Lecture VII: Regression Discontinuity Designs

Stanislao Maldonado

Universidad del Rosario
stanislao.maldonado@urosario.edu.co

Impact Evaluation
Universidad del Rosario
March 14th, 2017

1. Motivation

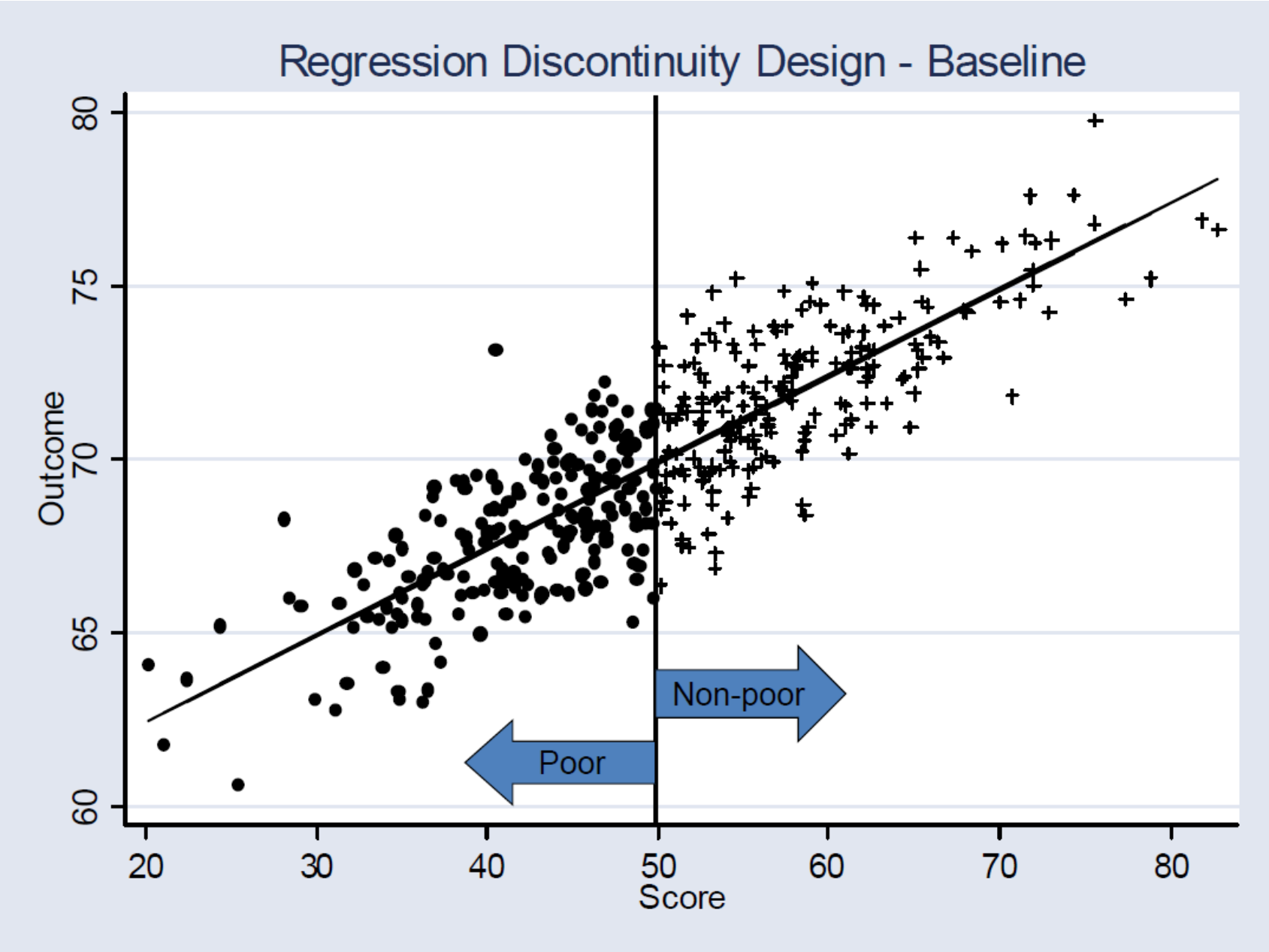
- Goal: to approximate a experimental design using observational data
- In absence of random assignment, causal effects can be estimated exploiting characteristics of the assignment rule
  - Example: fellowships programs, poverty programs based on scores, etc.
- LATE can be estimated at the discontinuity that determines which individuals are assigned to treatment and to control
- Examples: Angrist and Lavy (1999), Van der Klauuw (2002), Di Nardo and Lee (2005), among others

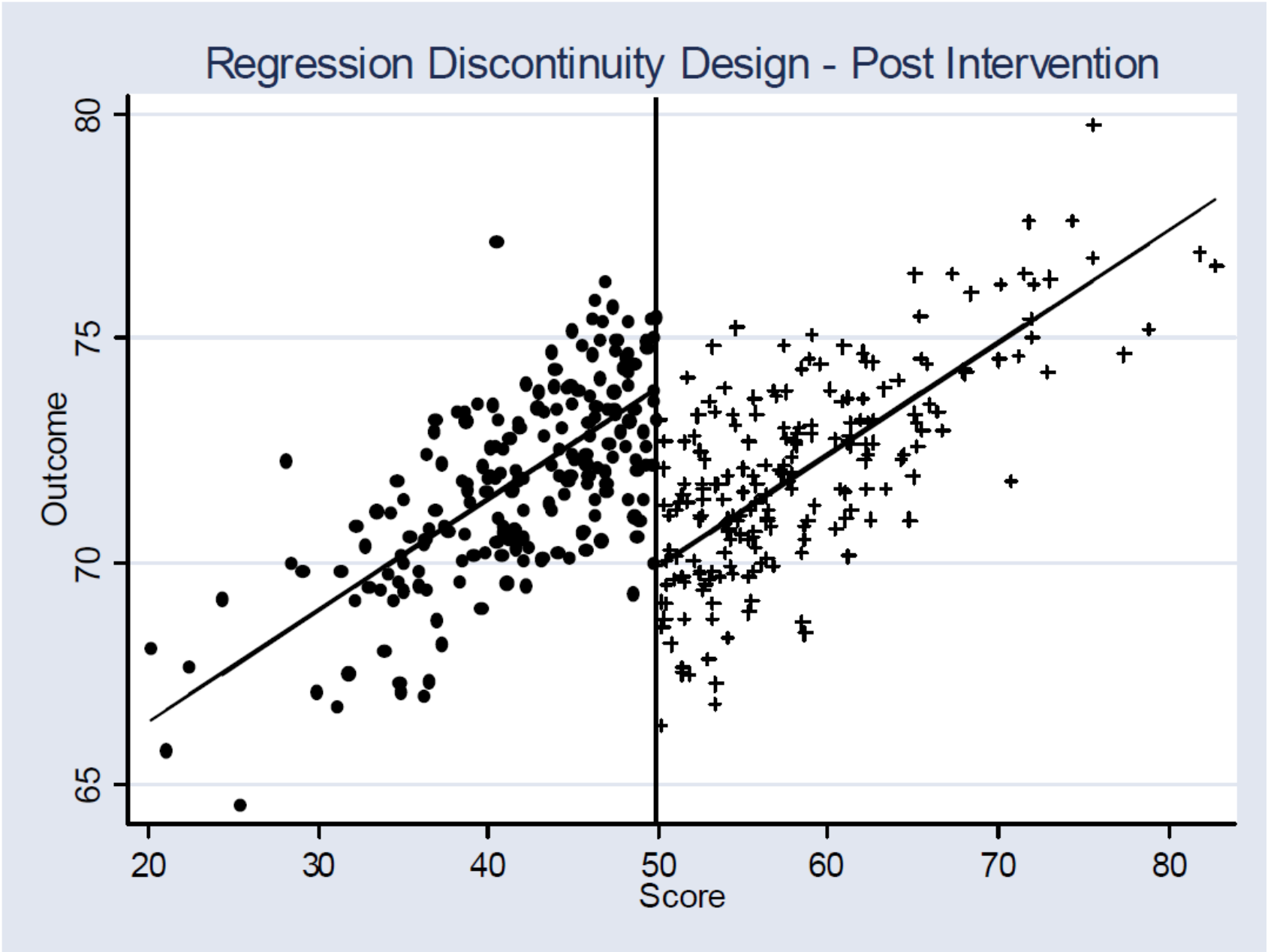
- When to use RDD?
  - The treated/non-treated can be ordered along a quantifiable dimension
  - This dimension can be used to compute a well-defined index or parameter
  - The index/parameter has a cut-off point for elegibility
  - The index value is what drives the assignment of a unit to the treatment (or to non-treatment)
- Intuitive explanation:
  - The treated units just above the cut-off point are very similar to the control units just below the cut-off point
  - We compare outcomes for units just above and below the cutoff point
  - This estimates the effect of the treatment for units AT the cut-off point, and may not be generalizable

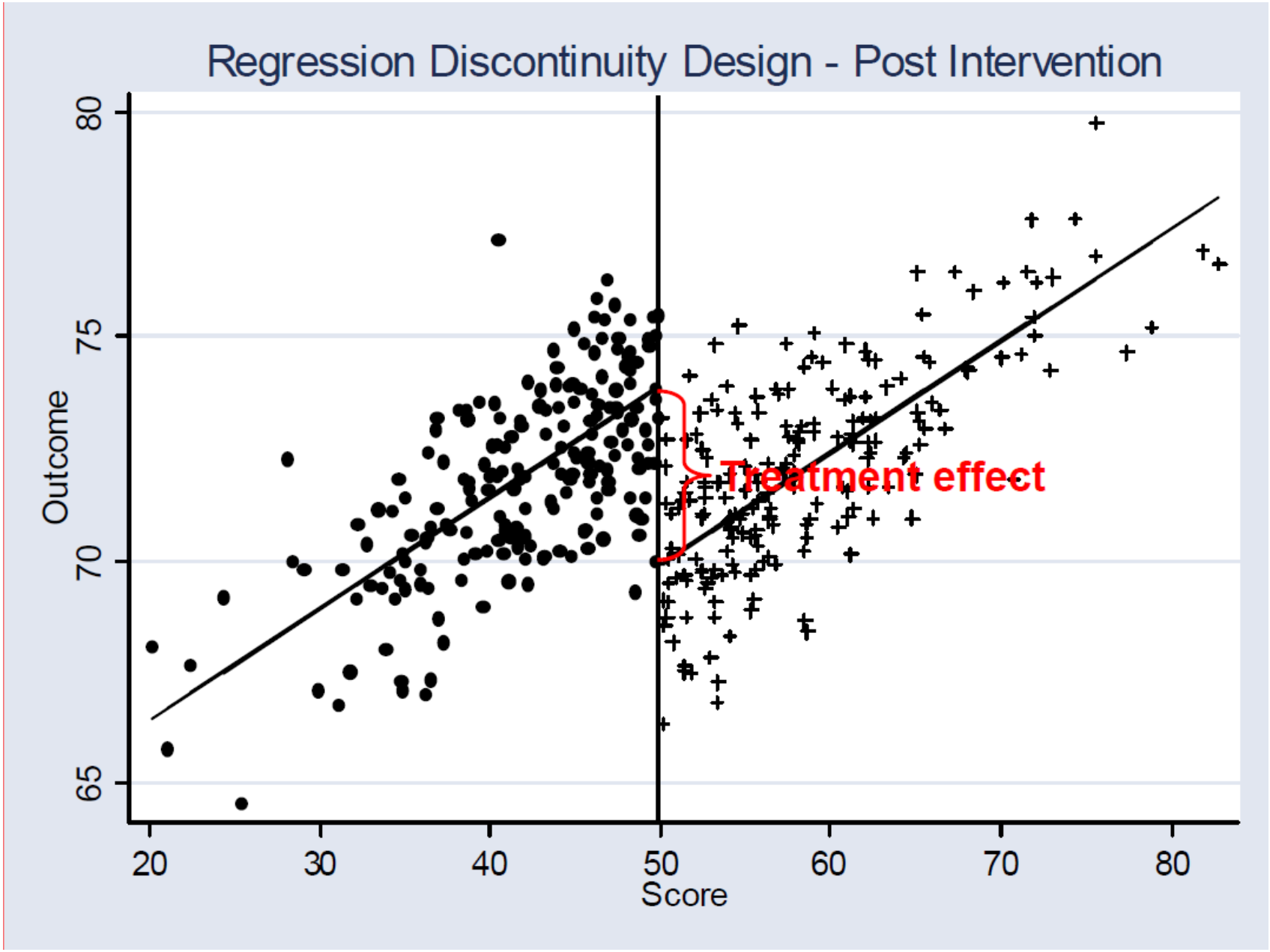
- Indexes are common in social programs:
  - Anti-poverty programs: targeted to households below a given poverty index
  - Pension programs: targeted to population above a certain age
  - Scholarships: targeted to students with high scores on standardized test
  - CDD Programs: awarded to NGOs that achieve highest scores

Example: effect of cash transfer on consumption

- Goal: Target transfer to poorest households
- Method:
  - Construct poverty index from 1 to 100 with pre-intervention characteristics
  - Households with a score $\leq$50 are poor
  - Households with a score $>$50 are non-poor
- Implementation: Cash transfer to poor households
- Evaluation:
  - Measure outcomes (i.e. consumption, school attendance rates) before and after transfer, comparing households just above and below the cut-off point


Regression Discontinuity Design - Baseline
Outcome
60
65
70
75
80
20
30
40
50
60
70
80
Score
Non-poor
Poor


Regression Discontinuity Design - Post Intervention
Outcome
80
75
70
65
20
30
40
50
60
70
80
Score


Regression Discontinuity Design - Post Intervention
Outcome
80
75
70
65
20
30
40
50
60
70
80
Score
Treatment effect

2. Treatment effects in RDD

- We already know:
  - $Y_i(1)$ and $Y_i(0)$ are the potential outcomes
  - $\beta_i = Y_i(1) - Y_i(0)$
  - D_i is the treatment status
- If assignment is determined by randomization and full compliance with treatment:

$$Y_i(1), Y_i(0) \perp\!\!\!\perp D_i \tag{1}$$

- The mean impact:

$$\mathbb{E}(\beta_i) = \mathbb{E}(Y_i(1)/D_i = 1) - \mathbb{E}(Y_i(0)/D_i = 0) \tag{2}$$

