Regression Discontinuity Design

Marcelo Coca Perraillon

University of Chicago

May 13 & 18, 2015

An updated version including nonparametric estimation and code is available here: https://clas.ucdenver.edu/marcelo-perraillon/teaching/health-services-research-methods-i-hsmp-7607

Plan

- Overview of RDD
- Meaning and validity of RDD
- Several examples from the literature
- Estimation (where most decisions are made)
- Discussion of Almond et al (low birth weight)
- Stata code and data for *all* examples will be available on Chalk. Email me if you have questions: mcoca@uchicago.edu

Basics

- Method developed to estimate treatment effects in non-experimental settings
- Provides causal estimates of treatment effects
- Good internal validity; some assumptions can be empirically verified
- Treatment effects are local (LATE)
- Limits external validity
- Relatively easy to estimate (like RCT)
- First application: Thistlethwaite and Campbell (1960)

Thistlethwaite and Campbell

- They studied the impact of merit awards on future academic outcomes
- Awards allocated based on test scores
- If a person had a score greater than c, the cutoff point, then she received the award
- Simple way of analyzing: compare those who received the award to those who didn't. (Why is this the wrong approach?)
- Confounding: factors that influence the test score are also related to future academic outcomes (income, parents' education, motivation)
- Thistlethwaite and Campbell realized they could compare individuals just above and below the cutoff point.

Validity

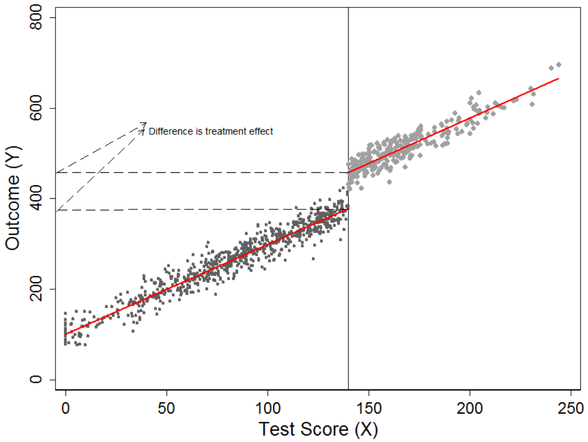
- Simple idea: assignment mechanism is known
- We know that the probability of treatment jumps to 1 if test score $> c$
- Assumption is that individuals cannot manipulate with precision their assignment variable (think about the SAT)
- Key word: *precision*. Consequence: comparable individuals near cutoff point
- If treated and untreated individuals are similar near the cutoff point then data can be analyzed as if it were a (conditionally) randomized experiment
- If this is true, then background characteristics should be similar near c (can be checked empirically)
- The estimated treatment effect applies to those near the cutoff point (limits external validity)

Validity

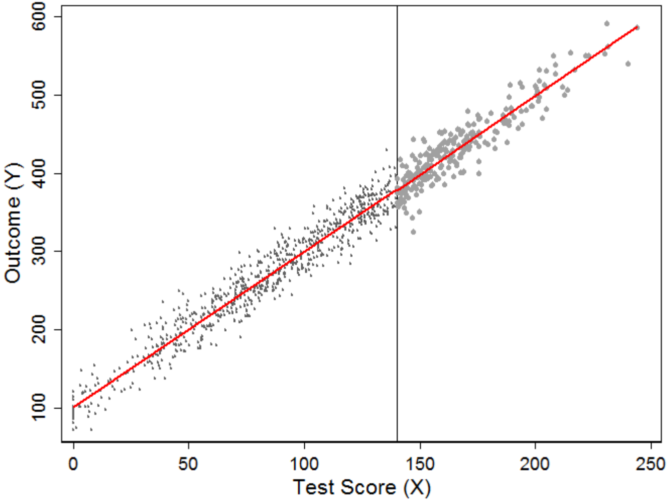
- Careful when you read that the validity depends on rule being "arbitrary" or assignment variable measured with error (e.g. Moscoe et al. 2015)
- Validity hinges on assignment mechanism being known and free of manipulation with precision or cutoff point in some way related to outcome of interest
- Manipulation example 1: Test with few questions and plenty of time
- Manipulation example 2: DMV test to get a driving license
- Example 3: Some mechanism makes cutoff point related to outcome (think biology: blood pressure). What if meassured with error?
- Example 4: Eligibility criteria to obtain some benefit (say, below income of 28K). Why? How could you verify assumptions?
- A comment on continuity
- Again: some manipulation is fine (you can always study harder, for example). Precision and lack of relation to outcome is the key to identify causal effects

Graphical Example

- Simulated data with $c = 140$
- `gen y = 100 + 80*T + 2*x + rnormal(0, 20)`



No effect



Outcome (Y)
100
200
300
400
500
600
0
50
100
150
200
250
Test Score (X)

