Applied Microeconometrics

Lecture 6b: RDD

Manuel Bagues

August 2022

Lecture Slides

Today: Regression discontinuity design

- Often certain rules are used to assign individuals to “treatments" which can be exploited for estimating causal effects
- The most notorious case are threshold rules that are based on some ex-ante variable, typically, correlated with the expected effectiveness of the treatment
- Example: Need-based Grant Programs for low income students (Fack and Grenet 2015)
  - there are discontinuities in the grant eligibility formula
- This ex-ante variable is called the running variable (or also, forcing or assignment variable).
- This threshold divides individuals into “treated" and “not treated". The idea in RDD design is to exploit the randomness of this threshold: we compare individuals just below (non-treated) and just above (treated).

Today: Regression discontinuity design

- Often certain rules are used to assign individuals to “treatments" which can be exploited for estimating causal effects
- The most notorious case are threshold rules that are based on some ex-ante variable, typically, correlated with the expected effectiveness of the treatment
- Example: Need-based Grant Programs for low income students (Fack and Grenet 2015)
  - there are discontinuities in the grant eligibility formula
- This ex-ante variable is called the running variable (or also, forcing or assignment variable).
- This threshold divides individuals into “treated" and “not treated". The idea in RDD design is to exploit the randomness of this threshold: we compare individuals just below (non-treated) and just above (treated).

Main requirements

Three conditions for consistent and precise RDD estimates:

1. Variation in the treatment status near the threshold is as good as random
   - If agents can anticipate the threshold they may try to manipulate the running variable. RD Design is invalid if individuals can precisely manipulate the assignment variable x_i in order to get (or to avoid) the treatment.
2. No other treatments at the same threshold
3. Sufficient number of observations around the threshold

- Selected threshold (c) of the running variable (x_i) divides individuals into “treated" and “not treated"

$$D_i = \begin{cases} 1 \text{ if } x_i > c \\ 0 \text{ if } x_i \leq c \end{cases}$$

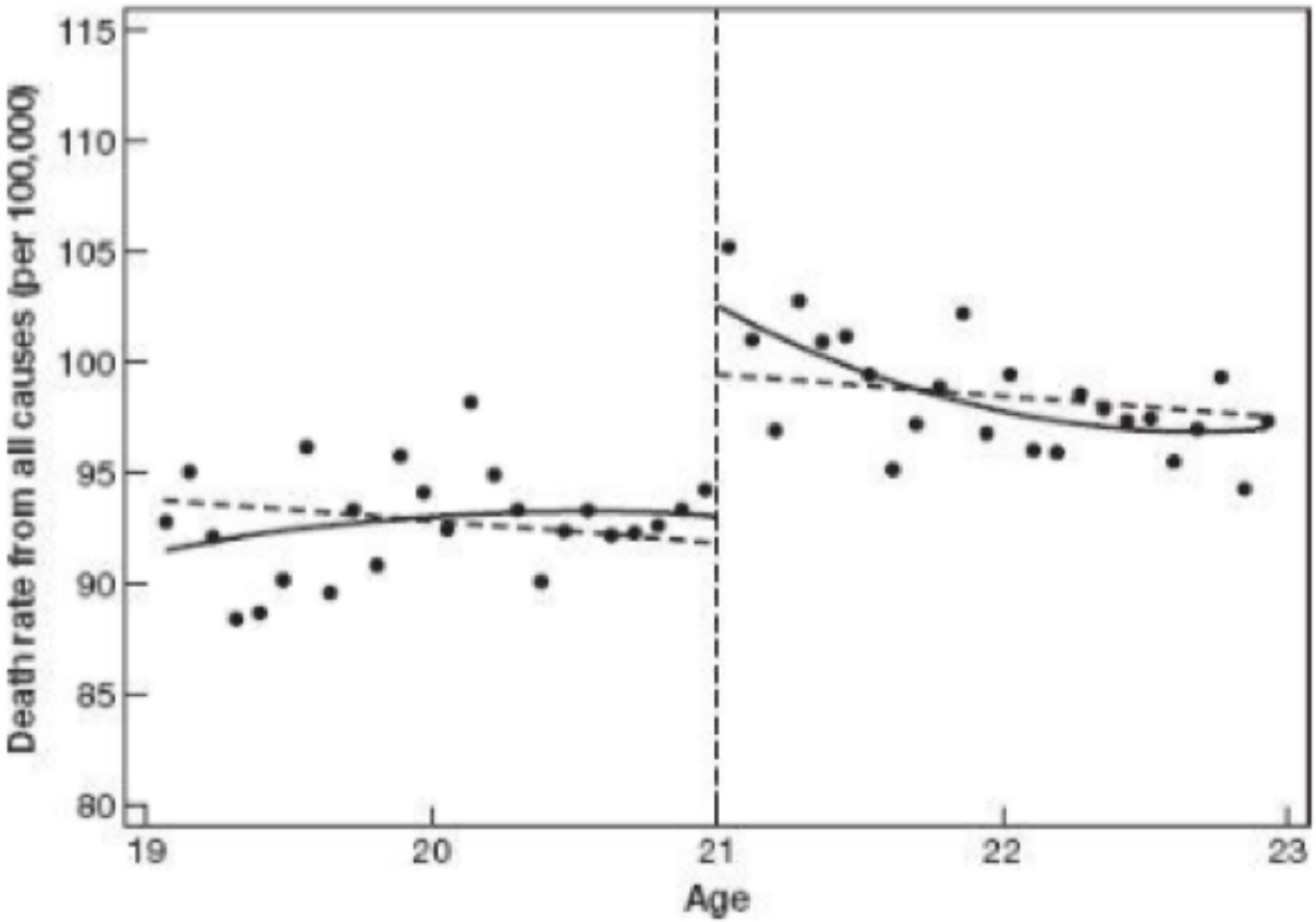
- The idea in RDD design is to exploit that being just below (non-treated) or just above (treated) the threshold is as good as random
- The RDD estimates the local average treatment effect for units around the threshold

