

Regression Discontinuity Designs

Erik Gahner Larsen

Advanced applied statistics, 2015



Agenda

- ▶ Experiments and natural experiments
- ▶ Regression discontinuity designs
- ▶ How to do it in R and STATA

Last week

- ▶ Causal inference and observational studies
- ▶ Use of matching methods to create overlap and balance between treatment and control group
- ▶ Estimate average treatment effects and average treatment effects on the treated

Natural experiments

“Katrina was an act of nature that was close to what a political scientist would have liked to have done if he or she could - intervening and changing the political makeup of several large U.S. cities such as New Orleans, Houston, and Atlanta.” (Morton and Williams 2010, 53)

How are natural experiments different from traditional experiments?

“However, unlike true experiments, the data used in natural experiments come from “naturally” occurring phenomena - actually, in the social sciences, from phenomena that are often the product of social and political forces. Because the manipulation of treatment variables is not generally under the control of the analyst, natural experiments are, in fact, observational studies.”(Dunning 2012, 16)

“Although natural experiments offer significant advantages, they do not possess key benefits of actual experiments and hence require careful theoretical and statistical work to make valid inferences.” (Sekhon and Titiunik 2012, 53)

“Without an experiment, a natural experiment, **a discontinuity**, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive.” (Sekhon 2009, 503)

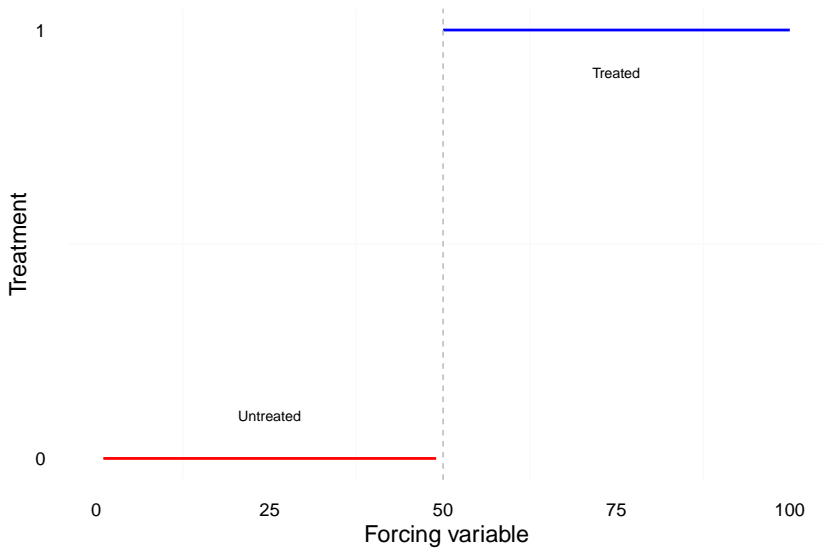
Do we need an overlap to make causal inferences?

- ▶ No, we can even utilize the lack of overlap
- ▶ A lack of overlap can sometimes be explained by a covariate that assign treatment to units: “Regression discontinuity is an approach for dealing with this extreme case of lack of overlap in which the assignment mechanism is clearly defined” (Gelman and Hill 2007, 212)
- ▶ So, today: Regression discontinuity designs (RDD)