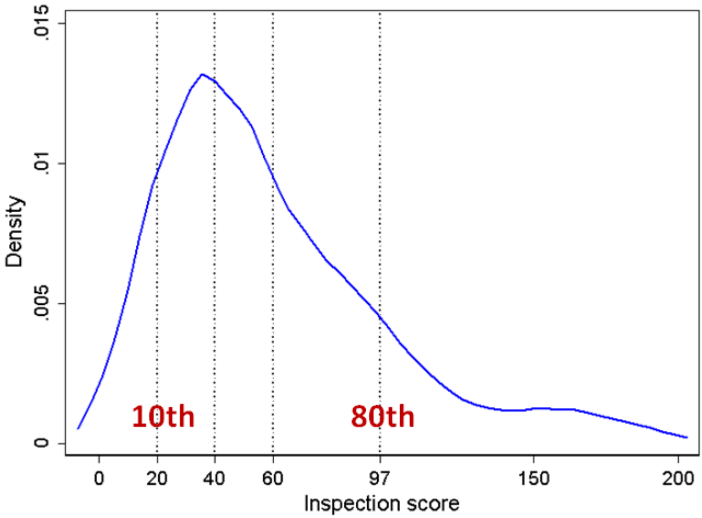
Sharp and fuzzy RDD

- Sharp RDD: Assignment or running variable completely determines treatment. A jump in the probability of treatment before and after cutoff point.
- Fuzzy RDD: Cutoff point increases the *probability* of treatment but doesn't completeley determines treatment.
- Which brings us back to the world of instrumental variables...
- Not used often but has a lot of potential
- Think of encouragement designs or imperfect compliance (like the Oregon study)

Examples from literature

- Almond et al. (2010): Assignment variable is birth weight. Infants with low birth weight ($< 1,500$ grams or about 3 pounds) receive more medical treatment.
- We'll talk more about this paper next class. Don't forget to read it!
- Lee, Moretti, Buttler (2004): The vote share (0 to 100 percent) for a candidate is a continuous variable. A candidate is elected if he or she obtains more than 50% of the votes. They evaluated voting record of candidates in close elections.
- CMS rates nursing homes using 1 to 5 stars. Overall stars are assigned based on deficiency data transformed into a points system. Outcome: new admissions six months after the release of ratings.

Assignment of stars based on scores



Examples from literature

- Anderson and Magruder (2012) and Lucas (2012): Yelp.com ratings have an underlying continuous score. Distribution determines cutoff points for 1 to 5 stars. Effect of an extra star on future reservations and revenue.
- Anderson et al. (2012): Young adults lose their health insurance as they age (older than 18 and in college but different after ACA). Age changes the probability of having health insurance (fuzzy design).

Estimation: Parametric

- Simplest case is linear relationship between Y and X

$$Y_i = \beta_0 + \beta_1 T_i + \beta_3 X_i + \epsilon_i$$

- $T_i = 1$ if subject i received treatment and $T_i = 0$ otherwise. You can also write this as $T_i = \mathbf{1}(X_i > c)$ or $T_i = \mathbb{1}_{[X_i > c]}$
- X is the assignment variable (sometimes called "forcing" or "running" variable)
- Usually centered at cutoff point
- $Y_i = \beta_0 + \beta_1 T_i + \beta_3 (X_i - c) + \epsilon_i$. Treatment effect is given by β_1.
- $E[Y|T = 1, X = c] = \beta_0 + \beta_1$ and $E[Y|T = 0, X = c] = \beta_0$.
- $E[Y|T = 1, X = c] - E[Y|T = 0, X = c] = \beta_1$.

Reminder on centering

- Centering changes the interpretation of the intercept:

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

- Compare to:

$$Y = \alpha_0 + \alpha_1 Age + \alpha_2 Edu$$

- $\beta_1 = \alpha_1$, $\beta_2 = \alpha_2$, but $\alpha_0 \neq (\beta_0 - \beta_1 65)$
- Useful with interactions:

$$Y = \alpha_0 + \alpha_1 Age + \alpha_2 Edu + \alpha_3 Age \times Edu$$

- Compare to:

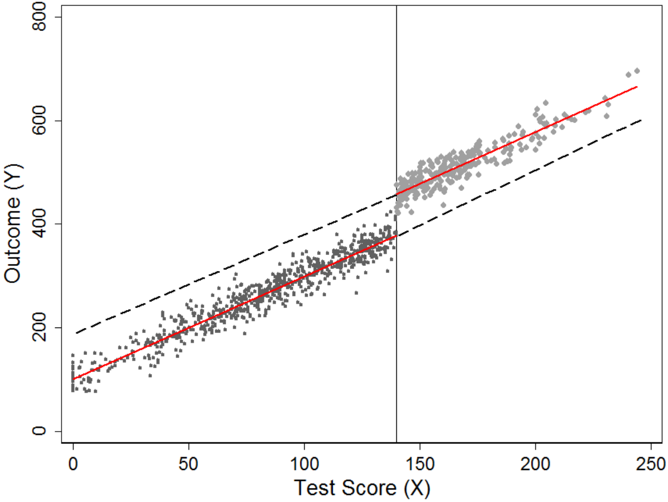
$$Y = \beta_0 + \beta_1(Age - 65) + \beta_2(Edu - 12) + \beta_3(Age - 65) \times (Edu - 12)$$

Extrapolation

- Note that the estimation of treatment effect in RDD depends on extrapolation
- To the left of cutoff point only non-treated observations
- To the right of cutoff point only treated observations
- What is the treatment effect at $X = 130$? Just plug in:
- $E[Y|T, X = 130] = \beta_0 + \beta_1 T + \beta_3(130 - 140)$

Extrapolation...