Example: Effect of the Minimum Legal Drinking Age (MLDA) on death rates

Carpenter and Dobkin (2009)

1. outcome variable y_i: death rate
2. treatment D_i: legal drinking status
3. running variable x_i: age
4. cutoff: MLDA transforms 21-year-olds from underage minors to legal alcohol consumers.

RDD estimation: Visual example



Note: we need an out-of-range prediction to get a counterfactual!

Example: Causal Effect of Incumbency

Lee, David (2008), 'Randomized experiments from non-random selection in U.S. House elections', Journal of Econometrics 142, 675Ð697.

- Does a Democratic candidate for a seat in the U.S. House of Representatives has an advantage if his party won the seat last time?
- Exploits the fact that and (previous) election winner is determined by rule: $D_i = 1\left(x_i \geq c\right)$ where x_i is the vote share margin of victory (the difference between Demogratic and Republican votes shares).
- Because D_i is a deterministic function of x_i, there should be no confounding factors other than x_i.

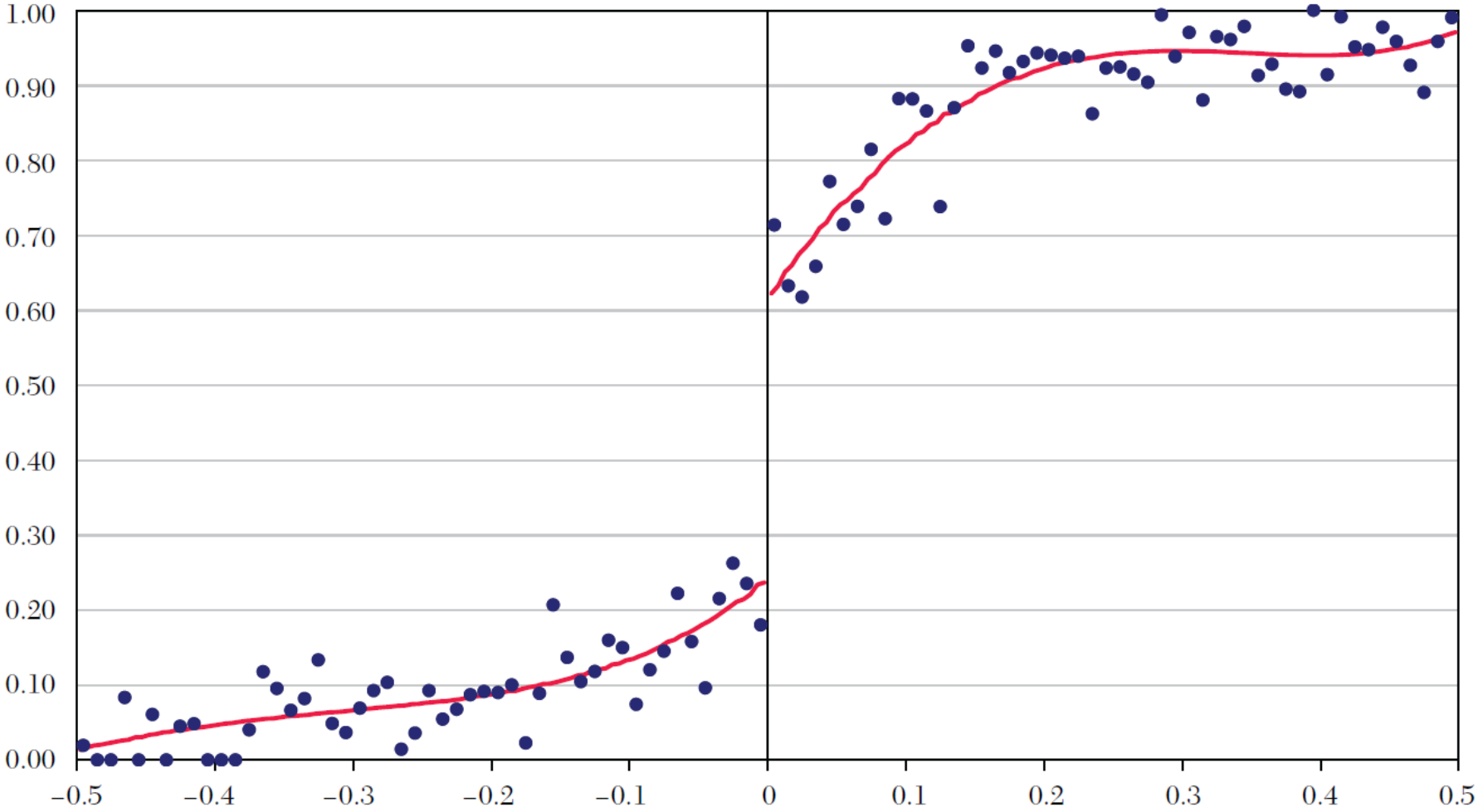


Figure 10. Winning the Next Election, Bandwidth of 0.01 (100 bins)

Other examples:

- Examples: (Treatment / Outcome variable / Running variable)
  - Merit award / Future academic outcomes / Test score
  - Attending a certain university / Future labor outcomes / Test score in entry exams
  - Length of maternity leave / Children cognitive ability / Date of policy change
  - R&D public subsidies / firms R&D investments / Project quality score
  - Survivor Benefit Program / Widow(er)s labor supply / Date of policy change
  - Legal status for immigrants / Crime rate / Application time

Road map for a RDD

- An RD paper typically starts showing that the set up is adequate for RD
- Main results should be visible with an RD plot (as the ones you saw earlier)
  - Stata command: *rdplot*
- The RD plot provides merely suggestive evidence
- The main estimates are then reported in a table.
- Several choices need to be made:
  1. How to control for the running variable
  2. Bandwidth
  3. Kernel

Road map for a RDD

1. Validity of the RDD set up
2. Graphical presentation (RD plots)
3. Estimation

1. Testing the validity of the RDD set up

1. Is this threshold relevant for anything else? (make sure you know the institutional context!)
   - example 1: impact of maternity leave extension on children's cognitive ability - cutoff date: Jan 1st
   - example 2: population thresholds (Eggers et al. 2018)
2. Predetermined characteristics should not exhibit discontinuities at the cut-off
3. Density function of the running variable should be continuous at the cut-off
   - McCrary test (`DCdensity`)
   - Cattaneo, Jansson and Ma (2017b): `rddensity`

Note: If there is some manipulation around the threshold you may still perhaps do a 'donut' RDD (e.g. Barreca et al. 2011)

2. RD Plots

RD plot:

- Divide the assignment variable into an (optimal) number of bins
- Plot the average value of the outcome variable for each bin
- RD plots typically include also a polynomial regression model (unrelated to the actual estimation)

Advantages:

- Can we see a jump at the threshold?
- Do we see jumps at other points?

3. Estimation

We estimate the following equation:

$$Y_i = \beta I[x_i > c] + f(x_i) + \epsilon_i$$

- $I[\cdot]$ takes value one if the running variable is above cutoff c
- $f(x_i)$ is a flexible function that is estimated separately above and below the threshold

Several things have to be decided:

- Functional form assumptions $[f(x_i)]$
- Bandwidth
- Kernel

Control for the running variable

- Traditionally:
  - Global polynomials of higher order
- Gelman and Imbens (2014) critique:
  - 'Why high order polynomials should not be used in regression discontinuity designs'
  - Instead use 1st or 2nd order
- Example:
  - 'Yuyu Chen, Avraham Ebenstein, Michael Greenstone and Hongbin Li (PNAS 2013) 'Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai river policy'