- RDD arises when:
  - Treatment status depends on an **observable** unit characteristic X
  - There exist a **known point** c in the support of X where the probability of participation changes discontinuously
- Let c be the discontinuity point, then:

$$Pr(D_i = 1/c^+) \neq Pr(D_i = 1/c^-) \tag{3}$$

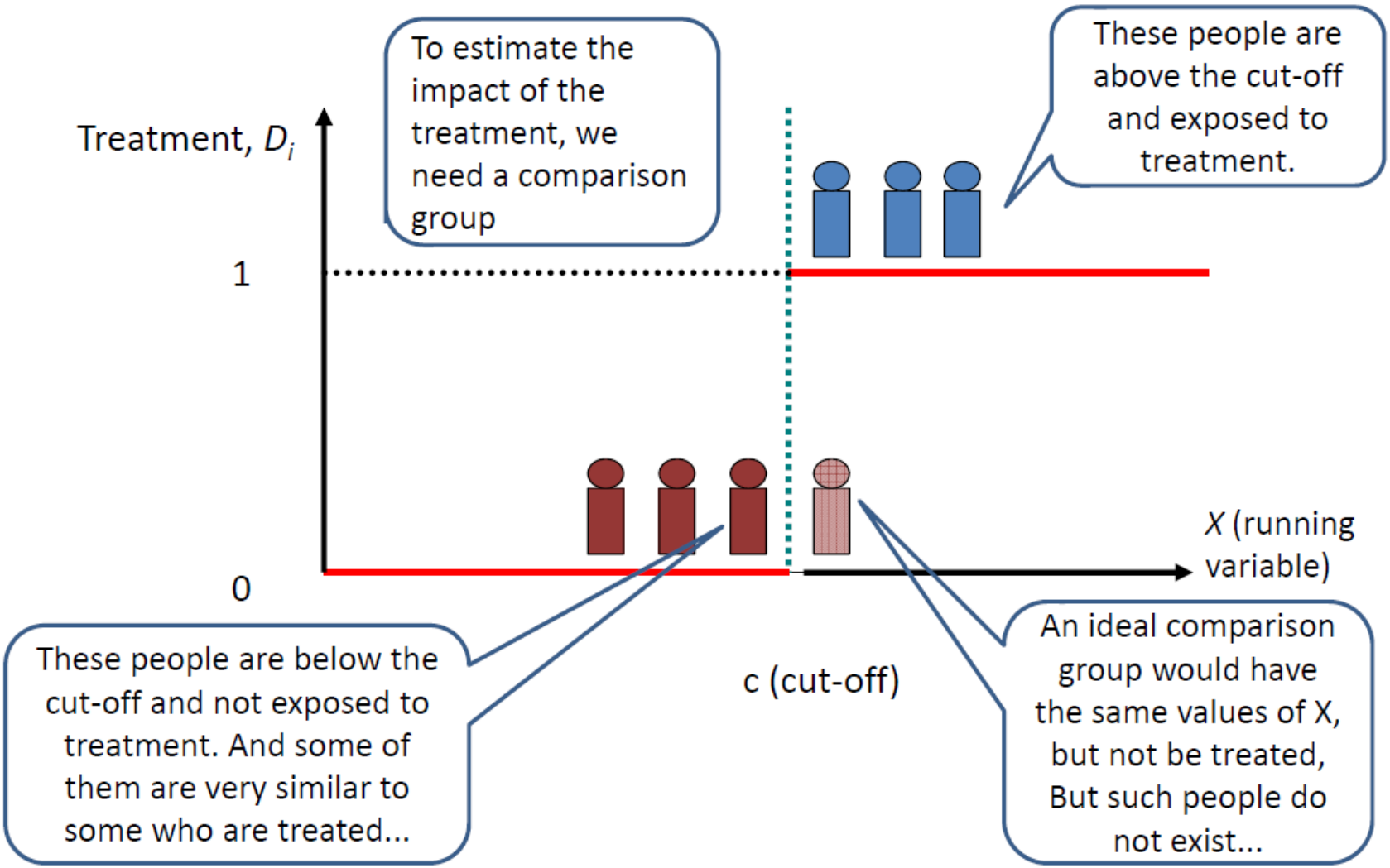
- WLOG:

$$Pr(D_i = 1/c^+) - Pr(D_i = 1/c^-) > 0 \tag{4}$$

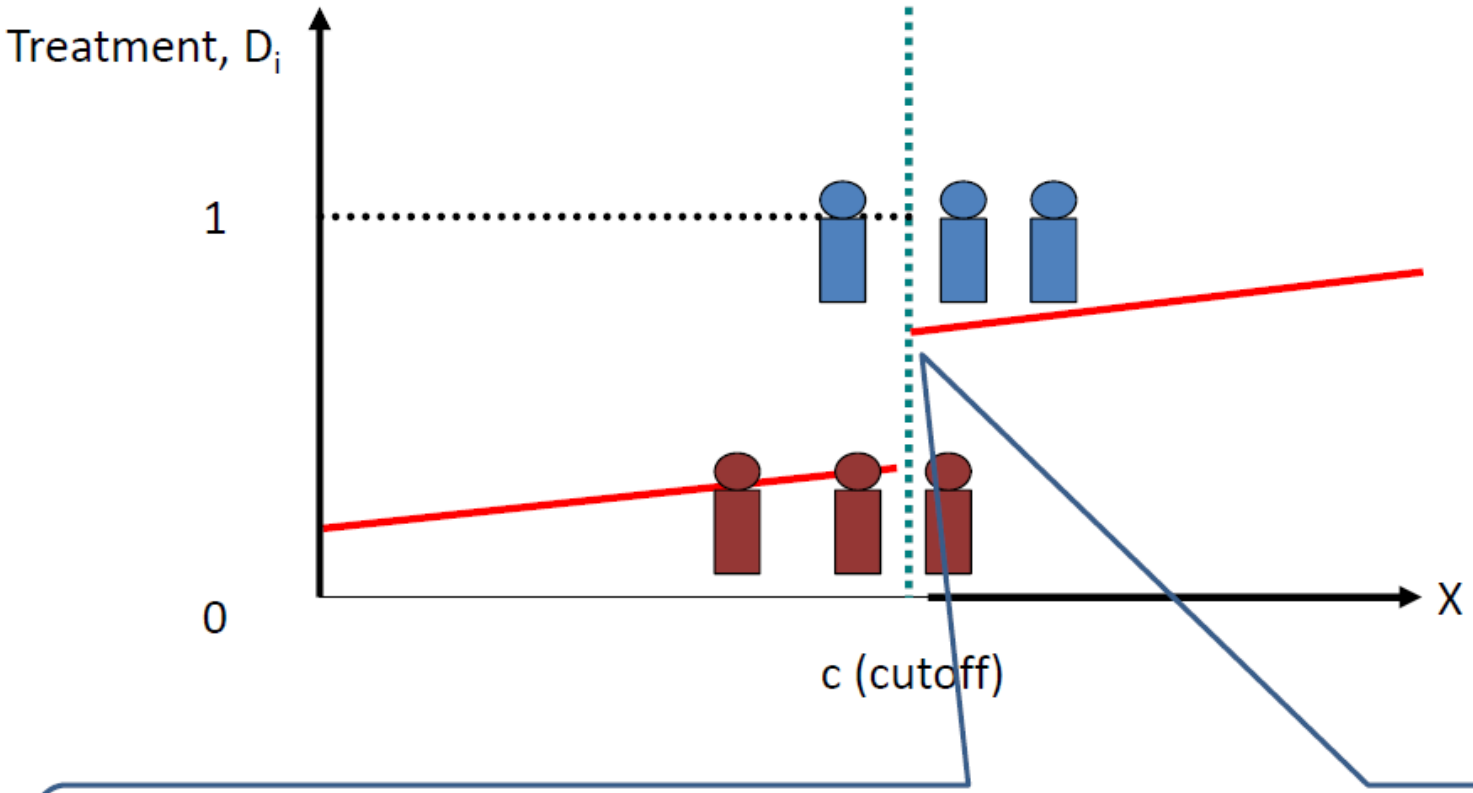
Sharp and Fuzzy Discontinuity

- Sharp discontinuity
  - The discontinuity precisely determines treatment
  - Equivalent to random assignment in a neighborhood
  - E.g. Social security payment depend directly and immediately on a person's age
- Fuzzy discontinuity
  - Discontinuity is highly correlated with treatment
  - Use the assignment as an IV for program participation
  - E.g. Rules determine eligibility but there is a margin of administrative error

Sharp RD


To estimate the impact of the treatment, we need a comparison group
These people are above the cut-off and exposed to treatment.
Treatment, D_i
1
0
X (running variable)
c (cut-off)
These people are below the cut-off and not exposed to treatment. And some of them are very similar to some who are treated...
An ideal comparison group would have the same values of X, but not be treated, But such people do not exist...

Fuzzy RD



Now treatment depends on whether *X* bigger than cut-off *c*, but this is not the only factor. There is a jump in the fraction who are treated as we cross the cut-off, *c*.

2.1 Sharp RDD

- Probability of treatment conditional on X steps from zero to one as X crosses the threshold c. Therefore:

$$D_i = 1(X \geq c) \tag{5}$$

- Observed outcome:

$$Y_i = Y_i(0) + D_i(X)\beta \tag{6}$$

- The difference of observed mean outcomes marginally above and below c is:

$$\begin{aligned}\mathbb{E}(Y_i/c^+) - \mathbb{E}(Y_i/c^-) &= \mathbb{E}(Y_i(0)/c^+) - \mathbb{E}(Y_i(0)/c^-) \\ &+ \mathbb{E}(D_i(X)\beta/c^+) - \mathbb{E}(D_i(X)\beta/c^-) \\ &= \mathbb{E}(Y_i(0)/c^+) - \mathbb{E}(Y_i(0)/c^-) + \mathbb{E}(\beta/c^+)\end{aligned}$$

- The mean treatment effect at c^+ is identified if the following condition is true:

Condition I: Continuity of potential outcomes

The mean value of $Y(0)$ conditional on X is a continuous function of X at c^+:

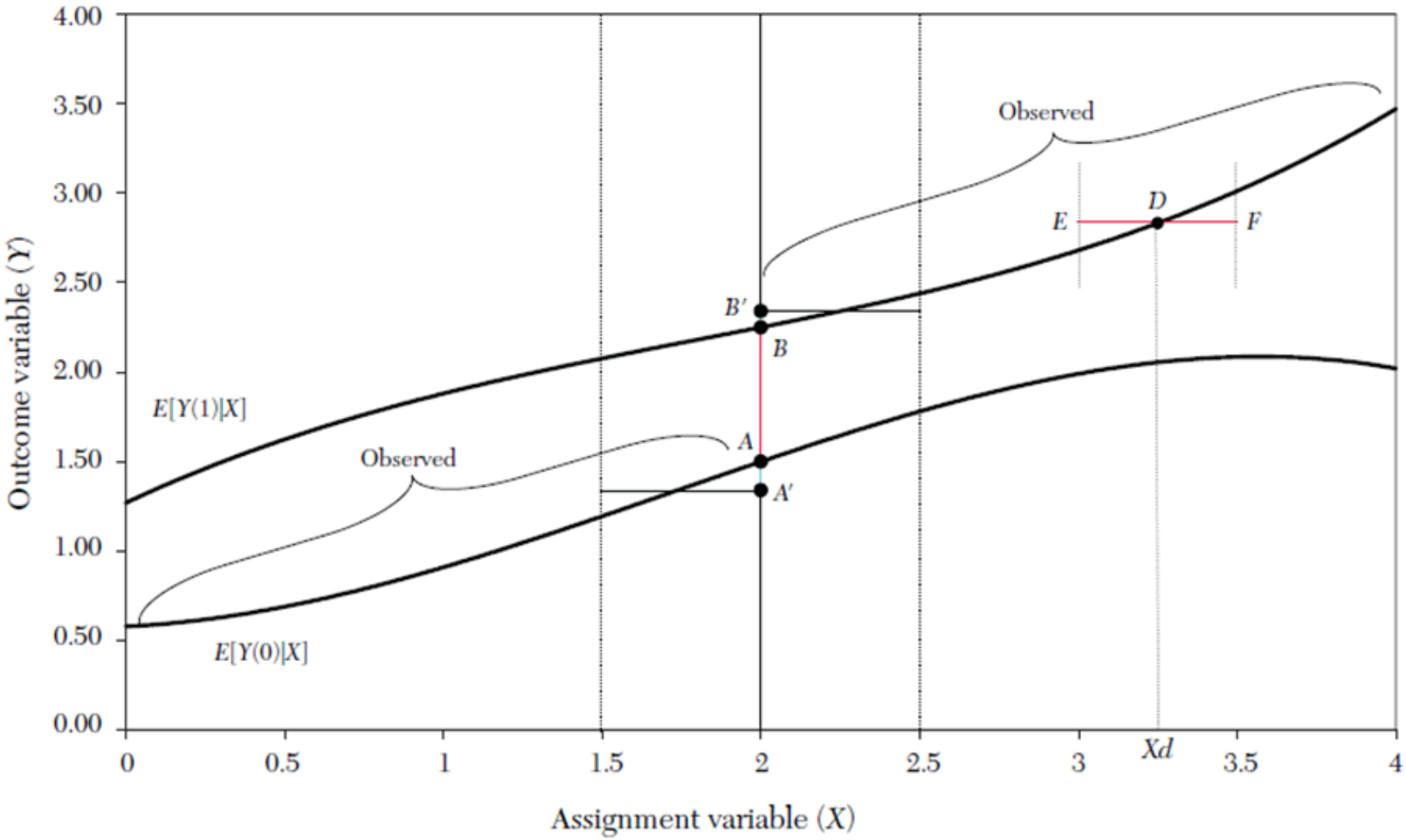
$$\mathbb{E}(Y_i(0)/c^+) = \mathbb{E}(Y_i(0)/c^-) \tag{7}$$

- Counterfactual world: no discontinuity takes place at the threshold for selection
- Then, we can compute:

$$\mathbb{E}(\beta/c^+) = \mathbb{E}(Y_i/c^+) - \mathbb{E}(Y_i/c^-) \tag{8}$$

- If sample is large enough: compute these expressions using data for subjects in a neighborhood of the discontinuity
- If sample is small: use some parametric assumptions about the regression curve away the discontinuity point

Continuity assumption



2.2 Fuzzy RDD

- Arises when there is no perfect compliance with the assignation rule
- Treatment status depends not only of X but some unobservable characteristics
- A new condition is needed:

Condition II

The triple $Y_i(1), Y_i(0), D_i(X)$ is stochastically independent of X in a neighborhood c

- Standard exclusion restriction in an IV setup: X affects the outcome only through its effect on the treatment D

- If Condition 2 is true:

$$\mathbb{E}(Y_i/c^+) - \mathbb{E}(Y_i/c^-) = \mathbb{E}(\beta/D_i(c^+) > D_i(c^-)).Pr(D_i(c^+) > D_i(c^-)) - \mathbb{E}(\beta/D_i(c^+) < D_i(c^-)).Pr(D_i(c^+) < D_i(c^-))$$

- Where:
  - RHS 1: average effect for compliers, times probability of compliance
  - RHS 2: average effect for defiers, times the probability of defiance

- Remember from your imperfect compliance class:
  - Always takers and never takers do not contribute because their potential treatment status does not change at the discontinuity
  - We need a strong monotonicity assumption for ruling out the defiers
- The additional assumption:

Condition III

Participation into the program is monotone around the discontinuity

- Then, the outcome comparison of subjects above and below the threshold gives:

$$\mathbb{E}(\beta/D_i(c^+) \neq D_i(c^-)) = \frac{\mathbb{E}(Y_i/c^+) - \mathbb{E}(Y_i/c^-)}{\mathbb{E}(D_i/c^+) - \mathbb{E}(D_i/c^-)} \tag{9}$$