Dashed lines are extrapolations



Counterfactuals

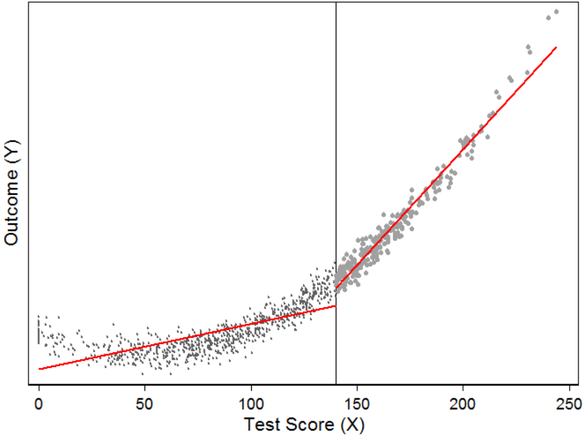
- The extrapolation is a counterfactual or potential outcome
- Each person i has two potential outcomes (Rubin's causal framework).
- $Y_i(1)$ denotes the outcome of person i if in the treated group
- $Y_i(0)$ denotes the outcome of person i if in the non-treated group
- Causal effect of treatment for person i is $Y_i(1) - Y_i(0)$
- Average treatment effect is $E[Y_i(1) - Y_i(0)]$
- Only one potential outcome is observed. In randomized experiments, one group provides the conterfactual for the other because they are comparable (exchangeable)
- Exchangeability (epi). Also called "selection on observables" or "no unmeasured confounders"

Counterfactuals, II

- In RDD the counterfactuals are conditional on X as in a conditionally randomized trial (think severity)
- We are interested in the treatment effect at $X = c$: $E[Y_i(1) - Y_i(0)|X_i = c]$
- Treatment effect is $lim_{x \to c} E[Yi|Xi = x] - lim_{x \leftarrow c} E[Yi|Xi = x]$
- Estimation possible because of the continuity of $E[Y_i(1)|X]$ and $E[Y_i(0)|X]$
- See Hahn, Todd, and Van der Klaauw (2001) for details
- The estimation of the treatment effect is based on extrapolation because of lack of overlap. Thefore, the functional relationship between X and Y must be correctly specified

Need to model relationship between X and Y correctly

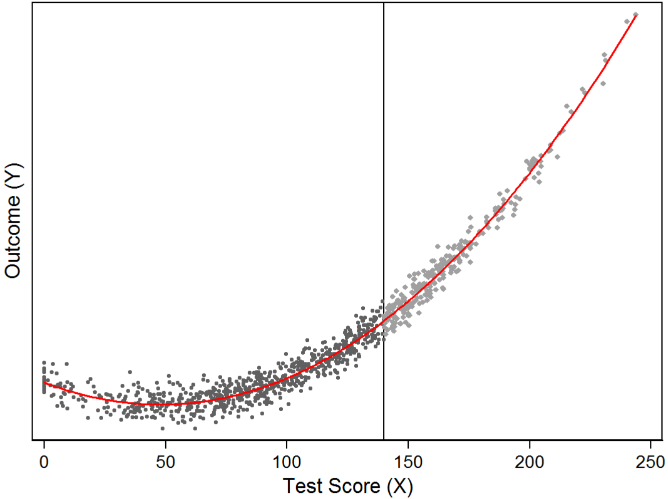
- What if nonlinear? Could result in a biased treatment effect if one assumes a linear model.



Other specifications

- More general: $Y_i = \beta_0 + \beta_1 T_i + \beta_3 f(X_i - c) + \epsilon_i$
- If $(X_i - c) = \tilde{X}_i$ then $Y_i = \beta_0 + \beta_1 T_i + \beta_3 f(\tilde{X}_i) + \epsilon_i$
- Most common form for $f(\tilde{X}_i)$ are polynomials
- Polynomials of order p:
  $Y_i = \beta_0 + \beta_1 T_i + \beta_2 \tilde{X}_i + \beta_3 \tilde{X}_i^{\,2} + \beta_4 \tilde{X}_i^{\,3} + \cdots + \beta_{p+1} \tilde{X}_i^{\,p} + \epsilon_i$
- More flexibility with interactions
- 2nd degree with interactions:
  $Y_i = \beta_0 + \beta_1 T_i + \beta_3 \tilde{X}_i + \beta_4 \tilde{X}_i^{\,2} + \beta_5 \tilde{X}_i \times T_i + \beta_6 \tilde{X}_i^{\,2} \times T_i + \epsilon_i$
- Question: Why not controlling for other covariates?

