Regression Discontinuity

NYU Wagner

Rajeev Dehejia

Today’s agenda

I. Methodological Overview

II. Discussion of Exemplary Papers

III. Practitioner’s Guide

IV. Stata examples

Methodological Overview

Important references

- RD methods were introduced by Thistlewaite and Campbell (1960). See Cook (2008) for an historical perspective.
- Recent applications in politics include analyses of the incumbency effect (Lee, 2008), electoral competition on policy (Lee, Moretti, and Butler 2004), the effect of gender on legislator behavior (Rehavi nd), the value of a seat in the legislature (Eggers and Hainmueller 2009), the effect of mayoral party ID on policy (Ferreria and Gyourko 2009).
- Recent important theoretical work has dealt with identification issues (Hahn, Todd, and Van Der Klaauw, 2001), optimal estimation (Porter, 2003), tests for validity of the design (McCrary, 2008), bandwidth selection (Imbens and Kalyanaraman 2011).
- General surveys include Lee and Lemieux (2009), Van Der Klaauw (2008), and Imbens and Lemieux (2008).

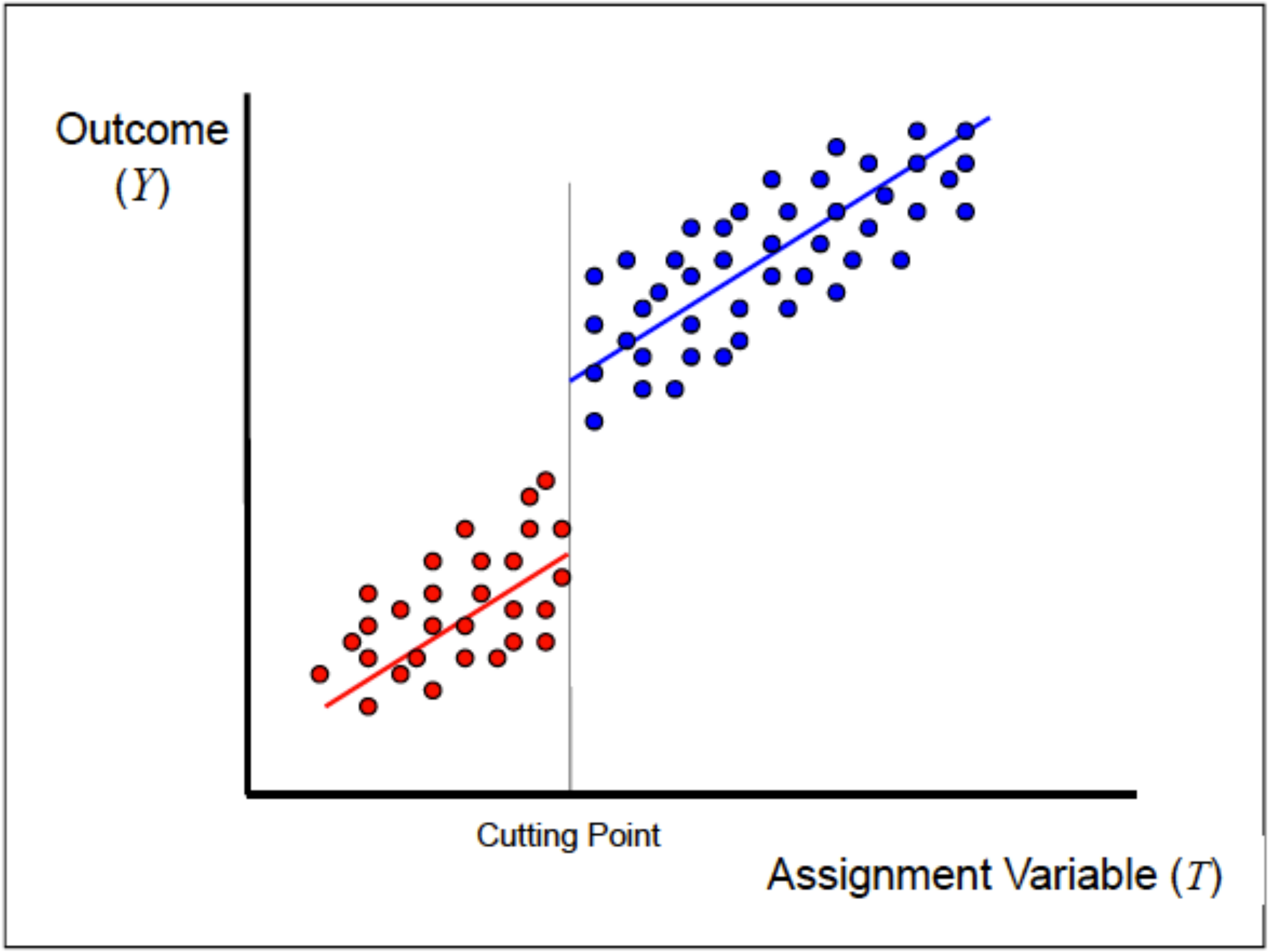
Current Events and Papers

- “Setting a Good Example? Examining Sibling Spillovers in Educational Achievement Using a Regression Discontinuity Design”- NBER Working Paper (2019); Krzysztof Karbownik, Umut Özek
- “The causal effects of R&D grants” Pietro Santoleri, Andrea Mina (19 July 2020)
- “Regression Discontinuity Model for TV Series” Arthur Charpintier (Dec 2020)- Note: Blog Post/hypothesis on Freakonometrics (more on this on R-bloggers)

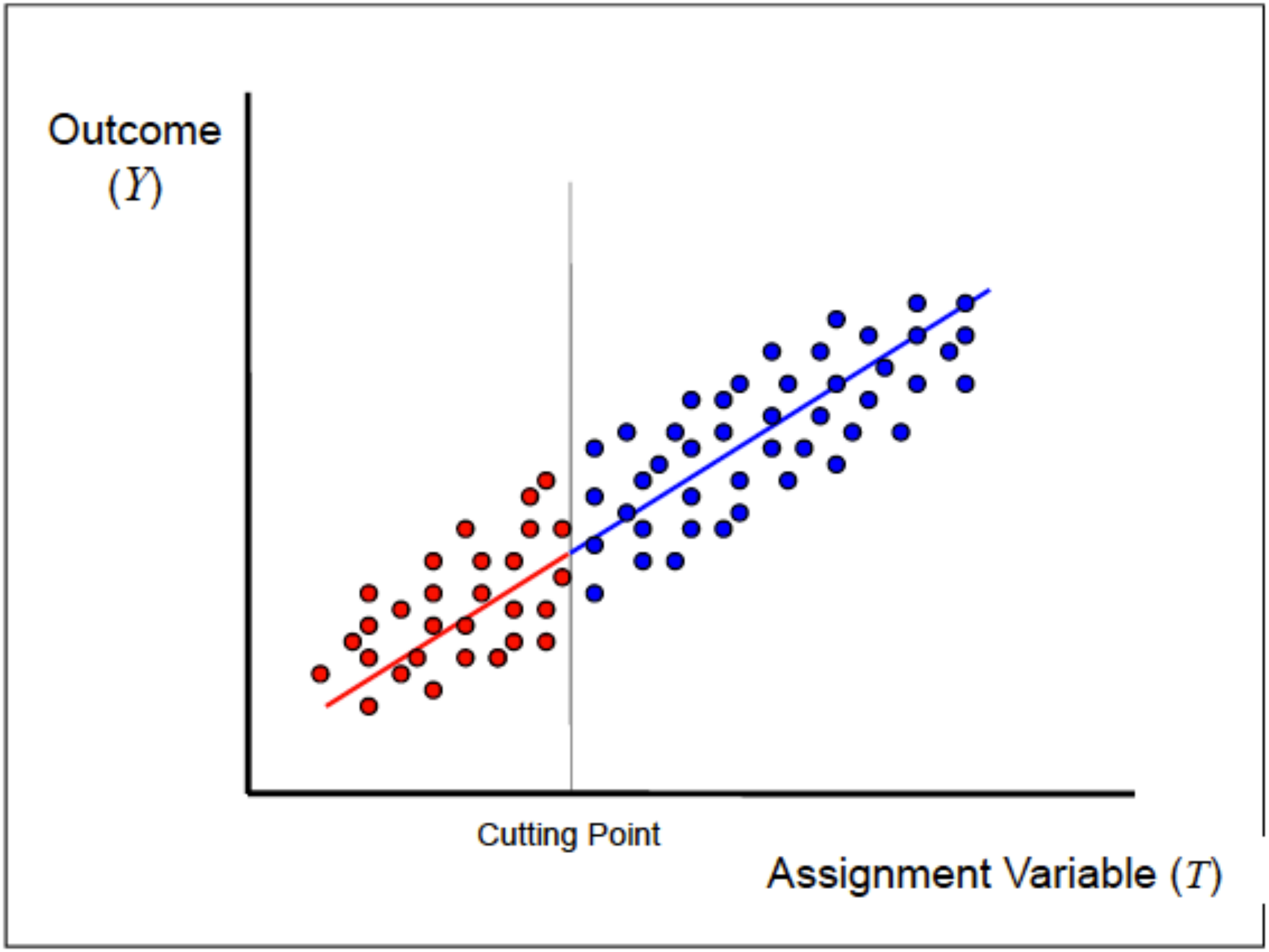
Regression discontinuity basics

- The basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of an assignment (or forcing) variable (the covariate X) being on either side of a fixed threshold.
  - Assignment to treatment by covariate value, assign all units with $X_i \geq c$ to treatment
- RD estimates the local average treatment effect (LATE) of the treatment at $X = c$
- RD is like a randomized experiment at the cutpoint $X = c$
- The RD design is generally regarded as having the greatest internal validity of all quasi-experimental methods. Its external validity is more limited, since the estimated treatment effect is local to the discontinuity.