The New York Times

Asia Pacific

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HEALTH | SPORTS | OPINION

AFRICA | AMERICAS | **ASIA PACIFIC** | EUROPE | MIDDLE EAST

Pollution Leads to Drop in Life Span in Northern China, Research Finds

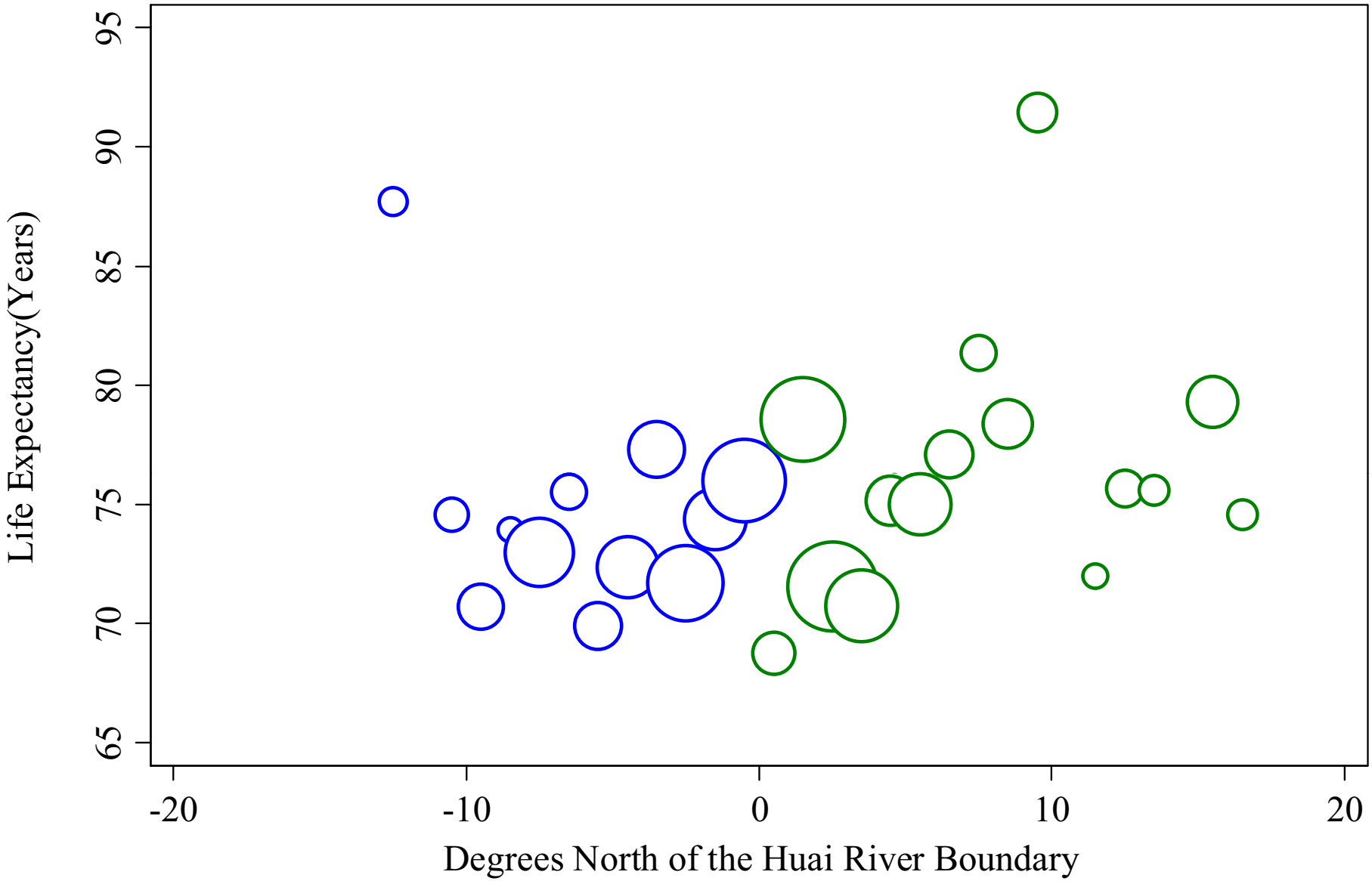
Proceedings of the National Academy of Sciences of the United States of America