- This recovers the mean impact of the treatment in those individuals in a neighborhood of who would switch their treatment status if the threshold for participation switched from just above their score to just below it
- It is an analogous of LATE
- Denominator RHS: proportion of compliers at the discontinuity

2.3 A regression framework for fuzzy RDD

- Under the assumptions above:

$$Y_i = g(X_i) + \beta D_i + \epsilon_i \tag{10}$$

- Where:
  - Y_i is the observed outcome
  - $g(X_i)$ is a polynomial in the score of X
  - D is a binary indicator that denotes actual exposure to treatment
  - $D = 1(X \geq c)$ is the side of the threshold on which each subject is located

2.4 Advantages and disadvantages of RDD

- Advantages:
  - RD yields an unbiased estimate of treatment effect at the discontinuity
  - It takes advantage of a known rule for assigning the benefit: No need to "exclude" a group of eligible households/ individuals from treatment
- Disadvantages:
  - Local average treatment effects
  - Power
  - Specification can be sensitive to functional form: make sure the relationship between the assignment variable and the outcome variable is correctly modeled, including non-linearities and interactions

3. Graphical presentation of RDD

- RDD provides a graphical transparent way of showing how the treatment effect is identified
- Standard way of graphing:
  - Divide the assignment variable into a number of bins in both sides of the discontinuity
  - Compute the average value of the outcome variable for each bin
  - Graph these values against the mid-points of the bins
- Advantages:
  - Simple way of visualizing what the functional form of the regression function looks like
  - Provide indication of the magnitude of the jump in the regression function
  - Allows for detection of unexpected comparable jumps

- Notice that: choice of the width of the bin matters!
  - If bins are too narrow, the estimation will be highly imprecise
  - If they are too wide, estimates may be biased

Choice of bins

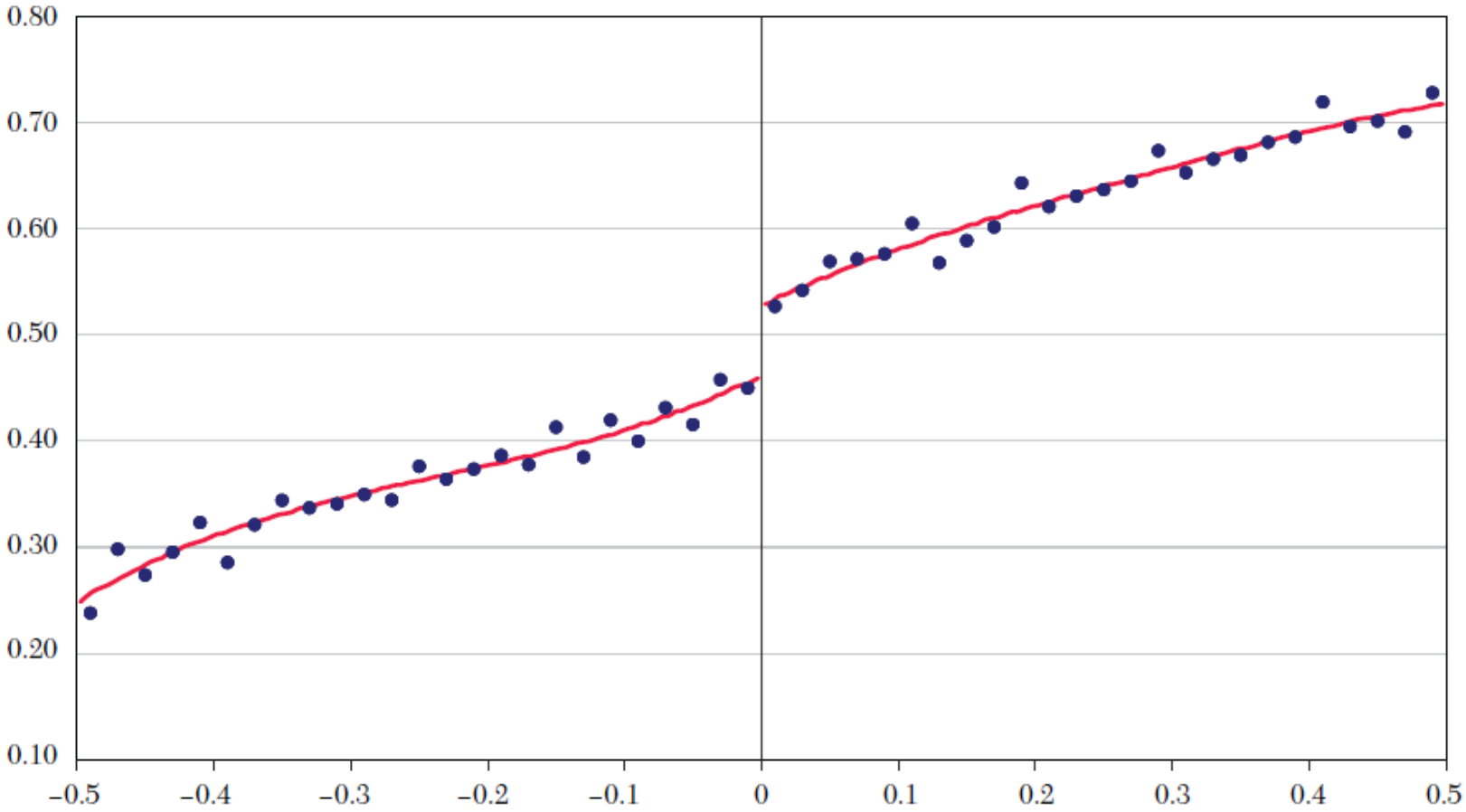


Figure 6. Share of Vote in Next Election, Bandwidth of 0.02 (50 bins)

Choice of bins

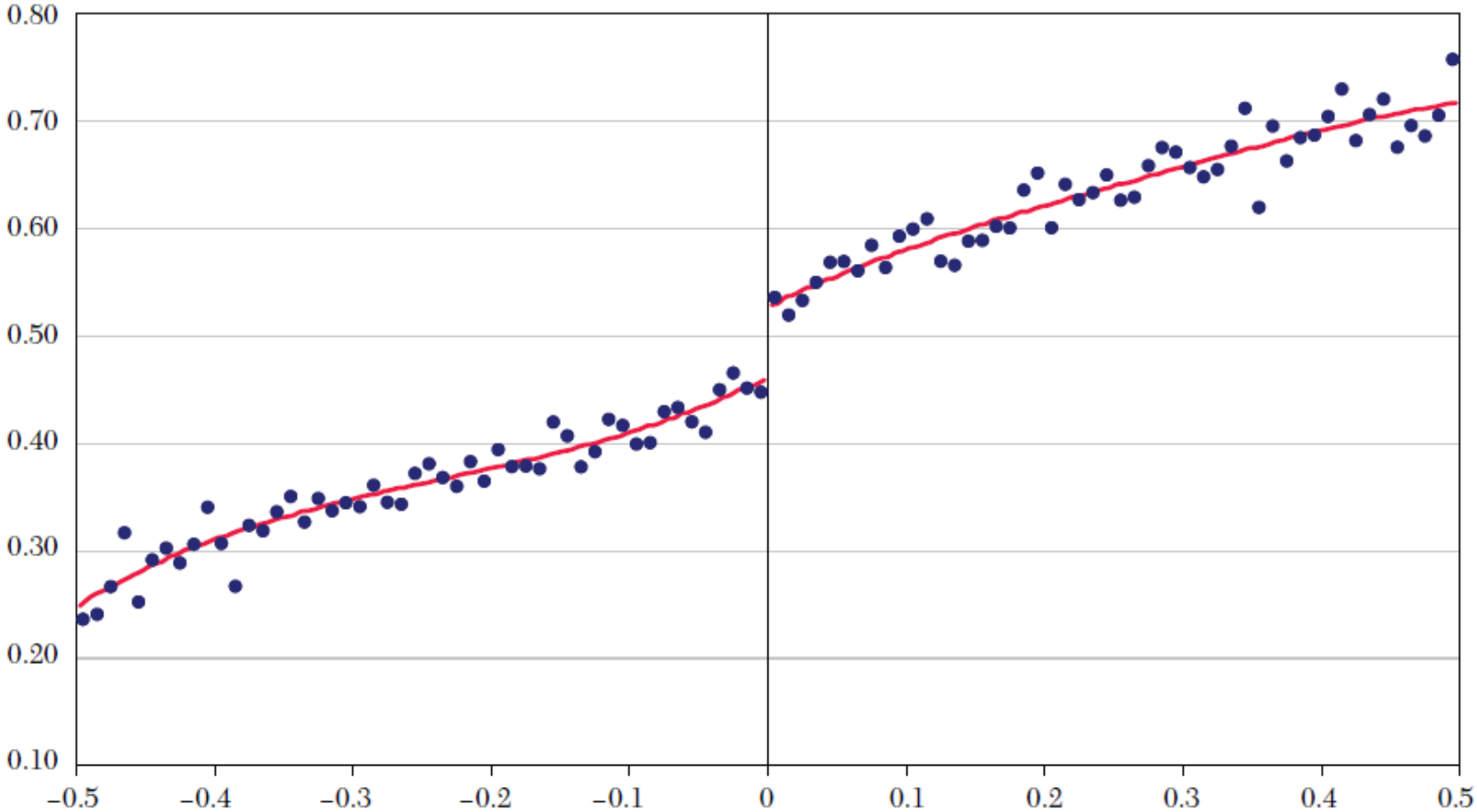


Figure 7. Share of Vote in Next Election, Bandwidth of 0.01 (100 bins)

Choice of bins

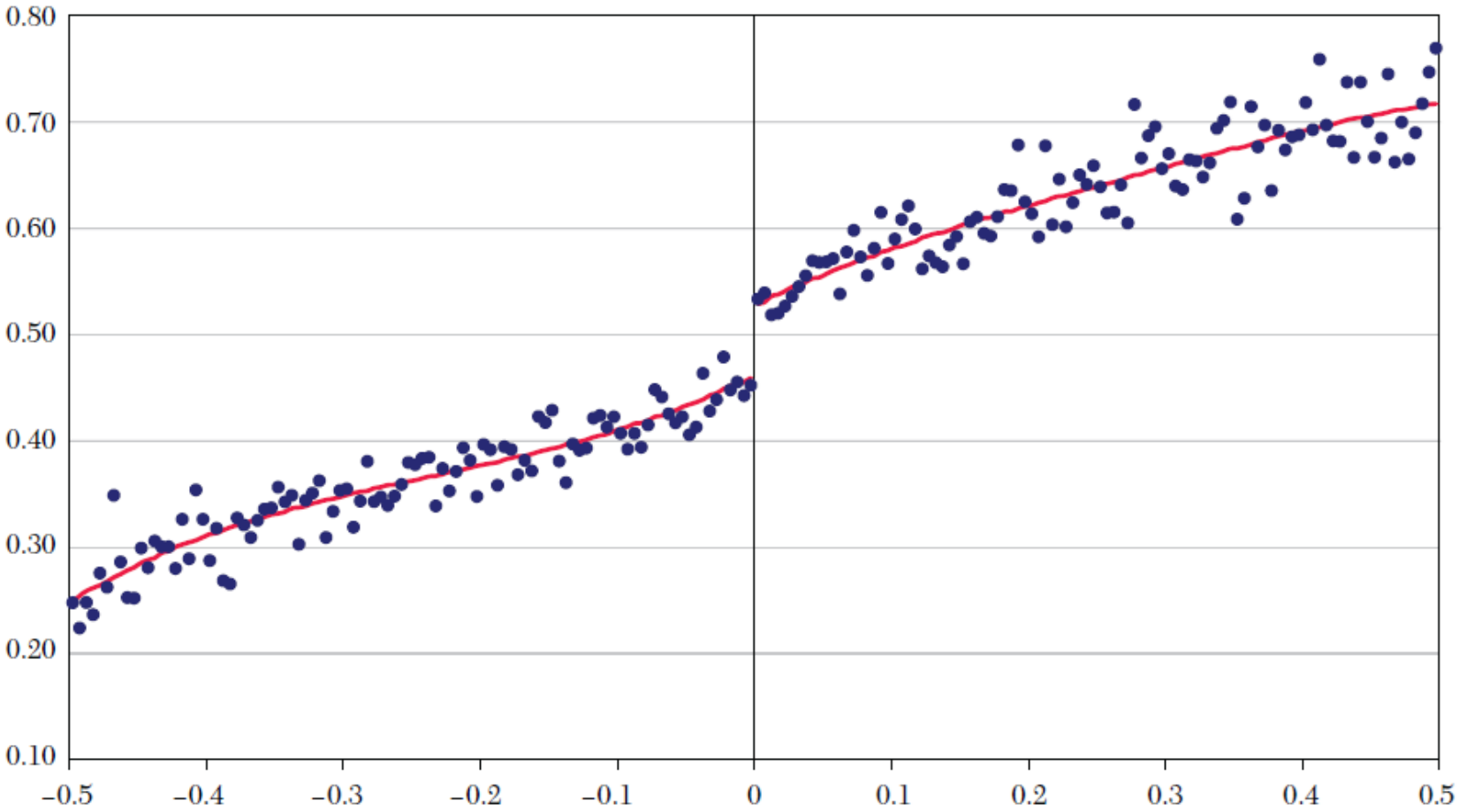


Figure 8. Share of Vote in Next Election, Bandwidth of 0.005 (200 bins)

4. Estimation issues

- Recall the basic regression model discussed in previous lecture:

$$Y = \beta D + g(X) + \epsilon \tag{11}$$

- Consequences of using an incorrect functional form are critical in RD since it can induce serious bias. Non-parametric approaches should be preferred
- Despite this, most of the applied papers report estimates from parametric models. Are all they wrong?
  - Nonparametric estimation is not the "solution" for RD functional issues. Parametric and non-parametric approaches should be seen as complements (Lee and Lemieux 2010)
  - Causal estimates based on high order parametric polynomials are misleading and should not be used (Gelman and Imbens 2014)

- Non-parametric regression is typically suggested when the true functional form is unknown
  - Problem: RDD poses a particular problem because it requires the estimation of a regression at the cutoff point (boundary problem)
- A simple parametric way to relax the linearity assumption is to include polynomial functions of X in a regression model
  - Problem: It provides global estimates of the regression function over all values of X