RD scatterplot: positive treatment effect



RD scatterplot: no treatment effect



RD setup formally

Start with the usual potential outcomes framework:

- $Y_i(1) - Y_i(0)$ = Unit level treatment effect
- $Y_i(1)$ = Outcome that would occur if the unit were exposed to treatment
- $Y_i(0)$ = Outcome that would occur if the unit were not exposed to treatment
- The familiar problem is that **we cannot observe the pair $Y_i(0)$ and $Y_i(1)$ simultaneously**

RD setup features and assumptions:

- T_i = binary treatment variable (i.e. $T = 1$ = *treated;* $T = 0$ = *not treated*)
- X_i = continuous assignment (or forcing) variable
- In a **sharp regression discontinuity** design:

$$\boldsymbol{T_i = 1\{X_i \geq c\}}$$

Where c is a cutoff such that all units above receive the treatment and all those below do not.

Defining the causal estimate

Under fairly weak assumptions, the effect of the treatment is identified by:

$$\begin{aligned}&\lim_{\varepsilon\downarrow 0} E\left[Y_i \middle| X_i = c+\varepsilon\right] - \lim_{\varepsilon\uparrow 0} E\left[Y_i \middle| X_i = c+\varepsilon\right] \\ &= \lim_{\varepsilon\downarrow 0} E\left[Y_{1i} \middle| X_i = c+\varepsilon\right] - \lim_{\varepsilon\uparrow 0} E\left[Y_{0i} \middle| X_i = c+\varepsilon\right]\end{aligned}$$

Which is: $\tau_{\mathrm{RD}} = \mathrm{E}[Y_{\mathrm{i}}(1) - Y_{\mathrm{i}}(0) \mid X_{\mathrm{i}} = c]$.

This is the **average treatment effect *at the cutoff point***, a very specific LATE.

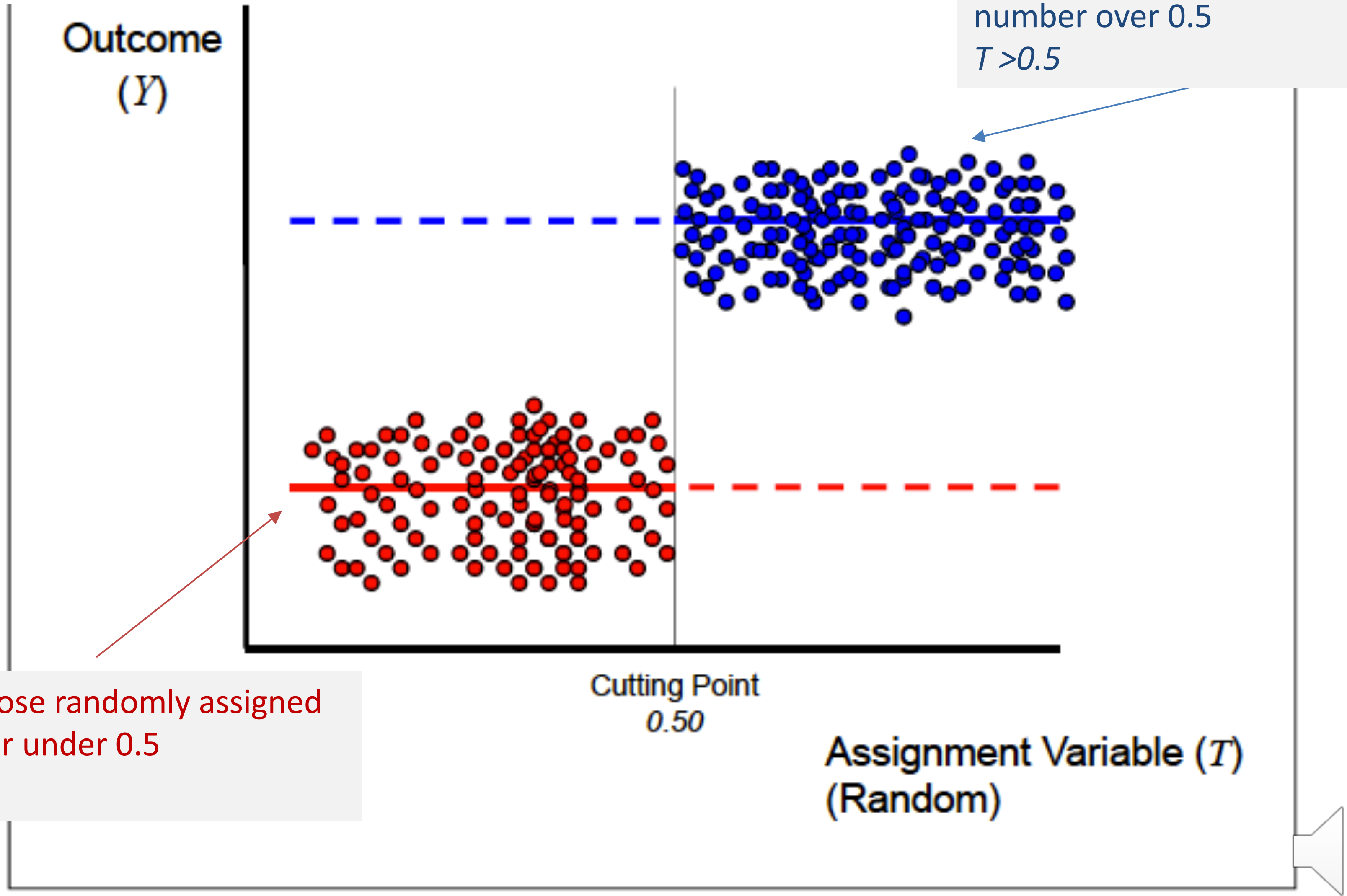
This inference is possible because the functions E[Y(1)|X] and E[Y(0)|X] are (assumed to be) continuous in all variables except running variable.

Some extrapolation is required because by design there are no units with X_i=c for whom we observe Y_i(0) .

RD compared to an experiment

- RD is often described as a “close cousin” of a randomized experiment or as a “local randomized experiment.”
- Coughey and Sekhon argue against this conceptualization, for reasons we will see later, but it worth understanding why the analogy is made
- Consider an experiment in which each participant is assigned a randomly generated number, v, from a uniform distribution over the range [0,1].
  - Units with $v \geq 0.5$ assigned to treatment; units with $v < 0.5$ assigned to control
- This randomized experiment can be thought of as an RD where the assignment variable $X = v$ and the cutoff is at 0.5.
  - Because the assignment variable is random, the curves $E[Y(0)|X]$ and $E[Y(1)|X]$ are flat. And we know that they are flat.
  - The ATE can be computed as the difference in the mean value of Y on either side of the cutoff.
  - Because the functions are flat everywhere, the “optimal bandwidth” is to use all the data

Experiment as RD



E.g. Those randomly assigned number over 0.5
T >0.5

E.g. Those randomly assigned number under 0.5
T < 0.5

RD compared to an experiment

- Note that there are **two main ways in which RD differs** from a randomized experiment in actuality
  - The functional form of E[Y(1)|X] and E[Y(0)|X] need not be flat (or linear or monotonic) and may not even be known.
  - It may be **possible for units to alter their assignment to treatment** by manipulating the forcing variable in a way that is not possible when it is assigned at random by the investigator.

Estimation basics

- We have now defined a causal effect as the difference of two functions at a point. How do we estimate that? There are 3 general approaches.

- Approach #1: Compare means
- Approach #2: OLS (with polynomials)
- Approach #3: Local Linear Regression

Estimation basics: Approach # 1, compare means

- In the data, **we never observe $\mathbf{E}[\boldsymbol{Y}(0)|\boldsymbol{X}=\boldsymbol{c}]$**
  - There are no units at the cutoff that don’t get the treatment, but in principle it can be approximated arbitrarily well by $E[Y(0)|X = c - \varepsilon]$.

- Therefore we estimate:

$$E[Y \mid X = c + \varepsilon] - E[Y \mid X = c - \varepsilon]$$

(Mean dependent variable observed for units above the cutoff) – (Mean dependent variable observed for units below the cutoff)

  - This is the difference in means for those just above and below the cutoff.

- This is a nonparametric approach. A great virtue is that it does not depend on correct specification of functional forms (e.g. linear, quadratic, etc.)

- Note that we said “in principle” we can estimate means arbitrarily close to the cutoff. In practice, this depends on having lots of data within ε of the cutoff. Suppose you don’t….

