Data Science for Economists

Regression Discontinuity

Kyle Coombs, adapted from Nick Huntington-Klein
Bates College | ECON/DCS 368

Table of contents

Prologue

Prologue

- We've just finished covering difference-in-differences, which is one way of estimating a causal effect using observational data
- DID is *very* widely applicable, but it relies on strong assumptions like parallel trends
- Today we'll cover another causal inference method that uses observational data: Regression Discontinuity
    - This method can sometimes be easier to defend
    - But it is rarer to find situations where it applies
    - There's also plenty of room for "snake oil" here as with all causal inference
- Today I'm intentionally using simulated data to illustrate concepts because this approach looks so different in different applications
- But the fundamentals are the same no matter what you're studying and I don't want that lost in the econometric sauce
- As always, there's a ton here and we're just scratching the surface

Questions

- Any questions?
- Feel free to interrupt with questions during lecture if needed

Attribution

- Most of these slides are borrowed from Nick Huntington-Klein slides on RDD and Raj Chetty's course of Big Data and Economics

Regression Discontinuity

Regression Discontinuity

- Regression discontinuity design (RDD) is currently the darling of the econometric world for estimating causal effects without running an experiment
- It doesn't apply everywhere, but when it does, it's very easy to buy the identification assumptions
- Not that it doesn't have its own issues, of course, but it's pretty good!

Regression Discontinuity

The basic idea is this:

- We look for a treatment that is assigned on the basis of being above/below a *cutoff value* of a continuous variable
- For example, if you get above a certain test score they let you into a "gifted and talented" program
- Or if you are just on one side of a time zone line, your day starts one hour earlier/later
- Or if a candidate gets 50.1% of the vote they're in, 40.9% and they're out
- Or if you're 65 years old you get Medicare, if you're 64.99 years old you don't
- Class size must be below 40 students, so there are small classes when a grade reaches 41, 81, 121, etc. students

We call these continuous variables "Running variables" because we *run along them* until we hit the cutoff

Line up in height order

1. Line up in height order

2. Those below 5'6" get a pill to increase their height

3. Those above 5'6" don't

4. We want to know the effect of the pill on height

- Can we compare the average height of the treated and untreated groups after a year?

Line up in height order

1. Line up in height order
2. Those below 5'6" get a pill to increase their height
3. Those above 5'6" don't
4. We want to know the effect of the pill on height

- Can we compare the average height of the treated and untreated groups after a year?
- Nope! Heights and the rate of growth is different for other reasons than the pill
- But what if we compared people right around 5'6"? They're basically the same, except for random chance

Running var triggers treatment

There is a relationship between an outcome (Y) and a running variable (X)

There is also a treatment that triffers if $X < c$, a cutoff.

- Let's do the wrong thing

1. Assign $Treatment = 1$ if running variable above c and $Treatment = 0$ if below
2. Regress $y = \beta_0 + \beta_1 Treatment + \varepsilon$
3. Get a biased estimate. Why?

Running var triggers treatment

There is a relationship between an outcome (Y) and a running variable (X)

There is also a treatment that triffers if $X < c$, a cutoff.

- Let's do the wrong thing

1. Assign $Treatment = 1$ if running variable above c and $Treatment = 0$ if below
2. Regress $y = \beta_0 + \beta_1 Treatment + \varepsilon$
3. Get a biased estimate. Why?

- The running variable is omitted, so we have endogeneity!

Endogenous running variables

- The key to RDD is that the running variable is *endogenous* - it's not randomly assigned
- For example, people with higher test scores are better equipped to succeed in academia as are people in gift-and-talented programs
- Similarly, you might wonder how Medicare affects health outcomes, but older people have worse health outcomes than young people
- School enrollment increases both education resources AND class sizes at cutoff points
- Shoot! Our treatment is endogenous! We have endogeneity if we omit the running variable from our model

Regression Discontinuity

- So what does this mean?
- If we can control for the running variable *everywhere except the cutoff*, then...
- We will be controlling for the running variable, removing endogeneity
- But leaving variation at the cutoff open, allowing for variation in treatment
- We focus on just the variation around the treatment, narrowing the range of the running variable we use so sharply that it's basically controlled for.
    - Then the effect of cutoff on treatment is like an experiment!

Regression Discontinuity

- The idea is that *right around the cutoff*, treatment is randomly assigned
- If you have a test score of 89.9 (not high enough for gifted-and-talented), you're basically the same as someone who has a test score of 90.0 (just barely high enough)
- So if we just focus around the cutoff, we remove endogeneity because it's basically random which side of the line you're on
- But we get variation in treatment!
- This specifically gives us the effect of treatment *for people who are right around the cutoff* a.k.a. a "local average treatment effect"
    - we still won't know the effect of being put in gifted-and-talented for someone who gets a 30

Terminology

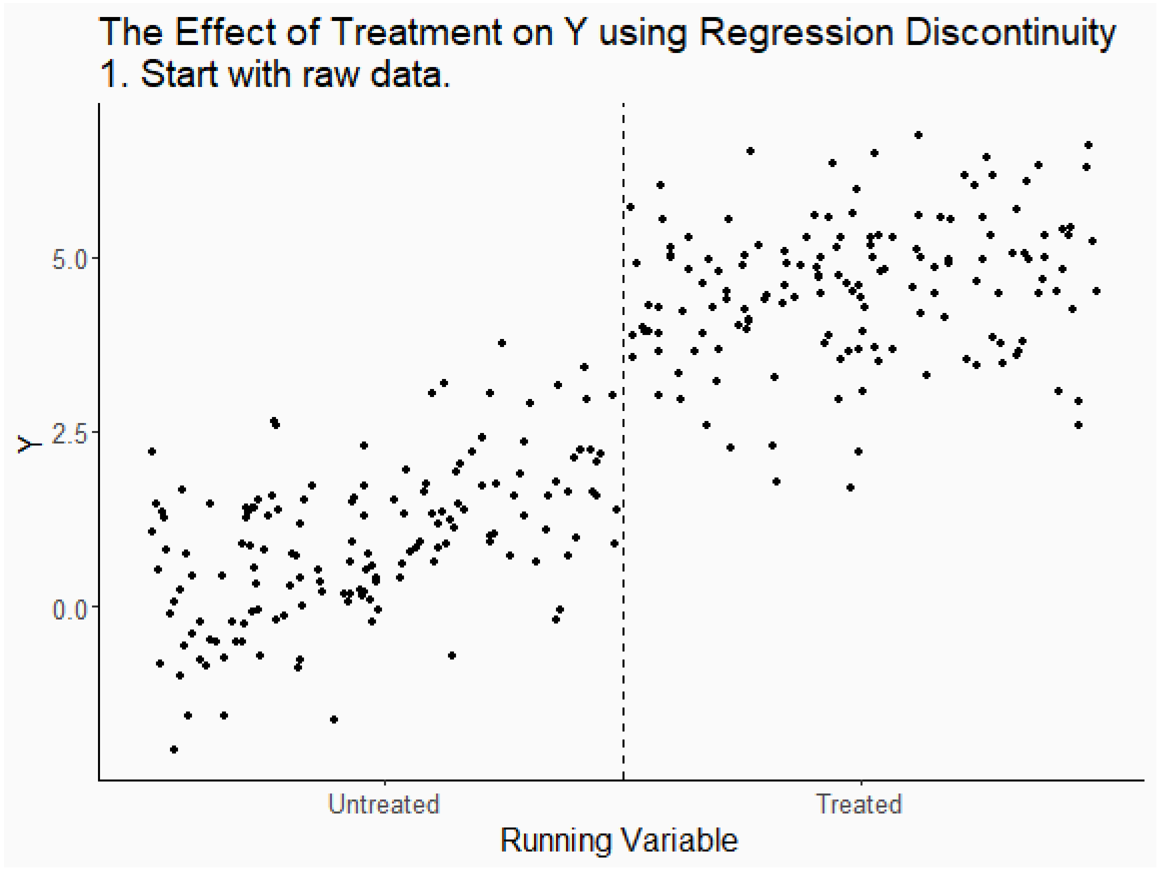
- Some quick terminology before we go on

