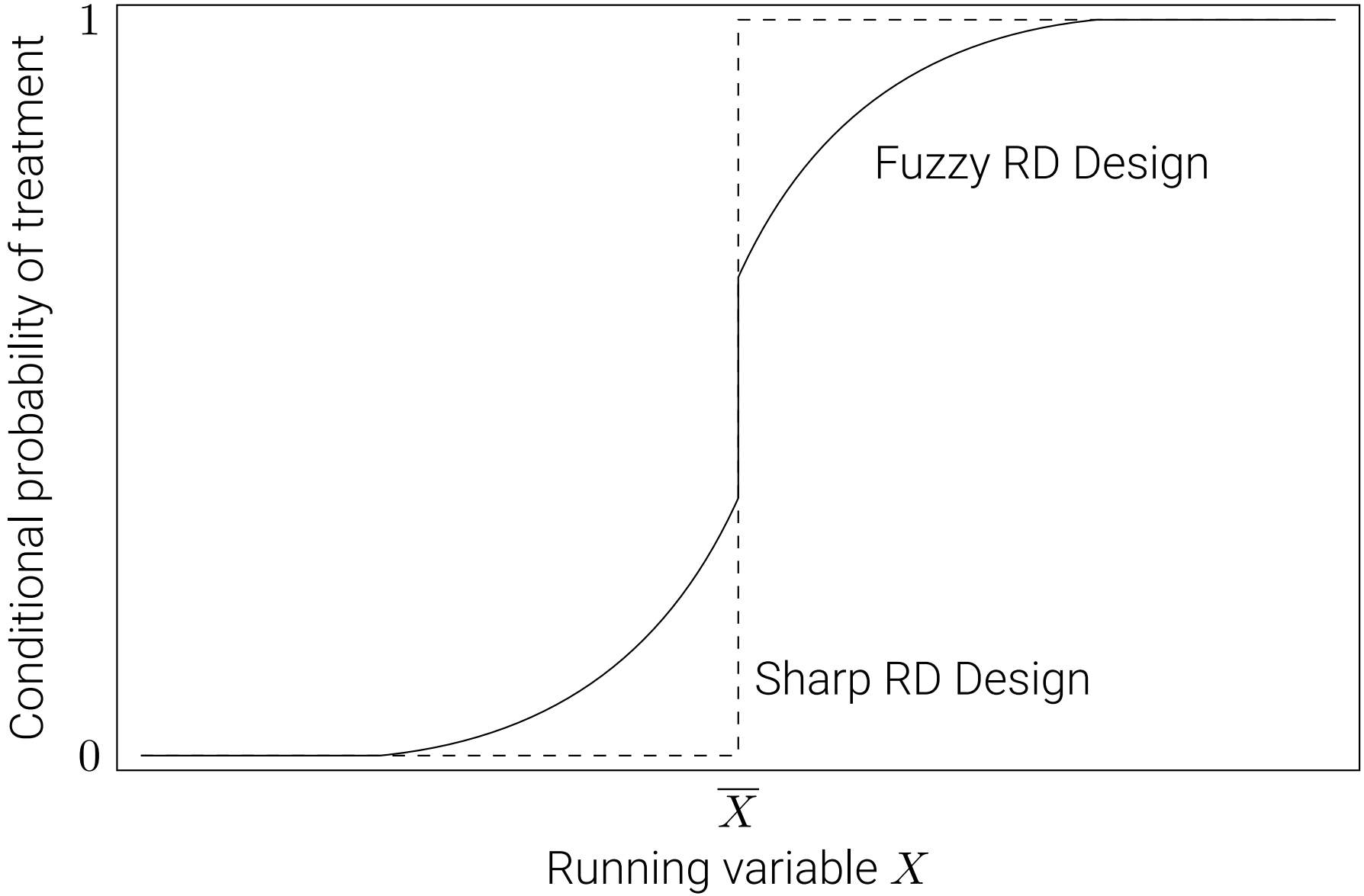


Figure: Sharp (dashed) vs. Fuzzy (solid) RDD

Some RDD issues

- **Common support**: We don't have units in treatment and control along the running variable which makes comparisons across the running variable impossible
- **Extrapolation**: Without common support, we extrapolate using models like regression and nonparametric methods by comparing units just below and above the cutoff to one another but this is sensitive to trends, bandwidths, and number of observations
- **Treatment effects**: We are estimating average treatment effects but only for people *at the cutoff* and that may not be informative of any other point on the running variable with extreme heterogeneity

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.

Extrapolation, common support and functional form

- Sharp designs create common support problems because there will literally *never* be a unit in treatment and control across the running variable
- This requires “extrapolation”; prediction beyond the support of the data (i.e., where treatment switches at cutoff)
- But since you’re predicting, modeling choices like **functional form** are key and that’s a structural assumption

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 (LATE) of the treatment (D) at the discontinuity (c_0)

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

Classes of identification assumptions

Independence assumption – largely associated with physical randomization, eliminates selection bias

- RCTs and A/B tests
- Matching and weighting (selection on observables)
- Instrumental variables

Classes of identification assumptions

Restricting potential outcomes – largely associated with RDD, DiD and synthetic control

- Regression discontinuity design – restricts $E[Y^0]$ and $E[Y^1]$ across some threshold
- Difference-in-differences – restricts $E[Y^0|D=1]$ to change over time in a restricted way
- Synthetic control – models Y^0 as a factor model

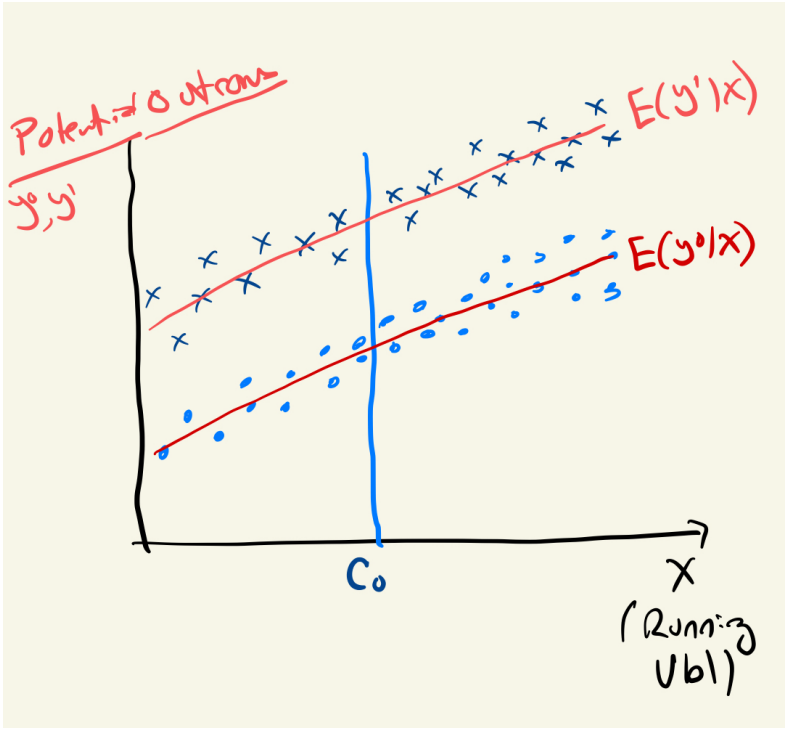
Smoothness as the identifying assumption

Smoothness of conditional expected potential outcome functions through the cutoff

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

- If population average *potential outcomes*, $E[Y^1]$ and $E[Y^0]$, are smooth functions of X across the cutoff, c_0, then expected potential average outcomes *won't* jump at c_0.
- Implies that the confounders are also evolving smoothly across the cutoff

Smoothness vs Treatment Effect



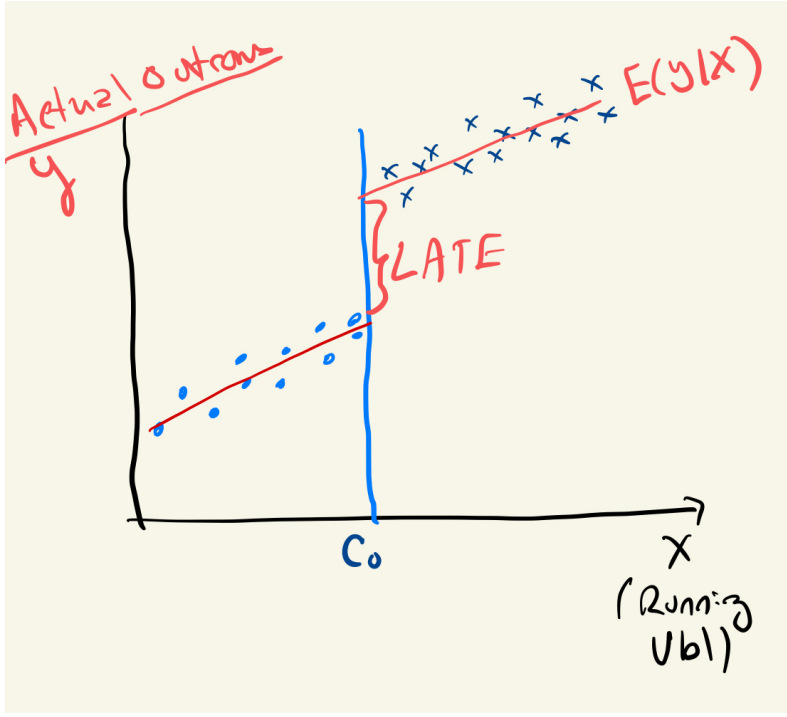


Figure: Smoothness of potential outcomes (left) vs estimation of LATE (right)

Discussion: Why is the left picture different from the right picture? Where did the two lines go?

Potential and observable outcomes

- Smoothness is about potential outcomes:
  - → Potential outcomes are on average smoothly changing across the threshold
  - → But remember: once we pass the threshold, we switch between potential outcomes
- Discontinuity is about realized outcomes:
  - → The cutoff *is the assignment mechanism*
  - → The cutoff switches between potential outcomes
  - → Therefore if there is a treatment effect, we can observe the *realized* outcomes jump at the cutoff
  - → If there is a treatment effect, it would be visible but it requires some extrapolation to see

Smoothness permits extrapolation

- Smoothness justifies the use of regression models to extrapolate missing potential outcomes from one side of the cutoff to the other (has a matching feel)
- Average causal effect is defined *at the cutoff*, but estimation uses data left and right *around the cutoff*
- Remember: identification and estimation are not the same thing – smoothness is what gives our estimates a causal interpretation

Roadmap

- Introducing Regression Discontinuity Design
  - Basic background
  - Identification Basics
  - Sharp Design
  - Smoothness and Identification
- Estimation
  - Local Regressions
  - Nonparametric estimation
  - Testing for violations
  - Data Visualization
- Examples
  - DWI and recidivism
  - Close election design
  - Fuzzy RDD

Approximate the functional form

Two ways to estimate the treatment effect at $X = c_0$

1. Use global and local regressions with $f(X_i)$ equalling a p^{th} order polynomial (highly sensitive to functional form)
2. Nonparametric kernel methods and local linear regressions (less sensitive)

Estimation with extrapolation

- We use *extrapolation* to estimate average treatment effects with the sharp RDD which is unbiased under *smoothness*
- Our statistical models predict expected conditional *counterfactuals* using data on *the other side of the cutoff*
- Keep in mind though: the actual aggregate causal effect is $Y_i^1 - Y_i^0$ at any point on X_i – not across $X = c_0$

Re-centering the running variable

- Assume a linear function

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

- People will often "re-center" by subtracting c_0 from X_i:

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

- This doesn't change the interpretation of the treatment effect; just the intercept.

Re-centering the running variable

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

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

  where $\alpha = \beta_0 - \beta_1 65$.

- All other coefficients, notice, have the same interpretation, except for the intercept.

Smooth but nonlinear expected potential outcomes

- Smoothness is an assumption about the behavior of the conditional expected potential outcomes as we move across the running variable and through the cutoff
- Does not imply *linear* evolution of expected potential outcomes though
- 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

Smooth but nonlinear expected potential outcomes

- 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:

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

- $f(X_i)$ models the counterfactual values of Y^0 and since we are extrapolating, we need an estimator that we think is extrapolating correctly
- You'll likely use higher order polynomial transformations of the running variable

Potential outcomes and nonlinear running variable

- But what if the potential outcomes aren't just nonlinear – the nonlinearities are different for $E[Y^1]$ than they are for $E[Y^0]$
- We can generalize the potential outcome expressions by allowing them to depend on the running variables, but in different ways depending on whether it is or is not treated

Potential outcomes and nonlinear running variable

- This will require saturated models in which you include them both individually and interacting them with D_i.

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

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

- Notice the treatment effect in the second line, and the intrinsic ATE when comparing the two equations, $E[Y_i^0 - Y_i^1|X_i]$

Interact running variable with treatment

- 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
- Interactions of treatment with running variable terms, $D \times X$, where X has be re-centered allows for the "two line" pictures we've been looking at
- The sole coefficient will yield the ATE at the cutoff and it obtains this estimating by extrapolating both sides to the other

Regression equation

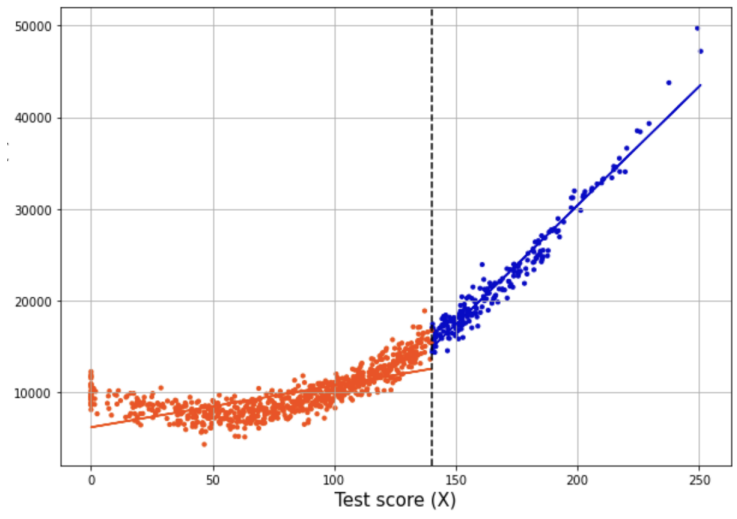
- Regression model you estimate is:

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

  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
- Look closely at how the beta terms aren't the same for $D = 0$ as they are for $D = 1$ (which corresponds to below and above the cutoff)
- But the parameter of interest, the treatment effect, is the coefficient at c_0 or $\widehat{\delta}$

Estimation without and with specifying nonlinear running variable



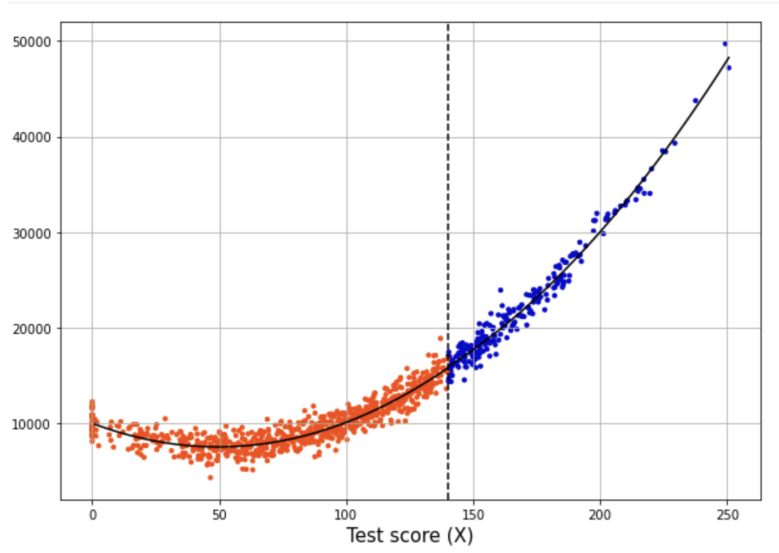


Figure: Spurious treatment effects with linear specification (left) versus 3rd order polynomial (right)

Look close: See how the lines don't touch on the left, but they do on the right?