Estimation basics: Approach # 2, OLS with polynomials

- The original RD design (Thistlewaite and Campbell 1960) was implemented by OLS:

$$Y = \alpha + \tau T + \beta X + \eta$$

  - Where τ is the causal effect of interest and *η* is an error term.
- This regression distinguishes the nonlinear and discontinuous jump from the smooth linear function.
- OLS with one linear term in *X* is seldom used anymore because the functional form assumptions are very strong.
- What are they?
  - *See Appendix of these slides for polynomial and logarithmic function review*
  - *Note: OLS Assumptions still apply to polynomial regressions because coefficients remain linear (that is, polynomials are applied to variables, not their coefficients)*

Estimation basics: Approach # 2, OLS with polynomials

- Suppose the underlying functions are nonlinear and maybe unknown. In particular, suppose you want to estimate

  $$Y = \alpha + \tau T + f(X) + \eta$$

  where $f(X)$ is a smooth nonlinear function of X.

- Perhaps the simplest way to approximate $f(X)$ is via OLS with polynomials in X. Common practice is to fit different polynomial functions on each side of the cutoff by including interactions between T and X.

- Modeling $f(X)$ with a ***p***th-order polynomial in this way leads to

  $$Y = \alpha + \beta_{01}X + \beta_{02}X^2 + \ldots + \beta_{0p}X^p +$$
  $$\tau T + \beta_1 TX + \beta_2 TX^2 + \ldots + \beta_p TX^p + \eta$$

- Centering X at the cutoff prior to running the regression ensures that the coefficient on T is the treatment effect.

  $$Y = \alpha + \beta_{01}X + \beta_{02}X^2 + \ldots + \beta_{0p}X^p +$$
  $$\tau T + \beta_1 T(X-c) + \beta_2 T(X-c)^2 + \ldots + \beta_p T(X-c)^p + \eta$$

Estimation basics: Approach # 2, OLS with polynomials

- Common practice, for whatever reason, seems to use a 4th order polynomial, though you should be sure that your results are robust to other specifications (more on this below).
- OLS with polynomials is a particularly simple way of allowing a flexible functional form of *X*. A drawback is that it provides global estimates of the regression function that use data far from the cutoff.
- There are many other ways, but the RD setup poses a couple of problems for standard nonparametric smoothers.
  - We are interested in the estimate of a function at a boundary point. (For why this is a problem, see Imbens and Lemieux.)
  - Standard nonparametric kernel regression does not work well here

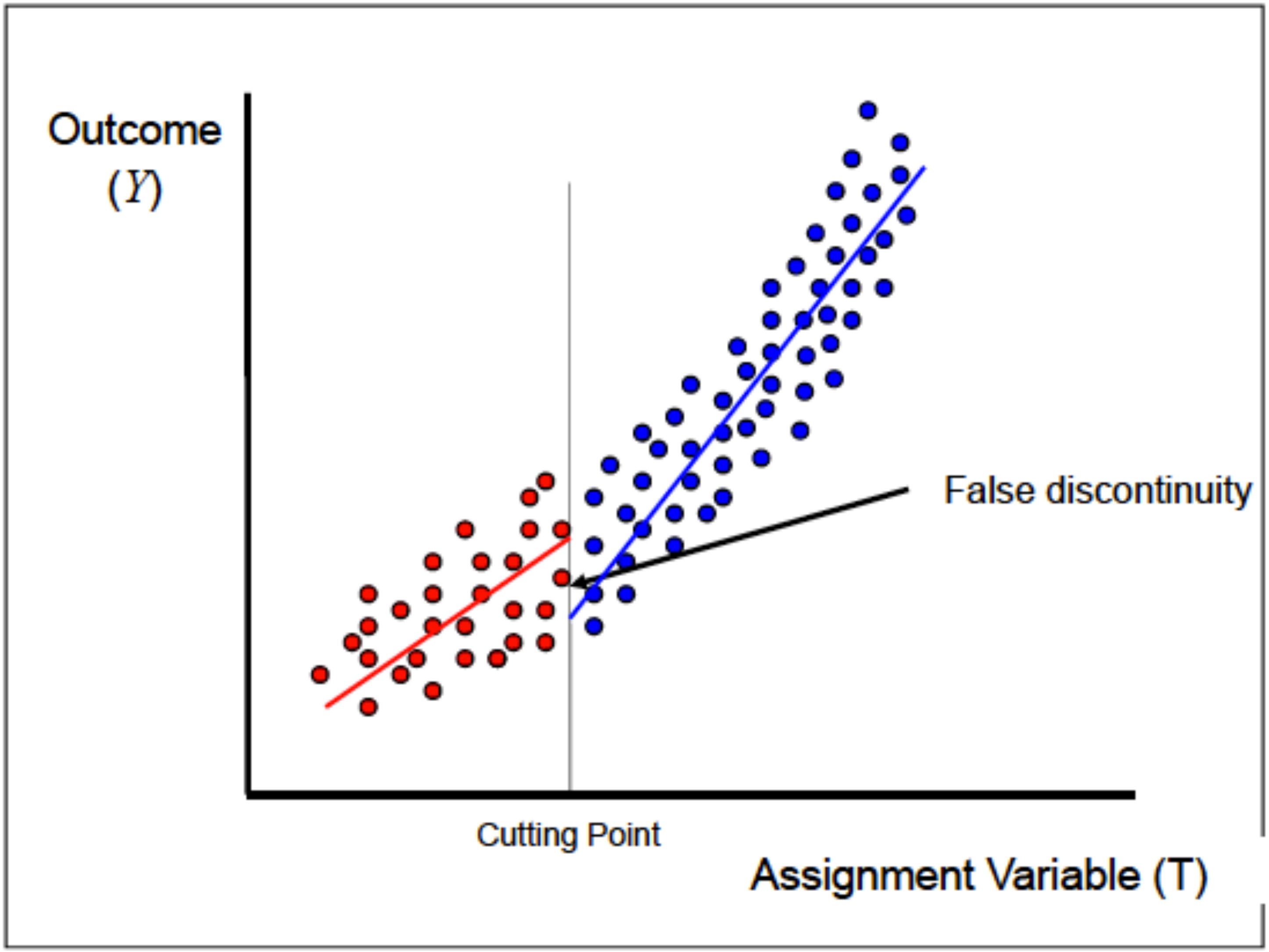
Estimation basics: Approach 3, Local Linear Regression

- Instead of locally fitting a constant function (e.g., the mean), fit linear regressions to observations within some bandwidth of the cutoff
- A rectangular kernel seems to work best (see Imbens and Lemieux) but optimal bandwidth selection is an open question
- A serious discussion of local linear regression is beyond the scope of this lecture. See, for example, Fan and Gijbels (1996)
- But really, we're just talking about running regressions on data near the cutoff.

RD pitfall: mistaking nonlinearity for discontinuity

- Consequences of using an incorrect functional form are potentially more severe for RD than for other methods we study
- Misspecification of the functional form may generate a bias in the treatment effect
- The most common situation of this type is when an unaccounted for nonlinearity in the conditional mean function is mistaken for a discontinuity
- Each of the 3 estimation methods deals with this issue in a different way