1. **Running Variable**: The continuous variable that triggers treatment, sometimes called the **forcing variable**
2. **Cutoff**: The value of the running variable that triggers treatment
3. **Bandwidth**: The range of the running variable we use to estimate the effect of treatment -- do we look at everyone within .1 of the cutoff? .5? 1? The whole running variable?

Regression Discontinuity

- A very basic idea of this, before we even get to regression, is to create a *binned scatterplot*
- And see how the bin values jump at the cutoff
- A binned chart chops the Y-axis up into bins
- Then takes the average Y value within that bin. That's it!
- Then, we look at how those X bins relate to the Y binned values.
- If it looks like a pretty normal, continuous relationship... then JUMPS UP at the cutoff X-axis value, that tells us that the treatment itself must be doing something!

Regression Discontinuity



Concept Checks

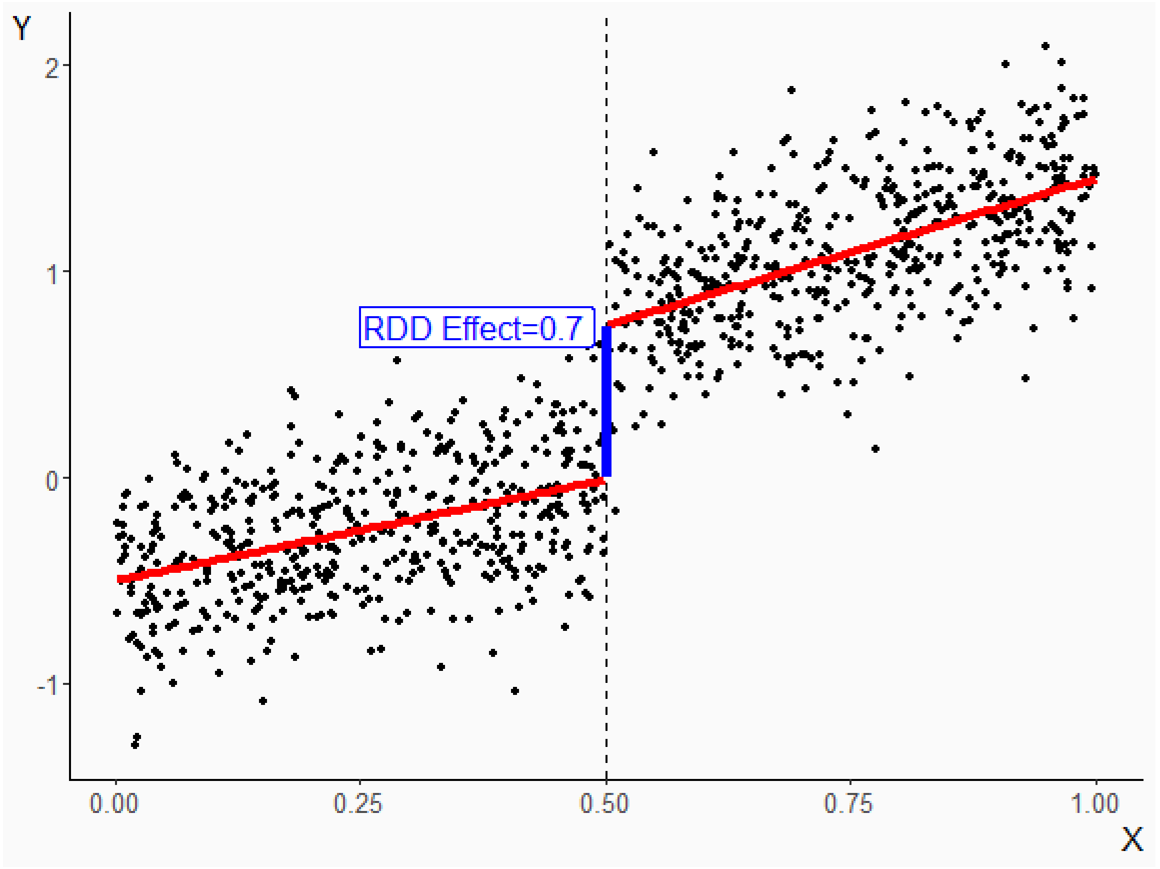
- Can you think of an example of a treatment that is assigned at least partially on a cutoff?
- Why is it important to look as narrowly as possible around the cutoff? What does this gain over comparing the entire treated and untreated groups?

Fitting Lines in RDD

- Looking purely just at the cutoff and making no use of the space *away* from the cutoff throws out a lot of useful information
- We know that the running variable is related to outcome, so we can probably improve our *prediction* of what the value on either side of the cutoff should be if we *use data away from the cutoff to help with prediction* than if we *just use data near the cutoff*, which is what that animation does
- We can do this with good ol' OLS.
- The bin plot we did can help us pick a functional form for the slope

Fitting Lines in RDD

- To be clear, producing the line(s) below is our goal. How can we do it?
- The true model I've made is an RDD effect of 0.7, with a slope of 1 to the left of the cutoff and a slope of 1.5 to the right



Regression in RDD

- First, we need to *transform our data*
- Need a "Treated" variable that's `TRUE` when treatment is applied (one side of cutoff)
- Then, we are going to want a bunch of things to change at the cutoff.
- This will be easier if the running variable is *centered around the cutoff*.[1]
- So we'll turn our running variable X into $X - cutoff$ and call that $XCentered$

```
cutoff = .5

df <- df %>%
  mutate(treated = X >= .5,
         X_centered = X - .5)
```

[1] If `(X)` is not centered, you can still back out the treatment effect, but you'll need to adjust standard errors and point estimates for the fact that the running variable is not centered. You save yourself a ton of time and headaches by just centering it.