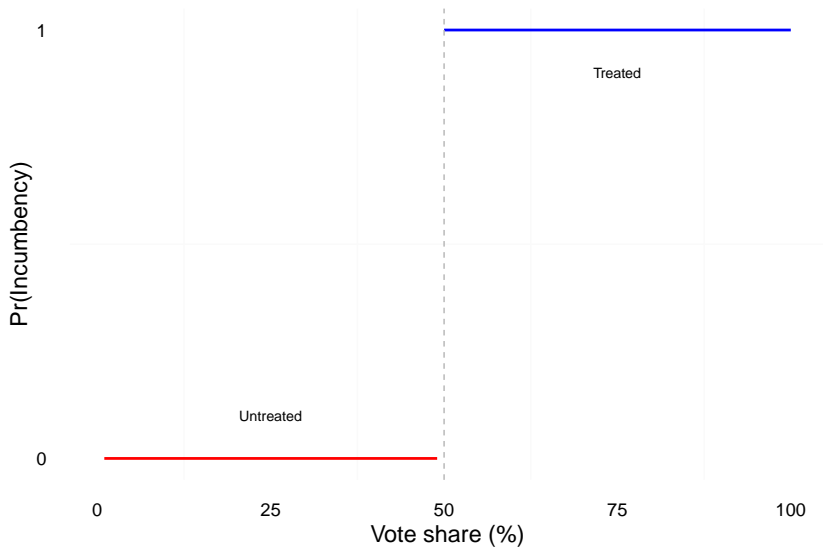
Regression discontinuity designs in a nutshell

- ▶ An identification strategy that creates treatment and control groups by utilizing instances where continuous variables are made into categorical variables due to clear cutoff values
- ▶ Units scoring above/below a cutoff point in assigned to the treatment/control group
- ▶ Treatment assignment is a deterministic function of a forcing variable (assignment by rule, not randomization)
- ▶ As-if randomness in the window around the cutoff point
- ▶ Treatment effect: Difference in outcome variable between units *just* above and *just* below the cutoff point

Forcing variable



Example: Vote share, two party system



“In a regression-discontinuity design, treatment assignment is determined by the value of a covariate, sometimes called a forcing variable, and there is a sharp discontinuity in the probability of receiving treatment at a particular threshold value of this covariate” (Dunning 2012, 64)

Sharp and fuzzy regression discontinuities

- ▶ Two types of RDD:
- ▶ Sharp (SRD): *All* units with a score above a cutoff is assigned to treatment
- ▶ Fuzzy (FRD): Propensity to be treated increases at cutoff point. Compliance with treatment is imperfect (not fully determined treatment assignment). Think of the ITT estimator.
- ▶ We will focus on sharp RDDs.

RDD in the Neyman-Rubin causal model

- ▶ We are (still) interested in potential outcomes to define causal effects
- ▶ For individual i , we have a potential outcome: Y_i
- ▶ Treatment: W_i
- ▶ Potential outcome given treatment status: $Y_i(W_i)$
- ▶ Two potential outcomes: $Y_i(1), Y_i(0)$
- ▶ Forcing variable: X_i (covariate, running variable)
- ▶ Cutoff value: c (discontinuity)
- ▶ Treatment assignment given by:

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

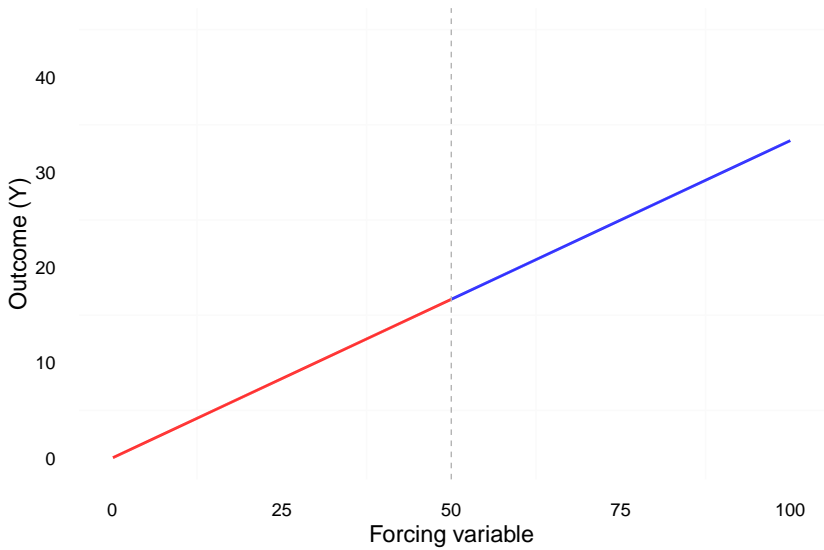
Average treatment effects

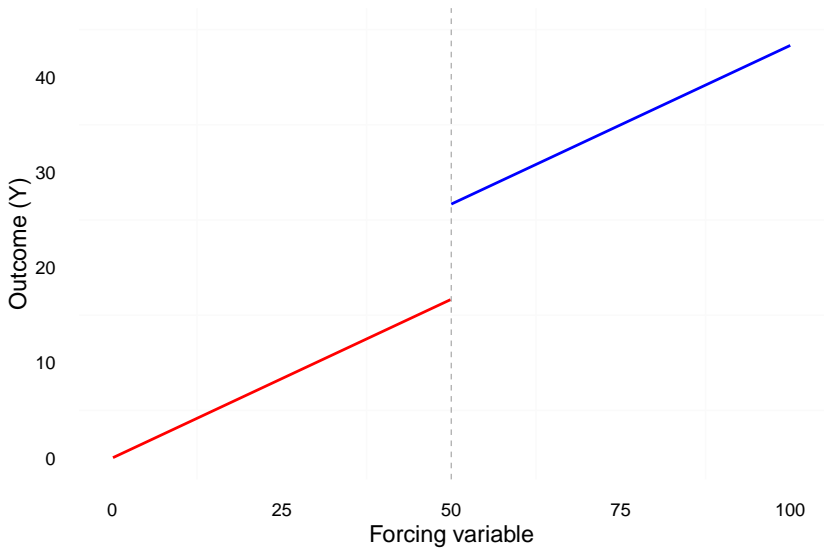
- ▶ FPCI: We only observe the outcome under treatment or control for each unit. We never observe both units for the same value of X_i .
- ▶ X_i is often correlated with Y_i . Simply comparing treatment and control does not provide causal estimates!
- ▶ We use the discontinuity in $E[Y_i|X_i]$ at the cutoff value $X_i = c$ to estimate the effect of W_i on Y_i .