Third degree polynomial. Actual model second degree polynomial (see Stata do file). However...



A note on higher order polynomials

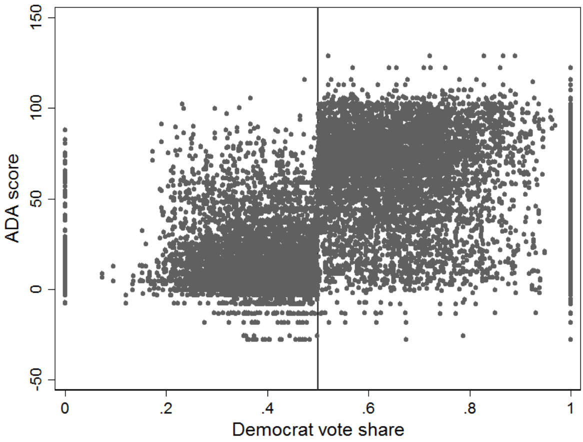
- We will see an example in which using higher order polynomials does not influence results
- In some cases, however, it may matter
- Gelman and Inbems (2014) subtle paper: "Why High-order Polynomials Should not be Used in Regression Discontinuity Designs"
- "We argue that estimators for causal effects based on [higher order polynomials] can be misleading, and we recommend researchers do not use them, and instead use estimators based on local linear or quadratic polynomials..."

Real dataset

- Data from Lee, Moretti, Buttler (2004)
- U.S. House elections (1946-1995)
- Forcing variable is Democratic vote share. If share > 50 then Democratic candidate is elected
- Outcome is a liberal voting score from the Americans for Democratic Action (ADA)
- Do candidates who are elected in close elections tend to moderate their congressional voting?
- "We find that the degree of electoral strength has no effect on a legislator's voting behavior"
- Data and code are on Chalk

Graph a bit messy (about 13,500 obs)

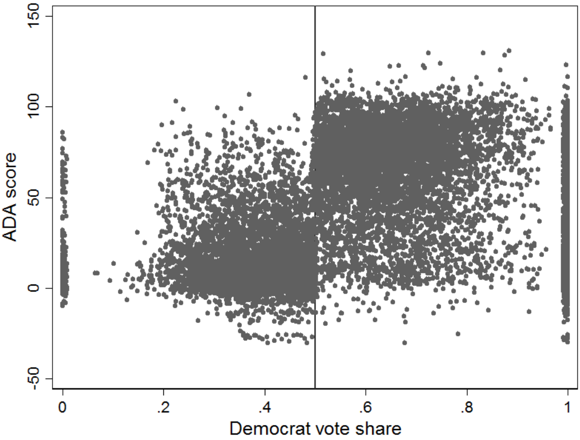
```
scatter score demvoteshare, msize(tiny) xline(0.5) ///
xtitle("Democrat vote share") ytitle("ADA score")
```



Good idea to add some “jittering”

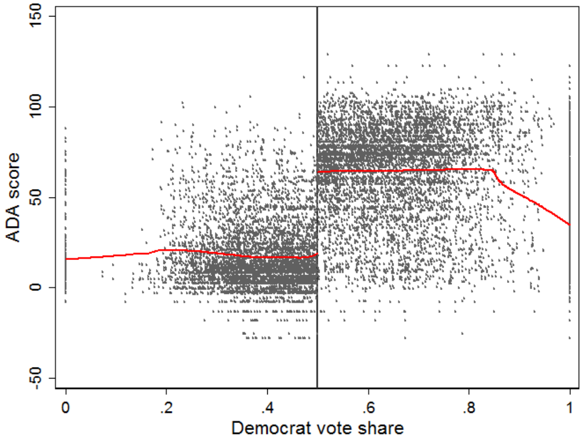
With the jitter option, it is easier to see where is the mass

```
scatter score demvoteshare, msize(tiny) xline(0.5) ///
xtitle("Democrat vote share") ytitle("ADA score") jitter(5)
```



Useful to "smooth" data with LOWESS

```
lowess score demvoteshare if democrat ==1, gen (lowess_y_d1) nograph bw(0.5)
lowess score demvoteshare if democrat ==0, gen (lowess_y_d0) nograph bw(0.5)
....
....
```