Varying Slope

- Let the slope vary to either side, i.e. fit a different regression on each side of the cutoff
- We can do this by interacting both slope and intercept with $Treated$!
  - β_1 estimates the intercept jump at treatment (RDD effect), β_3 is the slope change.[2]

$$Y = \beta_0 + \beta_1 Treated + \beta_2 XCentered + \beta_3 Treated \times XCentered + \varepsilon$$

```
etable(feols(Y ~ treated*X_centered, data = df,fitstat='N',vcov='HC1'))
```

```
##                         feols(Y ~ tr..
## Dependent Var.:                      Y
##
## Constant                 -0.01 (0.03)
## Treated                0.75*** (0.04)
## X_centered             0.98*** (0.09)
## Treated x X_centered   0.45*** (0.13)
## ____________________ ______________
## S.E. type              Heterosk.-rob.
## Observations                    1,000
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[2] Sometimes the change in slope is the effect of interest -- this is called a "regression kink" design, which measures how the relationship between `(X)` and `(Y)` changes at the cutoff.

Polynomial Terms

- We don't need to stop at linear slopes!
- Just like we brought in our knowledge of binary and interaction terms to understand the linear slope change, we can bring in polynomials too. Add a square maybe!
- Don't get too wild with cubes, quartics, etc. - polynomials tend to be at their "weirdest" near the edges, and we don't want super-weird predictions right at the cutoff. It could give us a mistaken result!
- A square term should be enough

Polynomial Terms

- How do we do this? Interactions again. Take *any* regression equation...

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$$

- And just center the X (let's call it XC, add on a set of the same terms multiplied by $Treated$ (don't forget $Treated$ by itself - that's $Treated$ times the interaction!))

$$Y = \beta_0 + \beta_1 XC + \beta_2 XC^2 + \beta_3 Treated + \beta_4 Treated \times XC + \beta_5 Treated \times XC^2 + \varepsilon$$

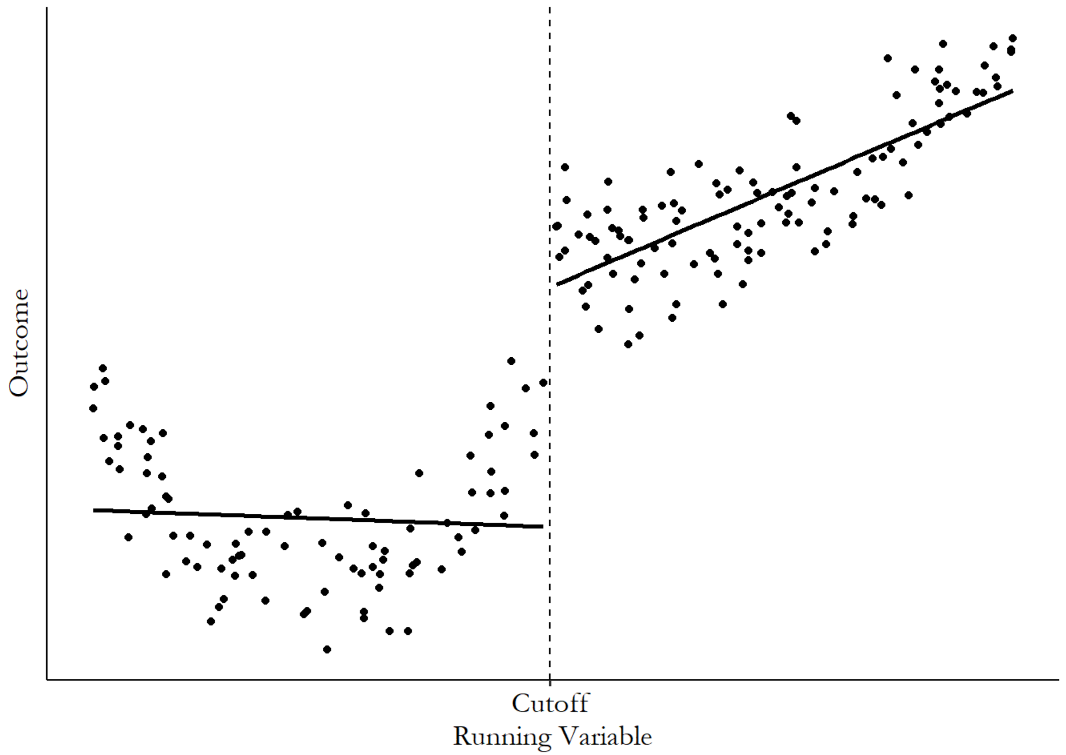
- The coefficient on $Treated$ remains our "jump at the cutoff" - our RDD estimate!

```
etable(feols(Y ~ X_centered*treated + I(X_centered^2)*treated, data = df,vcov='HC1'))
```