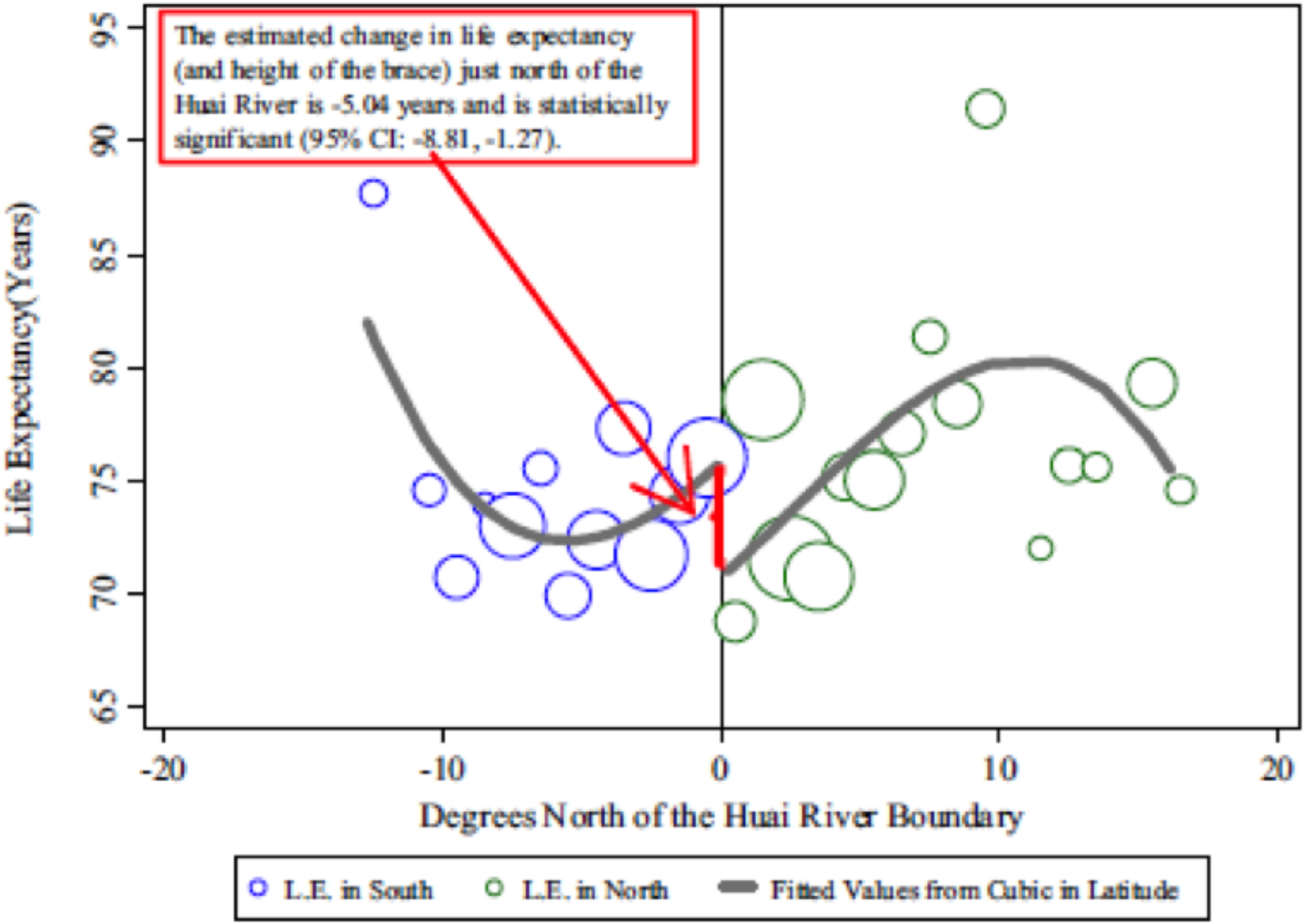
Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai River policy

Yuyu Chen[a,1], Avraham Ebenstein[b,1], Michael Greenstone[c,d,1,2], and Hongbin Li[e,1]

This paper's findings suggest that an arbitrary Chinese policy that greatly increases total suspended particulates (TSPs) air pollution is causing the 500 million residents of Northern China to lose more than 2.5 billion life years of life expectancy. The quasi-experimental empirical approach is based on China's Huai River policy, which provided free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. Using a regression discontinuity design based on distance from the Huai River, we find that ambient concentrations of TSPs are about 184 $\mu g/m^3$ [95% confidence interval (CI): 61, 307] or 55% higher in the north. Further, the results indicate that life expectancies are about 5.5 y (95% CI: 0.8, 10.2) lower in the north owing to an increased incidence of cardiorespiratory mortality. More generally, the analysis suggests that long-term exposure to an additional 100 $\mu g/m^3$ of TSPs is associated with a reduction in life expectancy at birth of about 3.0 y (95% CI: 0.4, 5.6).

Fig. 1. The cities shown are the locations of the Disease Surveillance Points. Cities north of the solid line were covered by the home heating policy.