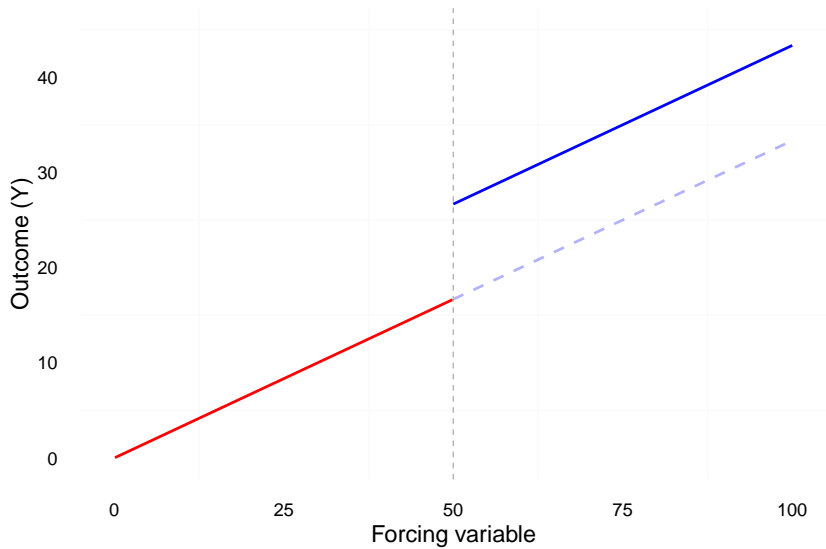
Average treatment effect

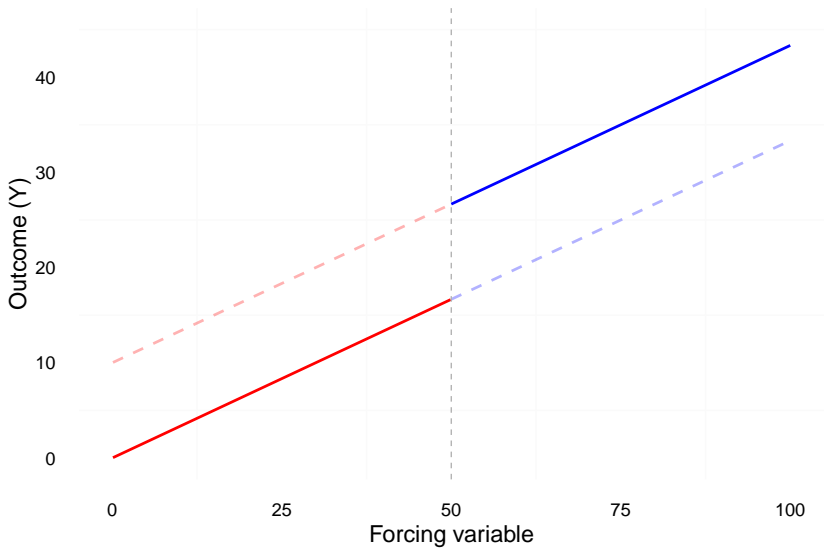
- ▶ Extrapolation
- ▶ The extrapolation is a counterfactual (potential outcome)
- ▶ Below cutoff point: only non-treated observations
- ▶ Above cutoff point: only treated observations
- ▶ Treatment effects are often local
- ▶ *Local* average treatment effect (LATE)
- ▶ Identified at the cutoff point

$$\tau = \lim_{x \downarrow c} E[Y(1)|X = c] - \lim_{x \uparrow c} E[Y(0)|X = c]$$









Bandwidth choice

- ▶ Bandwidth (h): the width of the window of observations used to estimate LATE
- ▶ Limit the sample:

$$c - h \leq X_i \leq c + h$$

- ▶ Lower bandwidth lowers the risk of bias but reduces the number of observations (Green et al. 2009; Robinson et al. 2009).

As-if randomness around cutoff point

- ▶ The procedure:

1. Decide on bandwidth
2. Treat the data as an experiment

- ▶ Problems:

1. Arbitrary bandwidth selection
2. Biased estimates if not random

- ▶ Suggestions:

1. Robustness tests (Imbens and Lemieux 2008, 622)
2. DID, matching

Examples of forcing variables

- ▶ The forcing variable can be: Vote share in an election, birthdate/age, income, population thresholds, score on a test, geographical distance, stock returns, classroom size, voter numbers
- ▶ Forcing variables are especially applicable in a rule based world, e.g. elections and policies (what we care about)
- ▶ What is a good forcing variable?

Robustness tests

- ▶ \uparrow Model specification choices \rightarrow \uparrow Researcher degrees of freedom
- ▶ Are the results sensitive to alternative model specifications?
- ▶ Functional form, nonparametric estimation, local-polynomial order
- ▶ Bandwidth
- ▶ Make all the robustness tests you can (appendix material)
- ▶ On a practical note: reproduce with different packages and statistical software

Covariates

- ▶ Covariates should not affect the estimate of the treatment effect
- ▶ Covariates should be random around the discontinuity
- ▶ Remember: as-if random
- ▶ Balance checks: Is there a discontinuity for other covariates at the cutoff point?
- ▶ What kind of balance checks? RDD estimates, and see the procedures introduced last week

Placebo tests

- ▶ Placebo cutoffs. Create other cutoff values and estimate the treatment effect
- ▶ Do we observe similar “jumps” as at the cutoff value?
- ▶ Similar LATE: Serious problem. Seriously.
- ▶ No LATE at placebo cutoffs: Less of a problem, but no guarantee (only absence of evidence)

Sharpness

- ▶ We want *sharp* discontinuities
- ▶ Estimates can be invalid if individuals can precisely manipulate the assignment (Lee and Lemieux 2010, 283)
- ▶ Lack of sharpness can result from:
 - ▶ measurement error in forcing variable
 - ▶ strategic behavior, manipulation of the running variable
 - ▶ Forcing variable might not be that “forcing”
 - ▶ FRD, estimate ITT, IV-estimation (next week)
 - ▶ Units might exploit the discontinuity (i.e. knowledge about cutoff among units can lead to biased effects)

Sorting

- ▶ Do units sort around the cutoff?
- ▶ The LATE will be biased if units (or other persons) strategically manipulate values to fall above or below cutoff
- ▶ Look at the distribution of X_i at the cutoff (do people sort to the right or left?)