```
##                               feols(Y ~ X_cent..
## Dependent Var.:                                Y
##
## Constant                        -0.0340 (0.0400)
## X_centered                      0.6990. (0.3700)
## Treated                       0.7677*** (0.0603)
## X_centered square               -0.5722 (0.7205)
## X_centered x Treated             0.7509 (0.5621)
## Treated x X_centered squared      0.5319 (1.076)
## ____________________________ __________________
## S.E. type                     Heteroskedas.-rob.
## Observations                               1,000
```

Fitting Quadratic Lines in RDD

- Sometimes it can be hard to tell if a quadratic (or higher-order) term is really necessary
- Visualizations can help!



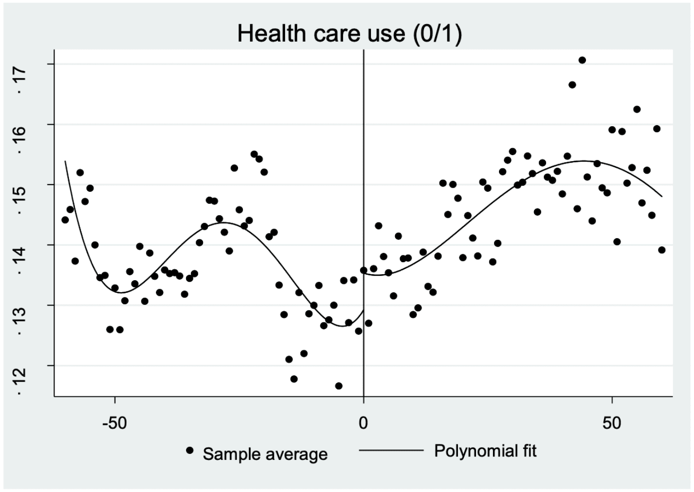
A linear slope makes the jump much bigger than it really is! (From the Effect Chapter 20)

Concept Checks

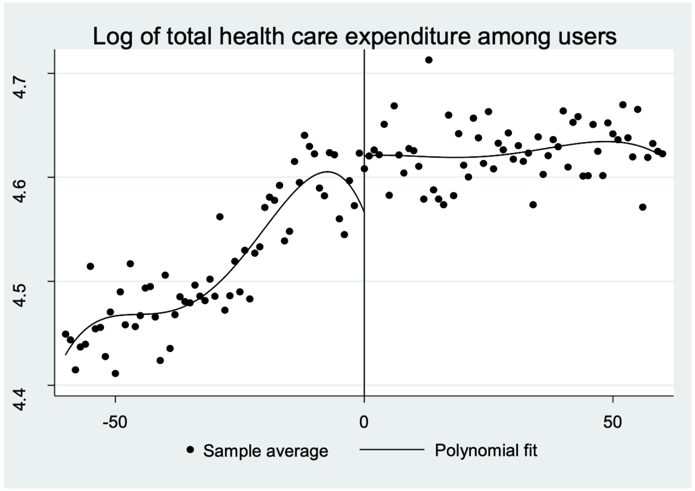
- Why might we want to use a polynomial term in our RDD model?
- What relationship are we assuming between the outcome variable and the running variable if we choose not to include $\boldsymbol{XCentered}$ in our model at all (i.e. a "zero-order polynomial")?

Careful with higher order polynomials

- Sometimes higher order polynoials can be a little too flexible and make it look like there's an effect where there isn't one
- "Overfitting" where your model too flexibly follows the data points can lie to you!



Health care use (0/1)



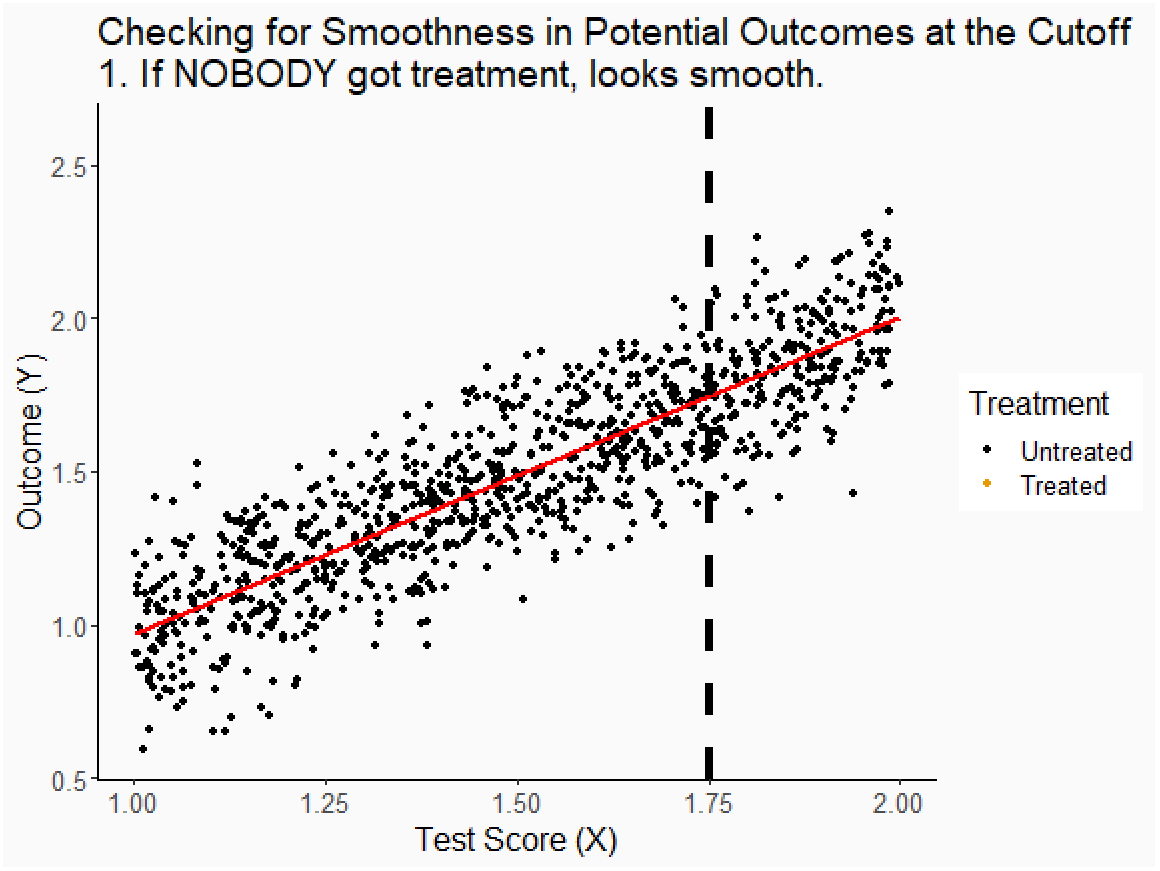
Log of total health care expenditure among users

Running variable is age with cutoff at age 20 (voting eligibility). Chang & Meyerhoefer (2020) on whether voting makes you sick via Andrew Gelman

Assumptions

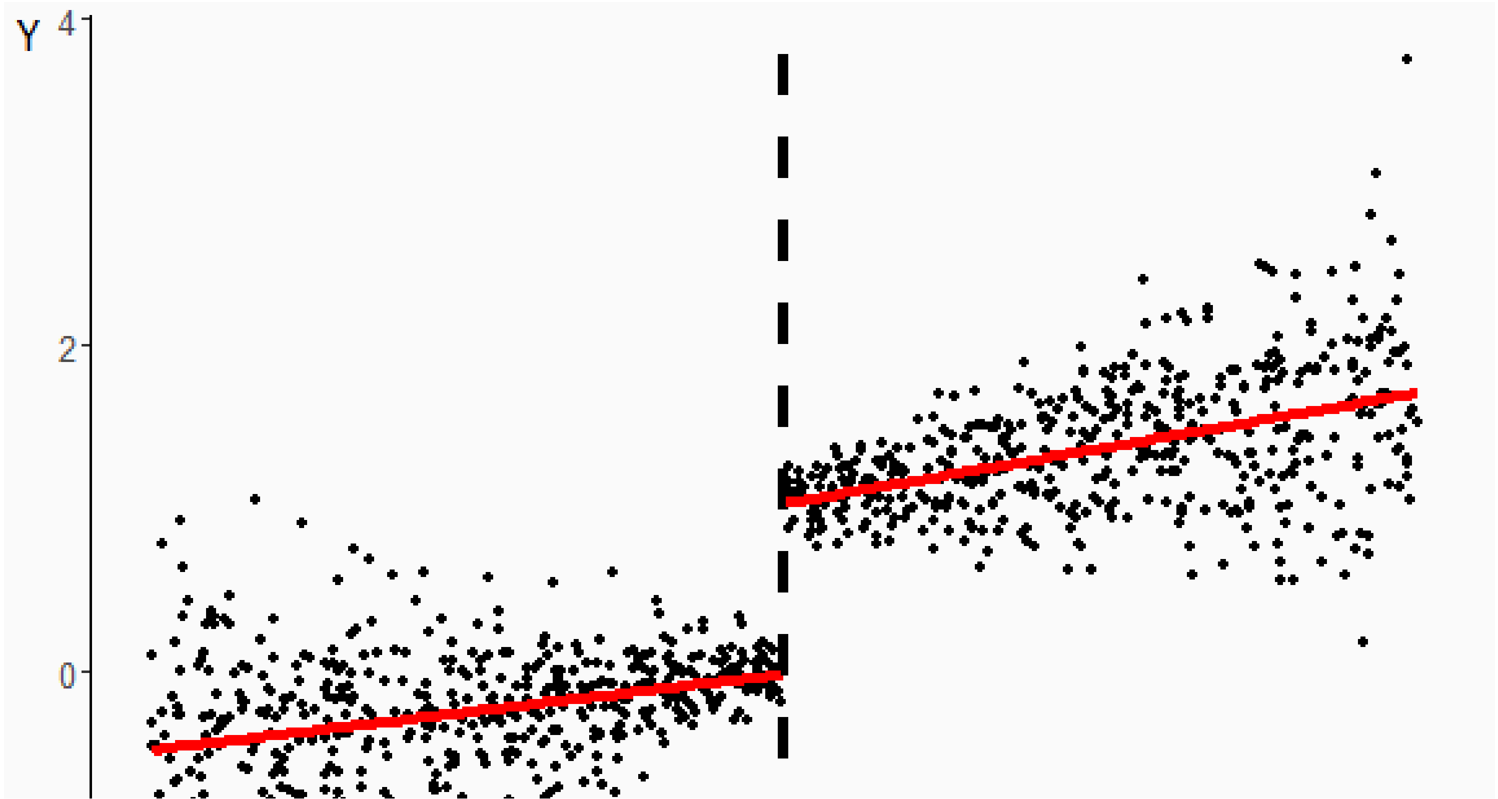
- There must be some assumptions lurking around here
- Some are more obvious (we use the correct functional form)
- Others are trickier. What are we assuming about the error term and endogeneity here?
- Specifically, we are assuming that *the only thing jumping at the cutoff is treatment*
- Sort of like parallel trends, but maybe more believable since we've narrowed in so far
- For example, if earning below 150% of the poverty line gets food stamps AND job training, then we can't isolate the effect of just food stamps
    - Or if the proportion of people who are self-employed jumps up just below 150% (based on *reported* income), that's endogeneity!
- The only thing different about just above/just below should be treatment

Graphically



Robust standard errors

- Oftentimes the error term is likely correlated with the running variable
- That means people tend to use "robust" standard errors, `vcov='HC1'` in R or `, r` in Stata
- Other times, it makes sense to clustered standard errors by the running variable



RDD Challenges

Other Difficulties

Assumptions, limitations, and diagnostics!

- Windows
- Granular running variables
- Manipulated running variables
- Fuzzy regression discontinuity
- How pros do it

Windows

- The basic idea of RDD is that we're interested in *the cutoff*
- The points away from the cutoff are only useful to help predict values at the cutoff
- Do we really want that full range? Is someone's test score of 30 really going to help us much in predicting Y at a test score of 89?
- So we might limit our analysis within just a narrow window around the cutoff, just like that initial animation we saw!
- This makes the exogenous-at-the-jump assumption more plausible, and lets us worry less about functional form (over a narrow range, not too much difference between a linear term and a square), but on the flip side reduces our sample size considerably

Windows

- Pay attention to the sample sizes, accuracy (true value .7) and standard errors!