Nonlinearity mistaken for discontinuity



Nonlinearity mistaken for discontinuity

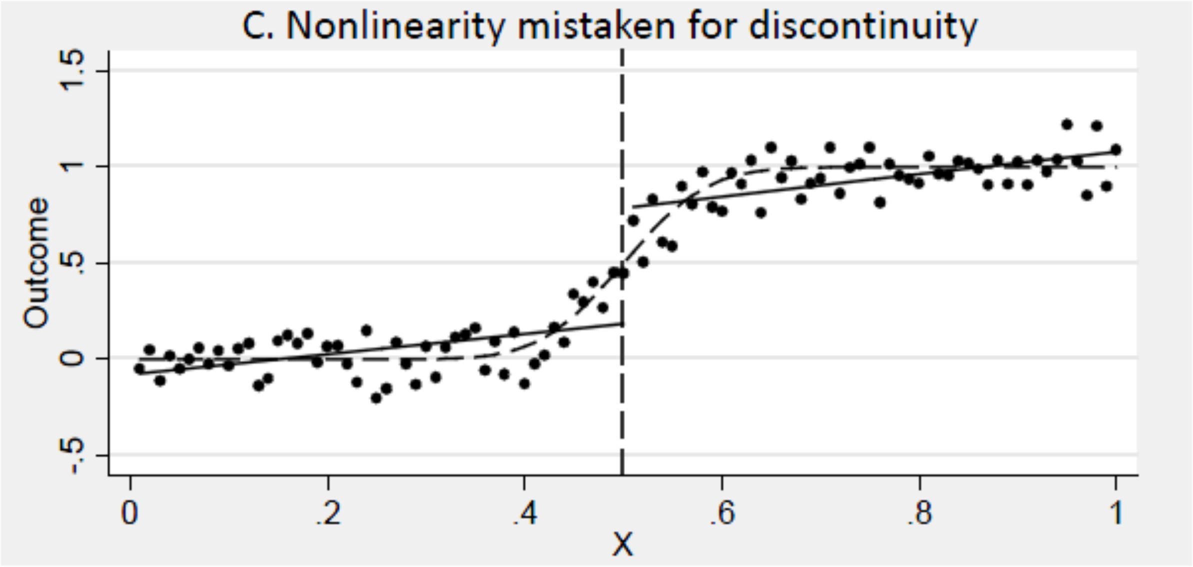
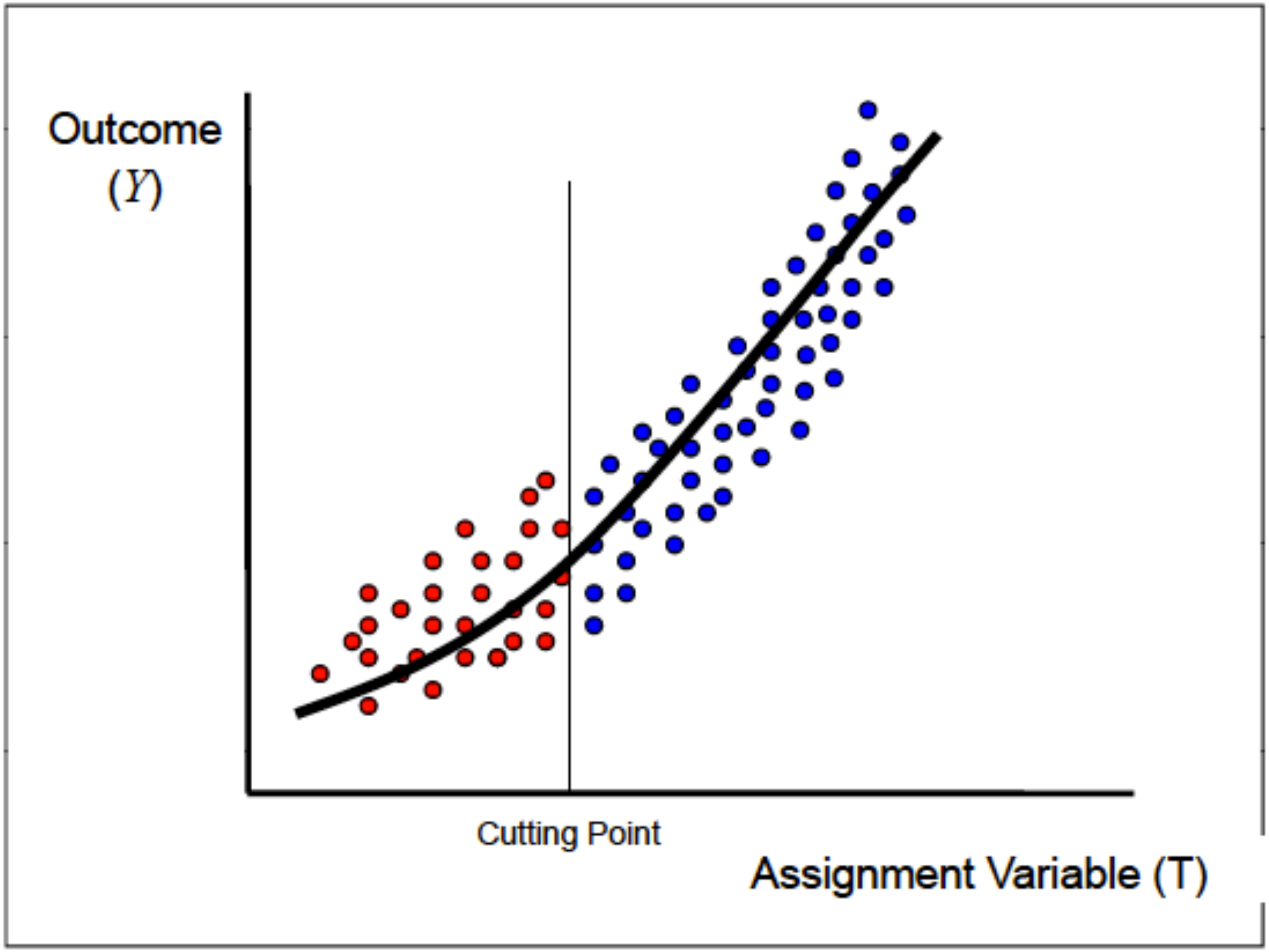


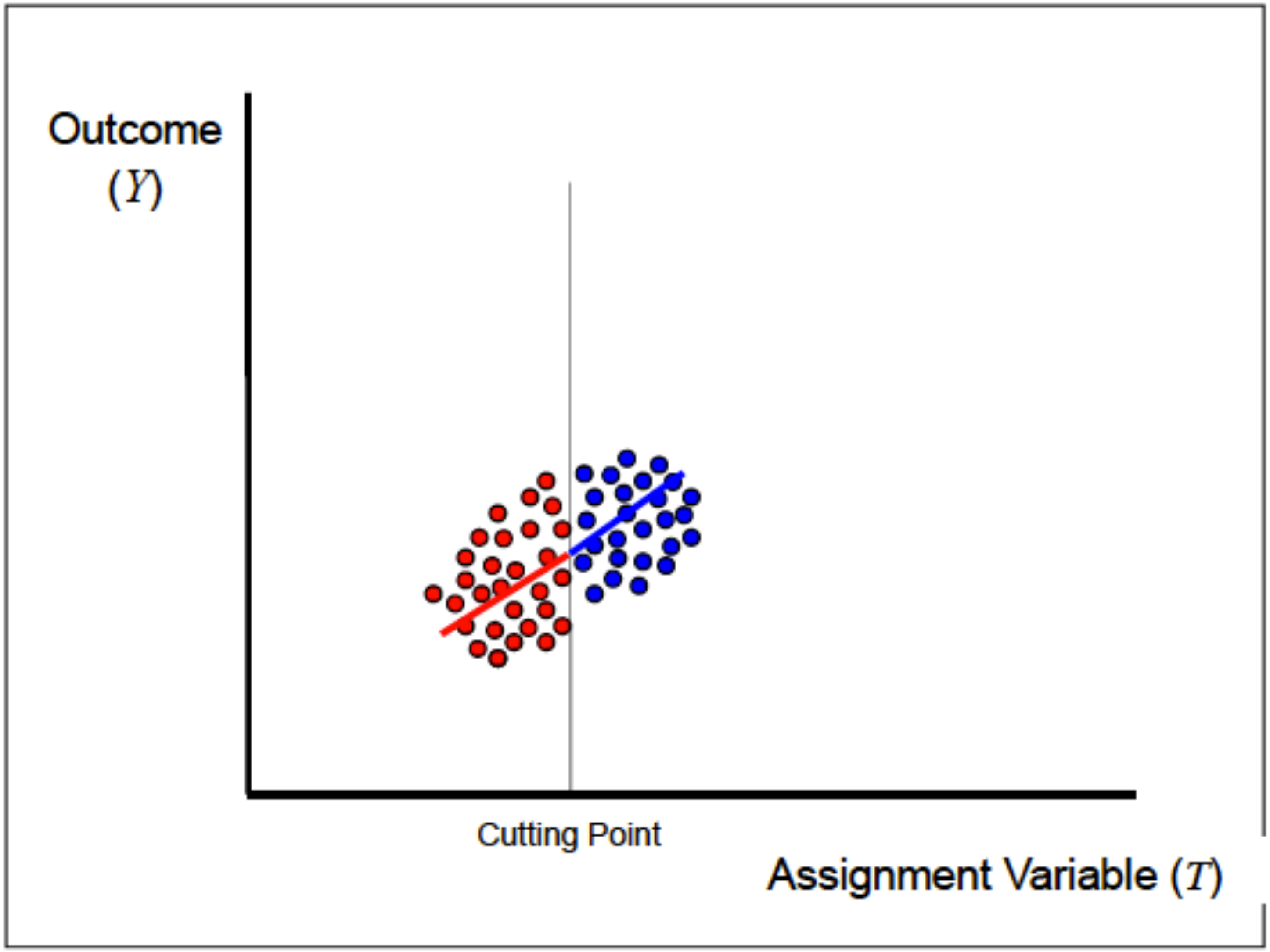
Figure from Angrist and Pischke (2008)

- A mis-specified functional form can lead to a spurious jump
- Check sensitivity to more flexible specifications