RDD recommendations (Eggers et al. 2015)

1. Consider theoretical mechanisms that could produce sorting around the discontinuity.
2. Evaluate balance on pretreatment covariates and especially on the lagged outcome variable, focusing on the presence or absence of substantively large imbalances in characteristics that might be related to the mechanisms that could produce sorting. These tests should employ the same specifications as those employed to estimate the effects of interest, and these specifications should account for the potential relationship between the running variable and the outcome variable.
3. Present estimates at a number of alternative bandwidths and specifications.

Example: Thistlethwaite and Campbell (1960)

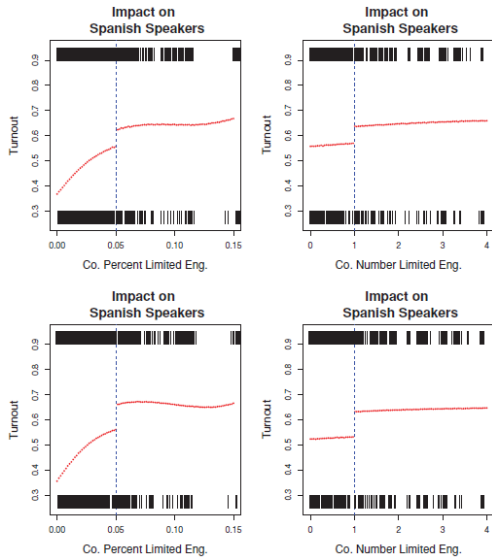
- ▶ First application of RDD
- ▶ The impact of merit awards (scholarships, W_i) on future academic outcomes (Y_i)
- ▶ Awards assigned based on SAT test scores (X_i)
- ▶ Score just above $c \rightarrow$ award
- ▶ What if we study the correlation between X_i and Y_i ?

Example: Thistlethwaite and Campbell (1960)

- ▶ Confoundedness: X_i is correlated with $Y_i(0)$ and $Y_i(1)$
- ▶ Higher score \rightarrow higher earnings, higher college grades
- ▶ Solution: Compare observations *just* above and *just* below the score cutoff point
- ▶ $E[Y_i(1)|X_i]$ and $E[Y_i(0)|X_i]$ are continuous functions of X_i at $X_i = c$
- ▶ Discontinuity in $E[Y_i|X_i]$ at $X_i = c$
- ▶ Treatment group ($W_i(1)$): Above cut-off and did receive award
- ▶ Control group ($W_i(0)$): Below cut-off and **did not** receive award
- ▶ Impact of award receipt for persons near the cutoff