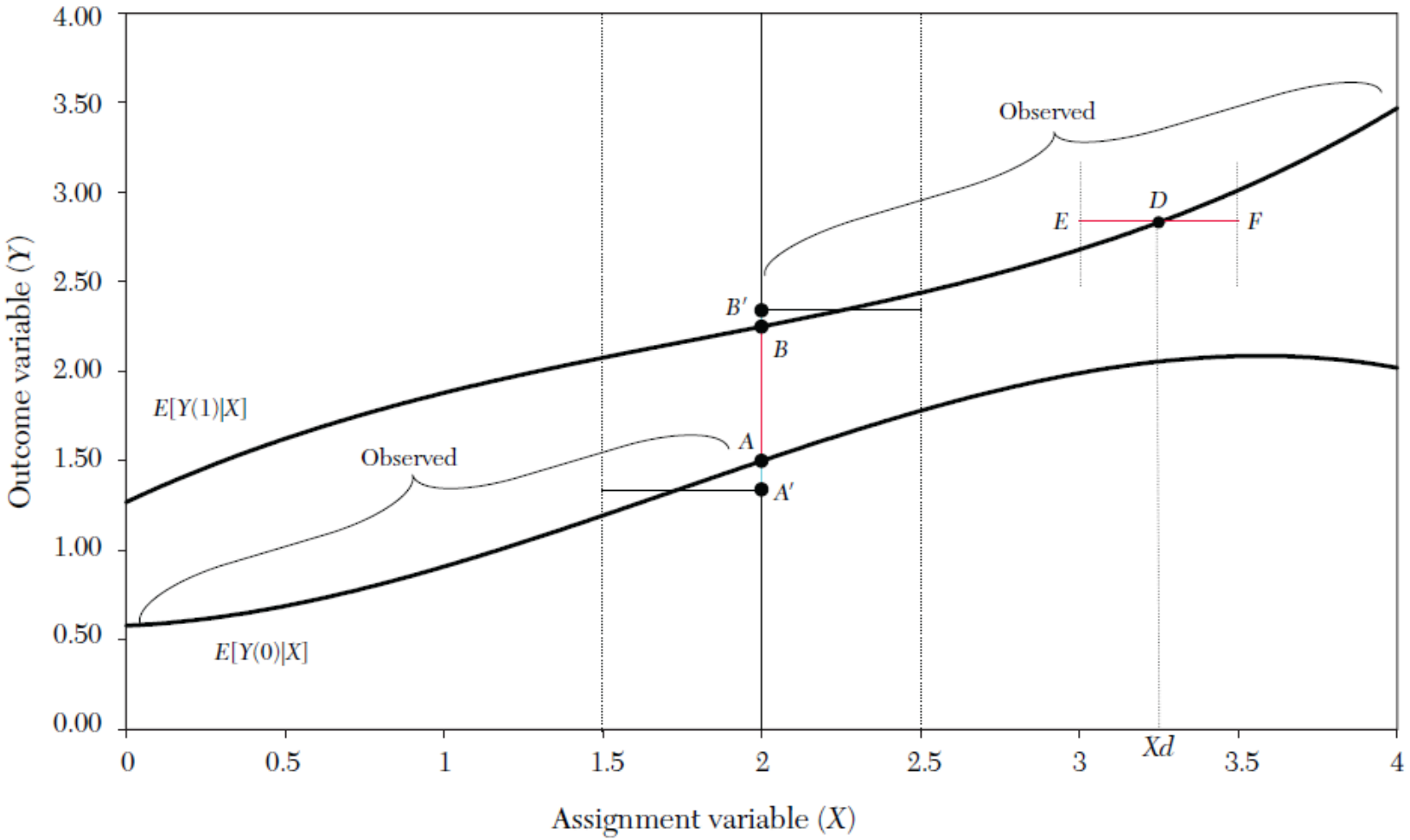
4.1 Non-parametric Sharp RDD estimation

- Consider the case of a Kernel regression:

$$\widehat{ATE}_{SRD} = \frac{\sum_{i \in R} K\left(\frac{X_i - c}{h}\right) Y_i}{\sum_{i \in R} K\left(\frac{X_i - c}{h}\right)} - \frac{\sum_{i \in L} K\left(\frac{X_i - c}{h}\right) Y_i}{\sum_{i \in L} K\left(\frac{X_i - c}{h}\right)} \quad (12)$$

- Recall that Kernel regression is a local method for estimating the regression function at a particular point. Unfortunately, it performs poorly at boundary points
- Consider point D in Figure 1. Applying a local averaging approach is problematic at the cutoff because only observations above/below the cutoff should be used

Bias in non-parametric RD estimates



- The best option is to compute the average value of Y in the bin just to right and just to the left of the cutoff, but that provides a biased estimate of the treatment effect (See the difference between B-A and B'-A' in the figure)
- Trade-off:
  - Bias can be reduced by reducing the bandwidth
  - Bandwidth has to be large enough to have enough observations to get precise estimates for the average of Y below and above the cutoff
- Local linear regression reduces bias compared to other nonparametric techniques (Hahn et al 2001)

4.2 Bandwidth Choice

- Choosing a bandwidth involves finding the optimal balance between precision and bias:
  - A large bandwidth yields more precise estimates as a larger number of observations can be used to estimate the treatment effect
  - A larger bandwidth is less likely to be accurate since observations far away from the discontinuity are included in the estimation of the treatment effect
- Two approaches have been advanced in the literature for the case of local linear regression:
  - Plug-in approach
  - Cross-validation approach

4.3 Parametric RDD estimation

- Again, a regression for each side of the discontinuity can be estimated:

$$Y = \alpha_l + g_l(X - c) + \epsilon \tag{13}$$

$$Y = \alpha_r + g_r(X - c) + \epsilon \tag{14}$$

  Where g_l and g_r are functional forms

- The treatment effect can be computed as the difference between the two regression intercepts, α_r and α_l
- A pooled regression can be used as a direct way of estimating the treatment effect:

$$\begin{aligned} Y =& \alpha_l + \tau D + g_l(X - c) + D[g_r(X - c) - g_l(X - c)] + \epsilon \\ =& \alpha_l + \tau D + g(X - c) + \epsilon \end{aligned}$$

Choosing the polynomial degree in parametric RDD

- Choosing the order of the polynomial regression plays the same role as choosing the bandwidth in the non-parametric case
- Although reporting several specifications for the polynomial degree is useful as a robustness check, a more formal criteria for choosing P is useful
- Black et al (2007) suggest using a generalized cross-validation procedure based on the **Akaike information criterion** (AIC):

$$AIC = N.ln(\widehat{\sigma}^2) + 2p \tag{15}$$

Where $\widehat{\sigma}^2$ is the mean squared error of the regression and p is the number of parameters in the regression model (order of the polynomial plus one for the intercept)

5. Specification tests

- Several tests have been suggested in order to evaluate the reliability of RDD:
  1. Testing quasi-randomness at the discontinuity
  2. Testing "non-manipulation" of the forcing variable (McCrary Test)
  3. Testing the continuity of the outcome conditional expectation
  4. Testing the sensitivity of results to different bandwidths and polynomial order

5.1 Specification Tests: Baseline Covariates

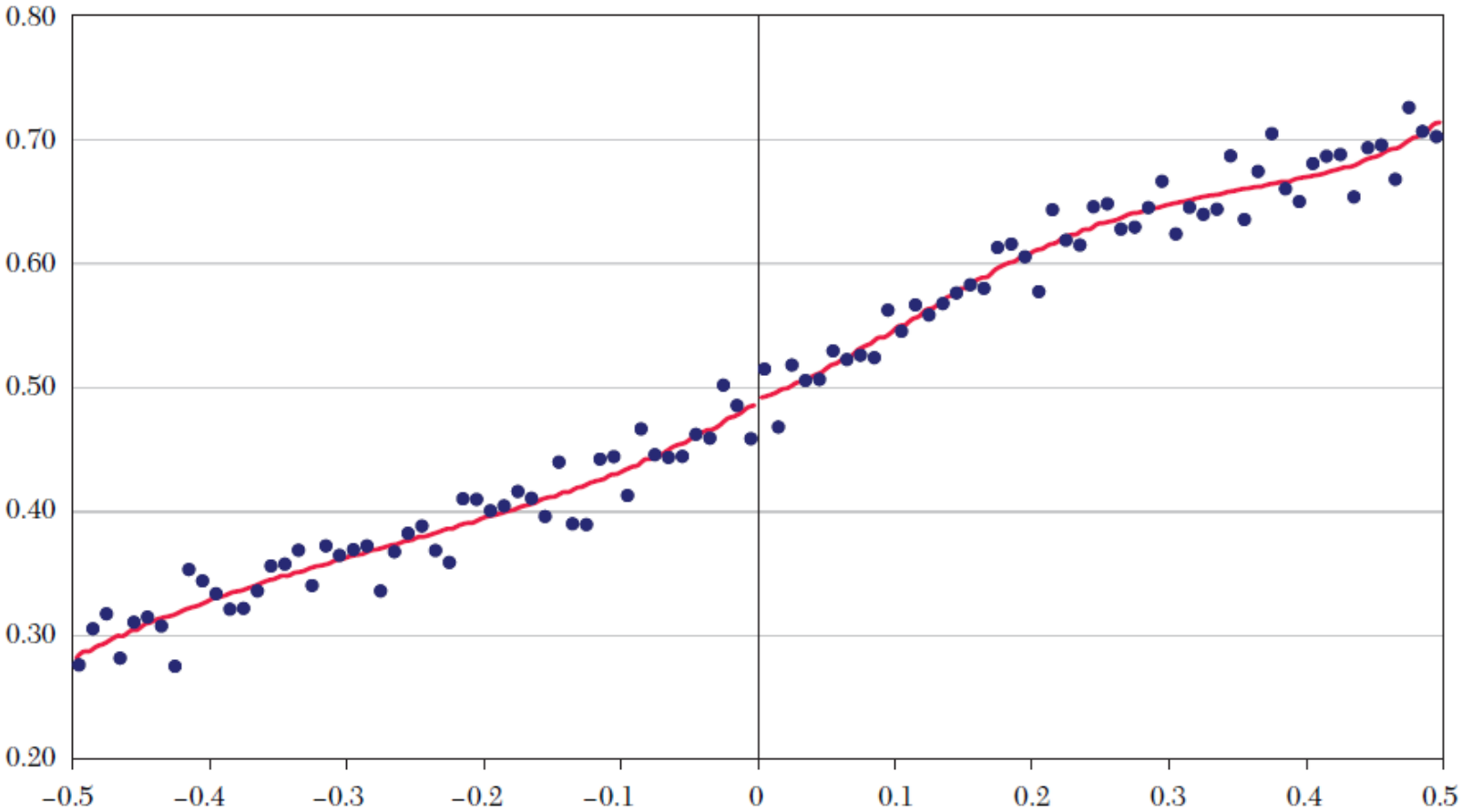


Figure 17. Discontinuity in Baseline Covariate (Share of Vote in Prior Election)

5.2 Specification Tests: McCrary Test

- McCrary (2008) proposes a simple two-step procedure to detect if there is a discontinuity in the assignment variable due to manipulation:
  1. Creates a finely gridded histogram (equally spaced bins) of the running variable
  2. Histogram is smoothed using a local linear regression (bins mid points used as regressors and the normalized counts of the number of observations falling into the bins are treated as the dependent variable. More weight is given to observations close to the cutoff)
- Example: democratic margin in US elections versus roll call voting in US house of representatives

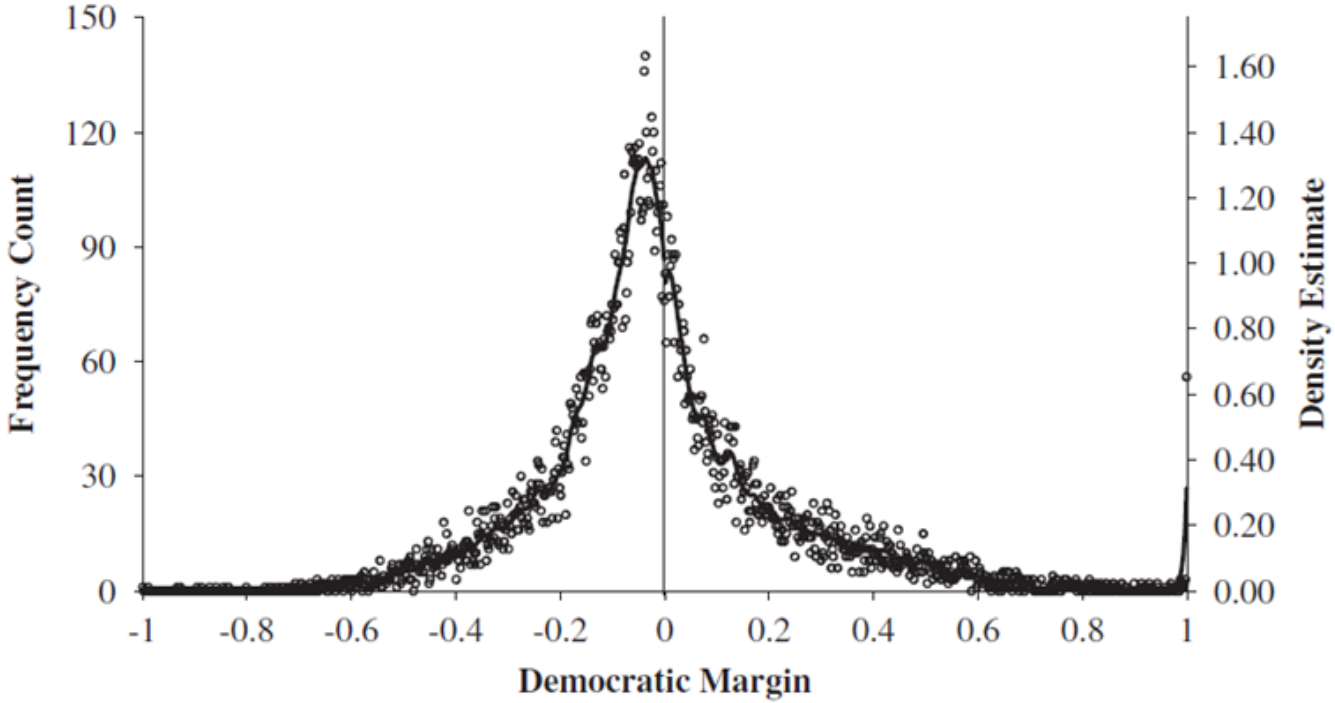


Fig. 4. Democratic vote share relative to cutoff: popular elections to the House of Representatives, 1900–1990.