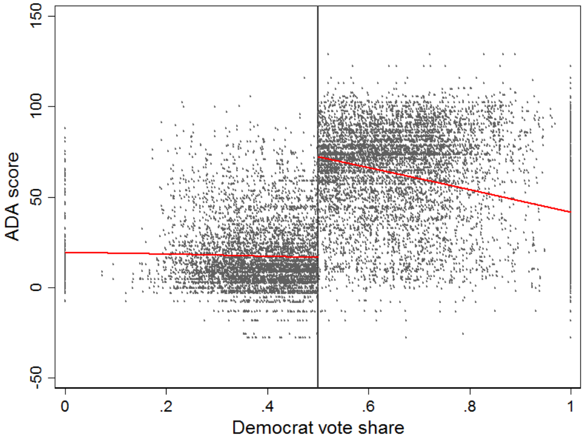


LOWESS

- LOcally WEighted Scatterplot Smoothing
- Non-parametric graphical method
- Computationally intensive (one regression per data point)
- For each data point, run a *weighted* linear regression (linear or polynomials on X) using all the observations within a window. Weights give more importances to observations close to data point
- Predicted y, $\hat{y}$, is then the "smoothed" (y_i, x_i) point

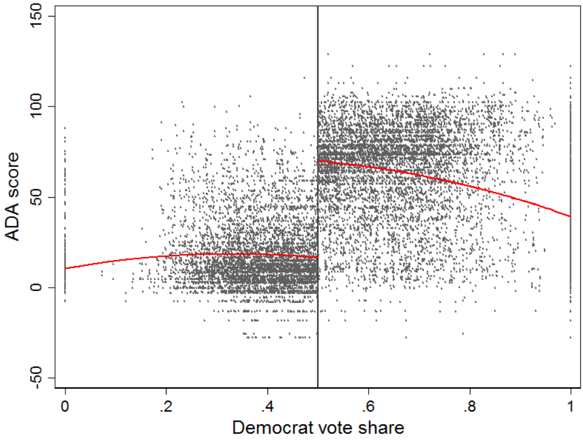
Parametric: Linear relationship

```
scatter score demvoteshare, msize(tiny) xline(0.5) xtitle("Democrat vote share") ///
 ytitle("ADA score") || lfit score demvoteshare if democrat ==1, color(red) || ///
 lfit score demvoteshare if democrat ==0, color(red) legend(off)
```



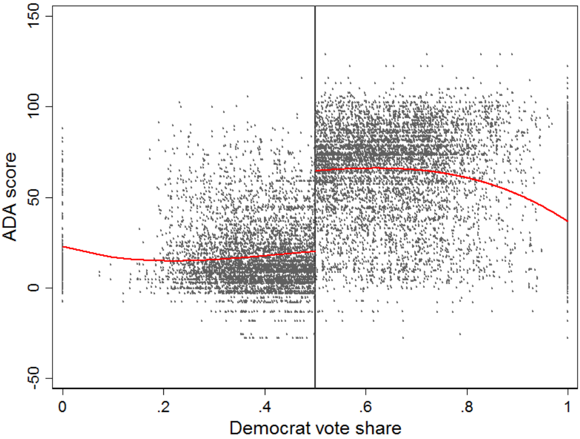
Quadratic

```
gen demvoteshare2 = demvoteshare^2
reg score demvoteshare demvoteshare2 democrat
predict scorehat0
```



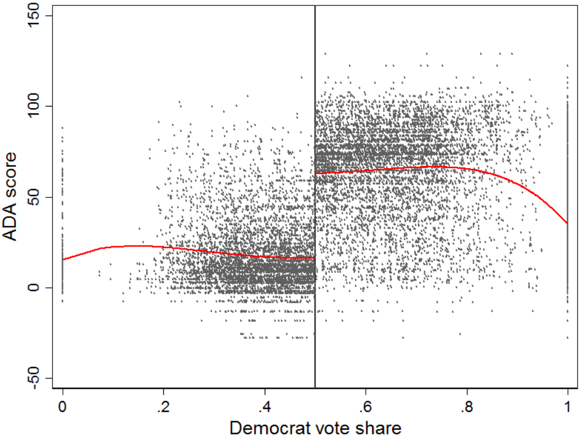
Third degree polynomial

```
gen demvoteshare3 = demvoteshare^3
reg score demvoteshare demvoteshare2 demvoteshare3 democrat
predict scorehat01
```



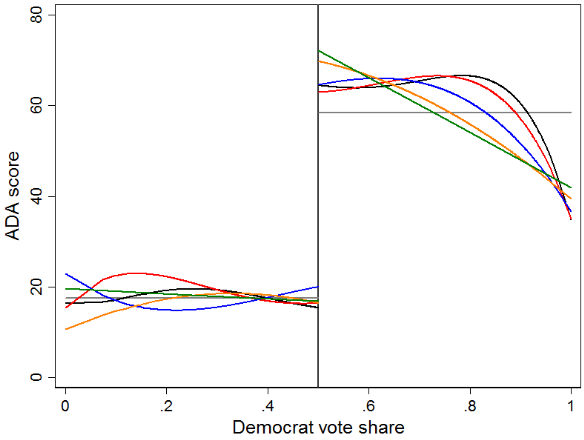
Fourth degree polynomial

```
gen demvoteshare4 = demvoteshare^4
reg score demvoteshare demvoteshare2 demvoteshare3 demvoteshare4 ///
    democrat
predict scorehat02
```