Example: Spanish-language ballots

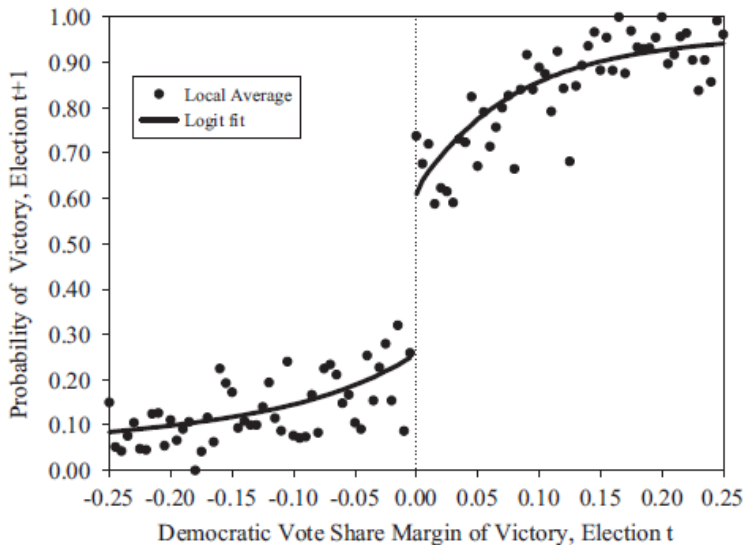
- ▶ Hopkins (2011): Translating into Votes: The Electoral Impacts of Spanish-Language Ballots
- ▶ Question: What is the impact of language assistance on voter turnout among Latino citizens?
- ▶ Why not simply study correlation between X_i and Y_i ?
- ▶ Forcing variable: U.S. counties with a language minority +10,000 citizens must provide Spanish-language ballots (Voting Rights Act)
- ▶ Forcing variable (X): Population size of the language minority in each county
- ▶ Cutoff value (c): 10,000. Treatment (W_i): presence of Spanish-language ballots.



Figur 1: Hopkins 2011

Example: Party incumbency advantage

- ▶ Lee (2008): Randomized experiments from non-random selection in U.S. House elections
- ▶ Question: How big is the party incumbency advantage?
- ▶ Design: RDD. Democratic vote share where democratic candidates almost won or lost the previous election.
- ▶ Data: Elections to the U.S. House of Representatives (1946 to 1998)



Figur 2: Lee 2008

- ▶ Grimmer et al. (2011): Gubernatorial and State House control is correlated with winning close elections for the U.S. House
- ▶ Caughey and Sekhon (2011): Covariate imbalances between near winners and losers.
- ▶ Eggers et al. (2015): No systematic evidence of sorting or imbalance around electoral thresholds.
- ▶ For more: See Hainmueller et al. (2015) and de la Cuesta and Imai (2016).

The validity of RDD

- ▶ Strong internal validity (Cook and Wong 2008; Dunning 2012; Berk 2010, 574)
- ▶ Less external validity (Imbens and Lemieux 2008, 622)
- ▶ Less conclusion validity
- ▶ Lower power
- ▶ Requires large N to estimate same effect

RDD in R and STATA

- ▶ Multiple packages for R (`rddtools`, `rdrobust` etc.)
- ▶ A lot of options (kernel density, bandwidth selection, local-polynomial order etc.)
- ▶ We will focus on the basics

RDD in R (rddtools)

```
# Load rddtools package  
library("rddtools")  
  
# Use data from Lee (2008)  
data(house)  
  
# Set outcome, forcing and cutoff variable  
house_rdd <- rdd_data(y=y, x=x, cutpoint=0, data=house)  
  
# Estimate RDD  
reg_para <- rdd_reg_lm(rdd_object=house_rdd)  
  
# Print results  
print(reg_para)
```

RDD in R (rddtools)

```
## Loading required package: AER
```

```
## Loading required package: car
```

```
## Loading required package: lmtest
```

```
## Loading required package: zoo
```

```
##
```

```
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
```

```
##
```

```
##      as.Date, as.Date.numeric
```

```
## Loading required package: sandwich
```