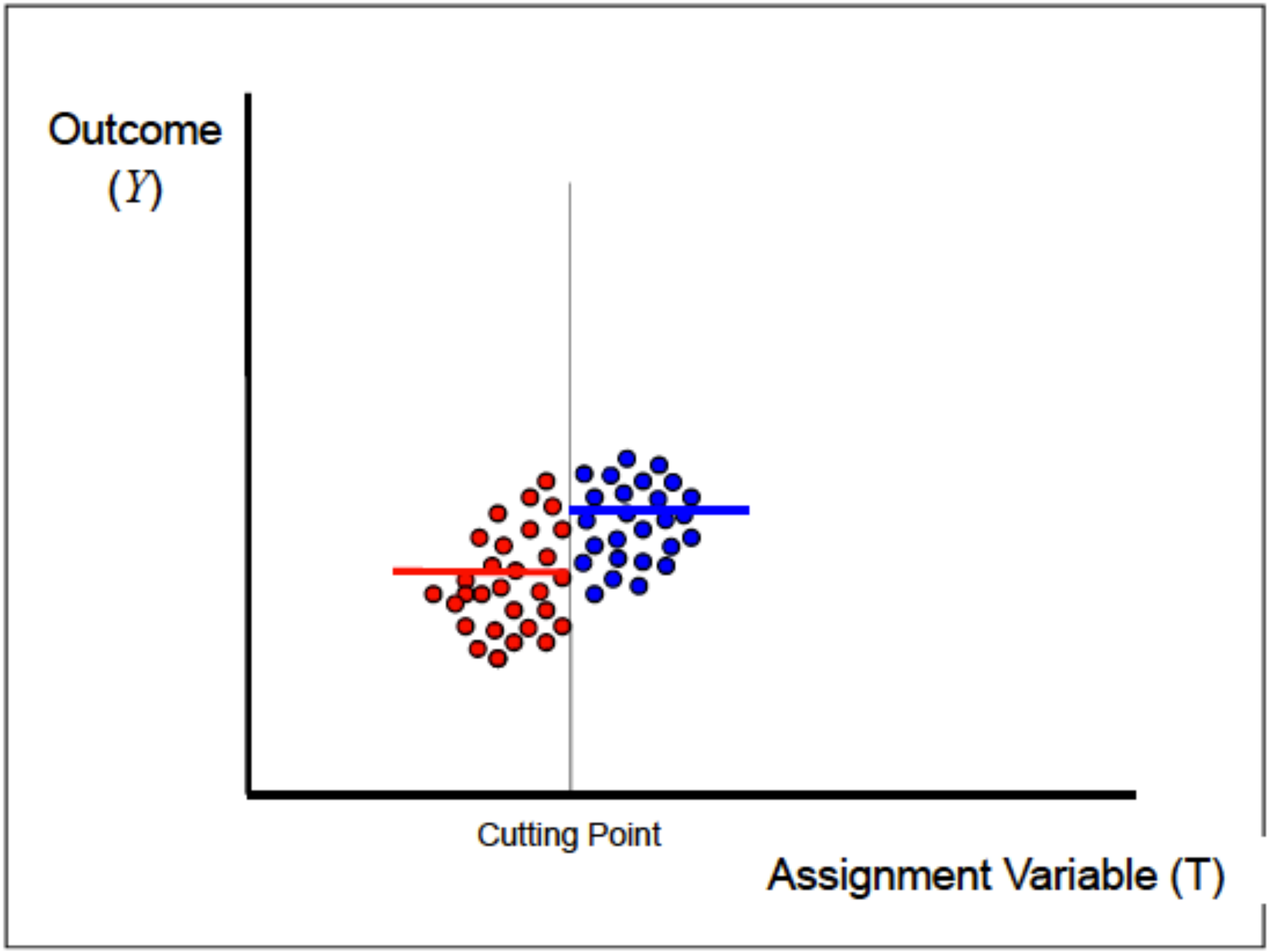
Polynomial regression (Approach 2)



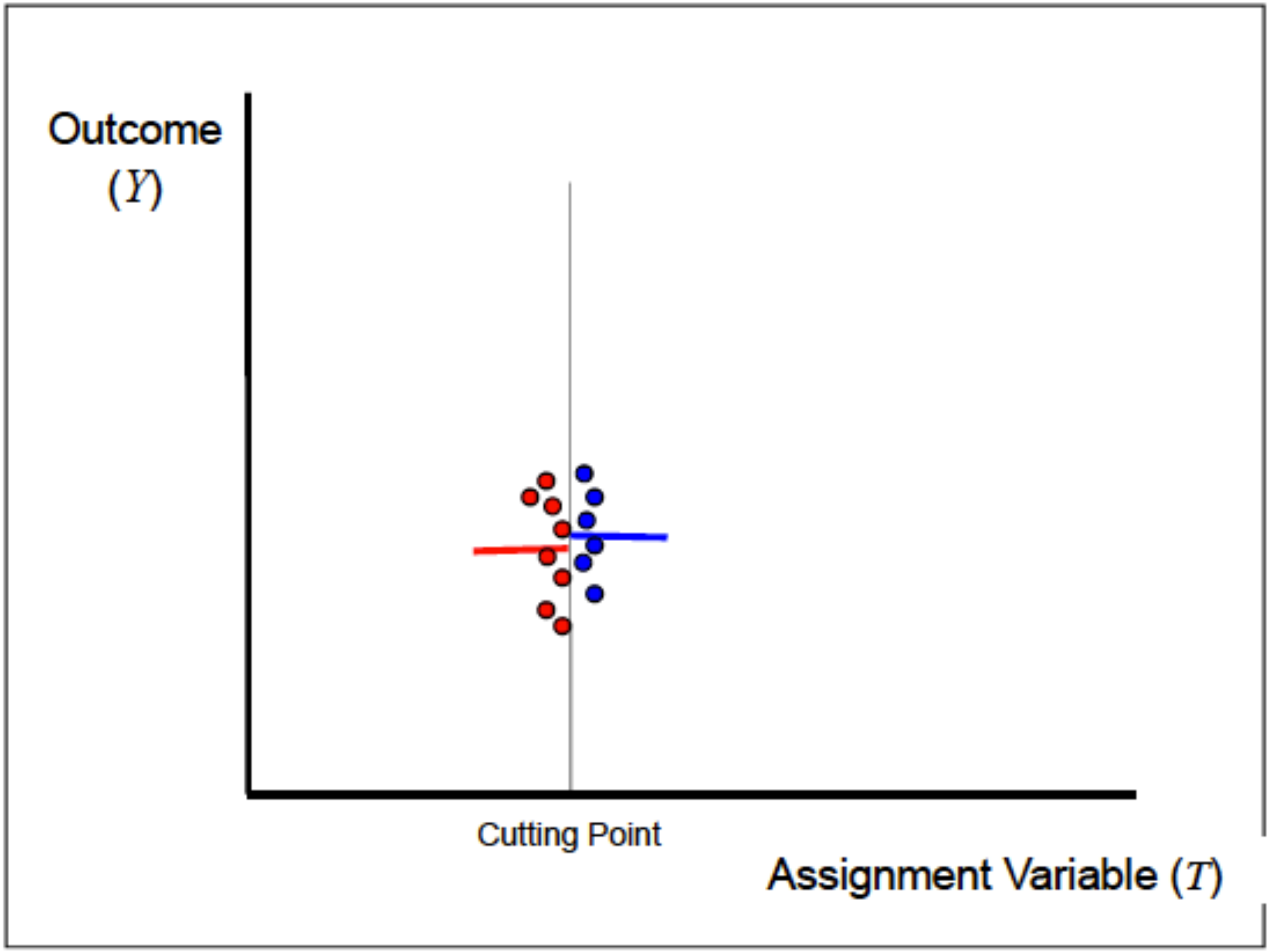
Local linear regression (Approach 3)



Compare means (Approach 1)



Compare means: smaller bandwidth (Approach 1?)



The role for covariates in RD

- In principle, covariates are not needed for identification in RD, but they can help reduce sampling variability in the estimator and improve precision
  - This is a standard argument which also supports inclusion of covariates in analyses of randomized trials
- Adding covariates should not affect the point estimate of the effect (very much). If it does, there is a problem.
- The wider the bandwidth the more important it may be to include covariates.

Graphical analysis

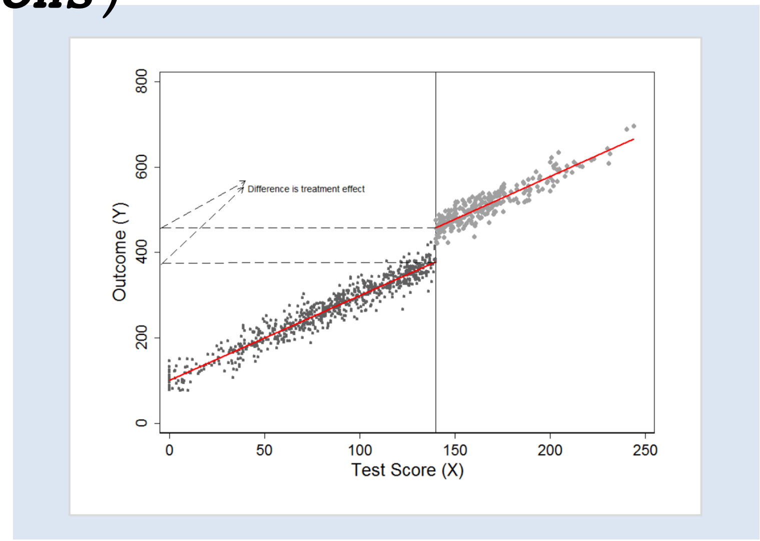
- **Graphical inspection is an integral part of any RD analysis**.
- 3 types of graphs should **always** be produced, where assignment variable is graphed against:
  1. The outcome
  2. Other covariates
  3. Density of cases
- 1 should show a discontinuity; 2 and 3 should show no discontinuity
- If you can't see the main result with such a simple graph, it's probably not there
- If you see a discontinuity in 2 or 3, be worried