Life Expectancy(Years)
65
70
75
80
85
90
95
-20
-10
0
10
20
Degrees North of the Huai River Boundary


The estimated change in life expectancy (and height of the brace) just north of the Huai River is -5.04 years and is statistically significant (95% CI: -8.81, -1.27).
Life Expectancy(Years)
65
70
75
80
85
90
95
-20
-10
0
10
20
Degrees North of the Huai River Boundary
L.E. in South
L.E. in North
Fitted Values from Cubic in Latitude

Different bandwidths (appendix table S9)

Table S9

Robustness checks of choice of functional form for latitude

	Linear & Controls	Quadratic & Controls	Cubic & Controls	Quartic & Controls	Quintic & Controls
	(1)	(2)	(3)	(4)	(5)
Panel 1: Impact of "North" on the Listed Variable, Ordinary Least Squares					
Life Expectancy (years)	-1.62	-1.29	-5.52**	-5.67**	-5.43*
	(1.66)	(1.68)	(2.39)	(2.36)	(2.94)
	[0.101]	[0.6]	[0.001]	[0.737]	[0.984]
	{757.1}	{758}	{746.8}	{748.2}	{750.2}

- Nowdays global polynomials are rarely used
- State of the art:
  - Local linear regression within a given window (bandwidth) of width h around the cutoff point, and a certain kernel

Bandwidth

- How to choose the bandwidth?
- Trade-off: the closer you get the better it is for identification, but the less data you have...
- Two steps to follow:
  1. Use existent statistical algorithms for selecting the “optimal bandwidth" (e.g.: Imbens-Kalyanaraman 2011, Calonico, Cattaneo and Titiunik 2014).
  2. Explore the robustness of results to different bandwidths
- In Stata, you may install `rdrobust` package to estimate RDD.
  - *net install rdrobust, from(https://sites.google.com/site/rdpackages/rdrobust/stata) replace*

Kernel

- Choice of kernel
  - rectangular vs. triangular
- In practice, usually it does not affect results

Control variables

- Similar to an RCT, control variables can be used to increase precision
- Standard errors become lower but the estimate is not expected to change significantly (otherwise it would mean that RDD is not valid)
- In order to deal with time invariant unobserved heterogeneity, you might also want to use the outcome variable in differences (known both as as "regression discontinuity-in-differences" or "difference in regression discontinuity")

Robustness

Make sure that your results are robust to:

- different functional form assumptions
- different bandwidths

Fuzzy Regression Discontinuity Design

- So far we have considered a sharp RDD, where the treatment status (D_i) is deterministic and discontinuous function of the running variable (x_i):
- When the increase in the probability of receiving the treatment does not increase from 0 to 1 when you cross the cutoff we have a fuzzy RDD.

$$P(D_i = 1) = \begin{cases} \overline{D} \text{ if } x_i \downarrow x_0 \\ \underline{D} \text{ if } x_i \uparrow x_0 \end{cases}$$

where $0 \leq \underline{D} < \overline{D} \leq 1$

- In sum, fuzzy RDD is an RDD without full compliance
- How do we deal with the lack of full compliance $\rightarrow$ combine RDD with instrumental variable

- Impact of the treatment on the outcome:

$$Y_i = \beta D_i + f(x_i) + \epsilon_i$$

- We cannot directly estimate this equation because the treatment is potentially endogenous: among people above or below the threshold, there is selection into the treatment
- The solution is to instrument D_i using $I[x_i > c]$ as an instrument, controlling for the running variable in a flexible way.
- Just one dychotomic instrument $\rightarrow$ Wald estimator

- First stage: impact of being above the threshold on the treatment:

$$D_i = \pi_1 I[x_i > c] + f(x_i) + \theta_1 i$$

- Reduced form: impact of being above the threshold on the outcome:

$$Y_i = \pi_2 I[x_i > c] + f(x_i) + \theta_2 i$$

- Wald estimator:

$$\beta_{IV} = \frac{\pi_2}{\pi_1}$$

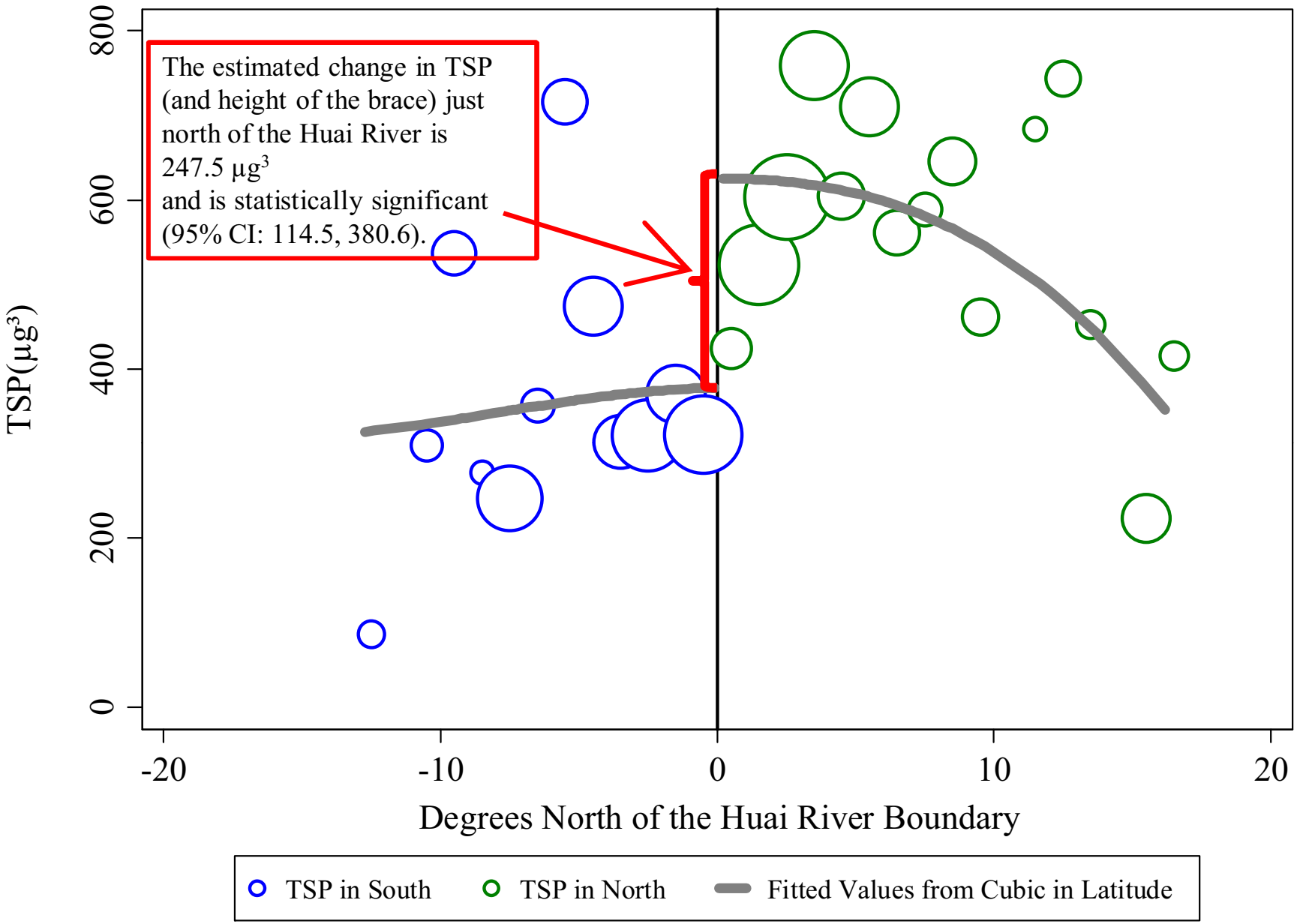
- In addition to the standard RD assumptions, this IV estimation requires an additional assumption:
  - Exclusion restriction: being above the threshold only affects the outcome variable through its impact on the treattment
- Local estimate: the fuzzy RDD estimate only identifies the impact of the treatment on compliers located around the threshold.