OLS estimation with 3rd order polynomial

```
# Fully interacted regression
all_columns = "+".join(dat.columns.difference(["D", 'y3']))
formula = 'y3 ~ D * ({})'.format(all_columns)

regression = sm.OLS.from_formula(formula, data = dat).fit()
print(regression.summary())
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                     y3   R-squared:                       0.976
Model:                            OLS   Adj. R-squared:                  0.975
Method:                 Least Squares   F-statistic:                     5675.
Date:                Tue, 18 Jan 2022   Prob (F-statistic):               0.00
Time:                        01:13:59   Log-Likelihood:                -8307.1
No. Observations:                1000   AIC:                         1.663e+04
Df Residuals:                     992   BIC:                         1.667e+04
Df Model:                           7
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept   9930.4618    148.763     66.754      0.000    9638.536    1.02e+04
D          -4214.5653   1.97e+04     -0.214      0.831   -4.29e+04    3.45e+04
x            -95.2554      8.504    -11.201      0.000    -111.944     -78.567
x2             0.9255      0.141      6.563      0.000       0.649       1.202
x3             0.0004      0.001      0.599      0.549      -0.001       0.002
D:x           92.5239    323.510      0.286      0.775    -542.319     727.367
D:x2          -0.5847      1.751     -0.334      0.739      -4.021       2.852
D:x3           0.0010      0.003      0.316      0.752      -0.005       0.007
==============================================================================
Omnibus:                        2.359   Durbin-Watson:                   1.943
Prob(Omnibus):                  0.307   Jarque-Bera (JB):                2.378
Skew:                           0.047   Prob(JB):                        0.305
Kurtosis:                       3.220   Cond. No.                     2.40e+09
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 2.4e+09. This might indicate that there are
strong multicollinearity or other numerical problems.
```

Comment about higher order polynomials

- If you don't have a lot of data, you will likely have to have very large bandwidths just to get the sample size up
- With a lot of data far from the cutoff, you'll likely overfit with a higher order polynomial series
- But higher order polynomials can have overfitting problems leading to poor prediction beyond the cutoff
- Gelman and Imbens (2018) caution against overfitting on these global regressions (i.e., quadratics)

Some new terms

- **Kernels** make a window and give you the shape of the window (e.g., triangular kernels weight the observations differently within the window)
- **Bandwidth** is the “length” of the window (small ones are tiny windows, bigger ones, bigger windows – think of a histogram)
- **Bins** are about the interval itself (a partition)

Local linear nonparametric regressions

- Least squares approaches models the counterfactual using functional forms which is parametric, but it can have poor predictive properties on counterfactuals above/below the cutoff
- Another way of approximating the running variable flexibly $f(X_i)$ is to use a nonparametric kernel
- Hahn, Todd and Van der Klaauw (2001) proposed "local linear nonparametric regressions" which is weighted least squares for a given bandwidth h and weights that vary by distance to the cutoff

Local linear nonparametric regressions

- 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.
- It's like a histogram (which weights each observation the same), but it's a regression with polynomial terms and the option for other weights than just uniform

Choices you have to make

1. Choose the bandwidth h
2. Choose the kernel $K(\cdot)$
3. Choose the polynomial ordering p

We have a broad set of writings and suggestions around each of these things, and the issues around choices is always subjective researcher bias, uncertainty and various forms of bias

Local polynomial estimation steps

1. Choose a polynomial order p and a kernel function $K(\cdot)$
2. Choose a bandwidth h
3. For observations above the cutoff, fit a WLS regression of Y on a constant and re-centered running variable terms with p polynomial terms and weight $K(\frac{X_i - c}{h})$ for each observation which is an estimate of the point$\mu_+ = E[Y_i^1 | X_i = c]$:

$$\widehat{\mu}_+ : \widehat{Y}_i = \widehat{\mu}_+ + \widehat{\mu}_{+,1}(X_i - c) + \widehat{\mu}_{+,1}(X_i - c)^2 + \cdots + \widehat{\mu}_{+,p}(X_i - c)^p$$

4. Repeat step 3 below the cutoff $\mu_- = E[Y_i^1 | X_i = c]$:

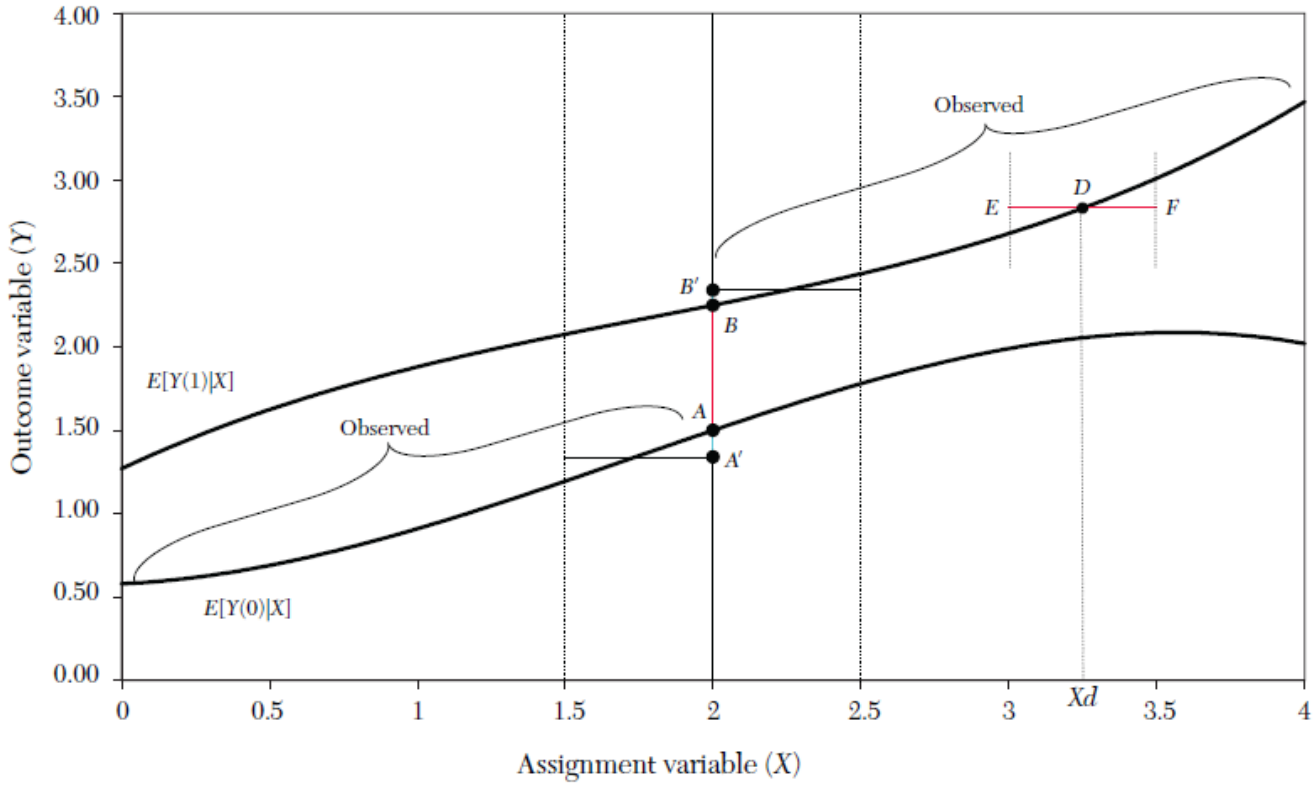
$$\widehat{\mu}_- : \widehat{Y}_i = \widehat{\mu}_- + .\widehat{\mu}_{-,1}(X_i - c) + \widehat{\mu}_{-,1}(X_i - c)^2 + \cdots + \widehat{\mu}_{-,p}(X_i - c)^p$$

5. Calculate the sharp RD point estimate as $\widehat{\delta} = \widehat{\mu}_+ - \widehat{\mu}_-$

Animation of a local linear regression

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

Boundary problems



"True" effect at cutoff is AB, but in this histogram of width h, we estimate $A'B'$ because of trends in the running variable; what can we do?

Local regressions with kernels

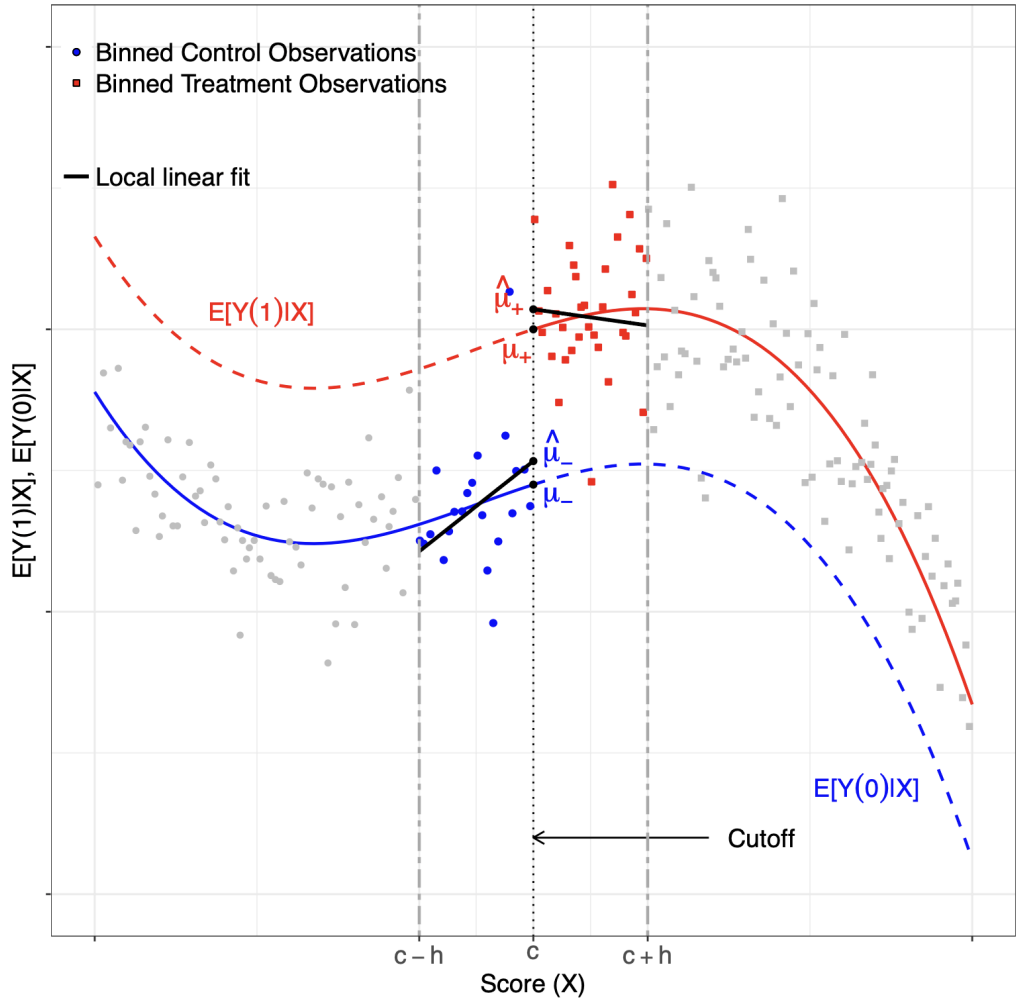


Figure: From Cattaneo, et al. (2019)

Estimation with kernels

The kernel function $K(\cdot)$ assigns non-negative weights to each re-centered observation based on the distance between each observation's running variable score X_i and the cutoff c and there are different kinds of kernel weights you can choose

Types of kernels

- **Rectangular** uniform weights equivalent to $E[Y]$ at a given bin on X
- **Triangular** draws a straight line from the threshold to the edge of the bandwidth and weights along the line
- **Epanechnikov** is similar but is more like a parabola

Estimation with kernels

- Cattaneo, et al. (2019) recommend using the triangular kernel because when you use it with a bandwidth that optimizes mean squared error, you can get a point estimate that is optimal.
- Triangular kernels assign zero weight to all observations outside bandwidth h interval and positive weights within it
- Weights are maximized at the cutoff and decline symmetrically and linearly as the value of the running variable gets further away

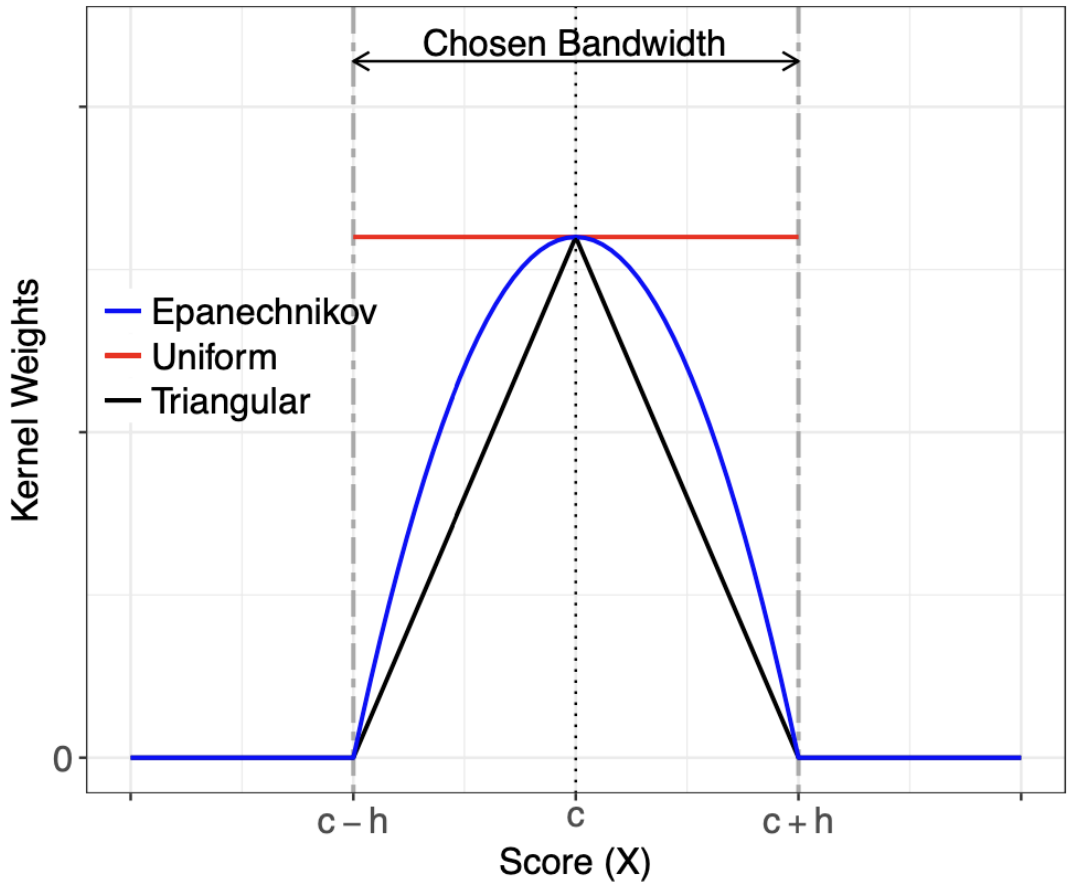


Figure: From Cattaneo, et al. (2019)