Mean (null model) to fifth degree polynomial

```
line scorehat04 demvoteshare if democrat ==1, sort color(gray) || ///
line scorehat04 demvoteshare if democrat ==0, sort color(gray) legend(off) ....
```



Parametric

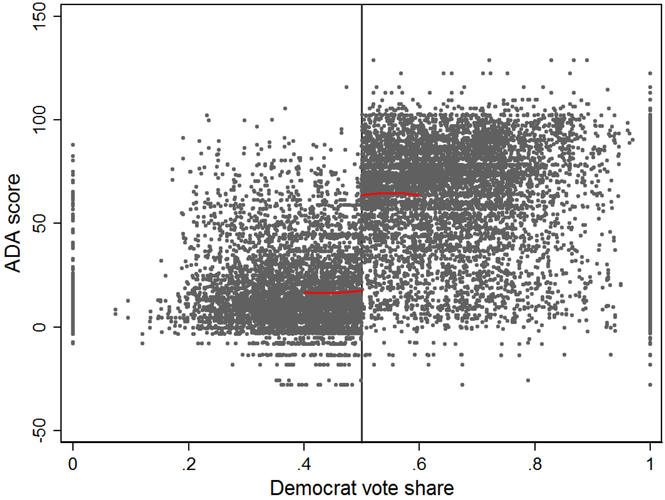
- Note that polynomials "smooth" the data (like LOWESS)
- We used *all* the data even though we want treatment effect at c
- But polynomials give weight to points away from c and tend to provide smaller SEs
- In other datasets, the choice of polynomial degree will matter (see Gelman and Inbems, 2014)
- Why not only use data close to c? Bias and variance trade-off

Restrict to a window

- Run a flexible regression like a polynomial with interactions (stratified) but don't use observations away from the cutoff. Choose a bandwidth around $X = 0.5$. Lee et al (2004) used 0.4 to 0.6.

```
reg score demvoteshare demvoteshare2 if democrat ==1 & ///
   (demvoteshare>.40 & demvoteshare<.60)
predict scorehat1 if e(sample)
reg score demvoteshare demvoteshare2 if democrat ==0 & ///
   (demvoteshare>.40 & demvoteshare<.60)
predict scorehat0 if e(sample)

scatter score demvoteshare, msize(tiny) xline(0.5) xtitle("Democrat vote share") //
  ytitle("ADA score") || ///
  line scorehat1 demvoteshare if democrat ==1, sort color(red) || ///
  line scorehat0 demvoteshare if democrat ==0, sort color(red) legend(off)
graph export lee3_1.png, replace
```


ADA score
150
100
50
0
-50
0
.2
.4
.6
.8
1
Democrat vote share

Limit to window, 2nd degree polynomial

```
gen x_c = demvoteshare - 0.5
gen x2_c = x_c^2
reg score i.democrat##(c.x_c c.x2_c) if (demvoteshare>.40 & demvoteshare<.60)

      Source |       SS       df       MS              Number of obs =    4632
-------------+------------------------------           F(  5,  4626) = 1153.29
       Model |  2622762.02     5  524552.404           Prob > F      =  0.0000
    Residual |   2104043.2  4626  454.829918           R-squared     =  0.5549
-------------+------------------------------           Adj R-squared =  0.5544
       Total |  4726805.22  4631   1020.6878           Root MSE      =  21.327

------------------------------------------------------------------------------
       score |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  1.democrat |    45.9283   1.892566    24.27   0.000     42.21797    49.63863
         x_c |   38.63988   60.77525     0.64   0.525     -80.5086    157.7884
        x2_c |   295.1723   594.3159     0.50   0.619    -869.9704    1460.315
             |
democrat#c.x_c |
           1 |   6.507415   88.51418     0.07   0.941    -167.0226    180.0374
             |
democrat#c.x2_c |
           1 |  -744.0247   862.0435    -0.86   0.388    -2434.041    945.9916
             |
       _cons |   17.71198   1.310861    13.51   0.000     15.14207    20.28189
------------------------------------------------------------------------------
```

So what should you do?