Example of Fuzzy RDD

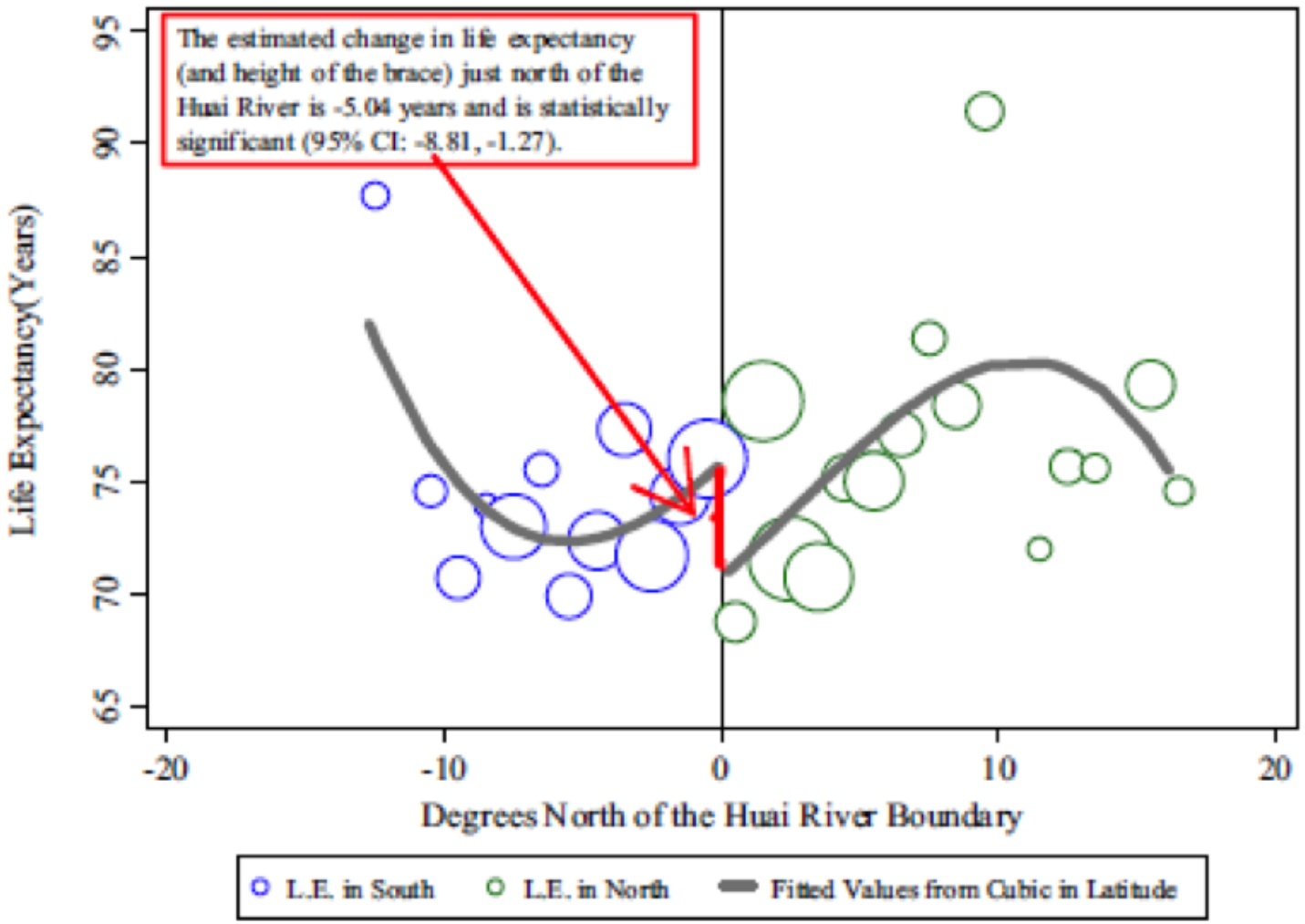
Chen et al. 2013,'Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai river policy'

- Main variables:
  - Outcome variable: life expectancy (in years)
  - Treatment: level of pollution (in 100 μg^3 of total suspended particles TSPs)
  - Running variable: distance to the Huai river (in degrees of latitude)
  - Instrument: being to the north of the Huai river
- For simplicity, let us give face value to the authors' estimates

First stage


The estimated change in TSP (and height of the brace) just north of the Huai River is 247.5 μg³ and is statistically significant (95% CI: 114.5, 380.6).
TSP(μg³)
0
200
400
600
800
-20
-10
0
10
20
Degrees North of the Huai River Boundary
TSP in South
TSP in North
Fitted Values from Cubic in Latitude

Reduced form



- Fuzzy RD estimate:

$$\beta_{IV} = \frac{\pi_2}{\pi_1} = \frac{-5.04}{2.47} = -2.04$$

- A 100 μg^3 increase in TSPs decreases life expectancy by around 2 years.
- Question 1: who are the compliers in this context?
- Question 2: can you think about some potential violation of the exclusion restriction?

Pseudo-RDDs

- Some authors have used close elections to estimate the impact of politician characteristics (e.g. gender, education, political party)
- Example: impact of Islamist Party majors on the education of women in Turkey
- Close elections allow to:
  - control for the characteristics of constituencies
  - control for the popularity of politicians among voters
- But these politicians are likely to differ in many other dimensions
  - depending on the underlying differences across groups
  - depending on the model of selection (e,g voters discrimination against certain candidates)
- Difficult to interpret unless we model the selection

RDD vs. other methods

RDD has several potential advantages and drawbacks relative to other "natural experiment" strategies such as difference-in-differences or instrumental variables:

- Main advantages:
  - Main identifying assumptions are testable
  - More transparent
- Possible drawbacks:
  - Often underpowered
  - Estimates the effect for a very local population
    - Ex. 1: worst student in university A vs. best in university B
    - Ex. 2: population threshold - open vs. closed lists in Spain)
    - Ex. 3: medical treatment for underweight babies
    - Always ask yourself why is there a threshold precisely there!

Summary

- Useful method to analyze the impact of treatment when the assignment varies discontinuously due to some rules! (test score, electoral results, income threshold, etc.)
- Only feasible when the number of observations around the threshold is sufficiently large
- Conditions for consistency:
  - No precise manipulation at the threshold. This may not hold if agents anticipate the threshold
  - Make sure there are no other treatments at the same threshold
- Usually graphical analysis is already very revealing
- Note that RDD provides information about the impact of the treatment only for individuals around the threshold. Depending on the context this might be a good thing or not.
  - Sometimes this might be different from the effect for other individuals (e.g.: entry to university)