1. Graphical analysis: The outcome

Assignment variable is graphed against the outcome

- Should show a discontinuity
- If you can’t see the main result with such a simple graph, it’s probably not there
- What might this look like?

```
Twoway (scatter yvar runningvar … code for
graphical design options)
```



2. Graphical analysis: Covariates

Assignment variable is graphed against other covariates

- Should show no discontinuity
- If you see a discontinuity, it means that you should be worried about systematic bias in the assignment variable (e.g. selection by covariates, non-compliance)
- What might this look like?
  - `Twoway (scatter` *`covar1 runningvar`* … ***`code for graphical design options`***`)`

3. Graphical analysis: Density

Assignment variable is graphed against the density of cases

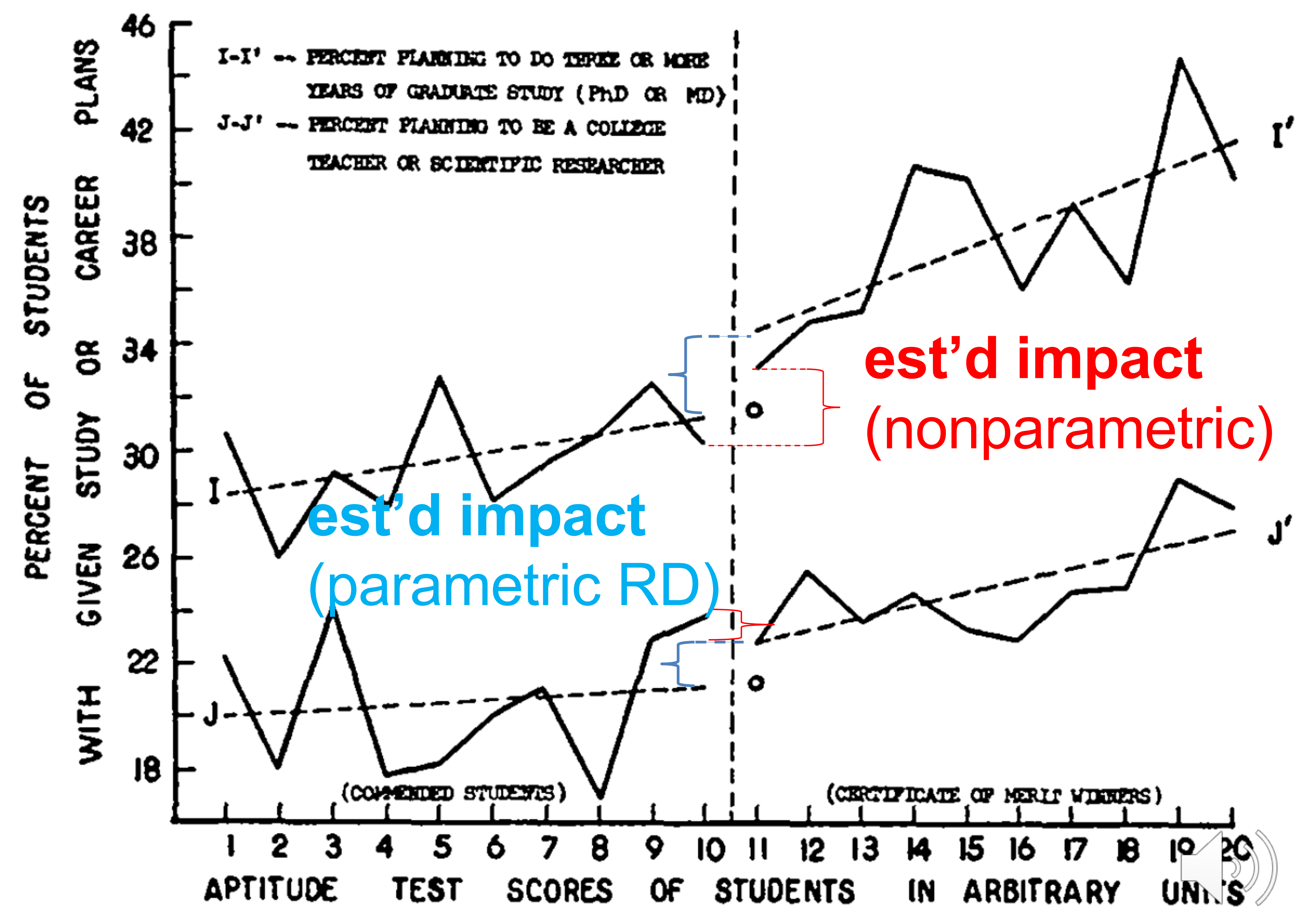
- Should show no discontinuity
- Here, you are testing for a discontinuity of cases in the running variable
- If there is a sharp increase in the number of observations either right above or right below the cut-point, it suggests that either the placement of the cut-point or the running variable itself has somehow been manipulated
- What might this look like?
  - `rddensity` *runningvar* … ***code for graphical design options****)*

Fuzzy RD

- Suppose the probability of treatment changes discontinuously at a threshold but not from 0 to 1. This is a situation for applying FRD.
  - Note that the "fuzziness" in FRD comes from the change in probability of treatment, not fuzziness about the threshold
- In sharp RD designs, the jump in the outcome at the cutoff is the estimate of the causal impact of the treatment. In a FRD design, the jump in the outcome is divided by the jump in the probability of treatment at the cutoff to produce the local Wald estimate (equivalent to a local IV estimate) of the causal impact. (In SRD, the jump is one, so the division is inconsequential, but this demonstrates the relationship between SRD and FRD.)
- The important point to remember is that fuzzy RD is numerically equivalent (and conceptually similar) to IV (see *Mostly Harmless* sec. 6.2). Think about the 'two stages'
- An additional graph is needed when doing FRD: probability of treatment by assignment variable should show a discontinuous probability at the threshold.

Example 1: impact of being a national Merit semi-finalist

- The original RD!
  - Campbell and Thistlewiate, 1960
- Criteria: be above a PSAT test score threshold
  - In a narrow range around the threshold, students should be close to identical (score differences mostly due to chance!)
  - Non-parametric RD: compare average outcomes in a narrow band on either side of the threshold
  - Parametric RD: fit a smooth of outcome as a function of test score (the “running variable”) and allow it to “jump” at the threshold (what they did)



est'd impact
(nonparametric)

est'd impact
(parametric RD)

Example 2
“Do Voters Affect or Elect Policies?”
by Lee, Moretti, and Butler (LMB)

Motivation

- Motivation: Two fundamentally different views of the role of elections
  - Convergence: Competition for votes drives candidates to seek middle ground policies, compromise (median voter theorem). Voters affect policy choices of politicians.
  - Divergence: Voters select candidates, who then enact their own preferred policies. Voters elect policies.
  - Which of the two we see in practice depends on whether politicians can make credible promises to implement policies that are not at their own bliss point (credible commitments are facilitated by repeat interactions)
- The goal of the paper is to examine which phenomenon is more empirically relevant for US politics, specifically voting in the House

Estimating framework

- The roll-call voting record of the representative in the district following election *t* is
  - $RC_t = (1-D_t)y_t + D_t x_t$
  - Where *y* is the Republican policy and *x* is the Democratic policy
  - D_t is the indicator for whether a democrat won. That is, only the winning candidate's policy is observable
  - RC_t = Roll-call voting partisanship record in time period *t* (0-100, where 100 is most liberal record)
- The expression can be transformed into:

$$(2) \quad RC_t = \text{constant} + \pi_0 P_t^* + \pi_1 D_t + \varepsilon_t$$

$$(3) \quad RC_{t+1} = \text{constant} + \pi_0 P_{t+1}^* + \pi_1 D_{t+1} + \varepsilon_{t+1},$$

where P^* is the underlying (true) popularity of party *D*.

- π_0 captures the effect of underlying policy preference (of voters in a district) on both *x* and *y* (some districts more liberal or conservative).
- It also allows an independent effect of party, π_1, as a shifter.

Estimating framework cont’d

- We cannot observe P^* so equation (2) cannot be directly estimated
- But suppose we could randomize D_t. Then D_t would be independent of P^*_t and ε_t. Then:

$$(4)\quad \underline{E\left[RC_{t+1}\middle|D_t=1\right]-E\left[RC_{t+1}\middle|D_t=0\right]=}\ \pi_0\left[P^{*D}_{t+1}-P^{*R}_{t+1}\right]$$

$$\underline{+\pi_1\left[P^D_{t+1}-P^R_{t+1}\right]=\gamma}$$

$$(5)\quad \underline{E\left[RC_t\middle|D_t=1\ \right]-E\left[RC_t\middle|D_t=0\right]}=\pi_1$$

$$(6)\quad \underline{E\left[D_{t+1}\middle|D_t=1\right]-E\left[D_{t+1}\middle|D_t=0\right]}=P^D_{t+1}-P^R_{t+1},$$

No π_0 because of random assignment

- where P^D_{t+1} is the probability of a Democrat at t+1 if there was a Republican at t
- Everything underlined in red can be estimated from the data.
- In fact, we don’t randomly assign, but rely on close elections for as-if randomization or RD.