```
##                          All    |X|<.25     |X|<.1    |X|<.05    |X|<.01
## Dependent Var.:            Y          Y          Y          Y          Y
##
## Treated              0.75***    0.77***    0.71***    0.61***      0.56
##                      (0.04)     (0.06)     (0.10)     (0.15)      (0.36)
## ---------------     --------   --------   --------   --------   --------
## S.E. type          Het.-rob. Het.-rob. Het.-rob. Het.-rob. Het.-rob.
## Observations           1,000        492        206         93         15
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Granular Running Variable

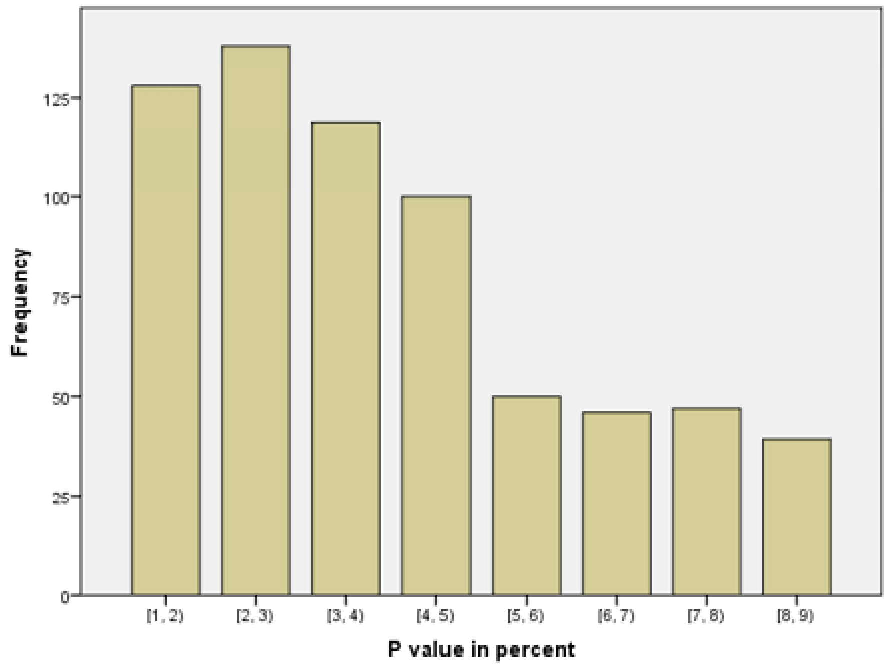
- We assume that the running variable varies more or less *continuously*
- That makes it possible to have, say, a test score of 89 compared to a test score of 90 it's almost certainly the same as except for random chance
- But what if our data only had test score in big chunks? i.e. I just know those earning "80-89" or "90-100"
    - Much less believable that groups only separated by random chance
- There are some fancy RDD estimators that allow for granular running variables
- But in general, if this is what you're facing, you might be in trouble
- Before doing an RDD, ask:
    - Is it plausible that someone with the highest value just below the cutoff, and someone with the lowest value just above the cutoff are only at different values because of random chance?