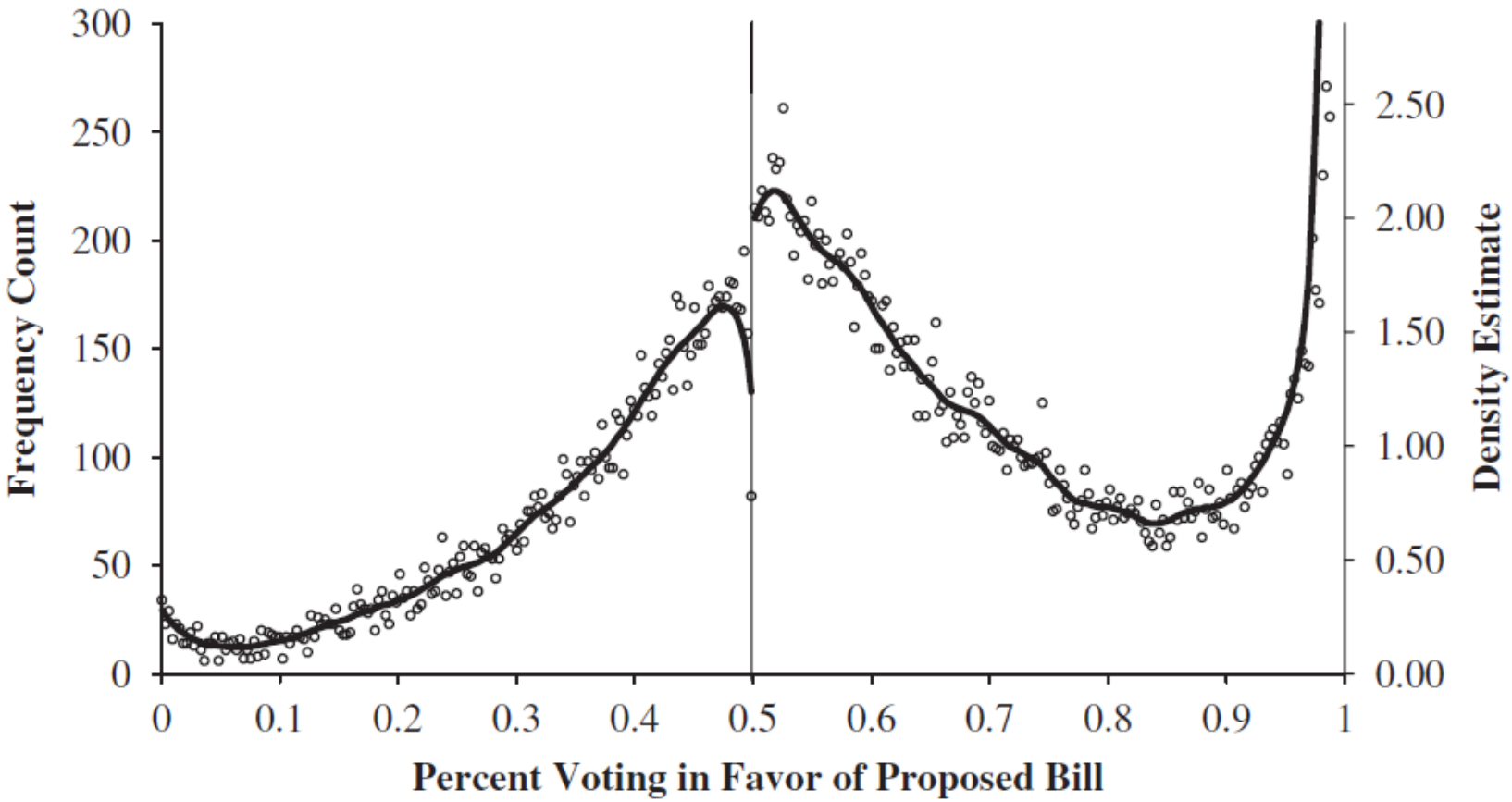
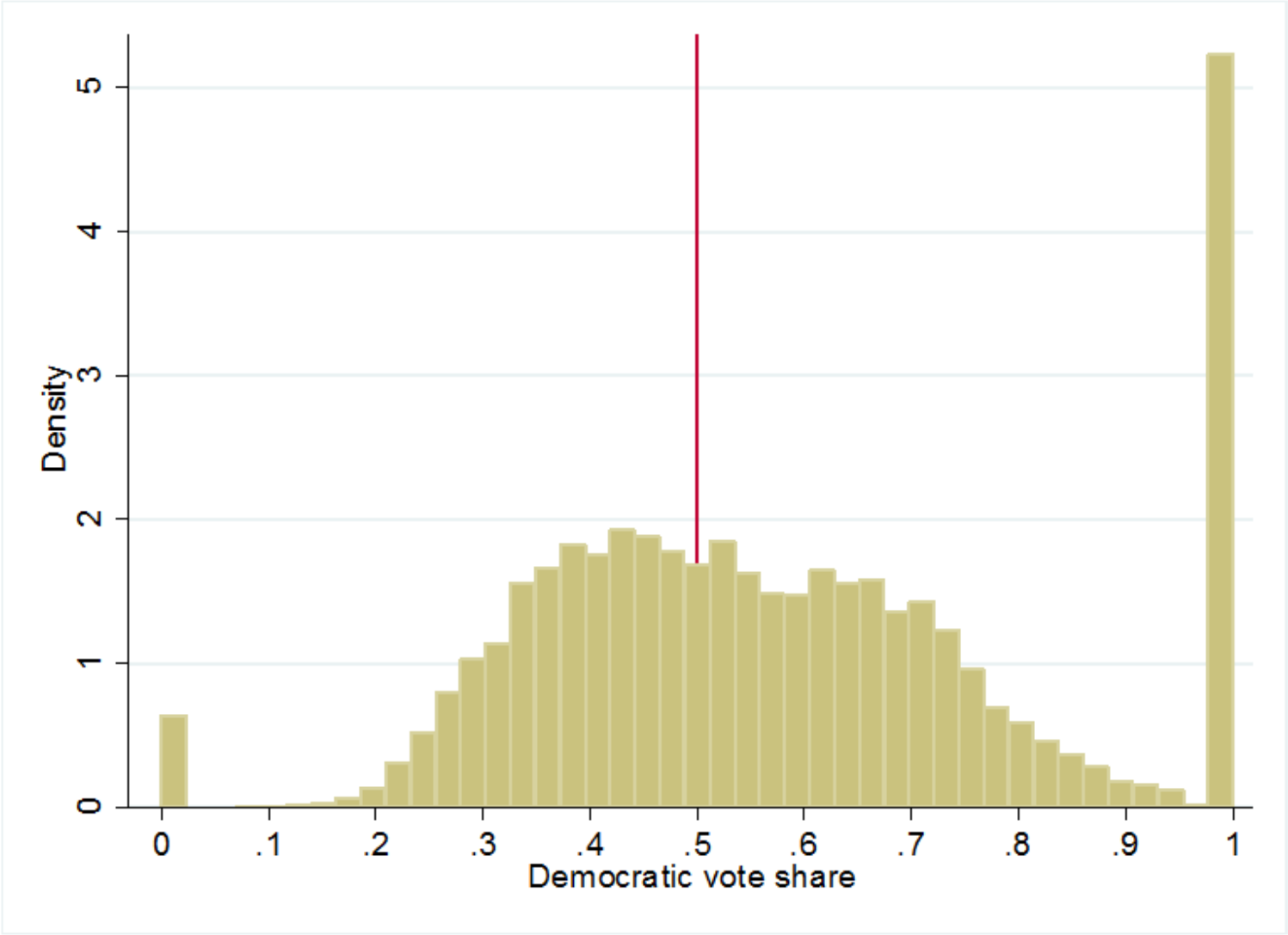
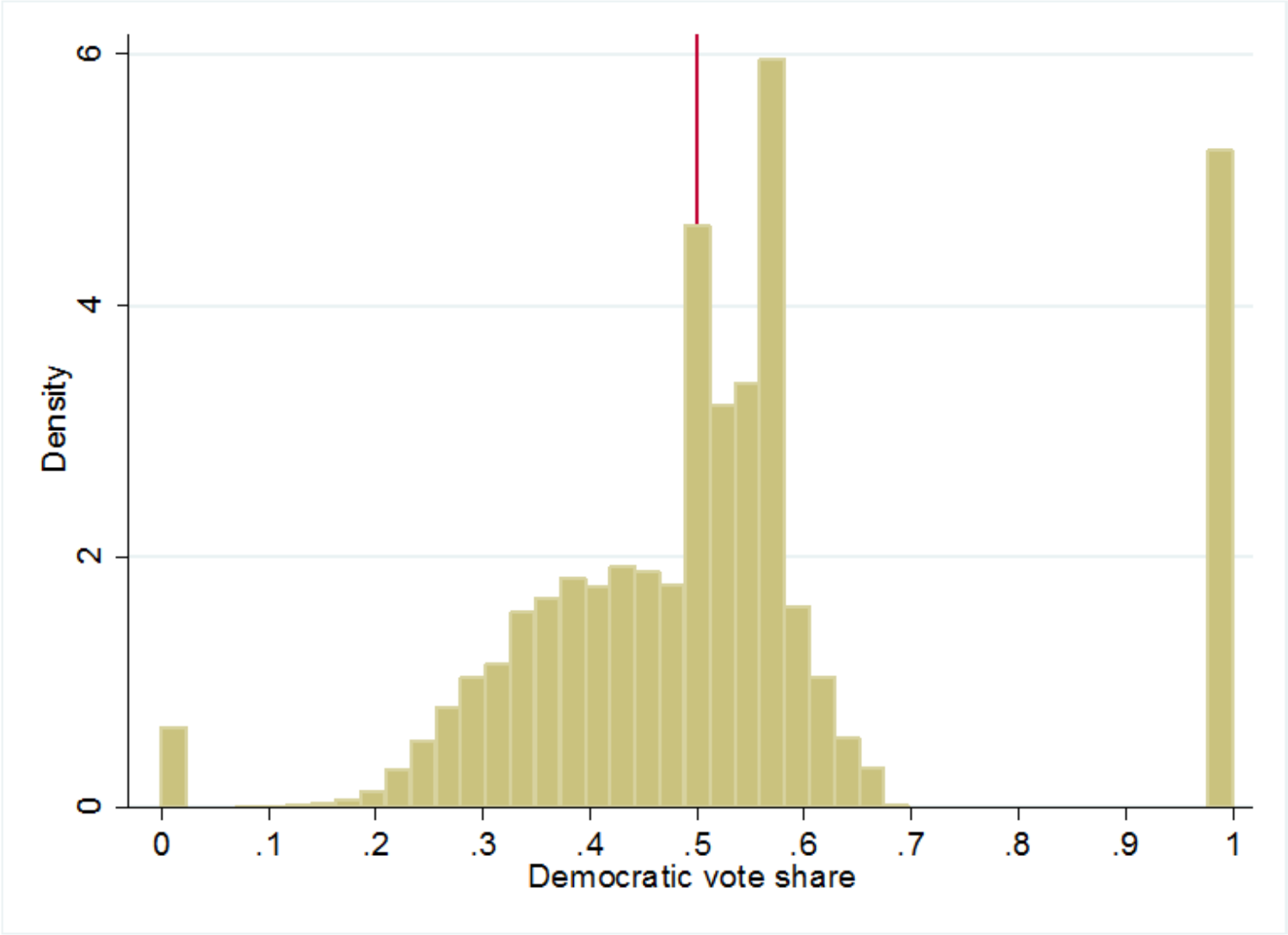


Fig. 5. Percent voting yeay: roll call votes, U.S. House of Representatives, 1857–2004.

Example of manipulation

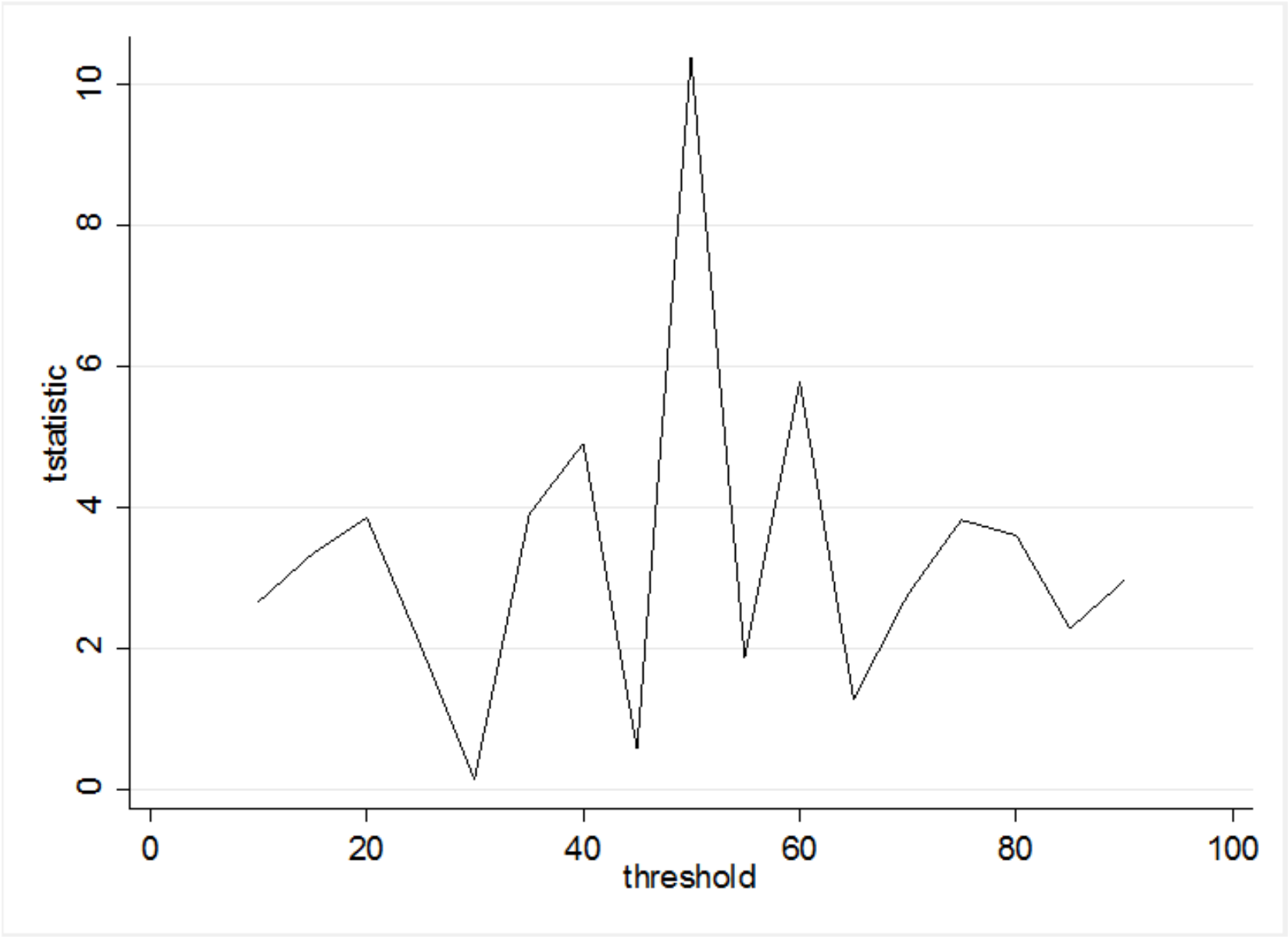


Example of manipulation



5.3 Specification Tests: Continuity of outcome conditional expectation

- Imbens and Lemiux (2008) propose to estimate the jump of the conditional expectation of Y at points of the forcing variable X different from that of the threshold
- An algorithm:
  1. Run a regression for different thresholds
  2. Using t-test for each regression
  3. t-statistic should be maximized at the threshold