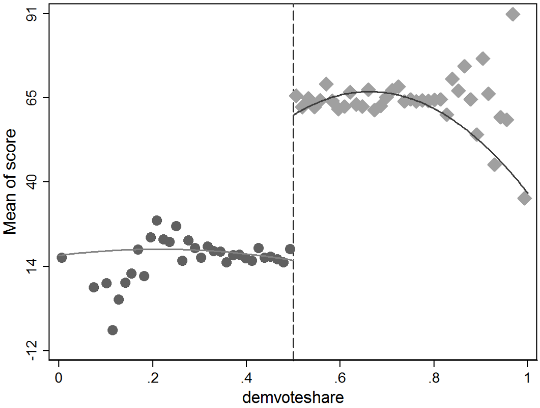
- Best case: Whatever you do gives you similar results (like in this example)
- Most common strategy is to restrict estimation to a window adjusting for covariates
- It used to be popular to use higher order polynomials
- Try different windows and present sensitivity analyses
- Balance should determine the size of window
- Try non-parametric methods

Nonparametric methods

- Paper by Hahn, Todd, and Van der Klaauw (2001) clarified assumptions about RDD and framed estimation as a nonparametric problem
- Emphasized using local polynomial regression instead of something like LOWESS
- "Nonparametric methods" means a lot of things in statistics
- In the context of RDD, the idea is to estimate a model that does not assume a functional form for the relationship between Y and X. The model is something like $Y_i = f(X_i) + \epsilon_i$
- A very basic method: calculate $E[Y]$ for each bin on X (think of a histogram)

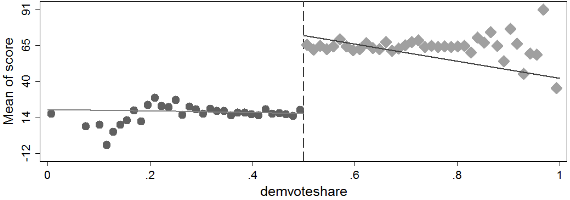
- Stata has a command to do just that: `cmogram`
- After installing the command (`ssc install cmogram`) type `help cmogram`. Lots of useful options
- Common way to show RDD data. See for example Figure II of Almond et al. (2010). To recreate something like Figure 1 of Lee et al (2004):

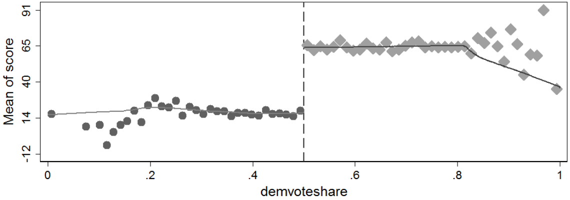
```
cmogram score demvoteshare, cut(.5) scatter line(.5) qfit
```



Compare to linear and LOWESS fits

```
cmogram score demvoteshare, cut(.5) scatter line(.5) lfit
cmogram score demvoteshare, cut(.5) scatter line(.5) lowess
```





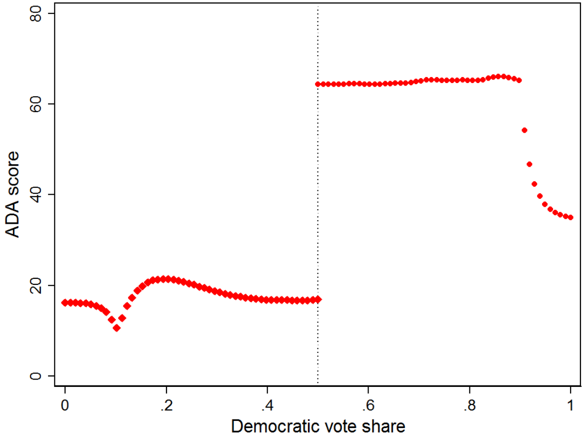
Local polynomial regression

- Hahn, Todd, and Van der Klaauw (2001) showed that one-side Kernel estimation (like LOWESS) may have poor properties because the point of interest is at a boundary
- Proposed to use instead a local linear nonparametric regression
- Stata's `lpoly` command estimates kernel-weighted local polynomial regression
- Think of it as a weighted regression restricted to a window (hence "local"). The Kernel provides the weights
- A rectangular Kernel would give the same result as taking $E[Y]$ at a given bin on X. The triangular Kernel gives more importance to observations close to the center
- Method sensitive to choice of bandwidth (window)

Local regression is a smoothing method

- Kernel-weighted local polynomial regression is a smoothing method

```
lpoly score demvoteshare if democrat == 0, nograph kernel(triangle) gen(x0 sdem0) bwidth(0.1)
lpoly score demvoteshare if democrat == 1, nograph kernel(triangle) gen(x1 sdem1) bwidth(0.1)
<omitted>
```