Why this works

- The "elect" (or party) component is $\pi_1\left[P_{t+1}^{D} - P_{t+1}^{R}\right]$
- π_1 is estimated by the difference in voting records between the parties at time *t* (from random assignment at *t*)
- The fraction of districts won by Democrats in *t*+1 in Democratic vs non-Democratic seats at *t* is an estimate of
$$\left[P_{t+1}^{D} - P_{t+1}^{R}\right]$$
- Because we estimate the total effect, *γ*, of a Democratic victory in *t* on RC_{t+1}, we can then net out the elect component to implicitly get the affect (competition) component
- Random assignment of D_t is crucial. Without it, equation (5) would reflect π_1 and that Democratic districts have more liberal policy preferences

Graphical estimate of equation 4

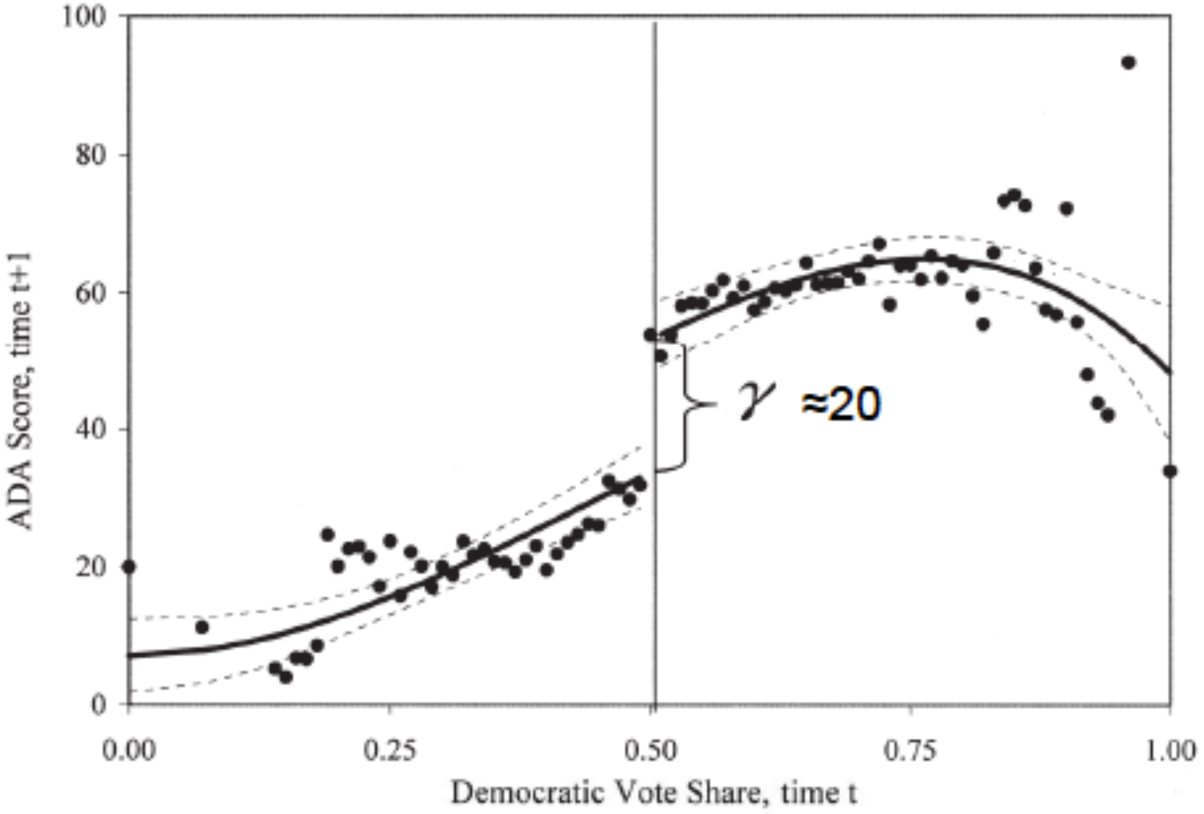


FIGURE I

Total Effect of Initial Win on Future ADA Scores: γ

This figure plots ADA scores after the election at time $t + 1$ against the Democrat vote share, time t. Each circle is the average ADA score within 0.01 intervals of the Democrat vote share. Solid lines are fitted values from fourth-order polynomial regressions on either side of the discontinuity. Dotted lines are pointwise 95 percent confidence intervals. The discontinuity gap estimates

$$\gamma = \underbrace{\pi_0(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Affect”}} + \underbrace{\pi_1(P^{*D}_{t+1} - P^{*R}_{t+1})}_{\text{“Elect”}}.$$

Graphical estimate of equation 5

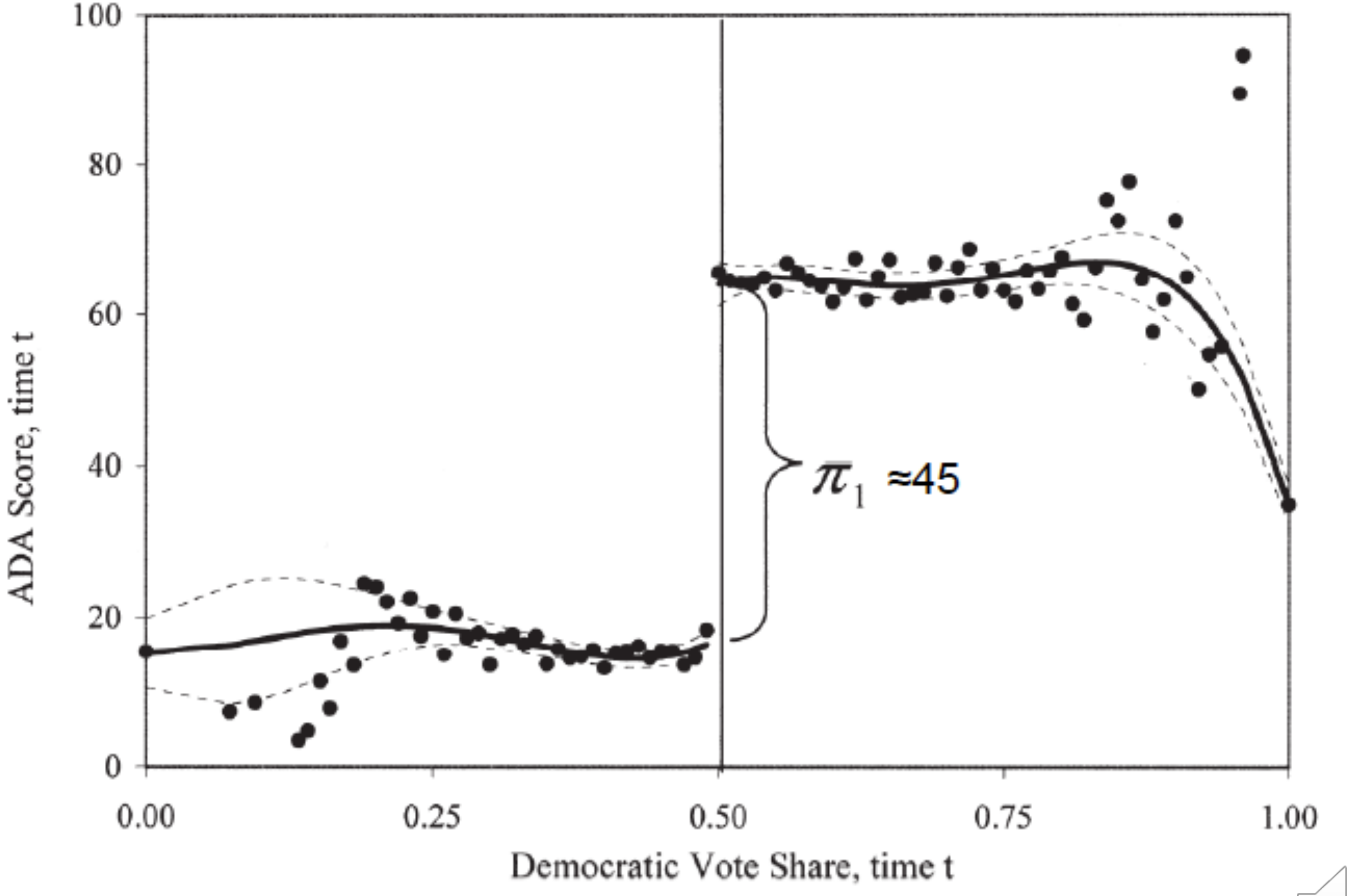
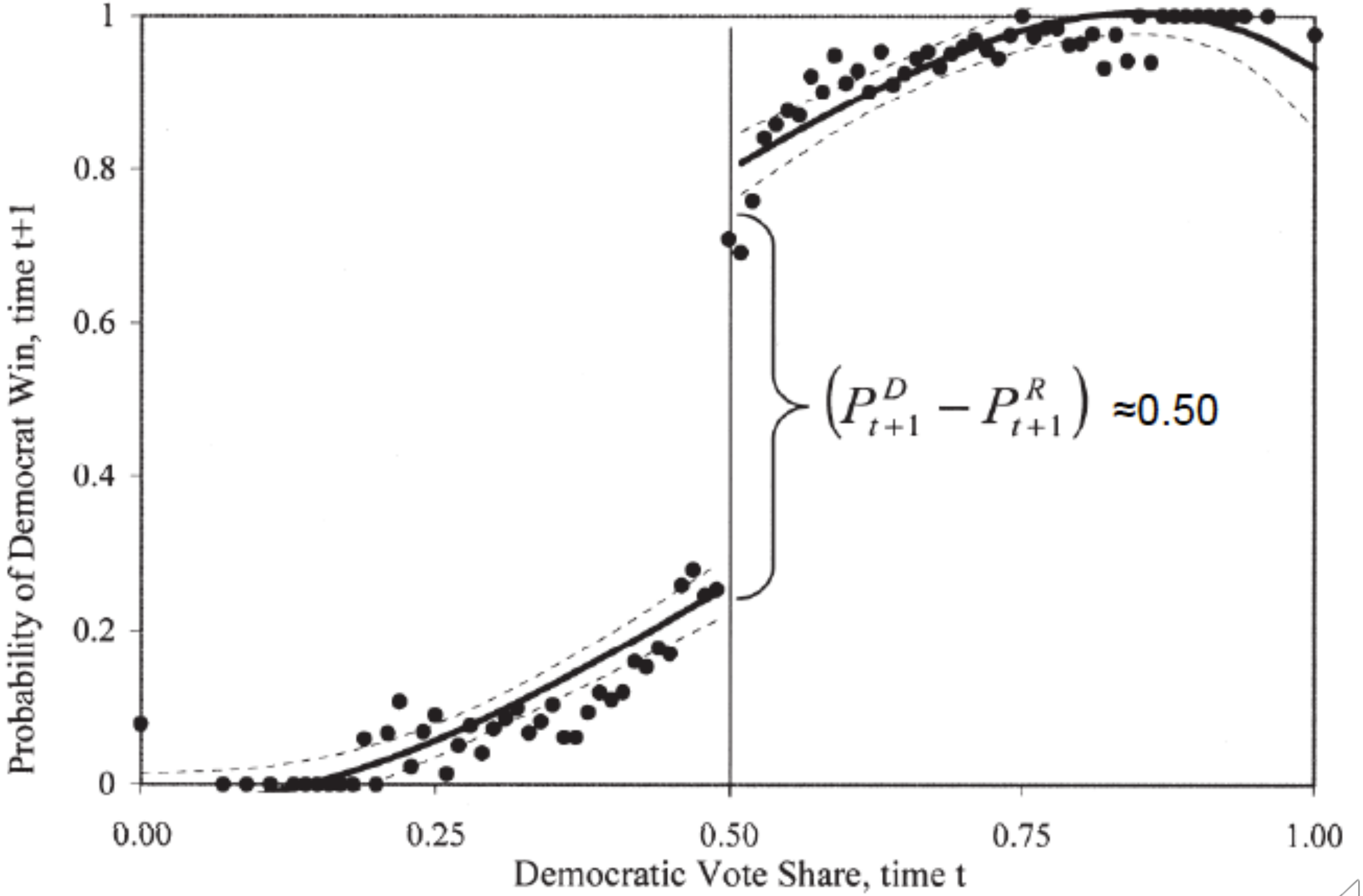