tstatistic
0
2
4
6
8
10
0
20
40
60
80
100
threshold

5.4 Specification Tests: Bandwidth Choice

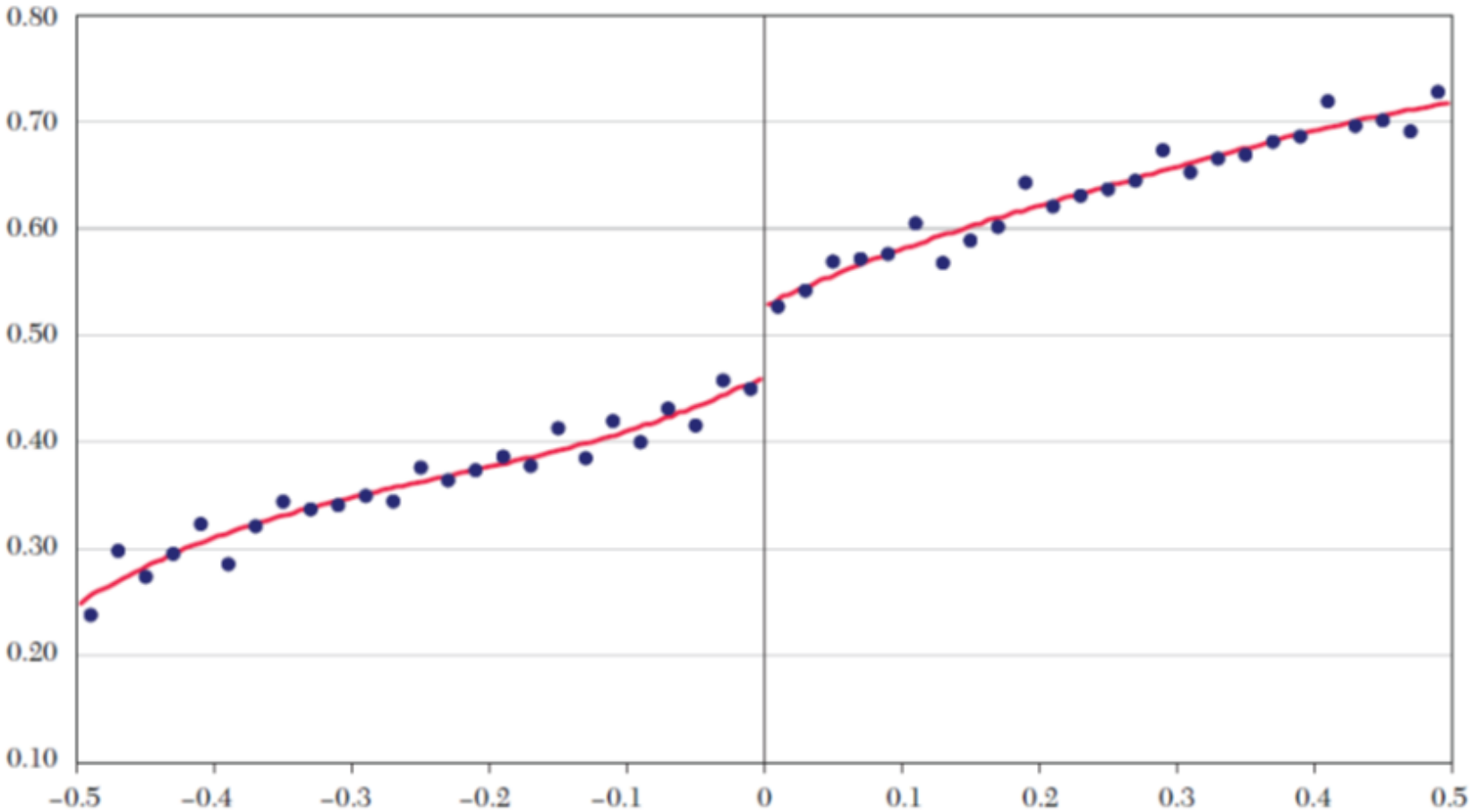


Figure 6. Share of Vote in Next Election, Bandwidth of 0.02 (50 bins)

5.4 Specification Tests: Bandwidth Choice

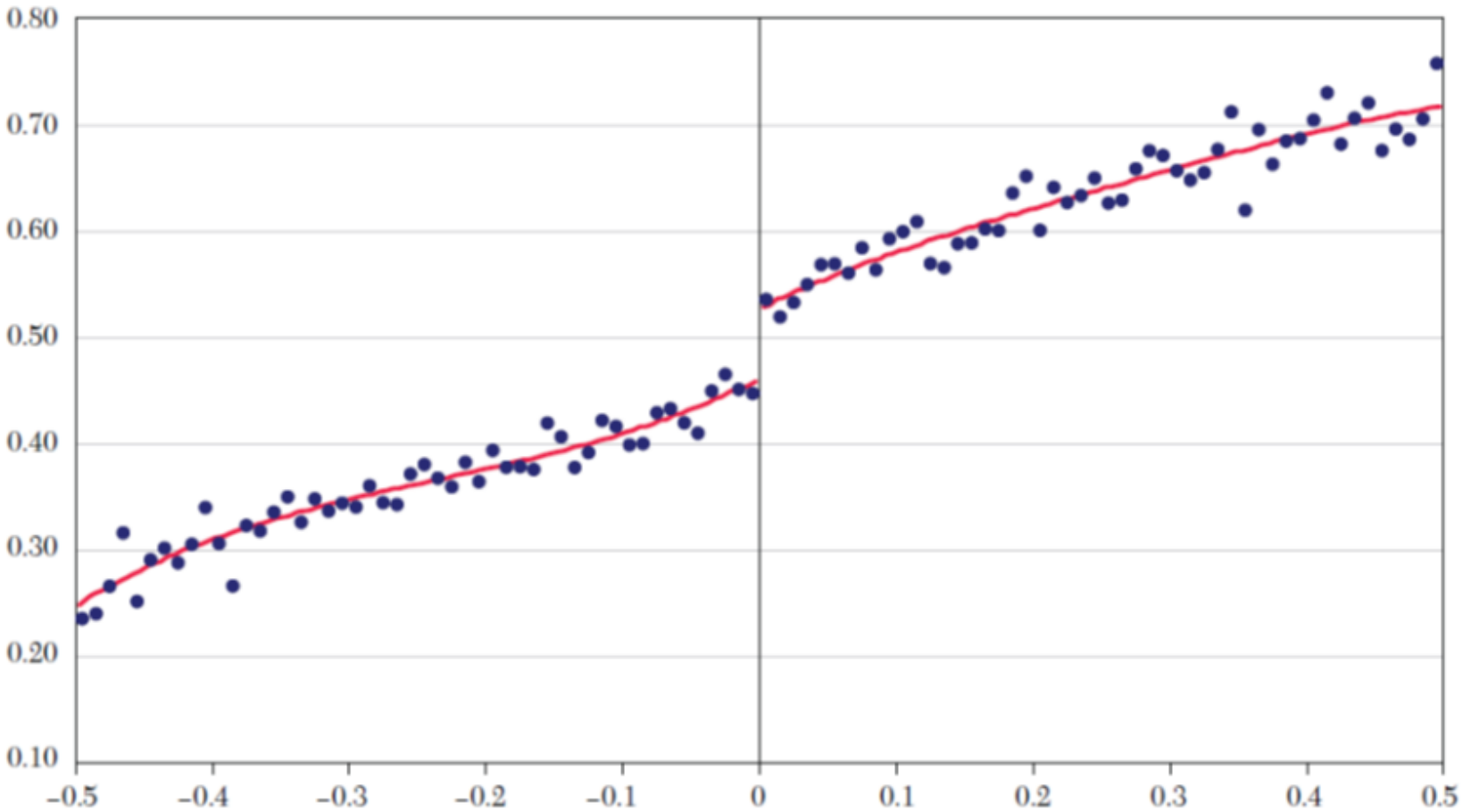


Figure 7. Share of Vote in Next Election, Bandwidth of 0.01 (100 bins)

5.4 Specification Tests: Bandwidth Choice

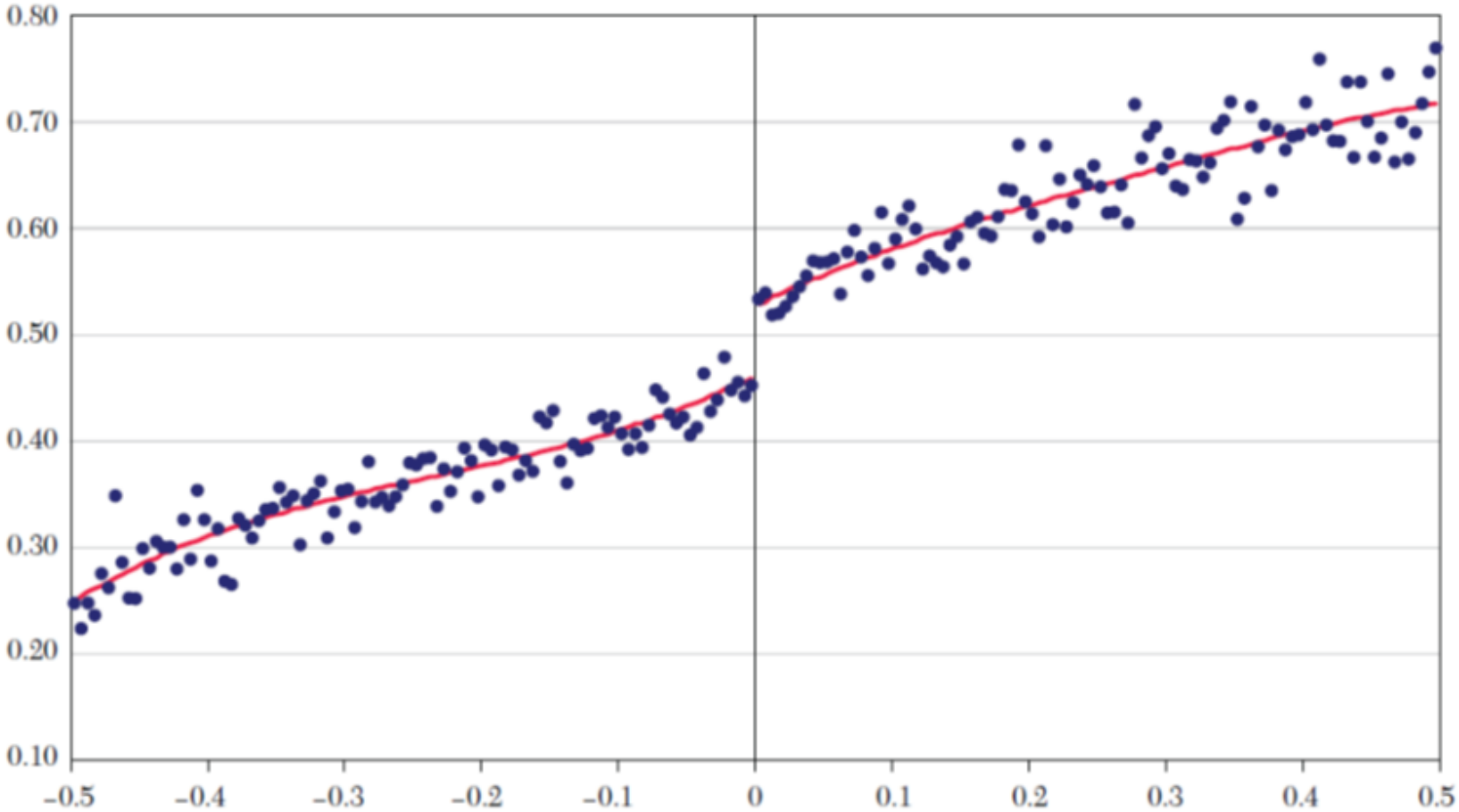
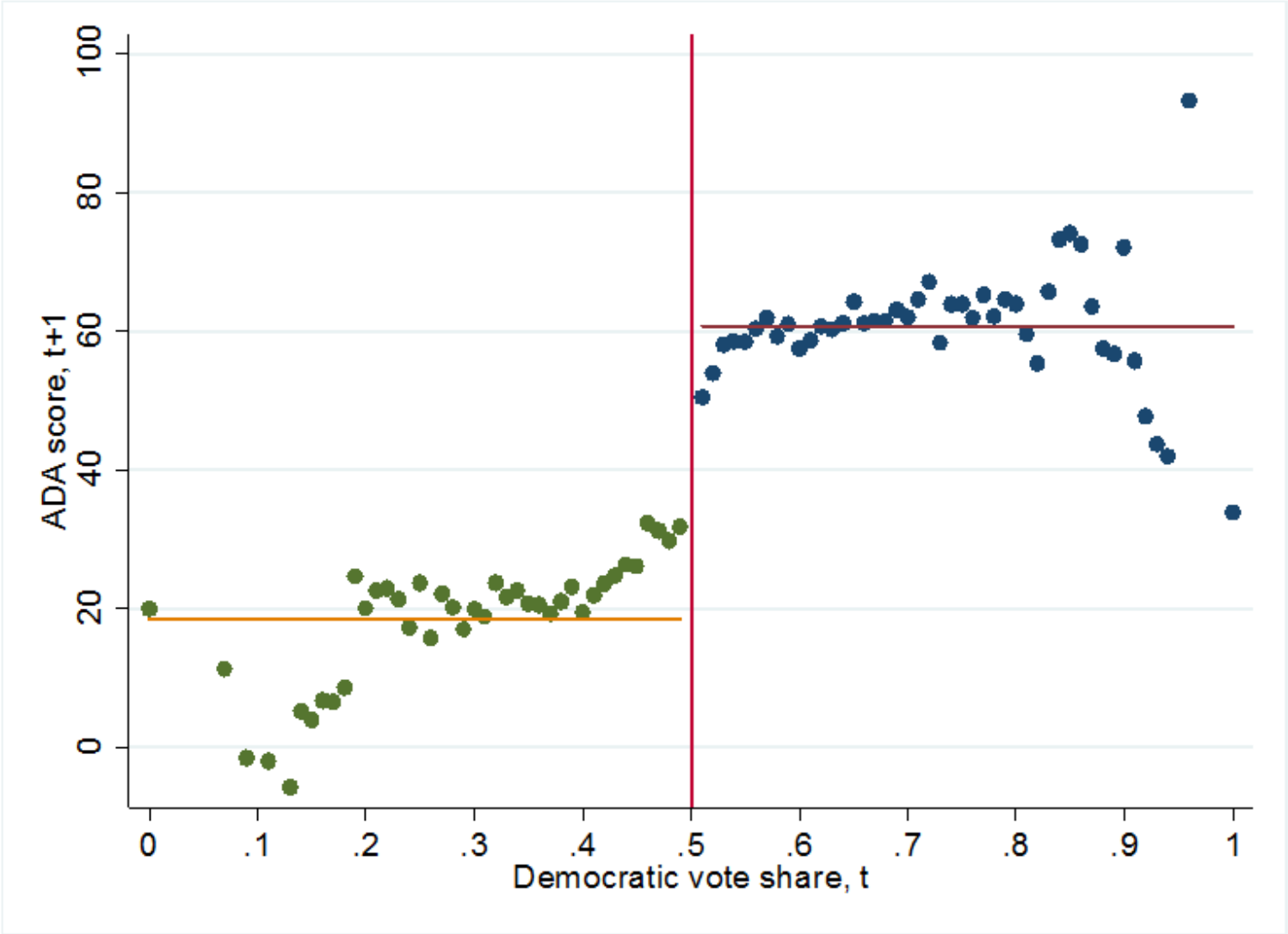
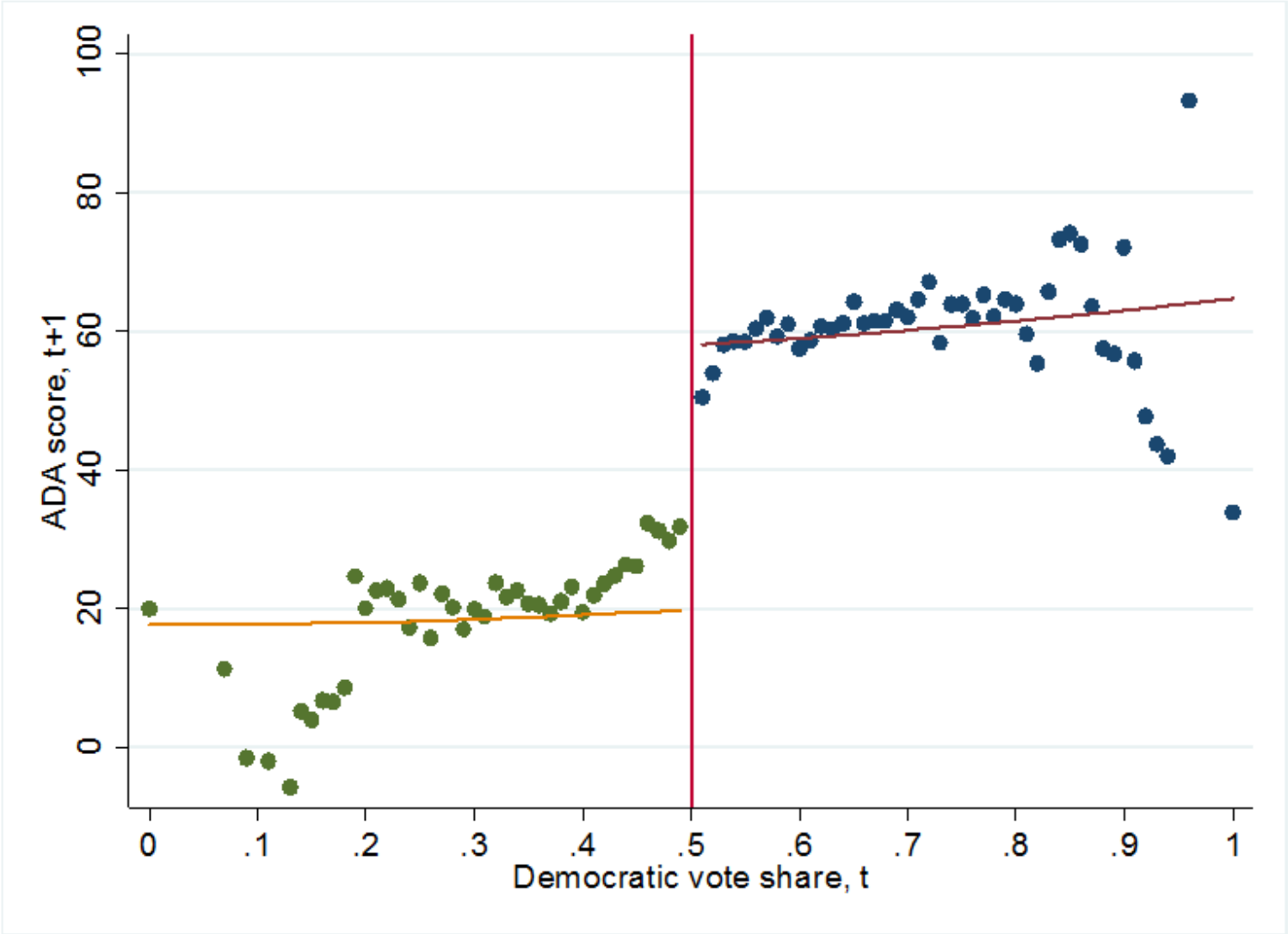