Treatment effect

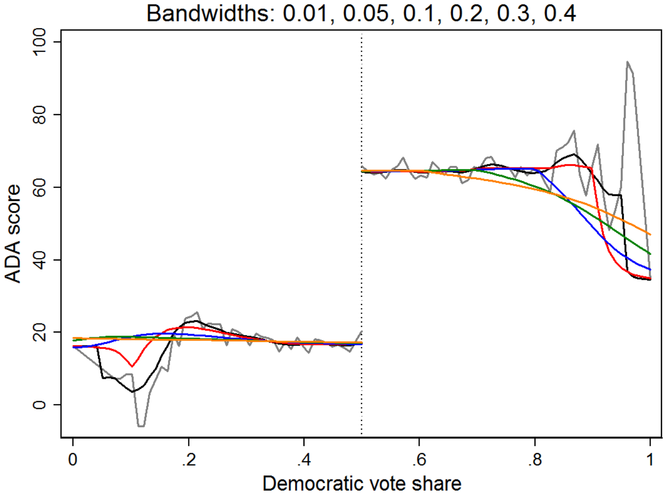
- We're interested in getting the treatment at $X = 0.5$

```
gen forat = 0.5 in 1
lpoly score demvoteshare if democrat == 0, nograph kernel(triangle) gen(sdem0)  ///
       at(forat) bwidth(0.1)
lpoly score demvoteshare if democrat == 1, nograph kernel(triangle) gen(sdem1)  ///
       at(forat) bwidth(0.1)
gen dif = sdem1 - sdem0
list sdem1 sdem0 dif in 1/1

     +-----------------------------------+
     |     sdem1        sdem0        dif |
     |-----------------------------------|
  1. | 64.395204   16.908821   47.48639 |
     +-----------------------------------+
```

Different windows

- What happens when we change the bandwidth?



Nonparametric

- With non-parametric methods in RDD came several methods to choose "optimal windows"
- In practical applications, you may want to check balance around that window
- Standard error of treatment effect can be bootstrapped
- Could add other variables to nonparametric methods but more complicated
- See Stata do file for examples using command `rdrobust`

Using rdrobust

```
. rdrobust score demvoteshare, c(0.5) all bwselect(IK)

Sharp RD Estimates using Local Polynomial Regression

        Cutoff c = .5 | Left of c  Right of c            Number of obs =      13577
----------------------+----------------------            Rho (h/b)     =      0.770
        Number of obs |      3535        3318            NN Matches    =          3
 Order Loc. Poly. (p) |         1           1            BW Type       =         IK
       Order Bias (q) |         2           2            Kernel Type   = Triangular
    BW Loc. Poly. (h) |     0.152       0.152
          BW Bias (b) |     0.197       0.197

--------------------------------------------------------------------------------------
                      | Loc. Poly.  Robust                                [Robust
                score |    Coef.   Std. Err.      z       P>|z|      95% Conf. Interval]
----------------------+---------------------------------------------------------------
         demvoteshare |   47.171      1.262    36.9043   0.000          44.1   49.047108
--------------------------------------------------------------------------------------

All Estimates. Outcome: score. Running Variable: demvoteshare.
--------------------------------------------------------------------------------------
               Method |    Coef.   Std. Err.      z       P>|z|      [95% Conf. Interval]
----------------------+---------------------------------------------------------------
         Conventional |   47.171     .98131    48.0692   0.000        45.247   49.093991
       Bias-Corrected |   46.574     .98131    47.4608   0.000         44.65   48.496943
               Robust |   46.574      1.262    36.9043   0.000          44.1   49.047108
--------------------------------------------------------------------------------------
```

Parametric or non-parametric?

- When would parametric or non-parametric or window size matter?
  - Small effect
  - Relationship between Y and X different away from cutoff
  - Functional form not well captured by polynomials (or other functional form)
- Parametric: can add random effects, clustering SEs,...
- But more important: What about if the outcome cannot be assumed to distribute normal?
- The curse and blessing of so many good RDD guides...
- With counts, for example, need to use Poisson or Negative Binomial models
- If conclusions are different, do worry

Marginal returns to medical care

- Big picture: is spending more money on health care worth it (in terms of health gained)?
- Actual research: is spending more money on low-weight newborns worth it in terms of mortality reductions? Compare marginal costs (dollars) to marginal benefits (mortality transformed into dollars).
- On jargon: In economics marginal = additional. So compare additional spending to additional benefit
- In IV language, the "marginal" patient is the "complier"
- RDD part used to estimate marginal benefits. Data from U.S Census birth 1983 to 2002
- Forcing variable is newborn weight. Cutoff point $c = 1,500$ grams (almost 3 lbs)

Data

- Did they use a fuzzy or sharp RDD?
- Related question: What is the "treatment"?
- What models did they use? And what was the outcome?

Estimating equation

- Their model is:

$$Y_i = \alpha_0 + \alpha_1 VLBW_i + \alpha_2 VLBW_i \times (g_i - 1500) + \alpha_3(1 - VLBW_i)(g_i - 1500) + \alpha_t + \alpha_s + \delta X_i' + \epsilon_i \quad (1)$$

- Change notation so $VLBW = T$ and $(g_i - 1500) = \tilde{X}$ and after doing some algebra the model is:

$$Y = \alpha_0 + \alpha_1 T + \alpha_3 \tilde{X} + (\alpha_2 - \alpha_3) T \times \tilde{X} + (\alpha_t + \alpha_s + \delta X') + \epsilon$$

- $(\alpha_t + \alpha_s + \delta X')$ are covariates

Covariates

- They compared means of covariates above and beyond cutoff point
- They found some differences (large sample) so they include covariates in the model
- They did a RDD-type analysis on covariates to see if they were "smooth" (no jump at VLBW cutoff)