Looking for Lumping

- Ok, now let's go back to our continuous running variables
- What if the running variable is *manipulated*?
    - Imagine you're a teacher grading the gifted-and-talented exam. You see someone with an 89 and think "aww, they're so close! I'll just give them an extra point..."
    - Or, a school knows that if they have 41 students in a grade, they have to hire another teacher. So they just... don't admit more students
- Suddenly, that treatment is a lot less randomly assigned around the cutoff!
- If there's manipulation of the running variable around the cutoff, we can often see it in the presence of *lumping*
    - i.e. if there's a big cluster of observations to one side of the cutoff and a seeming gap missing on the other side

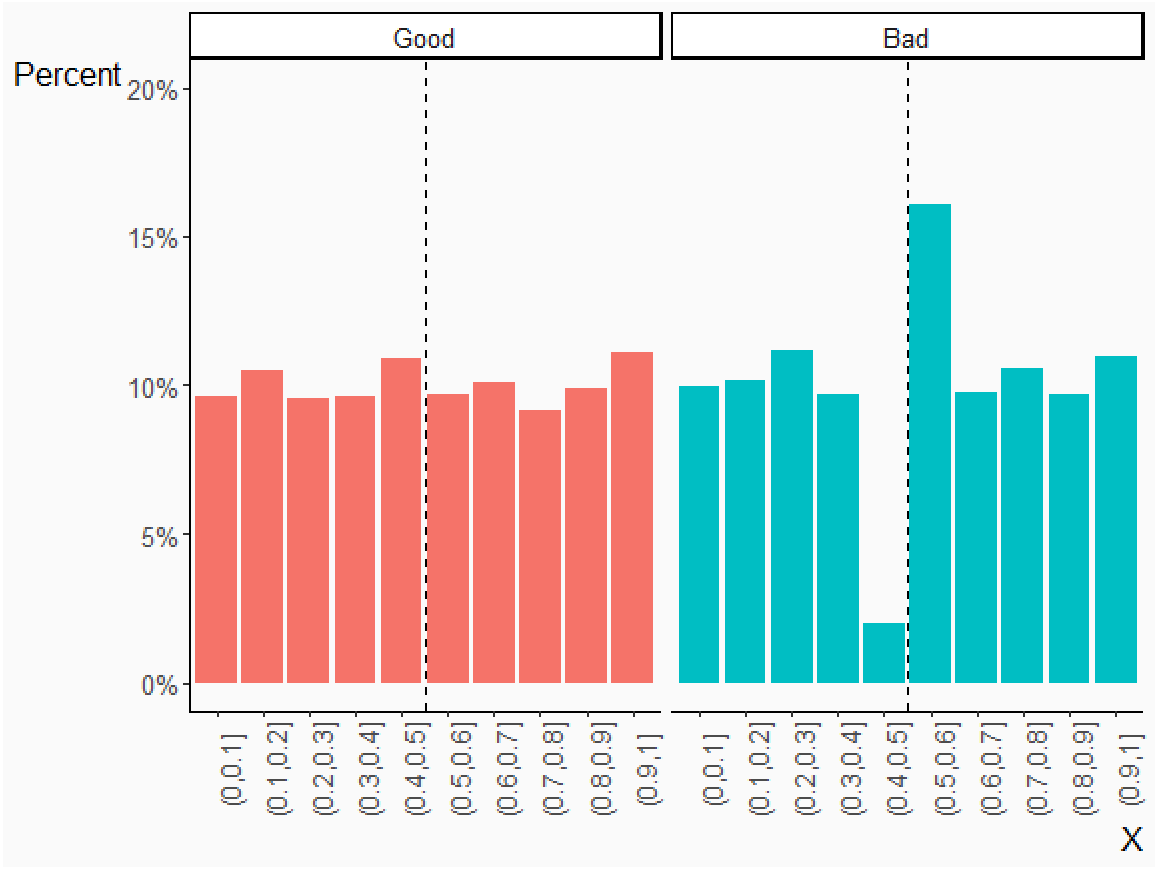
Looking for Lumping

- Here's an example from the real world in medical research - statistically, p-values *should* be uniformly distributed
- But it's hard to get insignificant results published. So people "p-hack" until they find significance and we have selection into publication based on $p < .05$.



Looking for Lumping

- We can look for lumping graphically by just plotting a binned histogram and looking for a jump in *number of observations* at the cutoff
- The first one looks pretty good. We have one that looks not-so-good on the right



McCrary Test

- A more formal way to check for manipulation is the McCrary test
- The null hypothesis that the running variable is continuous at the cutoff
- Intuitively, it assesses how likely the density of the running variable (number of observations) would be to occur if the running variable were continuous at the cutoff
- If really unlikely, we might reject that null, suggesting manipulation
- It can also be implemented easily with the **rddensity** package in R

rddensity() output

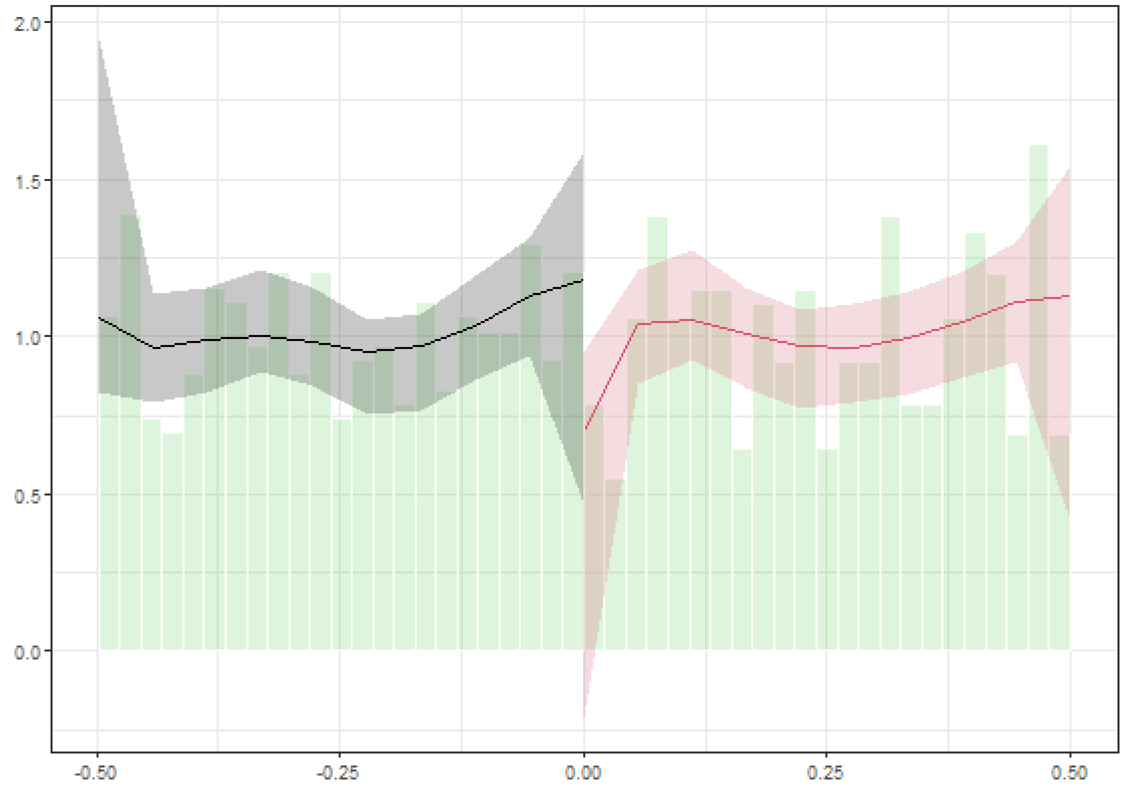
Null hypothesis is that the running variable is continuous at the cutoff (i.e. no manipulation) for various bandwidths/window lengths

```r
#library(rddensity) # Already loaed
mccrary ← rddensity(df$X_centered,c=0)
summary(mccrary)
```

```
##
## Manipulation testing using local polynomial density estimation.
##
## Number of obs =       1000
## Model =               unrestricted
## Kernel =              triangular
## BW method =           estimated
## VCE method =          jackknife
##
## c = 0                 Left of c           Right of c
## Number of obs         501                 499
## Eff. Number of obs    185                 161
## Order est. (p)        2                   2
## Order bias (q)        3                   3
## BW est. (h)           0.175               0.169
##
## Method                T                   P > |T|
```

rdplotdensity() implementation

```
rdplotdensity(mccrary,df$X_centered)
```



```
## $Estl
## Call: lpdensity
##
```