Polynomial order

- Simple difference in means (i.e., p order of zero) is like a histogram with uniform weights
- Suffers from what is called the “boundary problem” – the estimation of the true expected potential outcomes at the cutoff is biased with trends in the running variable
- But even after choosing kernel weights, we aren’t done as then there is the business of choosing polynomial order

Polynomial terms

- Two conceptual issues to keep in mind
  1. No polynomials has boundary problems, but
  2. Higher order polynomials, though, suffer from severe overfitting problems
- Local linear RD is the preferred method, but this is where we end up in the world of choosing the bandwidths, h, because that controls the width (and thus selects the units) of the neighborhood around the cutoff that will be used to fit the model

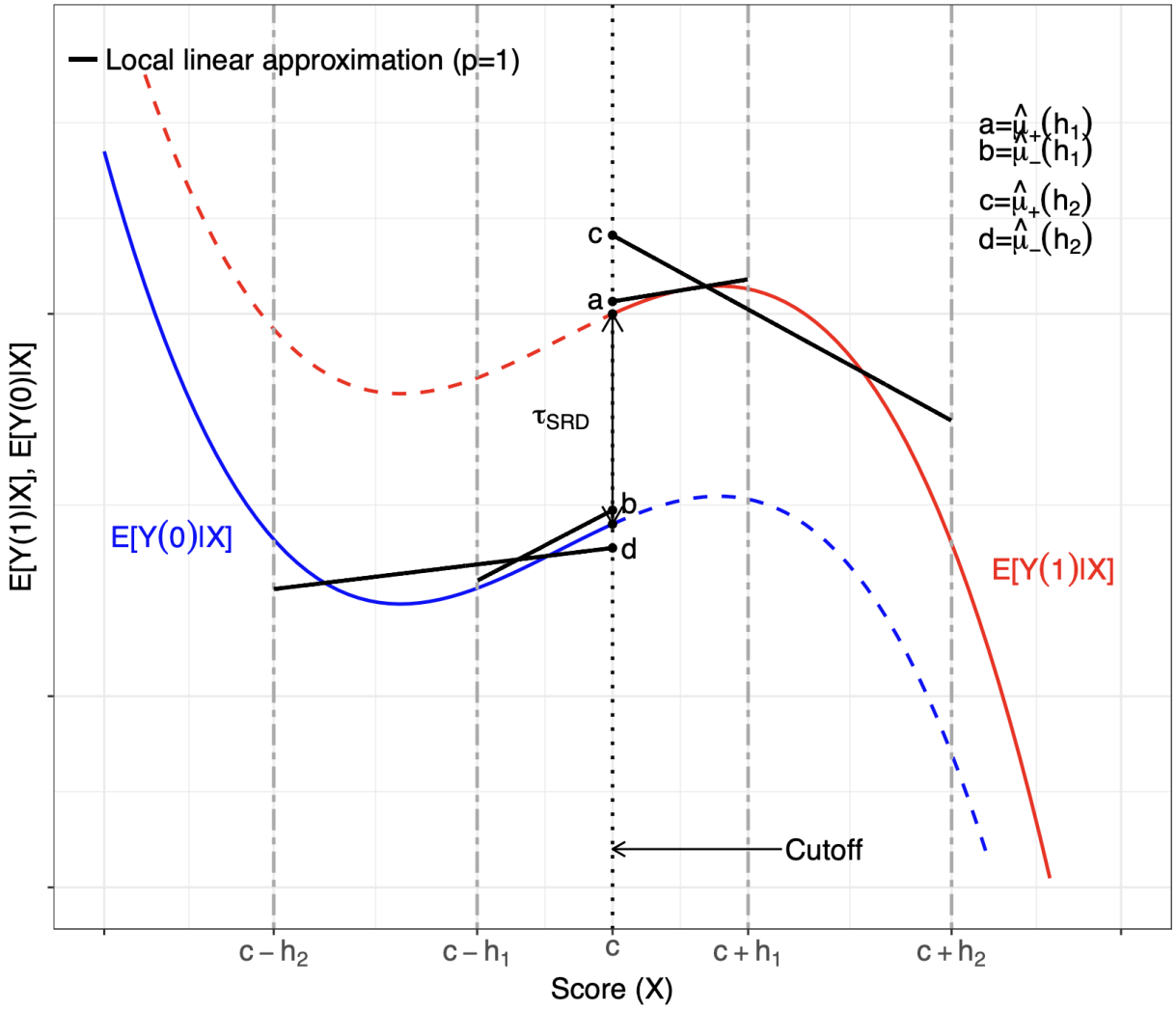


Figure: From Cattaneo, et al. (2019)

Bias and variance

- **Bias term**
  - → When we approximate the unknown functions with p, h and $K(\cdot)$, there's some approximation error because we do now actually know the true function
  - → Think about the earlier picture – when we used the larger bandwidth and p of zero, we came up short. Why? Because of the curvature of the functions we were approximating
- **Variance term**
  - → Variance depends on sample size and bandwidth h
  - → As number of observations near the cutoff falls, the contribution of the variance term to MSE grows and vice versa
  - → Variability of the the point estimator depends therefore on density at the cutoff (which gets back to why RD tends to be data intensive in the first place)

Optimal Bandwidths

- Most approaches have some balancing act between bias and variance that they're trying to address
- Minimizing the MSE of the local estimator, $\widehat{\delta}$, given a choice of p and $K(\cdot)$ has become the most popular since MSE is the sum of squared bias and variance

$$MSE(\widehat{\delta}) = Bias^2(\widehat{\delta}) + Variance^(\widehat{\delta})$$

- If you choose to minimize MSE, you are choosing h – hence "optimal bandwidths"

$$min_{h>0}\left(h^{2(p+1)}B^2 + \frac{1}{nh}V\right)$$

Optimal Bandwidths

- Solution to that minimization problem is h_{MSE} and is the MSE-optimal bandwidth choice

$$h_{MSE} = \left(\frac{V}{2(p+1)} B^2 \right)^{\frac{1}{(2p+3)}} n^{-1/(2p+3)}$$

  which directly addresses the bias-variance trade-off
- Optimal bandwidths that minimize MSE are proportional to that last term and therefore MSE-optimal bandwidths increase with V (more observations) and decrease with B (less observations)
- Hence why optimal bandwidths are “data driven” and automated which takes away some of the subjective decisions researchers must make

Implementation with software

- You have choices for implementing this – manually (see Cattaneo, et al. (2019) section 4.2.4, or with packages like `rdrobust`
- Very flexible – choose kernels (e.g., triangular), choose polynomials, choose number of bandwidths h
- But remember choosing h is not advisable bc of what we just said, so there is a separate package called `rdbwselect` which selects the MSE-optimal bandwidth for the local estimator (but you still choose p and $K(\cdot)$)

Implementation with software

- Tons of options with `rdbwselect` – different kernels, even different bandwidths left and right of the cutoff
- Once you use it, you can pass it on to `rdrobust` in a second stage, or
- Just use `bwselect` within the syntax of `rdrobust` itself (we will review this with our Hansen exercise later)
- All of this can be incorporated into plotting too with `rdplot`

Inference

- Asymptotic 95% confidence intervals for an RD point estimator will be too small if they ignore the bias and variance terms

$$CI = \left[(\widehat{\delta} - B) \pm 1.96 \cdot \sqrt{V} \right]$$

- Bias term arises because the local polynomial approach is a non-parametric approximation of the unknown potential outcome regression functions
- So different strategies are used to address this

Corrections

- **Conventional** confidence intervals assume that the polynomial gave an exact approximation of the potential outcomes, but as this can't be verified, it isn't considered credible

$$CI_{us} = \left[\widehat{\delta} \pm 1.96 \cdot \sqrt{V}\right]$$

- If the approximation error is non-trivial, inference based on the conventional approach will be incorrect, over-rejection of the null of zero treatment effects
- Cattaneo, et al. (2019) strongly discourage you from using this

Standard bias correction confidence intervals

- **Bias correction** confidence intervals adjust for bias by estimating the bias term B with an estimator $\widehat{B}$

$$CI_{bc} = \left[(\widehat{\delta} - \widehat{B}) \pm 1.96 \cdot \sqrt{V} \right]$$

- Bias term, as we said earlier in the MSE-optimal slides, depends on trends in the running variable for that unknown potential outcome function and whether that's captured by the polynomials you chose
- But it can have poor performance in applications because the bias estimation step isn't incorporated in the variance term which can introduce coverage distortions in practice

Robust bias correction confidence intervals

- **Robust bias correction** confidence intervals are based on the bias correction procedure. which estimated the bias term with $\widehat{B}$ and includes a new asymptotic variance that incorporates the bias correction step

$$CI_{rbc} = \left[(\widehat{\delta} - \widehat{B}) \pm 1.96 \cdot \sqrt{V_{bc}}\right]$$

- Because this new variance term incorporates the extra variability from that first bias estimation step, it's larger for the same bandwidth

Different confidence intervals

Table: Local polynomial confidence intervals (from Cattaneo, et al. (2019)

	Centered at	**Standard error**
Conventional: CI_{us}	$\widehat{\delta}$	$\sqrt{\widehat{V}}$
Bias-corrected: CI_{bc}	$\widehat{\delta} - \widehat{B}$	$\sqrt{\widehat{V}}$
Robust bias-corrected: CI_{rbc}	$\widehat{\delta} - \widehat{B}$	$\sqrt{\widehat{V_{bc}}}$

Inference – clustering and honesty

- Historically, people would cluster standard errors along the running variable (going back to early work by Lee), but recent work warns against this
- Kolesár and Rothe (2018) provide extensive theoretical and simulation-based evidence that clustering on the running variable has high over-rejection problems
- Propose two alternative confidence intervals that achieve correct coverage in large samples – called "honest" – which is available in R (`RDHonest`) at https://github.com/kolesarm/RDHonest

Evaluating violations of smoothness

- Smoothness isn't *directly* verifiable because it involves counterfactuals for each unit along X_i which don't exist bc of switching equation
- Doesn't stop us! People tend to use various ingenious deductions involving "placebos"
- People tend to want at least indirect evidence for smoothness since we don't have "the science of physical randomization" to ensure smoothness holds in our data

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
- In other words, we are looking for evidence of people choosing their value of X_i so as to get just barely get into the treatment
- **Example** an unusual bunching of reviews at the stars "just below" the cutoff on Yelp

Manipulation might violation smoothness

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.

If the individuals who manipulated their score had different potential outcomes, it could create a gap even though it’s just sorting

Manipulation is testable

Finally – a testable prediction

- Noom, weight loss app, wants to keep people using the app
- When number of skipped logging of meals and weighing exceeds some threshold, a coach privately texts them
- We want to know the effect of coaches on retention

Manipulation along the running variable

- Manipulation may occur if:
  1. Users can choose logging of meals and/or weighing (yes)
  2. Know about the cutoff score (probably not)
  3. Have enough time to change their behavior (yes)
- If all three, then we will tend to see “bunching” of people not logging around the cutoff
- Justin McCrary (2008 article) thought of this and suggested a “density test” to see if you could reject a null of smooth density at the cutoff

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 (needs multiple units in other words per 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