Figure 8. Share of Vote in Next Election, Bandwidth of 0.005 (200 bins)

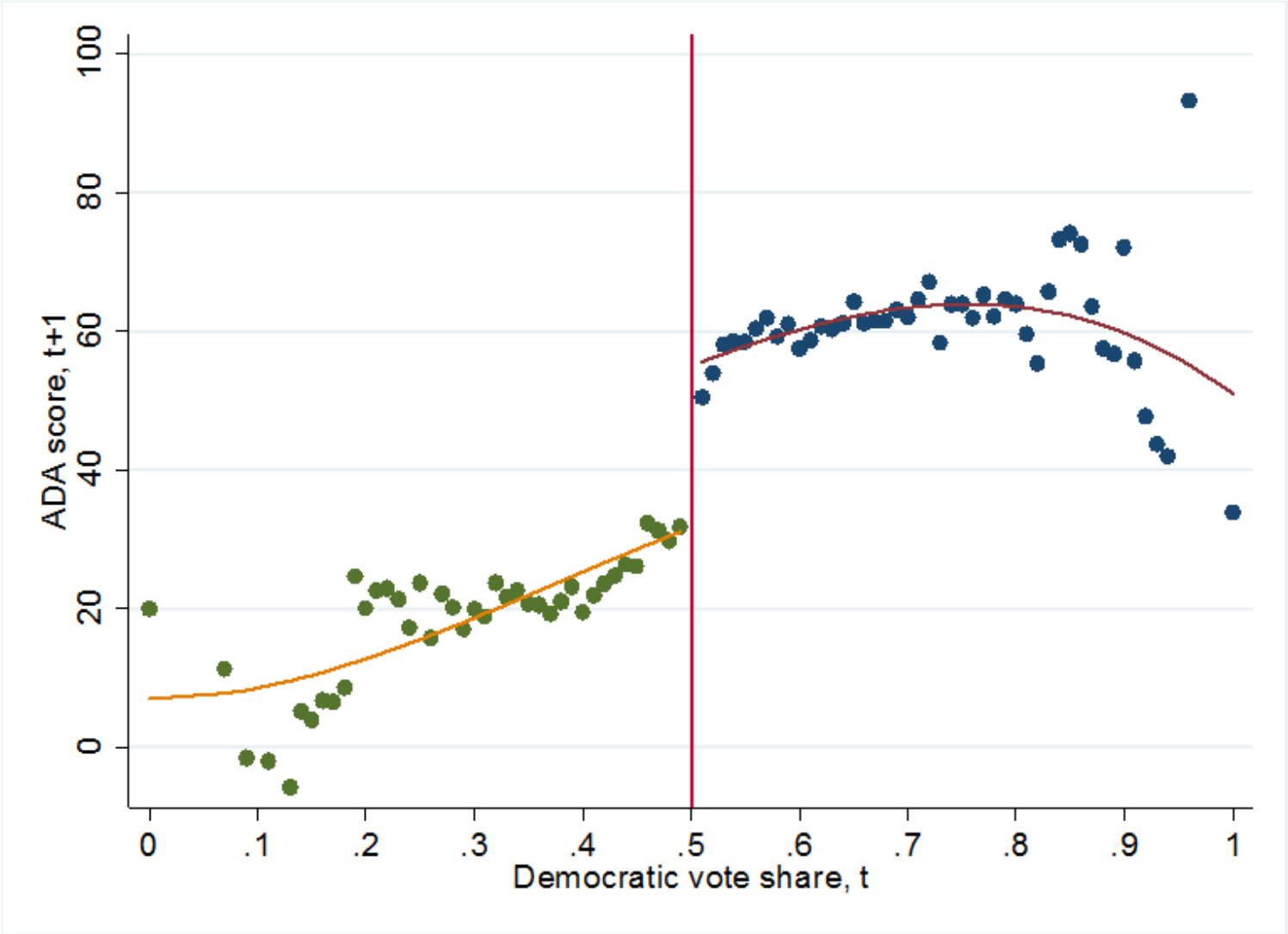
5.4 Specification Tests: Polynomial degree 1



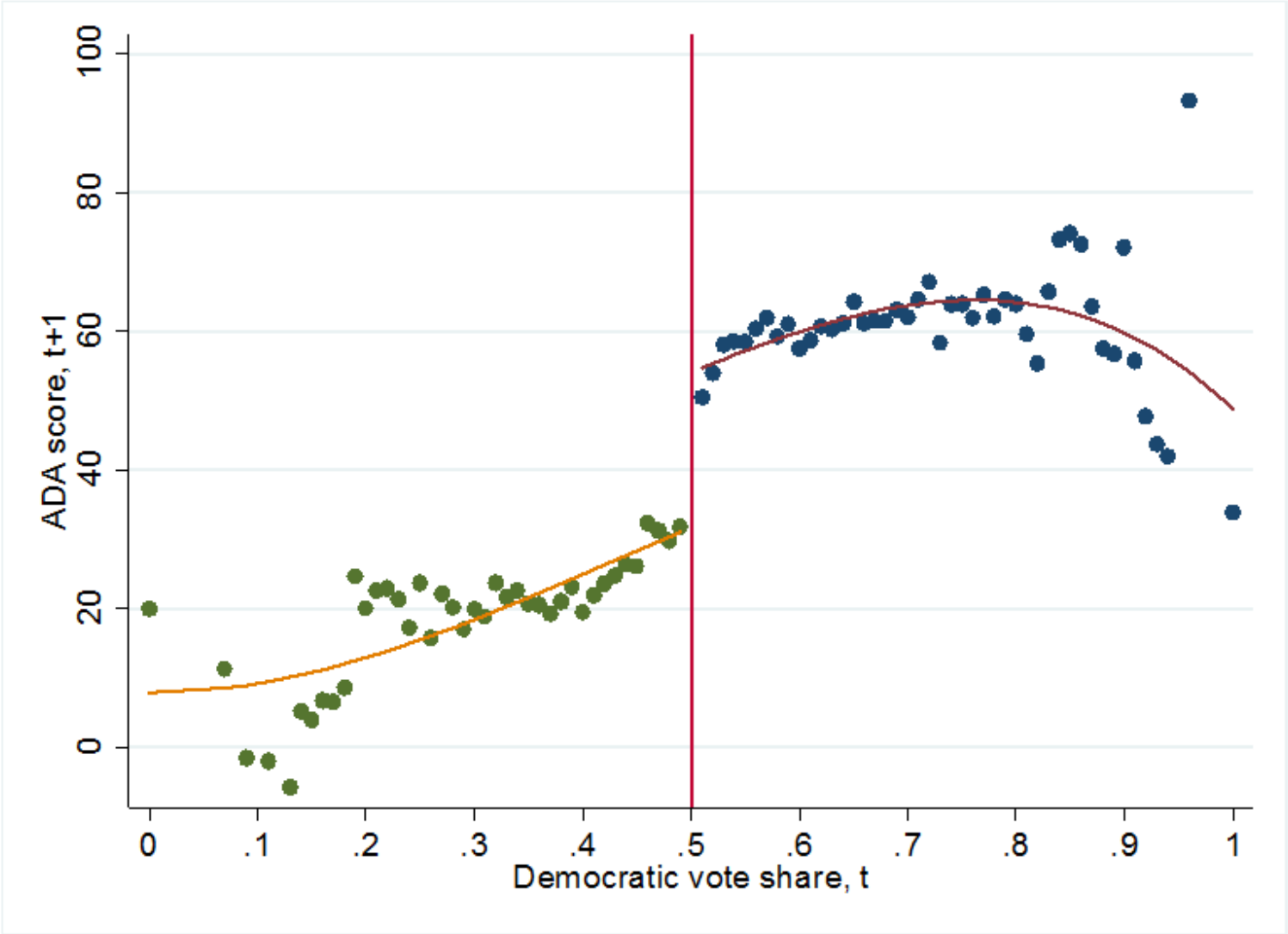
5.4 Specification Tests: Polynomial degree 2



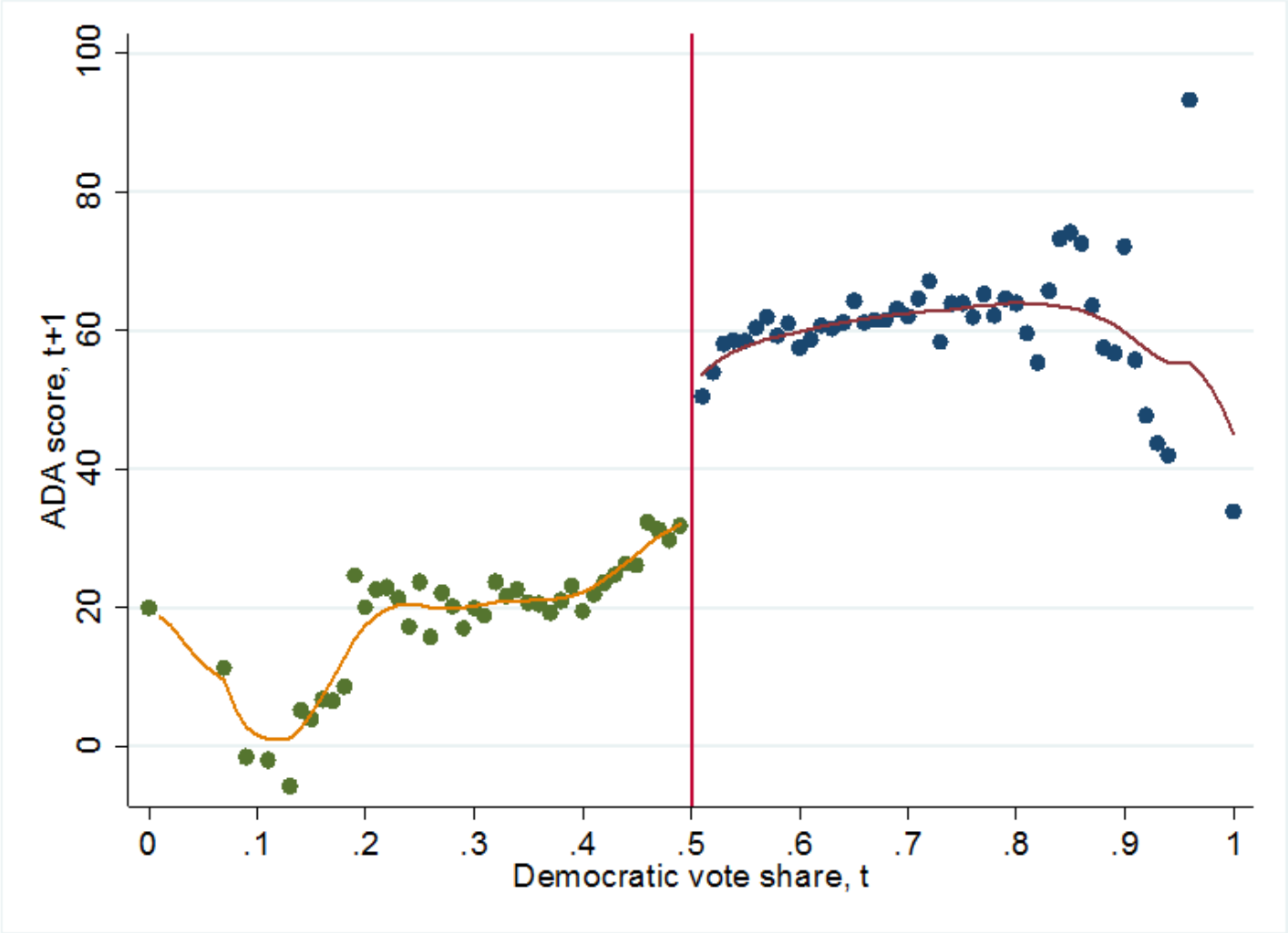
5.4 Specification Tests: Polynomial degree 3