Looking for Lumping

- Another thing we can do is do a "placebo test"
- Check if variables *other than treatment or outcome* vary at the cutoff
- We can do this by re-running our RDD but switching our outcome with another variable
- If we get a significant jump, that's bad! That tells us that *other things are changing at the cutoff* which implies some sort of manipulation (or just super lousy luck)
- If all placebo tests are passed, that's good news, but doesn't mean prove zero manipulation

Concept Checks

- Why does using a narrow window make the effect estimate noisier?
- Intuitively, why would we be skeptical that a regression discontinuity run on a very granular running variable is valid?
- Why might bunching of observations on one side of the cutoff be a sign of manipulation?

Fuzzy Regression Discontinuity

Fuzzy Regression Discontinuity

- So far, we've assumed that you're either on one side of the cutoff and untreated, or the other and treated
- What if it isn't so simple? What if the cutoff just *increases* your chances of treatment?
- For example, what if 30% of schools with fewer than 40 students make smaller classrooms for whatever reason
    - It can get more complicated than this -- it always can
- This is a "fuzzy regression discontinuity" (yes, that does sound like a bizarre Sesame Street episode)
- Now, our RDD will understate the true effect, since it's being calculated on the assumption that we added treatment to 100% of people at the cutoff, when really it's 70%. So we'll get roughly only about 70% of the effect

Fuzzy Regression Discontinuity

- We can account for this with a model designed to take this into account
- Specifically, we can use something called two-stage least squares (or Wald instrumental variable estimator) to handle these sorts of situations
- Basically, two-stage least squares estimates how much the chances of treatment go up at the cutoff, and scales the estimate by that change
- So it would take whatever result we got on the previous slide and divide it by 0.7 (the increased in treated share) to get the true effect

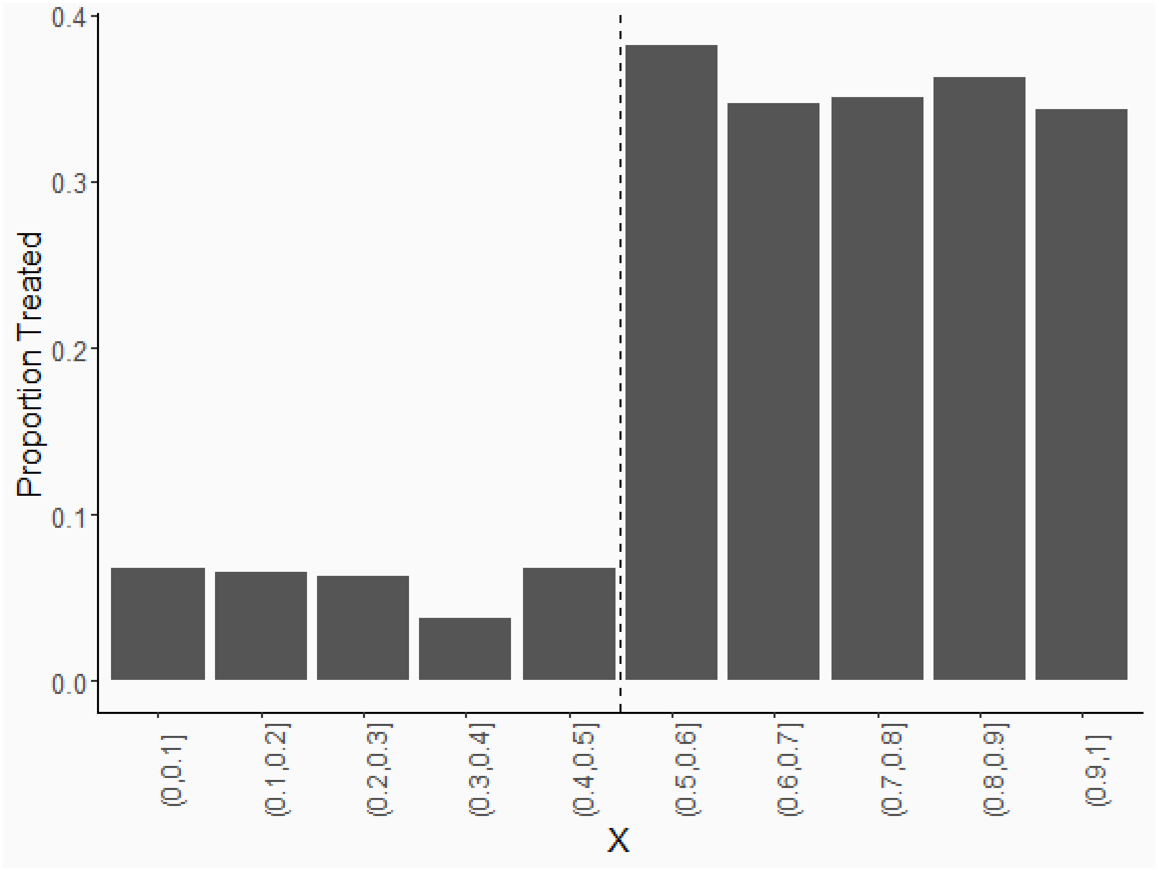
Fuzzy Regression Discontinuity

First let's make some fake data:

```
set.seed(1000)
df ← tibble(X = runif(1000)) %>%
  mutate(treatassign = .05 + .3*(X > .5)) %>%
  mutate(rand = runif(1000)) %>%
  mutate(treatment = treatassign > rand) %>%
  mutate(Y = .2 + .4*X + .5*treatment + rnorm(1000,0,0.3)) %>% # True effect .5
  mutate(X_center = X - .5) %>%
  mutate(above_cut = X > .5)
```

Fuzzy Regression Discontinuity

- Notice that the y-axis here isn't the outcome, it's "proportion treated"



Fuzzy Regression Discontinuity

- We can perform this using the instrumental-variables features of `feols`
- The first stage is the interaction between the running variable and whether treated regressed on the interaction of the running variable and the "sharp" cutoff
- `feols(outcome ~ controls | XC*treated ~ XC*above_the_cutoff)`