Simulations of density tests

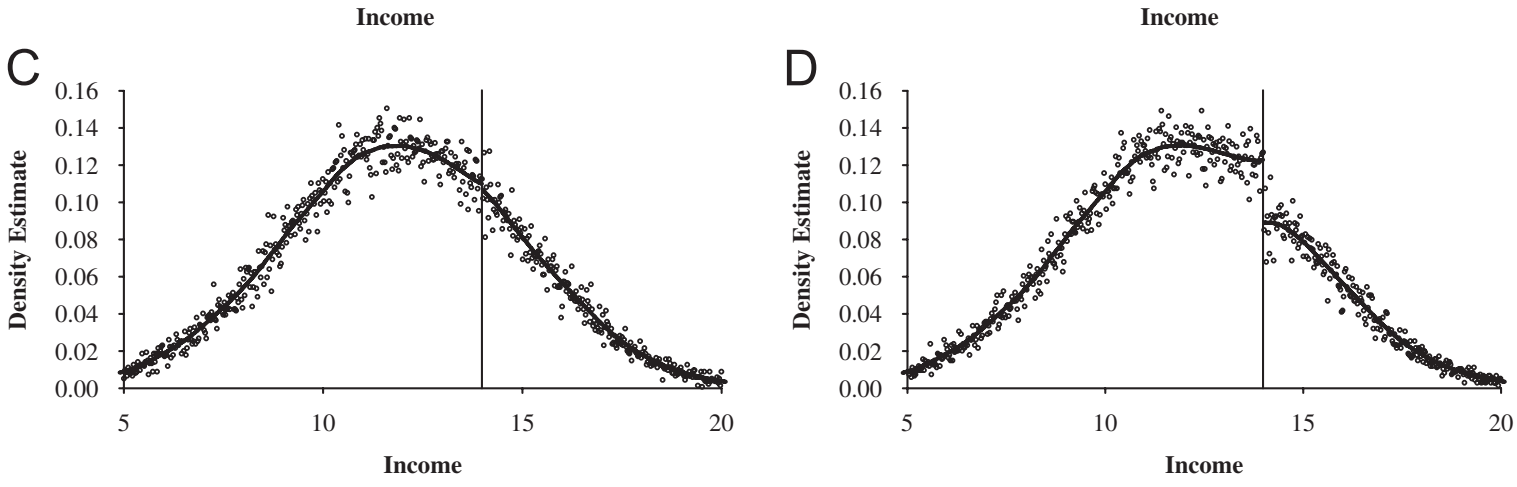
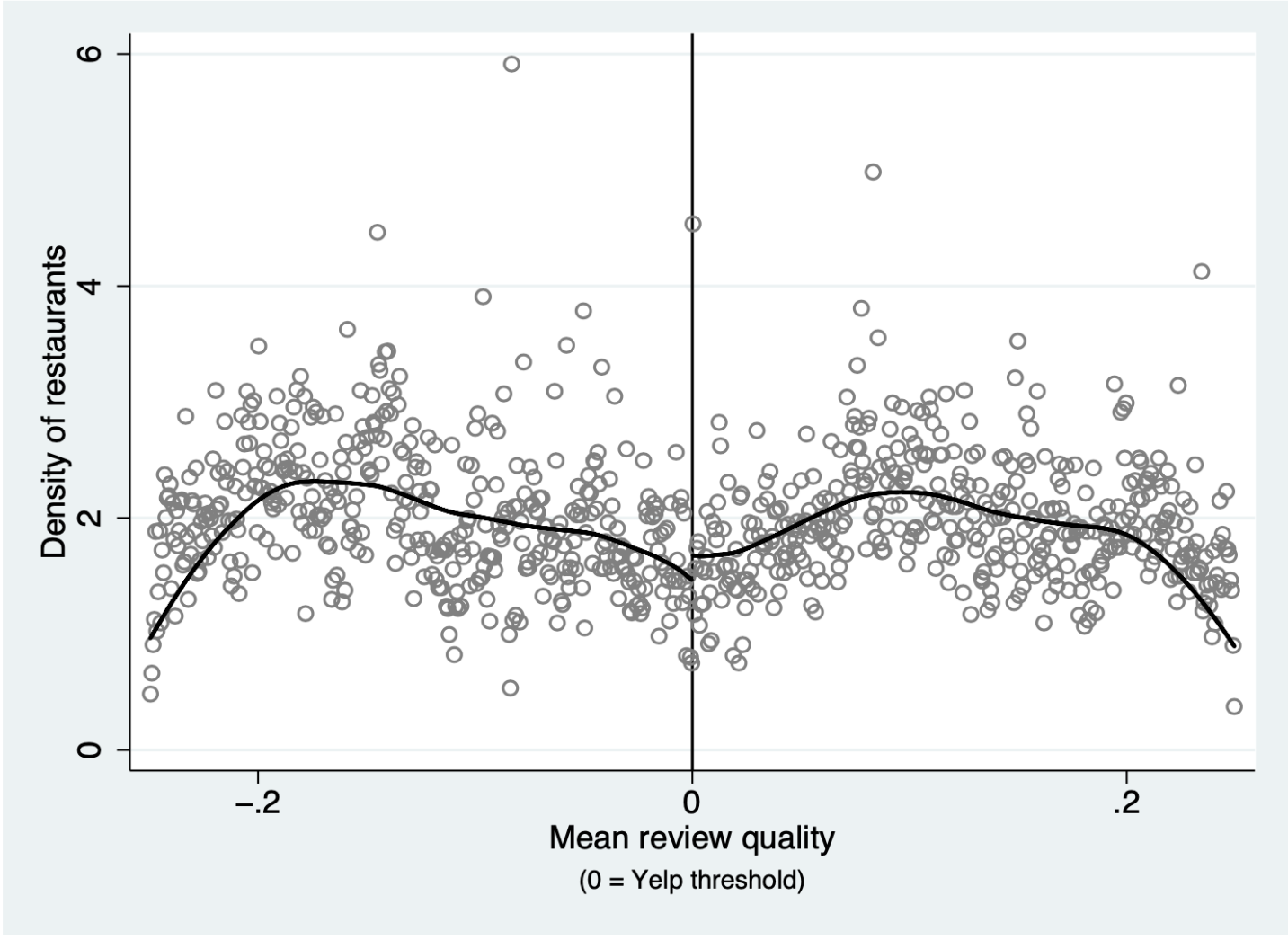


Figure: From McCrary (2008). Left shows failing to reject. Right shows rejection of the null.

Yelp density tests

Fig. 4: Empirical Density of Restaurants



Does NICU save premature babies' lives?

- What is the causal effect of heightened medical care for premature babies on infant mortality?
- Babies in NICU have lower mortality even if they weren't in the NICU
- Answer: use 1500 grams as the cutoff and large administrative data from hospitals

Heaping problem

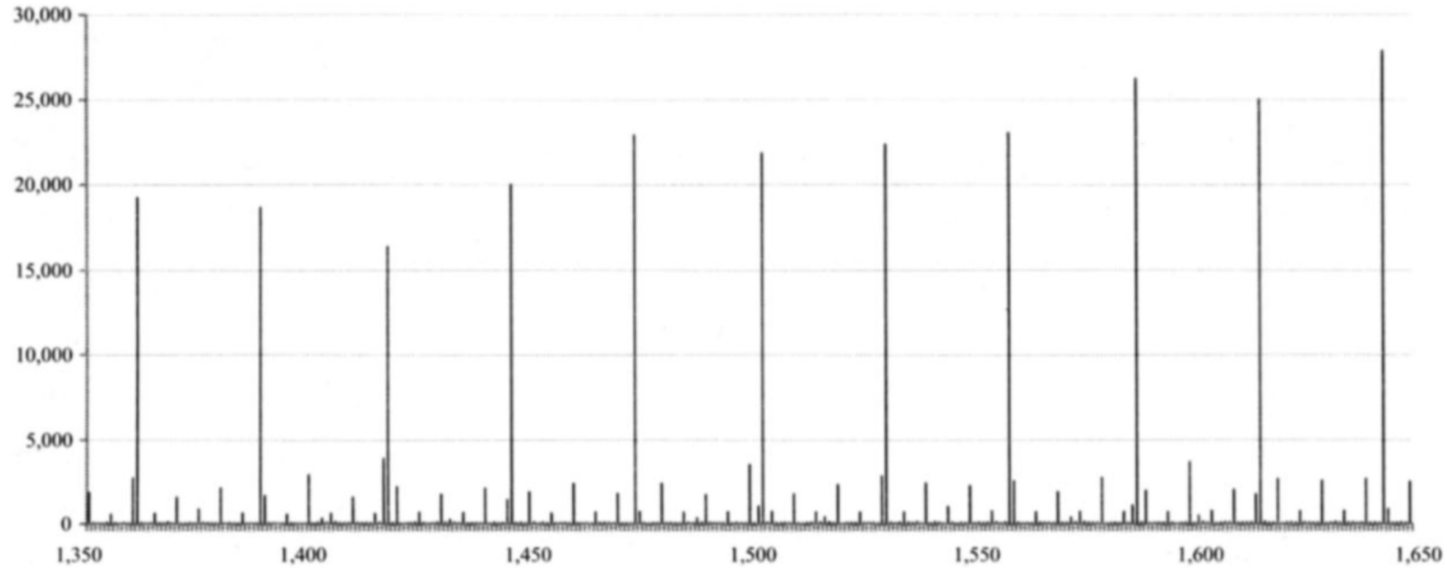


FIGURE I
Frequency of Births by Gram: Population of U.S. Births between 1,350 and 1,650 g
NCHS birth cohort linked birth/infant death files, 1983–1991 and 1995–2003, as described in the text.

Figure: Huge number of babies born with birth weights (grams) at non-random spaced intervals

Why heaping at birth?

- **H1**: Babies naturally born every 100 grams. Unlikely. Babies can't choose their birthweight, nor can parents. Seems more likely births are uniformly distributed around some arbitrary cutoff
- **H2**: Someone is rounding *the running variable itself*
  - → 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

Heaping can lead to spurious failure to reject

- Density tests are not designed to detect heaping because of low power around the cutoff
- 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 encouraged to use their eyes, as well as density tests

Rounding stars vs rounding reviews

- Remember: in the Yelp example, the actual score is the mean stars across all reviews
- The running variable, which is smooth and continuous probably (you'll need to check) is what Yelp uses to assign whole stars and half stars
- Manipulation in Yelp's example isn't the rounding to a star – it's the rounding of the underlying score

Non-random heaping

Strange patterns invited scrutiny

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.”

Why did density tests fail? Not enough power because you need to evaluate *between* the heaps.

Donut hole RDD

- Estimates should not logically be sensitive to the observations at the cutoff – if it is, then smoothness may be violated
- Drop units in the vicinity of the cutoff and re-estimate the model (called “donut hole”)
- Reanalyzing the birthweight mortality data, effects were 50% smaller than previously reported

Confounders at cutoff

- Examples of confounders at 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

- **Question**: is there a reason that the potential outcomes, which are based on observable covariates, should jump at the cutoff?
- We can't check for potential outcome jumps, but we can check for covariates jumping
- Only works with *observable* covariates though

Data visualization

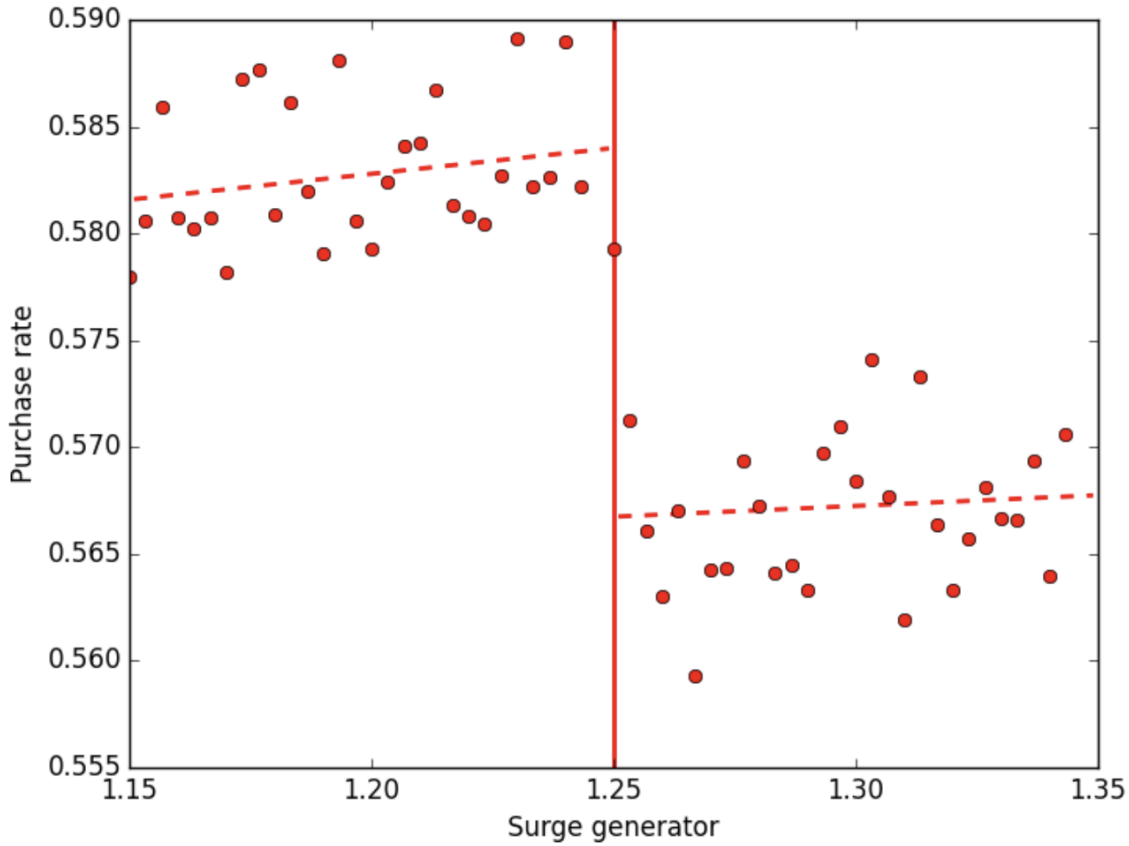
- Eyeball tests are very common in RDD studies
- Even if your main results are all parametric, you'll still want to present at least some nonparametric style pictures
- Let's review some of typical graphs

Visualizing Outcomes

1. **Outcome by running variable, (X_i):**
   - → Construct bins and average the outcome within bins on both sides of the cutoff (lots of options)
   - → 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”