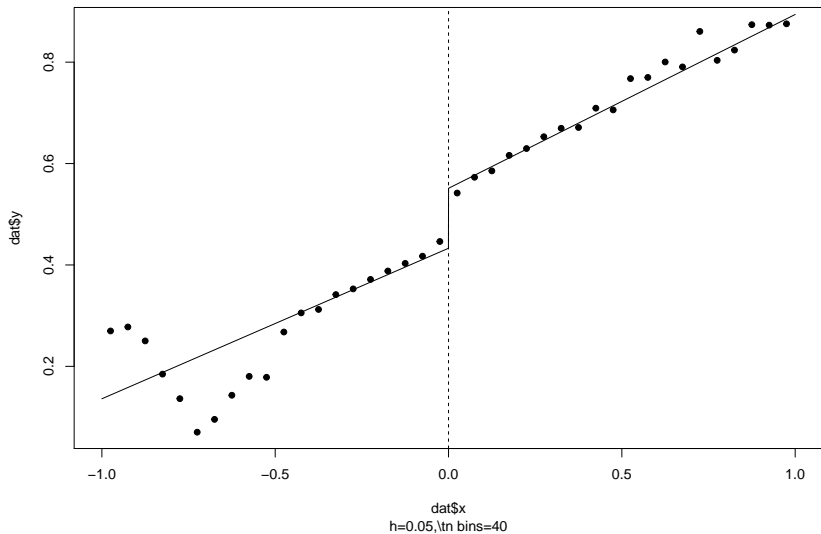
```
## Loading required package: survival
```

```
## Loading required package: np
```

```
## Nonparametric Kernel Methods for Mixed Datatypes (version 0.6
```

RDD in R (rddtools)

```
plot(reg_para)
```



RDD in R (rddtools)

```
# Restrict sample to bandwidth area
```

```
bw_ik <- rdd_bw_ik(house_rdd)
```

```
reg_para_ik <- rdd_reg_lm(rdd_object=house_rdd, bw=bw_ik)
```

```
reg_para_ik
```

```
## ### RDD regression: parametric ###
```

```
## Polynomial order: 1
```

```
## Slopes: separate
```

```
## Bandwidth: 0.2938561
```

```
## Number of obs: 3200 (left: 1594, right: 1606)
```

```
##
```

```
## Coefficient:
```

```
## Estimate Std. Error t value Pr(>|t|)
```

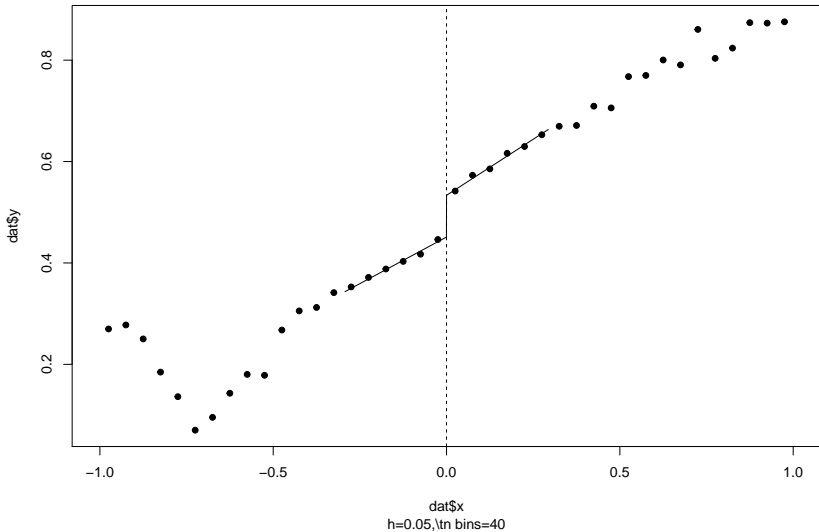
```
## D 0.0823378 0.0080236 10.262 < 2.2e-16 ***
```

```
## ---
```

```
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


RDD in R (rddtools)