FIGURE IIa
Effect of Party Affiliation: π_1

Graphical estimate of equation 6



$\left(P^{D}_{t+1} - P^{R}_{t+1}\right)$ ≈0.50

FIGURE IIb

Effect of Initial Win on Winning Next Election: $(P^D_{t+1} - P^R_{t+1})$

Statistical results

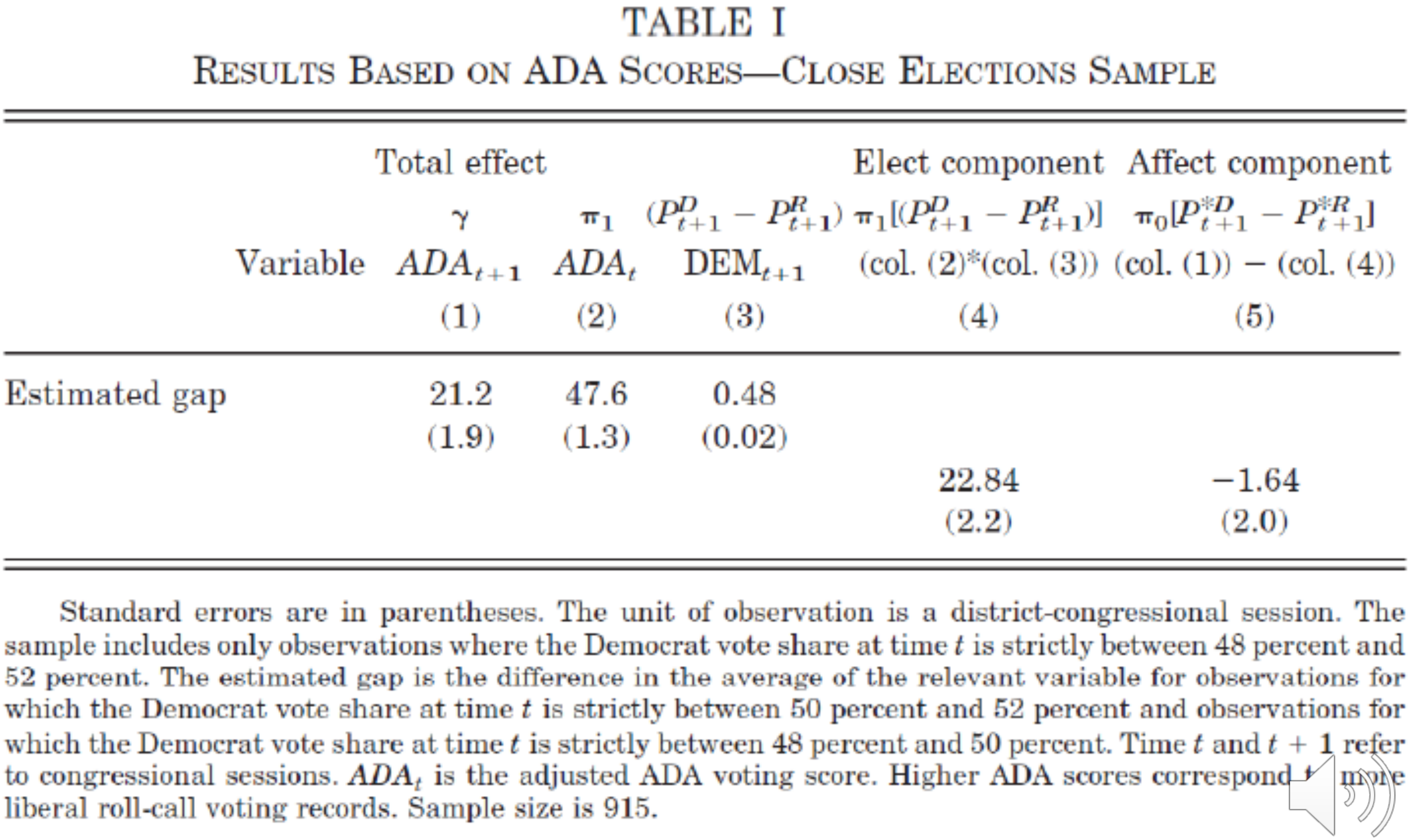


TABLE I

RESULTS BASED ON ADA SCORES—CLOSE ELECTIONS SAMPLE

	Total effect			Elect component	Affect component
	γ	π_1	$(P^D_{t+1} - P^R_{t+1})$	$\pi_1[(P^D_{t+1} - P^R_{t+1})]$	$\pi_0[P^{*D}_{t+1} - P^{*R}_{t+1}]$
Variable	ADA_{t+1}	ADA_t	DEM_{t+1}	(col. (2)*(col. (3))	(col. (1)) − (col. (4))
	(1)	(2)	(3)	(4)	(5)
Estimated gap	21.2	47.6	0.48		
	(1.9)	(1.3)	(0.02)		
				22.84	−1.64
				(2.2)	(2.0)

Standard errors are in parentheses. The unit of observation is a district-congressional session. The sample includes only observations where the Democrat vote share at time t is strictly between 48 percent and 52 percent. The estimated gap is the difference in the average of the relevant variable for observations for which the Democrat vote share at time t is strictly between 50 percent and 52 percent and observations for which the Democrat vote share at time t is strictly between 48 percent and 50 percent. Time t and $t + 1$ refer to congressional sessions. ADA_t is the adjusted ADA voting score. Higher ADA scores correspond to more liberal roll-call voting records. Sample size is 915.

Robustness etc.

- No other observable district attributes change to the discontinuity
- Results hold for many other measures of representative voting records
- Results robust to allowing for various sorts of district heterogeneity
- Results (small “affect” component) stable over time

Additional Resources

- Please see this thorough and clear guide on RD designs by MDRC. Specifically, look at the "Checklist for Researchers Conducting Retrospective RD Analysis" on page 88

Stata Package & Command Options

- `ted`
- `rdbwselect`
- `cmogram`
- `lpoly`
- `rdrobust`