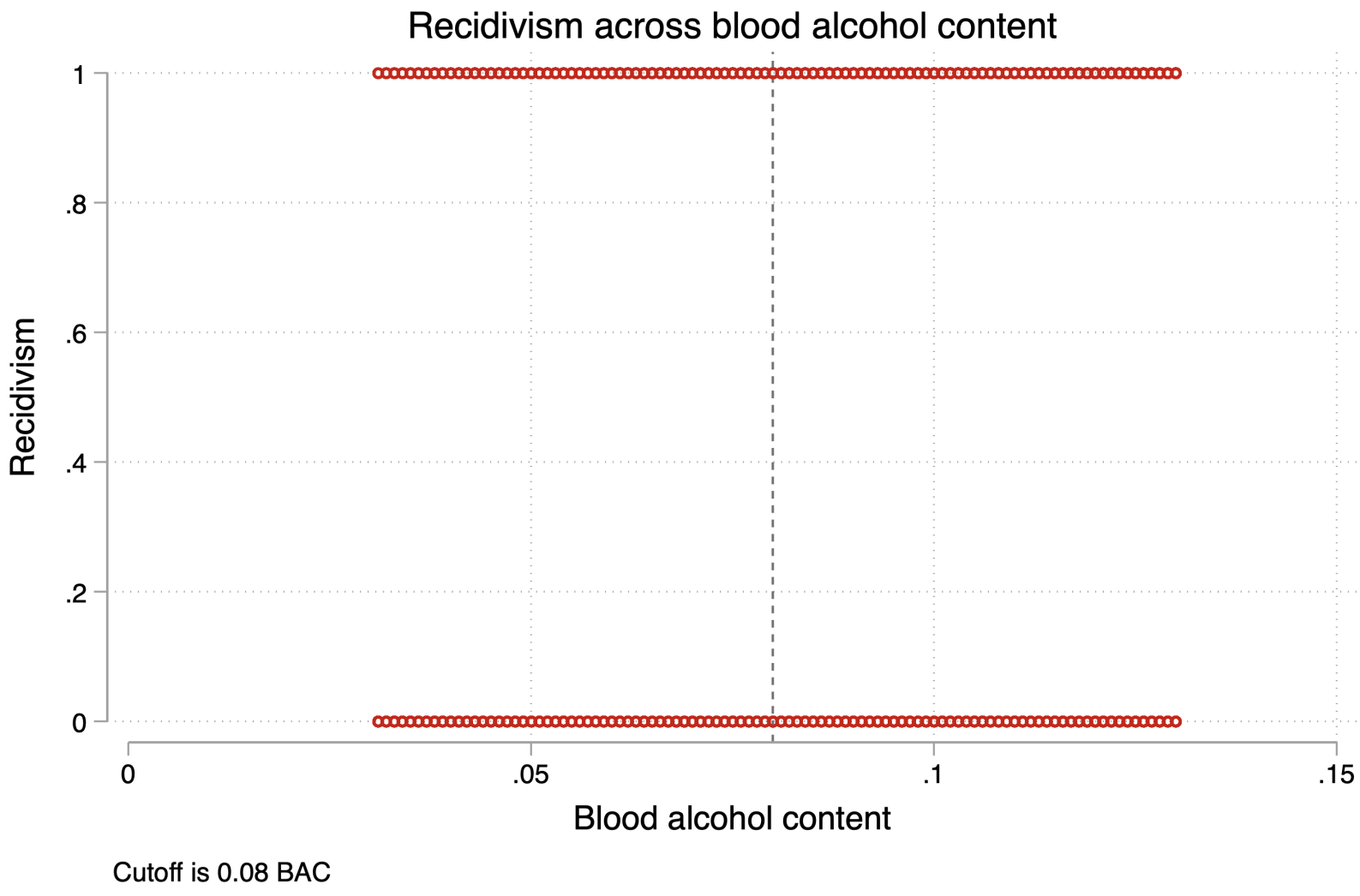
Visualizing Outcomes

Figure 4: Example of purchase rate changes at price discontinuity



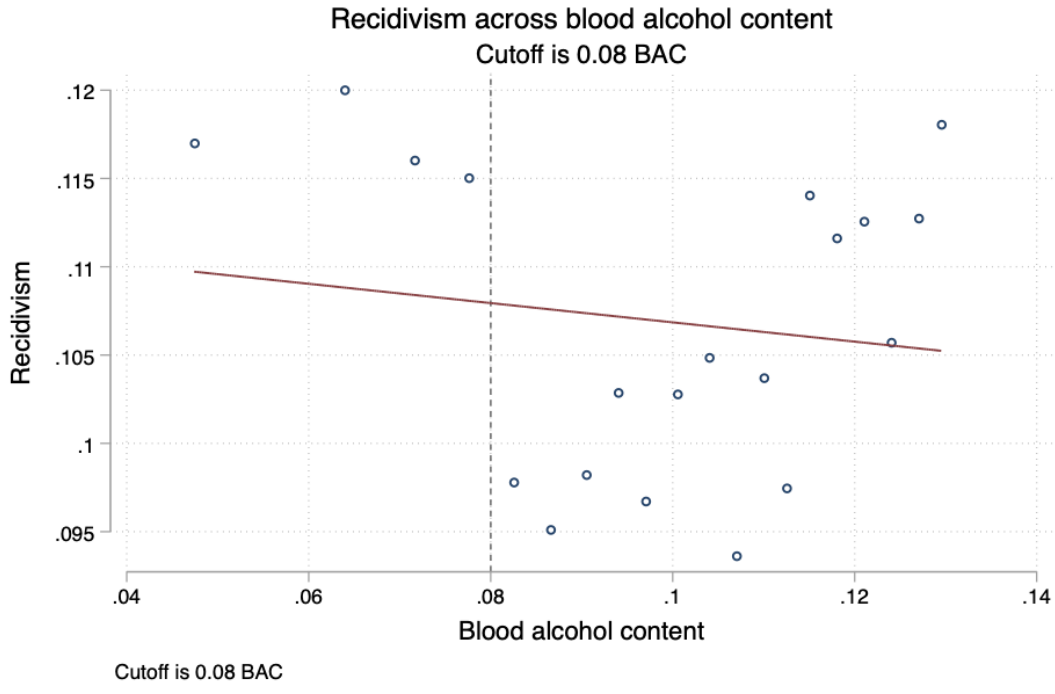
Note: This figure illustrates how purchase rates vary as a function of the surge generator over the range 1.15x to 1.35x. The vertical line when the surge generator equals 1.25 identifies the point at which the surge price changes from 1.2x to 1.3x.

Scatter without binning



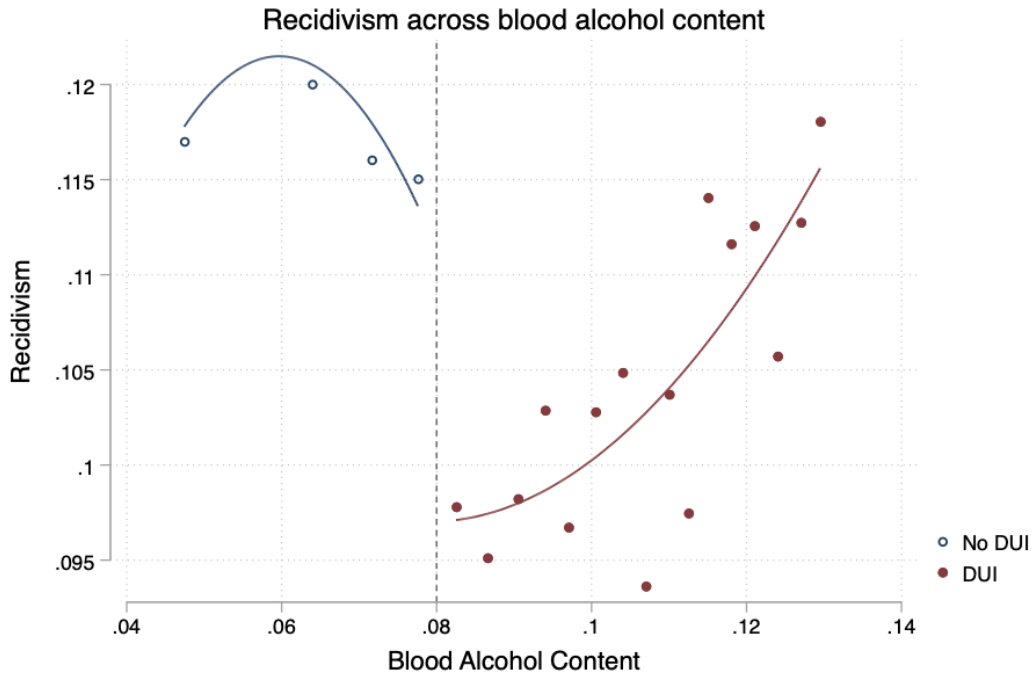

```
scatter recidivism bac1
```

Binscatter with linear fit



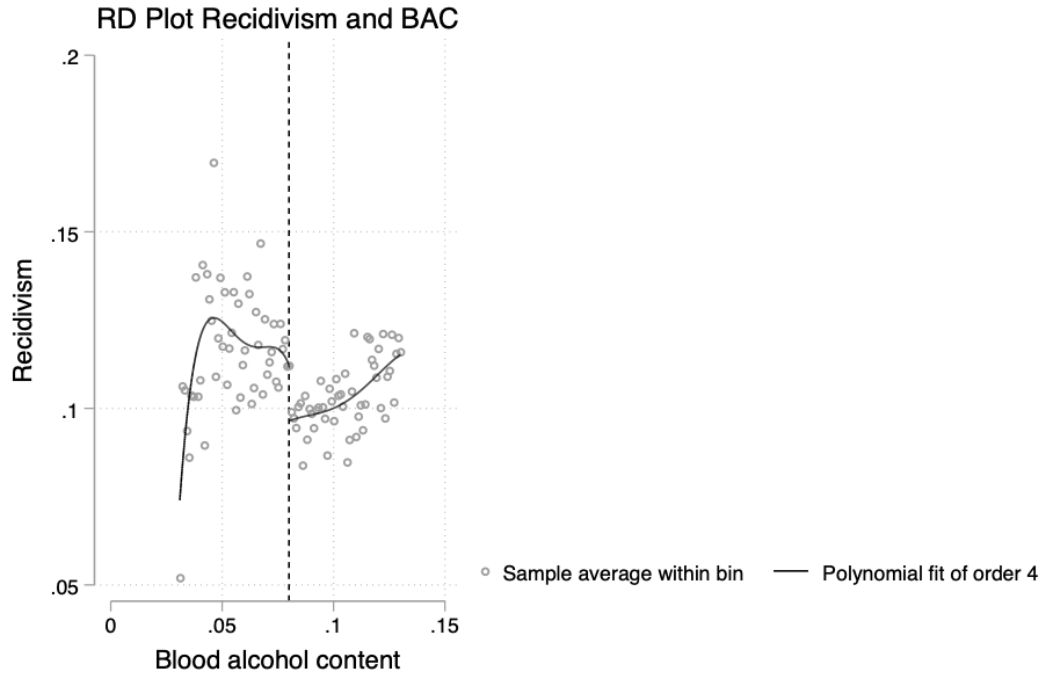

```
binscatter recidivism bac1 if bac1>0.03 & bac1<0.13
```

Binscatter with quadratic fit on each side




```
binscatter recidivism bac1 if bac1>0.03 & bac1<0.13,
line(qfit) by(dui)
```

Binscatter with quadratic fit on each side

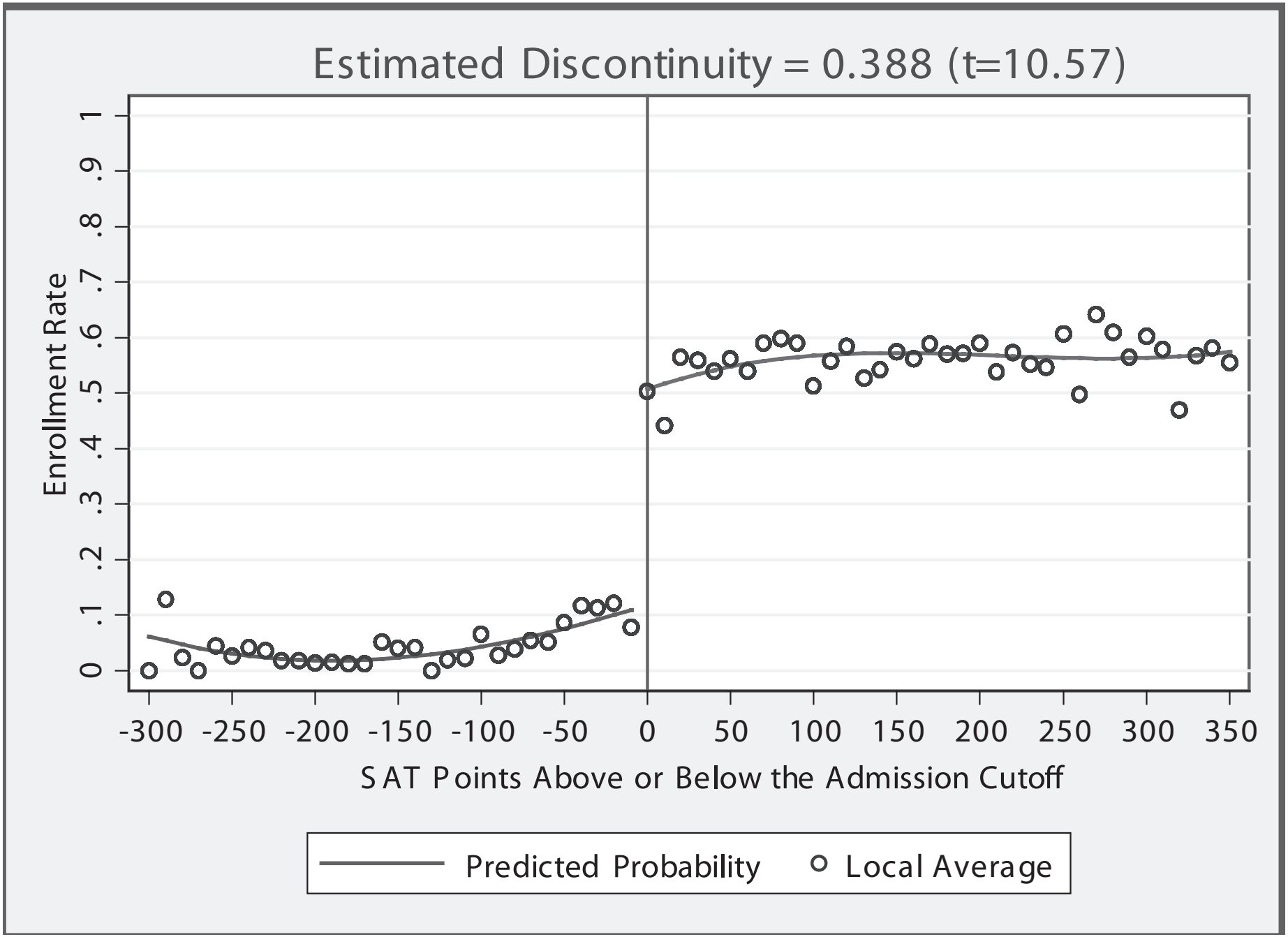


`rdplot`

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”)

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



McCrary Density

3. **Density of the running variable**
   - → First start off by plotting the number of observations in each bin.
   - → Investigate for discontinuity or heaping near the cutoff
   - → More formalized density tests are also useful so that you can conduct hypothesis tests