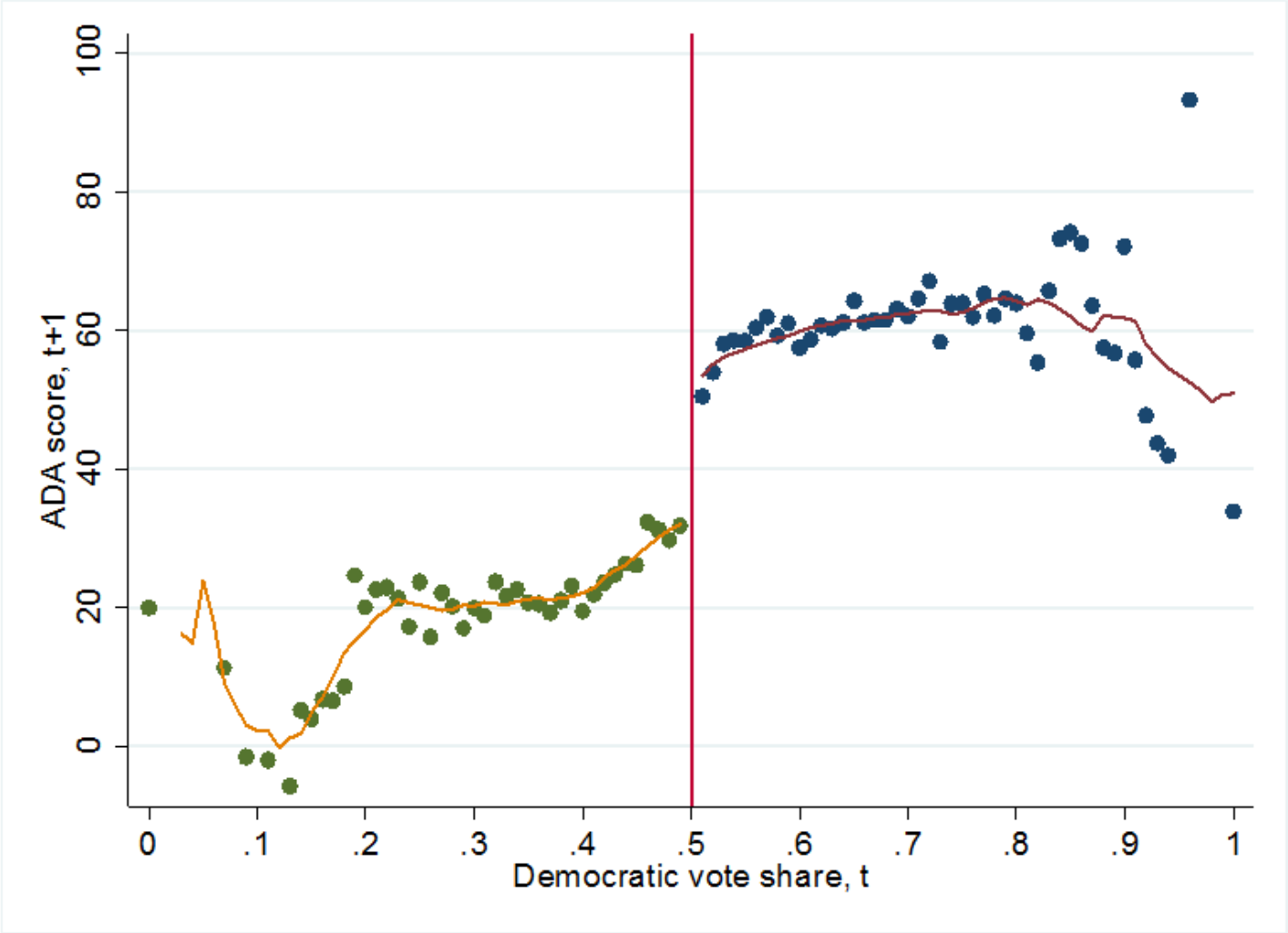
5.4 Specification Tests: Polynomial degree 4



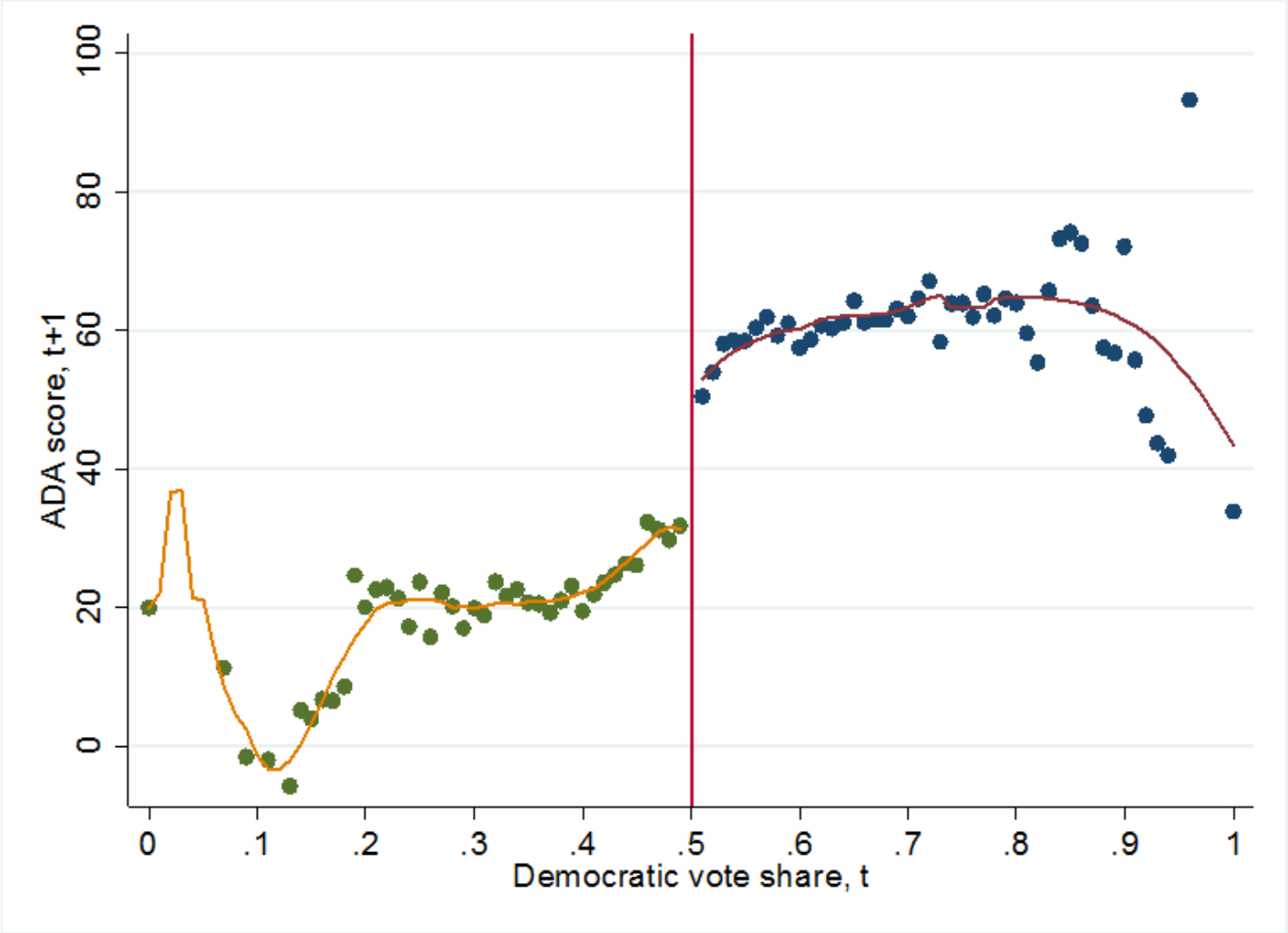
5.4 Specification Tests: Local linear regression (Triangular Kernel (1))



5.4 Specification Tests: Local linear regression (Rectang. Kernel (1))



5.4 Specification Tests: Local linear regression (Rectang. Kernel (3))



6. Evaluating RDD as a non-experimental estimator

- How good is RDD to recover experimental estimates?
- Several authors have used LaLonde's (1986) approach to evaluate the performance of RD designs to recover causal effects:
  - Buddelmeyer and Skoufias (2003): PROGRESA program in Mexico
  - Black, Galdo and Smith (2007): "Kentucky Working Profiling and Reemployment Services" experiment
  - Green et al (2009): Experimental data of experiments on voter mobilization in Michigan

7. RDD in practice!

- Checklist for RDD (Lee and Lemiux 2010):
  1. Show distribution of running variable to assess possibility of manipulation
  2. Present main RDD graphs using binned local averages
  3. Graph benchmark polynomial specification
  4. Explore sensitivity of results to the order of polynomial and bandwidth
  5. Conduct RDD analysis on covariates
  6. Explore sensitivity of results to the inclusion of covariates

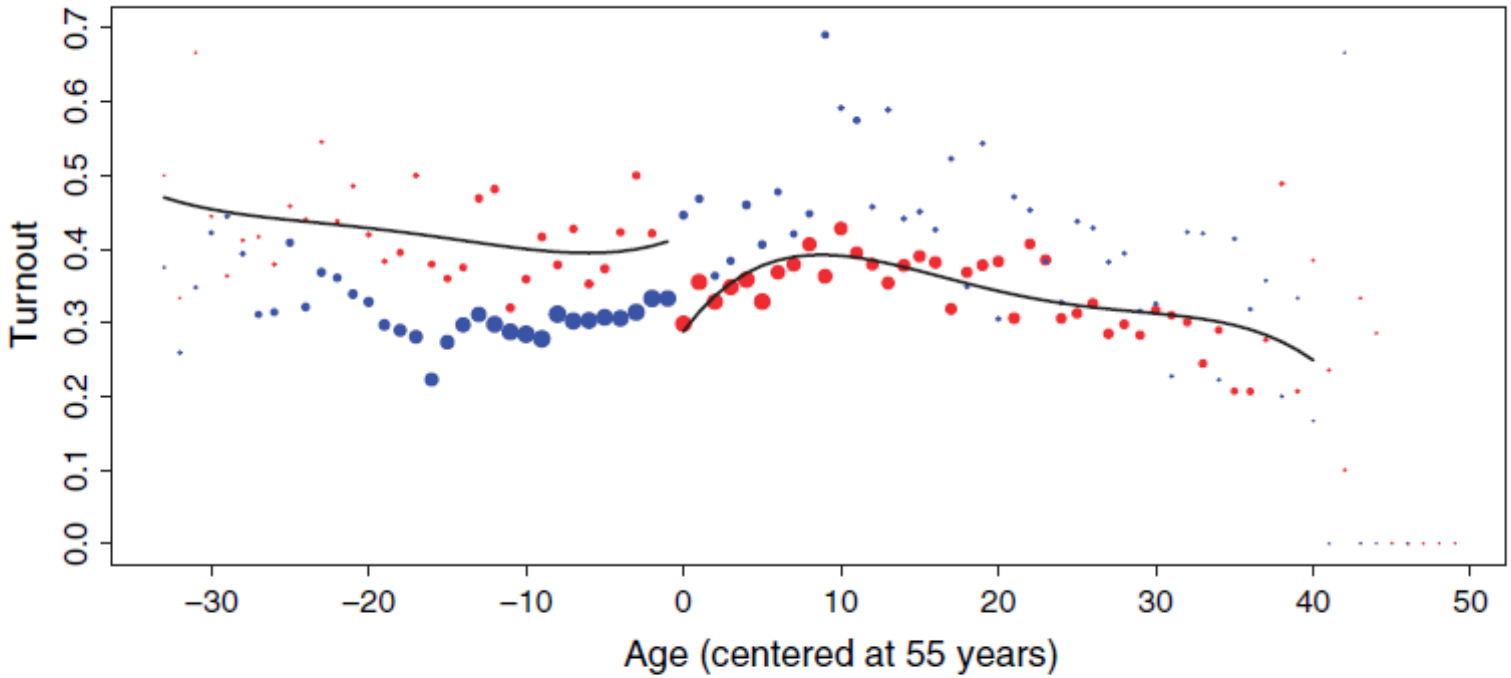


Fig. 1 Illustration of RD using age as a forcing variable. The red circles depict average voting rates among observed voters, grouped by year of age, which has been rescaled so that zero (age 55) is the point of discontinuity. The blue circles depict average voting rates among counterfactual voters. The red circles to the left of the age cutoff (where age equals 0) represent the treatment group, which received the experimental mailings. The red circles to the right of the cutoff represent the control group, which received no experimental mailings. The size of the circles is proportional to the number of observations in each age group.

Table 1 Comparison between RD estimates and experimental benchmarks, full sample

	Full Sample							
	Benchmark				*RD*			
	(1a)	*(1b)*	*(1c)*	*(1d)*	*(2a)*	*(2b)*	*(2c)*	*(2d)*
Neighbors	9.65**	10.43**	10.26**	10.0**	1.77	7.36**	10.42**	10.33**
treatment	(0.75)	(0.97)	(0.99)	(1.11)	(1.89)	(2.70)	(3.51)	(4.28)
(robust SE)								
Age		0.13**	0.039**	0.45**	−0.17**	0.66**	1.26**	1.79**
		(0.023)	(0.036)	(0.0026)	(0.042)	(0.14)	(0.31)	(0.55)
Age^2		−0.0054**	0.0016	−0.0058**		−0.025**	−0.074**	−0.12**
		(0.0012)	(0.0014)	(0.0026)		(0.0040)	(0.019)	(0.055)
Age^3			−0.00054**	−0.00072**			0.00088**	0.0029
			(0.000059)	(0.000083)			(0.00033)	(0.0020)
Age^4				0.0000097**				−0.000026
				(0.0000028)				(0.000024)
Age × Neighbors		−0.051	−0.0055	−0.032	0.032	−0.48	−0.71	−1.75
		(0.057)	(0.09)	(0.11)	(0.15)	(0.47)	(1.06)	(1.99)
Age^2 × Neighbors		−0.0031	−0.0014	0.0016		0.038**	0.13	0.083
		(0.0028)	(0.0035)	(0.0065)		(0.017)	(0.088)	(0.277)
Age^3 × Neighbors			−0.000084	−0.000016			0.00016	−0.0072
			(0.00015)	(0.00021)			(0.0020)	(0.014)
Age^4 × Neighbors				−0.0000040				−0.000066
				(0.0000071)				(0.00024)
Observations	30,038	30,038	30,038	30,038	14,717	14,717	14,717	14,717
Squared error					67.73	6.97	0.18	0.11
MSE					71.31	14.26	12.50	18.43

Note. Dependent variable is voter turnout. Age has been centered so that it is zero at the point of discontinuity (55 years). Table entries are least squares regression estimates and robust SEs. Squared error is the squared difference between the estimates in columns 2a–d and the benchmark estimate of 10.0. MSE is mean squared error, defined as squared error plus the square of the SE.

*$p < .10$, **$p < .05$ (two-tailed test).