Fuzzy Regression Discontinuity

- (the true effect of treatment is .5 - okay, it's not perfect)

```
predict_treatment <- feols(treatment ~ X_center*above_cut, data = df)
without_fuzzy <-feols(Y ~ X_center*treatment, data = df)
fuzzy_rdd <- feols(Y ~ 1 | X_center*treatment ~ X_center*above_cut, data = df)
etable(predict_treatment, without_fuzzy, fuzzy_rdd,
  dict=c('above_cutTRUE'='Above Cut','treatmentTRUE'='Treatment'))
```

```
##                      predict_trea..  without_fuzzy      fuzzy_rdd
## Dependent Var.:           treatment              Y              Y
##
## Constant               0.06. (0.04) 0.41*** (0.01) 0.41*** (0.03)
## X_center               0.004 (0.12) 0.40*** (0.04) 0.45*** (0.12)
## Above Cut            0.31*** (0.05)
## X_center x Above Cut   -0.04 (0.17)
## Treatment                           0.45*** (0.03) 0.48*** (0.12)
## X_center x Treatment                   0.07 (0.10)   -0.25 (0.48)
## ____________________ ______________ ______________ ______________
## S.E. type                       IID            IID            IID
## Observations                  1,000          1,000          1,000
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Concept Checks

- If the treatment variable is fuzzily assigned, do we underestimate with a sharp RDD?

Concept Checks

- If the treatment variable is fuzzily assigned, do we underestimate with a sharp RDD?
- Yes, but how do we know that, if our treatment variable is fuzzily assigned, we will *underestimate* the effect if we just run a regular RDD, rather than overestimate it?

How professionals do it

How professionals do it

- We've gone through all kinds of procedures for doing RDD in R already using regression
- But often, professional researchers won't do it that way because it's a bit too easy to mess up details
- Instead, they use packages like **rdrobust** (available in R, Stata, and Python) and written by a team of econometricians
- It abstracts the tedious stuff, like bandwidth selection and standard errors, and gives you loads of customization options for your RDD
- In general, packages like these written by experts who are well-published in discussing a method are a good idea to try

RDrobust

- There are three major functions in RD robust:

1. `rdrobust()` - the main estimation approach, it returns info about the regression and you can customize a variety of complex RD stuff
2. `rdplot()` - a plotting function that shows the jump at the cutoff and let's you customize much of the complexities
3. `rdbwselect()` - a bandwidth selection tool that helps you pick the best bandwidth for your RDD

Basics of **rdrobust**

- We can specify an RDD model by just telling it the dependent variable Y, the running variable X, and the cutoff c.
- We can also specify how many polynomials to use with `p`, defaults to 1
    - (it applies the polynomials more locally than our linear OLS models do - a bit more flexible)
- Use `c` to specify the cutoff (no need to center the running variable manually)
- Pick the bandwidth with `h` or use a data-driven technique with `rdbwselect()`
- Including a `fuzzy` option to specify actual treatment outside of the running variable/cutoff combo
- And many other options
- But output is pretty nasty, so you'll need to do some work to get it into a readable format

rdrobust

```r
summary(rdrobust(df$Y, df$X, c = .5))
```

```
## Sharp RD estimates using local polynomial regression.
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  488          512
## Eff. Number of Obs.             135          162
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.152        0.152
## BW bias (b)                   0.229        0.229
## rho (h/b)                     0.666        0.666
## Unique Obs.                     488          512
##
## ============================================================================
```

rdrobust

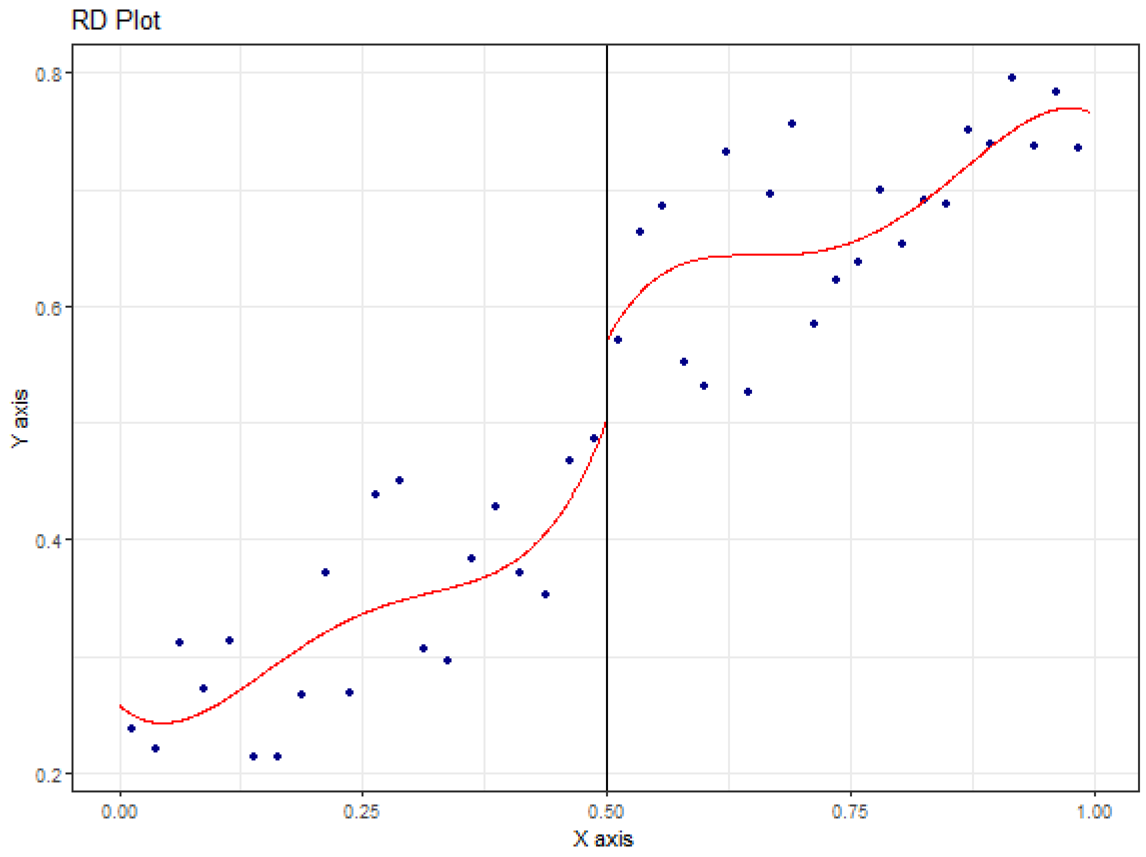
```
summary(rdrobust(df$Y, df$X, c = .5, fuzzy = df$treatment))
```

```
## Fuzzy RD estimates using local polynomial regression.
##
## Number of Obs.                 1000
## BW type                       mserd
## Kernel                   Triangular
## VCE method                       NN
##
## Number of Obs.                  488          512
## Eff. Number of Obs.             117          154
## Order est. (p)                    1            1
## Order bias  (q)                   2            2
## BW est. (h)                   0.141        0.141
## BW bias (b)                   0.206        0.206
## rho (h/b)                     0.685        0.685
## Unique Obs.                     488          512
##
## First-stage estimates.
```

rdrobust

- We can even have it automatically make plots of our RDD! Same syntax

```
rdplot(df$Y, df$X, c = .5)
```



That's it!

- That's what we have for RDD
- Go explore the regression discontinuity activity on class sizes

Next lecture: Bootstrapping