Density of the running variable

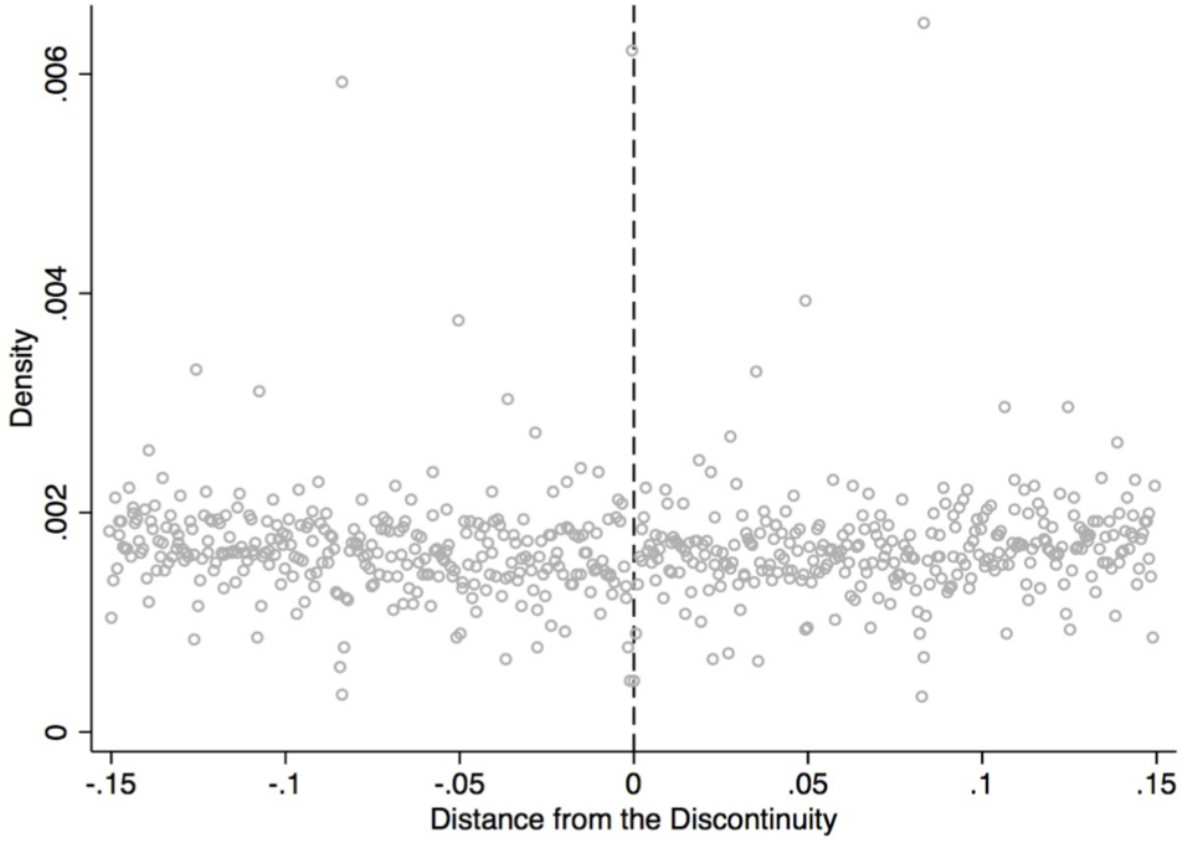


Figure 8: McCrary test: distribution of average ratings near rounding thresholds

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

Visualizing balance

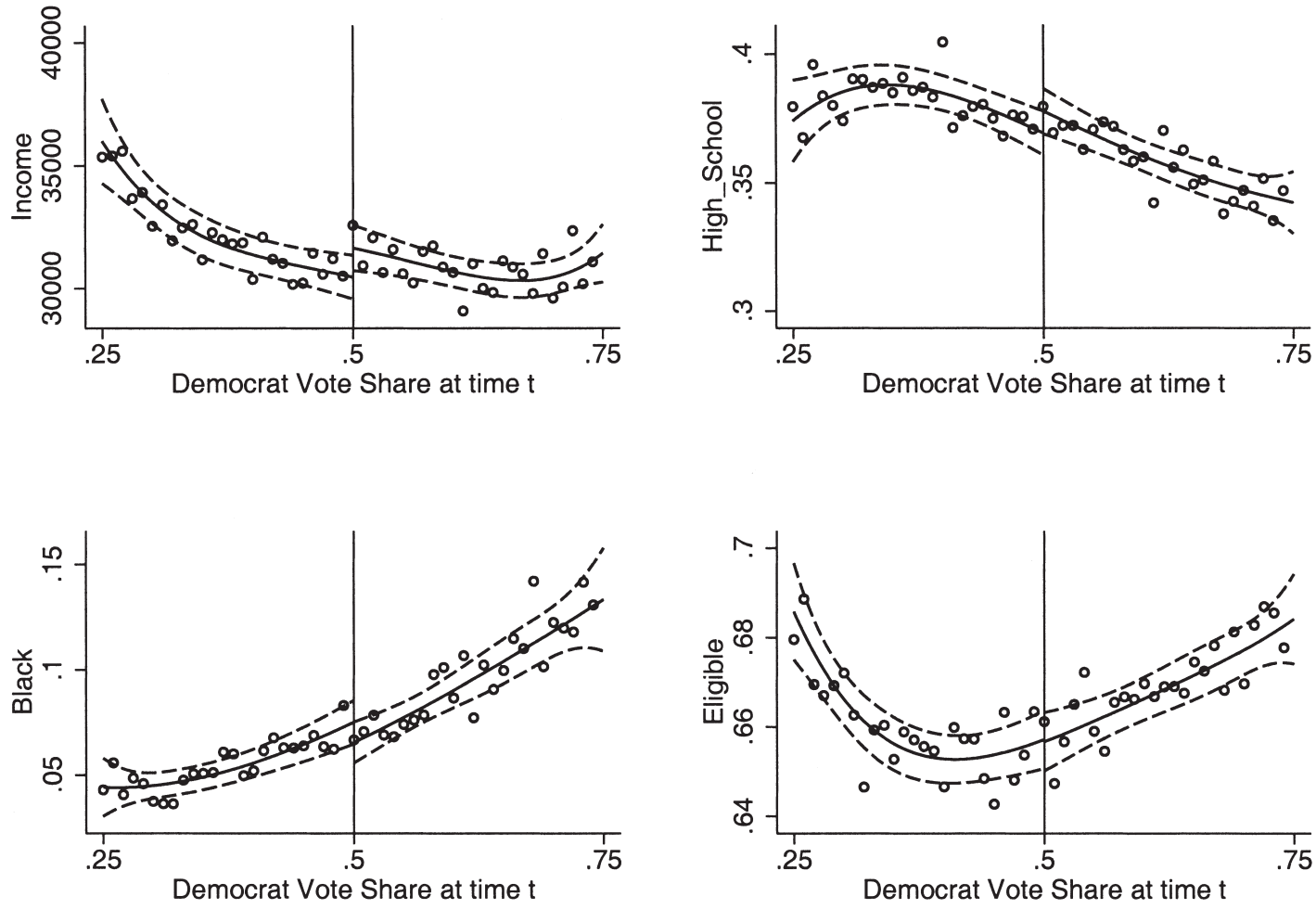


Figure: Panels refer to district characteristics: real income, percentage with high-school degree, black, eligible to vote. Circles represent the average characteristic within intervals of 0.01 in Democratic vote share.The dotted line represents the 95 percent confidence interval.

Discussion: Pros

- RDD is viewed as very credible among observational designs; for some reason people feel the smoothness assumption is easier to defend
- It may be because you only have to defend the exogeneity of the treatment at c_0 since you're essentially arguing the potential outcomes wouldn't have jumped there in counterfactual
- Rewards people who have access to large datasets bc as N grows, the mass at the cutoff should as well, giving you shorter windows for estimation and therefore lower bias and lower variance

Discussion: Caveats

- Under extreme heterogeneous treatment effects, keep in mind what you are and are not identifying
- You're identifying the average causal effect for those units flipped over into treatment bc their $X_i > c_0$
- If their treatment effects are profoundly different than anywhere else (even opposing sign), then you're ability to infer above treatment effects elsewhere is limited

Going forward

- Steve Tadelis at Berkeley Haas when he worked at eBay found a natural experiment which enabled him to run country wide RCTs on paid search advertising
- Oftentimes these quirks can bring solid discussion to hard questions
- What questions in industries you've observed do you think could be answered that maybe haven't been?

Roadmap

- Introducing Regression Discontinuity Design
  - Basic background
  - Identification Basics
  - Sharp Design
  - Smoothness and Identification
- Estimation
  - Local Regressions
  - Nonparametric estimation
  - Testing for violations
  - Data Visualization
- Examples
  - DWI and recidivism
  - Close election design
  - Fuzzy RDD

American Economic Review 2015, 105(4): 1581–1617
http://dx.doi.org/10.1257/aer.20130189

Punishment and Deterrence: Evidence from Drunk Driving†

By Benjamin Hansen*

I test the effect of harsher punishments and sanctions on driving under the influence (DUI). In this setting, punishments are determined by strict rules on blood alcohol content (BAC) and previous offenses. Regression discontinuity derived estimates suggest that having a BAC above the DUI threshold reduces recidivism by up to 2 percentage points (17 percent). Likewise having a BAC over the aggravated DUI threshold reduces recidivism by an additional percentage point (9 percent). The results suggest that the additional sanctions experienced by drunk drivers at BAC thresholds are effective in reducing repeat drunk driving. (*JEL* I12, K42, R41)

Since the National Highway Traffic and Safety Administration began recording fatal traffic accident data in 1975, drunk driving was a factor in 585,136 traffic fatalities.[1] To put that magnitude in perspective, 725,347 murders occurred in the United States over a similar window. However, given that drunk driving is a very different crime than murder and other crimes, and it is often closely linked with addiction and substance abuse, what little we do know about preventing other crimes may not apply for drunk drivers. Understanding whether punishments and sanctions are effective in reducing drunk driving is crucial to determine the appropriate combination of enforcement and punishment that can maximize social welfare. To that end this paper offers quasi-experimental evidence on the effectiveness of blood alcohol content (BAC) thresholds, a primary policy tool used in

Criminal deterrence

- Becker 1968 asserted a downward sloping demand for crime through two channels:
  1. Higher probability of arrest
  2. Punishment conditional on arrest and conviction
- Debates about whether these punishments work and how
  1. Deterrence – need variation in p or punishment
  2. Incapacitation – Locking them up isn't deterrence and is expensive

Evidence for deterrence

- Difficult to disentangle deterrence from incapacitation
- Drago, et al. (2009) exploited a natural experiment in Italy from a collective pardon where those who were pardoned had to finish their old sentence plus anything new – creating quasi-random variation in sentences for the same offnse (small deterrence effects)
- But few and far between – let's consider Hansen (2015) who finds evidence for deterrence with drunk driving using RDD

Background

- US has a minimum age of drinking at age 21
- Earlier study showed that this was associated with higher mortality due to what appeared to be drunk driving traffic fatalities
- Washington DWI stops (n=512,964), running variable is blood alcohol content, cutoff is 0.08

TABLE 1—PUNISHMENTS FOR DUI CONVICTION BASED ON BAC AND PRIOR OFFENSES

	1st offense		2nd offense		>=3rd offense	
	DUI	Agg. DUI	DUI	Agg. DUI	DUI	Agg. DUI
BAC	[0.08, 0.15]	(0.15, 1]	[0.08, 0.15]	(0.15, 1]	[0.08, 0.15]	(0.15, 1]
Min. penalty	$865.50	$1,120.50	$1,120.50	$1,545.50	$1,970.50	$2,820.50
Max. penalty	$5,000	$5,000	$5,000	$5,000	$5,000	$5,000
Min. jail time	24 hours	48 hours	30 days	45 days	90 days	150 days
Min. home release	14 days*	28 days*	60 days**	90 days**	120 days**	150 days
License susp./ revok. period	90 days^{+}	365 days^{++}	2 years^{++}	900 days^{++}	3 years^{++}	4 years^{++}
SR-22 insurance	Yes	Yes	Yes	Yes	Yes	Yes

Note: This table outlines the Washington statutes on sanctions and punishments depending on the BAC measured.

** Mandatory

* In lieu of jail time

$^{++}$Revocation

$^{+}$Suspension

Figure: Punishments for repeat DWI offense (Hansen 2015)

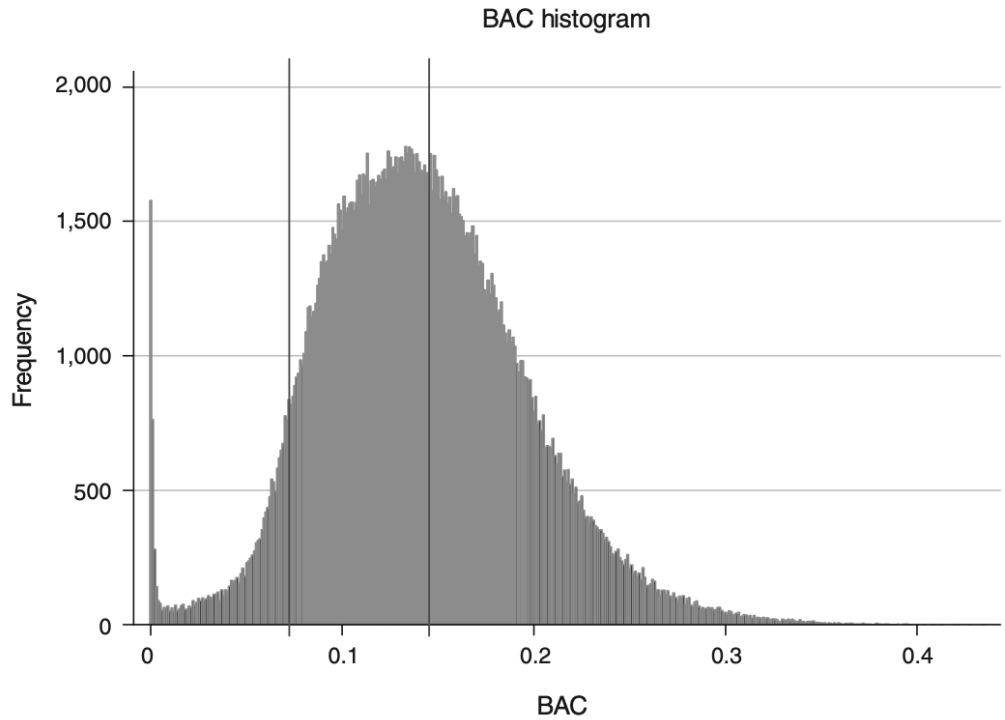


FIGURE 1. BAC DISTRIBUTION

Notes: Based on administrative records from the Washington State Impaired Driver Testing Program, 1999–2007. The histogram height on the vertical axis is based on frequency of observations, with BAC on the horizontal axis. The vertical black lines represent the two legal thresholds at 0.08 and 0.15. The bin width is 0.001, the original precision used on the breathalyzers.

Figure: Manipulation histogram along the running variable (Hansen 2015)

TABLE 2—REGRESSION DISCONTINUITY ESTIMATES FOR THE EFFECT OF EXCEEDING BAC THRESHOLDS ON PREDETERMINED CHARACTERISTICS

	Driver demographic characteristics				Police ex ante information	
Characteristics	Male (1)	White (2)	Age (3)	Accident (4)	Prior (5)	PBT (6)
Panel A. DUI threshold						
DUI	0.007	0.002	–0.165	−0.004	0.039	−0.0007
	(0.005)	(0.005)	(0.167)	(0.004)	(0.071)	(0.0005)
Mean (at 0.079)	0.792	0.852	34.9	0.089	0.139	0.090
Controls	No	No	No	No	No	No
Observations	95,111	95,111	95,111	95,111	95,111	60,485
Panel B. Agg. DUI threshold						
AGG DUI	−0.001	0.003	0.049	0.003	0.008	0.0002
	(0.001)	(0.003)	(0.123)	(0.004)	(0.006)	0.001)
Mean (at 0.149)	0.785	0.866	35.3	0.145	0.170	0.143
Controls	No	No	No	No	No	No
Observations	146,626	146,626	146,626	146,626	146,626	76,153

Notes: This table contains regression discontinuity based estimates of the effect of having BAC above the legal thresholds on predetermined characteristics. Panel A focuses on the estimated effect of BAC above the DUI threshold, while panel B focuses the Aggravated DUI threshold. All regressions have a bandwidth of 0.05 and use a rectangular kernel for weighting. Based on data from the 1999–2007 Washington State Impaired Driver Program. Standard errors are in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Figure: Placebos on exogenous demographics (Hansen 2015)

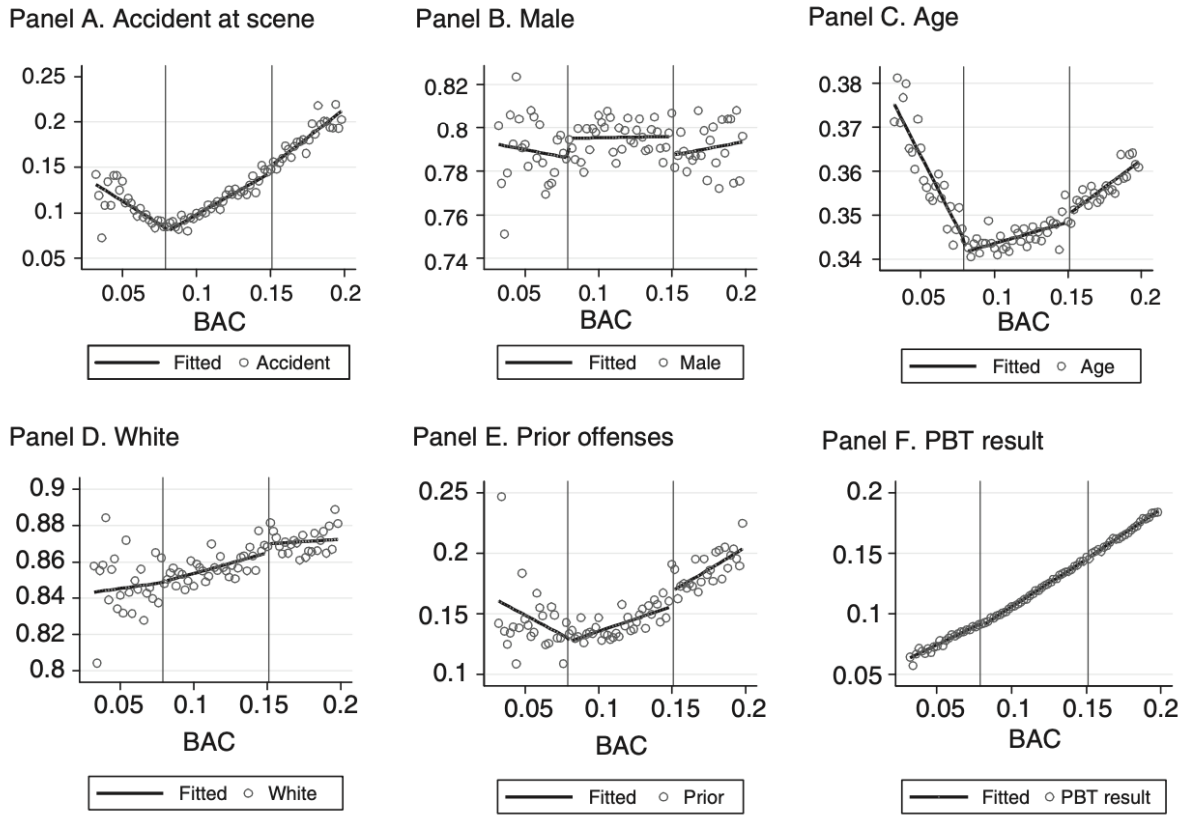


FIGURE 2. BAC AND CHARACTERISTICS

Notes: Based on administrative records from the Washington State Impaired Driver Testing Program, 1999–2007. Points represent the averages, with fitted values based on local linear models in black lines. The vertical black lines represent the two legal thresholds at 0.08 and 0.15. The bin width is 0.002.