```
plot(reg_para_ik)
```



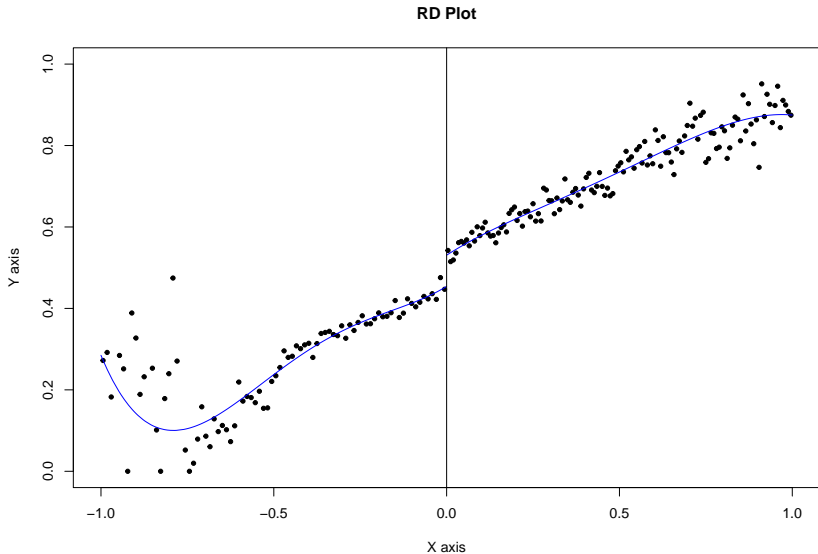
RDD in R (rdrobust)

```
# Load rdrobust package  
library("rdrobust")  
  
# Estimate effect  
rdrobust(house$y, house$x)
```

```
## Call:  
## rdrobust(y = house$y, x = house$x)  
##  
## Summary:  
##  
## Number of Obs 6558  
## NN Matches    3  
## BW Type       mserd  
## Kernel Type   Triangular  
##  
##              Left    Right  
## Number of Obs 2740   3818  
## Order Loc Poly (p) 1     1  
## Order Bias (q)   2     2
```

RDD in R (rdrobust)

```
rdplot(house$y, house$x)
```



RDD in STATA (rdrobust)

- ▶ `rdrobust` is also available in STATA
- ▶ Can be installed via `ssc install` but the easy way is just to go through:

```
. findit rdrobust
```

- ▶ RDD estimate:

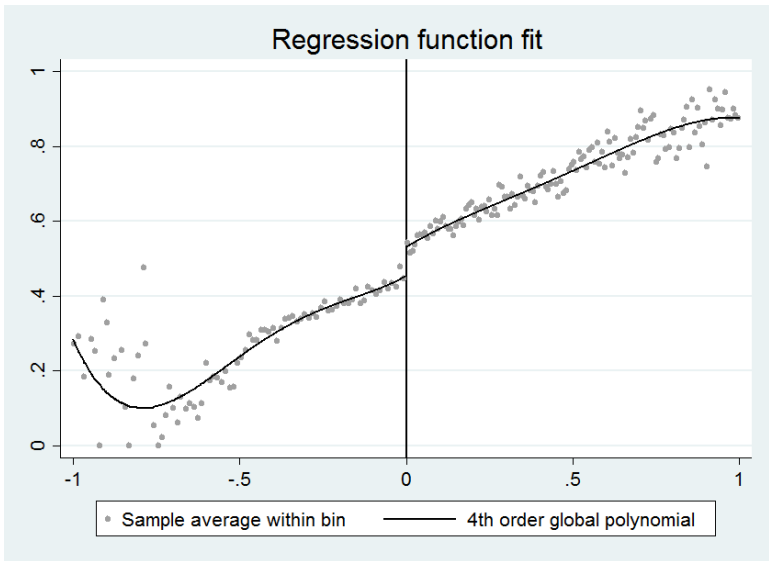
```
. rdrobust y x, c(0) bwselect(IK) all
```

- ▶ RDD plot:

```
. rdplot y x, c(0) bwselect(IK)
```

WARNING!

STATA plot on next slide



Figur 3: RDD Stata plot

Conclusion

- ▶ Three components: Outcome, forcing variable, cutoff
- ▶ Validity depends upon the quality of the design and data
- ▶ Do graphs. Do a lot of graphs.
- ▶ Further reading: Skovron and Titiunik (2015): A Practical Guide to Regression Discontinuity Designs in Political Science (see also the supplementary material for a list of examples on RDD studies)

Schedule

- ▶ Next week: IV-estimation
- ▶ Tomorrow: Matching in R and STATA
- ▶ Mandatory Assignment 4. *Good luck.*