Figure: Falsification figures on exogenous characteristics (Hansen 2015)

TABLE 3—REGRESSION DISCONTINUITY ESTIMATES FOR THE EFFECT OF EXCEEDING THE 0.08 BAC THRESHOLD ON RECIDIVISM

	All tested drivers (1)	No prior tests (2)	At least one prior test (3)
Panel A. BAC $\in$ [0.03, 0.13]			
DUI	−0.021***	−0.017***	−0.053***
	(0.004)	(0.004)	(0.015)
Mean	0.103	0.093	0.172
Controls	Yes	Yes	Yes
Observations	95,111	82,626	12,485
Panel B. BAC $\in$ [0.055, 0.105]			
DUI	−0.019***	−0.018***	−0.038**
	(0.005)	(0.005)	(0.018)
Mean	0.103	0.093	0.172
Controls	Yes	Yes	Yes
Observations	49,396	43,070	6,326

Notes: This table contains regression discontinuity based estimates of the effect of having BAC above the DUI threshold on recidivism for all drivers, those with no prior test, and drivers with at least one prior test. Panel A contains estimates with a bandwidth of 0.05 while Panel B has a bandwidth of 0.025, with all regressions utilizing a rectangular kernel for weighting. Controls include indicators for county, year, race, gender, and age of the offender. Based on administrative records from the Washington State Impaired Driver Testing Program, 1999–2007. Standard errors are in parentheses.

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Figure: Evidence for deterrence on recidivism (Hansen 2015)

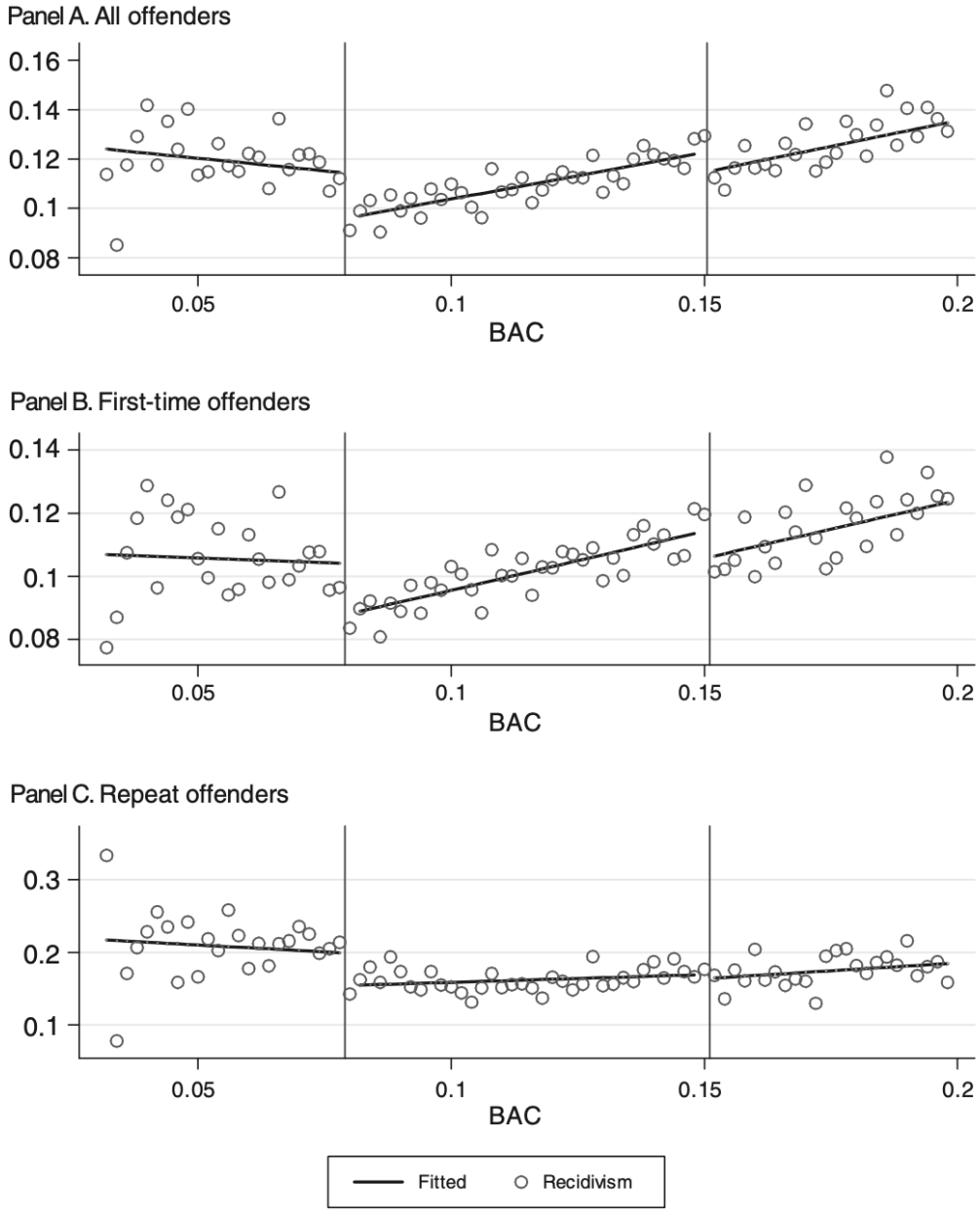


FIGURE 3. BAC AND RECIDIVISM

Notes: Based on administrative records from the Washington State Impaired Driver Testing Program, 1999–2007. Points represent the averages, with fitted values based on local linear models in black lines. The vertical black lines represent the two legal thresholds at 0.08 and 0.15. The bin width is 0.002.

Replicating pieces of Hansen's paper

- Let's work through an example doing this
- In Code is a word document walking us through a set of exercises focusing on each of these figures and tables
- We'll extend it a little with robust RDD methods

Concluding remarks

- In many ways, it's a surprising method for being so popular and so powerful
- Be wary of all the subjective researcher biases that can creep in – choosing h, choosing p, choosing kernels, choosing binned pictures
- You'll need to have strong visual effects is my experience to get buy in, which is why those issues are so critical
- Needs data, an open eye, and luck

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

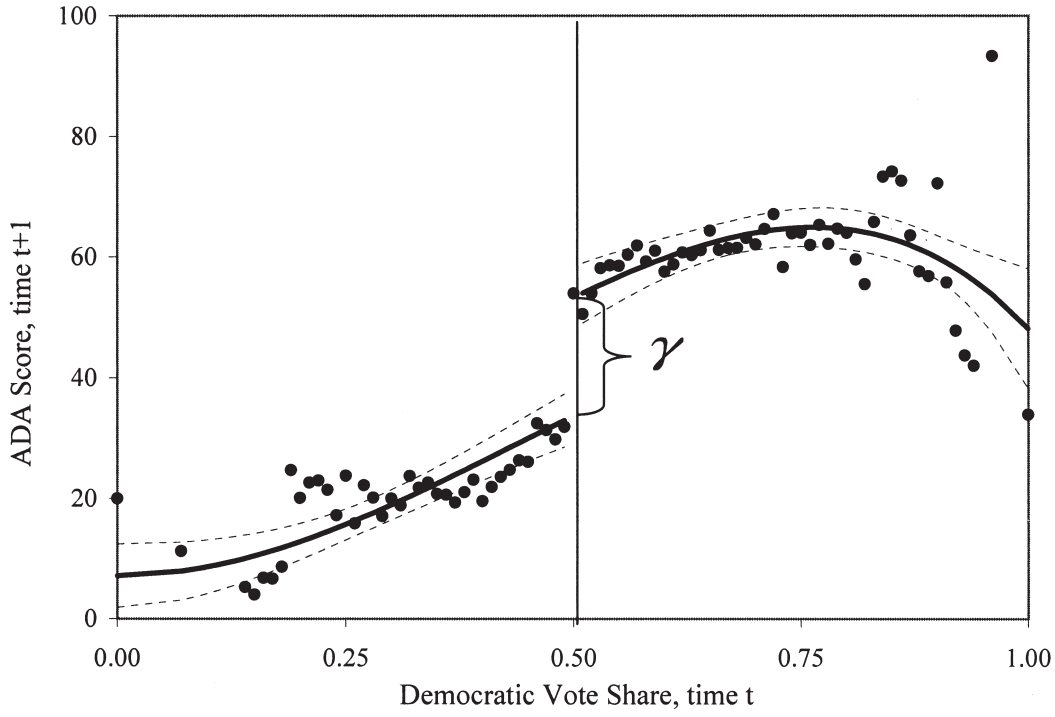


FIGURE I

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

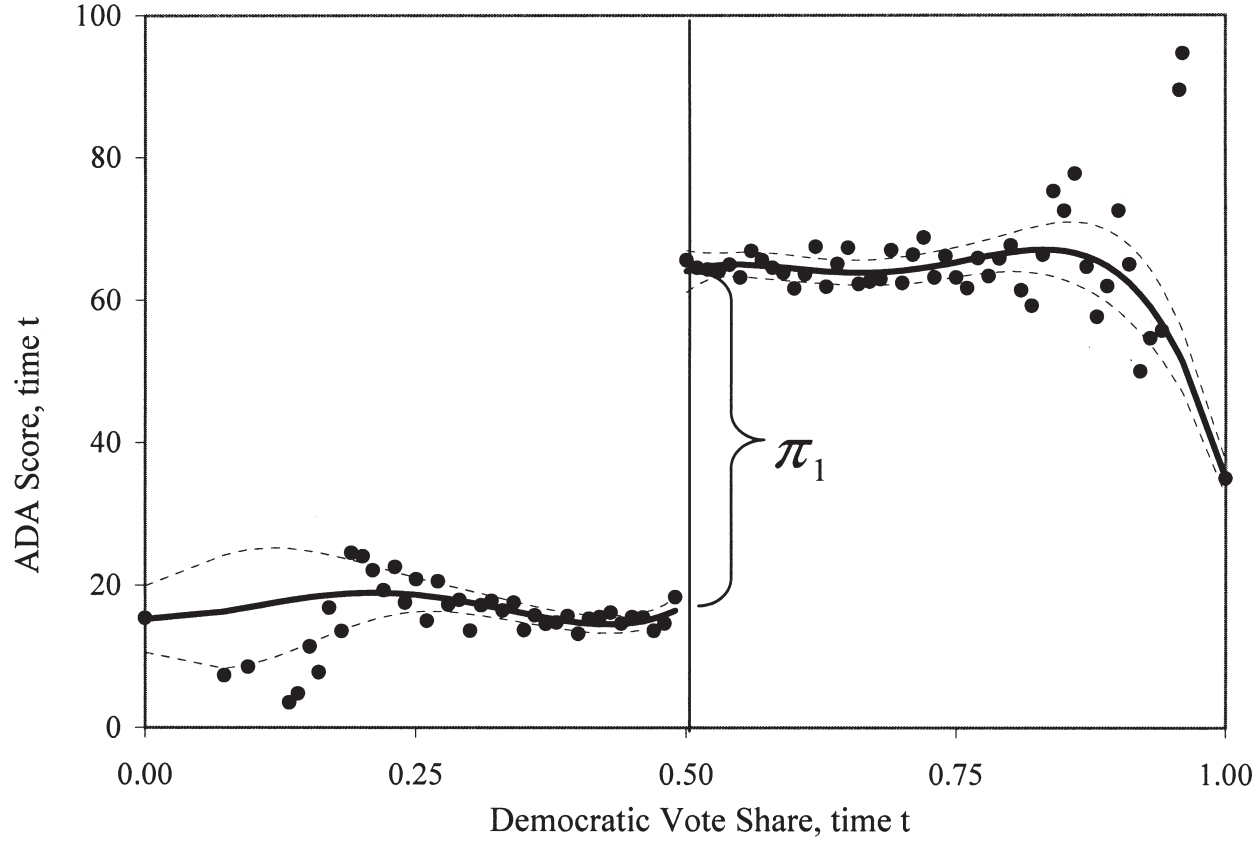


FIGURE IIa
Effect of Party Affiliation: π_1

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

Incumbency advantage

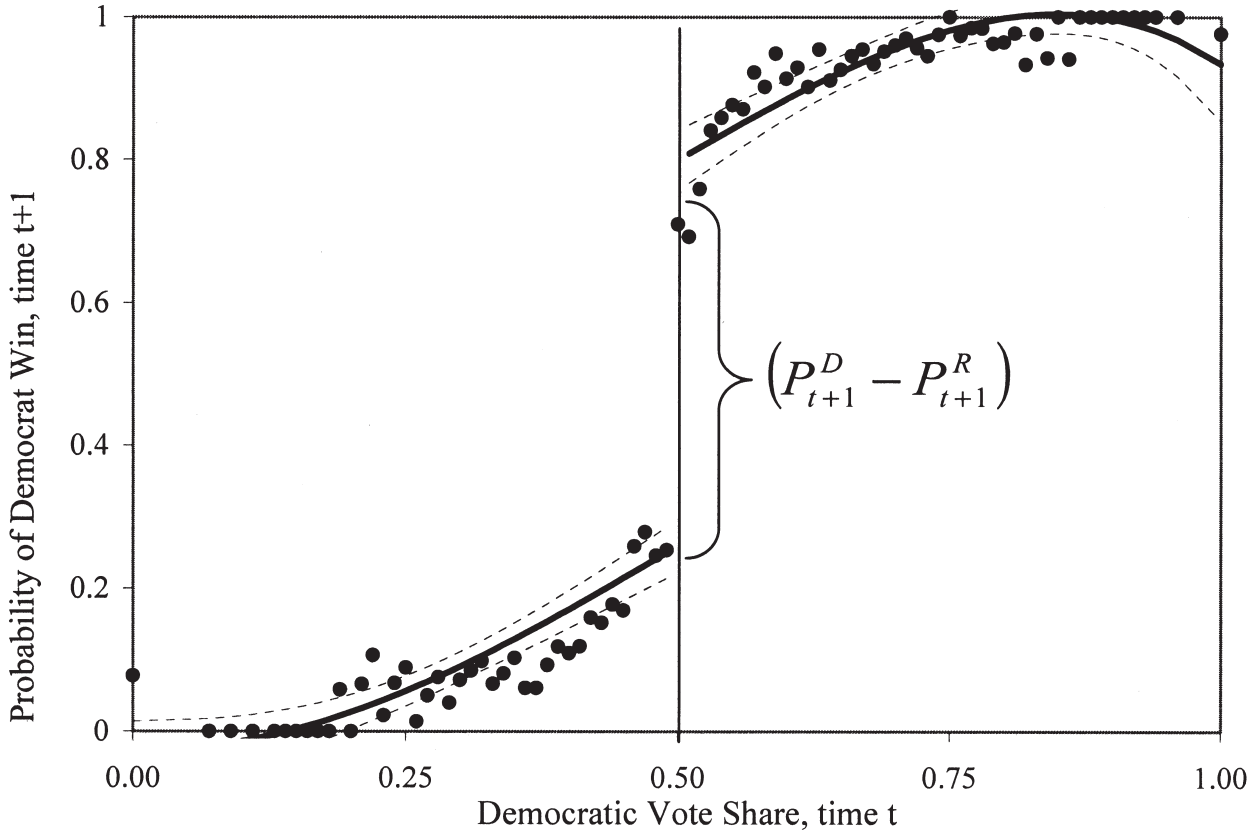


FIGURE IIb
Effect of Initial Win on Winning Next Election: $(P^D_{t+1} - P^R_{t+1})$

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

Noncompliance

- Fuzzy RDD is an approach taken when the treatment assignment was not entirely deterministic, implies overlap with selection bias
- Given the treatment assignment probabilities were non zero below the threshold (and maybe less than 1 above), you need IV
- Fuzzy RDD is in other words an instrumental variables design where the instrument is the cutoff and all earlier discussion about IV carries forward

Probability of treatment jumps at discontinuity

Probabilistic treatment assignment (i.e. “fuzzy RDD”)

The probability of receiving treatment changes discontinuously at the cutoff, c_0, but need not go from 0 to 1

$$lim_{X_i \to c_0} Pr(D_i = 1 | X_i = c_0) \neq lim_{c_0 \leftarrow X_i} Pr(D_i = 1 | X_i = c_0)$$

Examples: Incentives to participate in some program may change discontinuously at the cutoff but are not powerful enough to move everyone from non participation to participation.

Deterministic (sharp) vs. probabilistic (fuzzy)

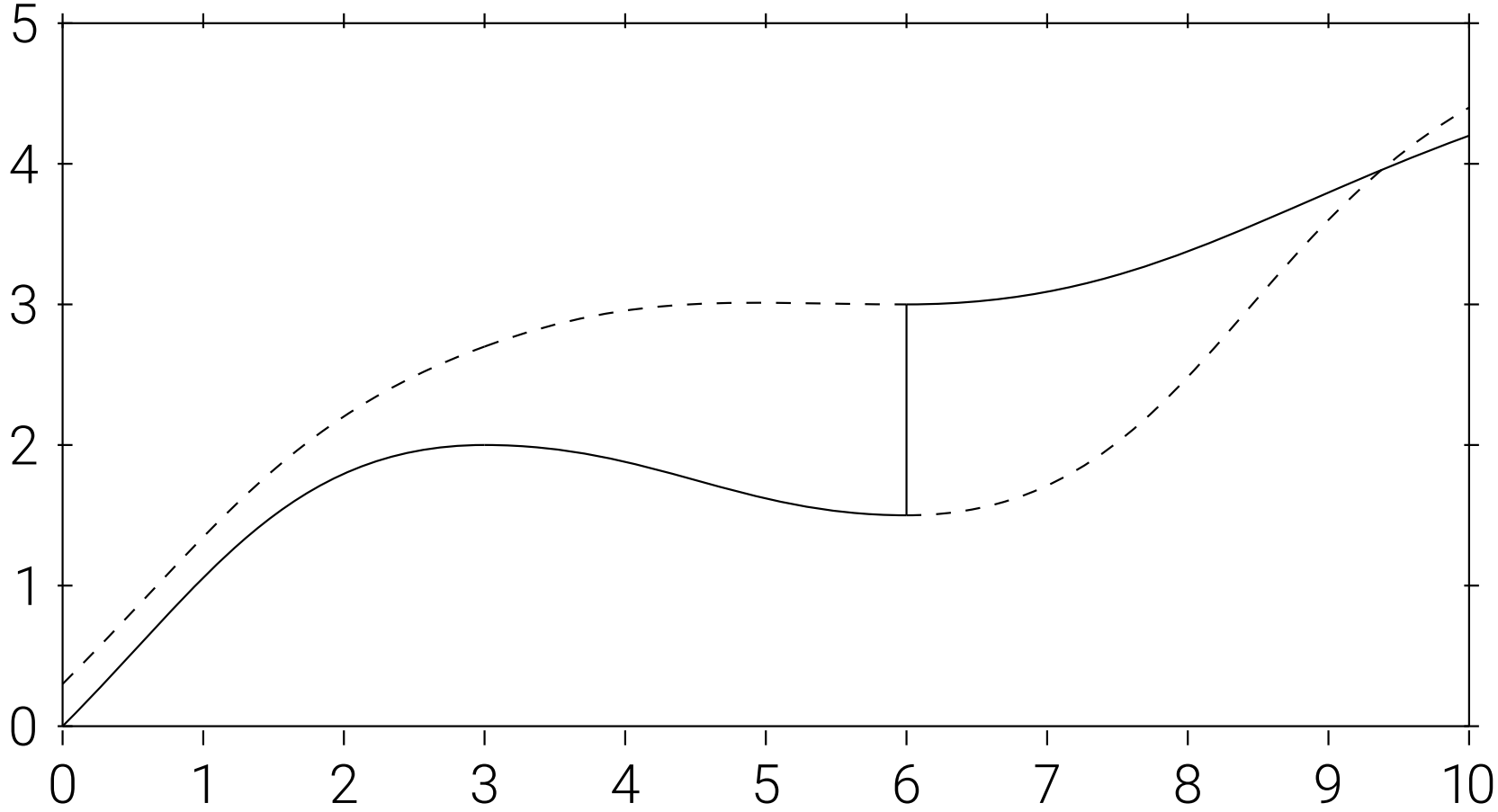
- In the sharp RDD, D_i was *determined* by $X_i \geq c_0$
- In the fuzzy RDD, the *conditional probability* of treatment *jumps* at c_0.
- The relationship between the conditional probability of treatment and X_i can be written as:

$$P[D_i = 1|X_i] = g_0(X_i) + [g_1(X_i) - g_0(X_i)]Z_i$$

where $Z_i = 1$ if $(X_i \geq c_0)$ and 0 otherwise.

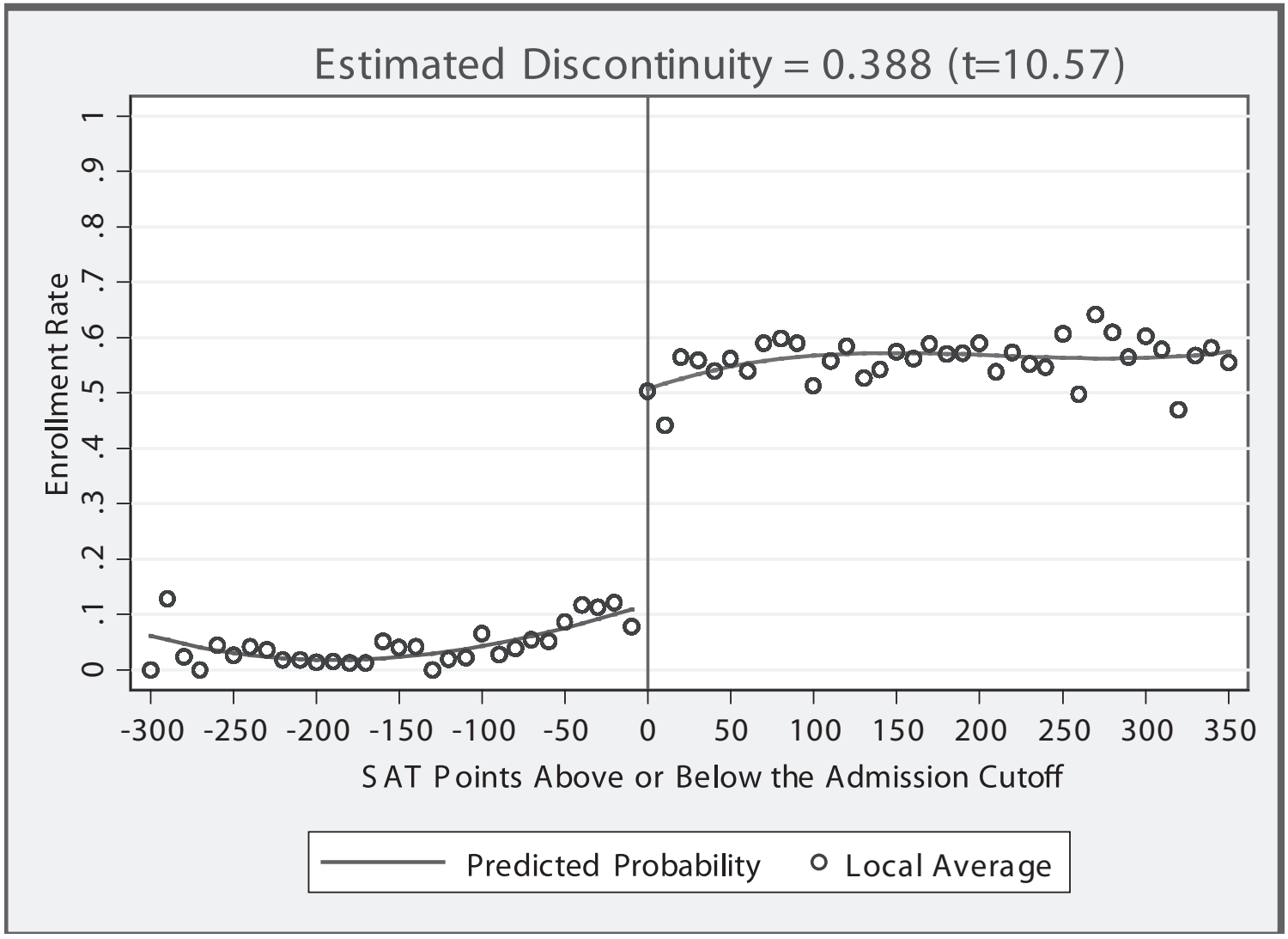
Visualization of identification strategy (i.e. smoothness)

- $E[Y^0|X]$ and $E[Y^1|X]$ for $D = 0, 1$ are the dashed/solid continuous functions
- $E[Y|X]$ is the solid which jumps at $X = 6$



Hoekstra flagship school

FIGURE 1.—FRACTION ENROLLED AT THE FLAGSHIP STATE UNIVERSITY



Instrumental variables

- As said, fuzzy designs are numerically equivalent and conceptually similar to IV
  - → "Reduced form" Numerator: "jump" in the regression of the outcome on the running variable, X.
  - → "First stage" Denominator: "jump" in the regression of the treatment indicator on the running variable X.
- Same IV assumptions, caveats about compliers vs. defiers, and statistical tests that we will discuss in next lecture with instrumental variables apply here – e.g., check for weak instruments using F test on instrument in first stage, etc.

Wald estimator

Wald estimator of treatment effect under Fuzzy RDD

Average causal effect of the treatment is the Wald IV parameter

$$\delta_{\text{Fuzzy RDD}} = \frac{lim_{X \to c_0} E[Y|X = c_0] - lim_{c_0 \leftarrow X} E[Y|X = c_0]}{lim_{X \to c_0} E[D|X = c_0] - lim_{c_0 \leftarrow X} E[D|X = c_0]}$$

RDD's Relationship to IV

- Center X it's equal to zero at c_0 and define $Z = \mathbf{1}(X \geq 0)$
- The coefficient on Z in a regression like

```
. reg Y Z X X2 X3
```

  is the reduced form discontinuity, and

```
. reg D Z X X2 X3
```

  is the first stage discontinuity
- Ratio of discontinuities is estimate of $\delta_{\text{Fuzzy RDD}}$
- Simple way to implement is IV

```
. ivregress 2sls Y (D=Z) X X2 X3
```

First stage relationship between X and D

- One can use both Z_i as well as the interaction terms as instruments for D_i.
- If one uses only Z_i as IV, then the it is a "just identified" model which usually has good finite sample properties.
- In the just identified case, the first stage would be:

$$D_i = \gamma_0 + \gamma_1 X_i + \gamma_2 X_i^2 + \cdots + \gamma_p X_i^p + \pi Z_i + \varepsilon_{1i}$$

where π is the causal effect of Z on the conditional probability of treatment.

- The fuzzy RD reduced form is:

$$Y_i = \mu + \kappa_1 X_i + \kappa_2 X_i^2 + \cdots + \kappa_p X_i^p + \rho\pi Z_i + \varepsilon_{2i}$$

Fuzzy RDD with varying Treatment Effects - Second Stage

- As in the sharp RDD case one can allow the smooth function to be different on both sides of the discontinuity.
- The second stage model with interaction terms would be the same as before:

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

- Where $\tilde{x}$ are now not only normalized with respect to c_0 but are also fitted values obtained from the first stage regression.

Fuzzy RDD with Varying Treatment Effects - First Stages

- Again one can use both Z_i as well as the interaction terms as instruments for D_i
- Only using Z the estimated first stages would be:

$$\begin{aligned} D_i &= \gamma_{00} + \gamma_{01}\tilde{X}_i + \gamma_{02}\tilde{X}_i^2 + \cdots + \gamma_{0p}\tilde{X}_i^p \\ &\quad + \pi Z_i + \gamma_1^*\tilde{X}_i Z_i + \gamma_2^*\tilde{X}_i^2 Z_i + \cdots + \gamma_p^* Z_i + \varepsilon_{1i} \end{aligned}$$

- We would also construct analogous first stages for $\tilde{X}_i D_i$, $\tilde{X}_i^2 D_i$, $\ldots$, $\tilde{X}_i^p D_i$.

Limitations of the LATE

- Fuzzy RDD has assumptions of all standard IV framework (exclusion, independence, nonzero first stage, and monotonicity)
- As with other binary IVs, the fuzzy RDD is estimating LATE: the local average treatment effect for the group of *compliers*
- In RDD, the compliers are those whose treatment status changed as we moved the value of x_i from just to the left of c_0 to just to the right of c_0
- Means we can use Medicare age cutoff to estimate the effect of public insurance on mortality (LATE) and still not know the effect of public insurance on mortality (ATE)