Lecture 7: Regression Discontinuity Designs

Chris Walters

University of California, Berkeley and NBER

Introduction

- The **regression discontinuity design** (RD) is a common research design in contemporary applied research
- RD methods can be applied when a researcher has specific information about the rules determining the treatment of interest
- This lecture describes the RD framework and implementation
- See Imbens and Lemieux (2008) and Lee and Lemieux (2010) for overviews

Sharp RD: Basic Setup

- Consider a setting with a binary treatment $D_i \in \{0, 1\}$, and potential outcomes $Y_i(1)$ and $Y_i(0)$
- Suppose the treatment is a deterministic and discontinuous function of an observed covariate R_i, such that

$$D_i = 1\{R_i > c\}.$$

- R_i is called the **running variable** or **forcing variable**
- This is a **sharp RD** because the probability of treatment switches from zero to one at the threshold
- Example: Scholarship awarded to students who score above a test score threshold (Thistlethwaite and Campbell, 1960)

Sharp RD: Basic Setup

- We get to observe $Y_i(1)$ when $R_i > c$ and $Y_i(0)$ when $R_i \leq c$
- Basic idea of the RD design: Compare observations just above and just below the threshold to infer treatment effect
- Intuitively, the treatment may be as good as randomly assigned for individuals in the neighborhood of $R_i = c$, so comparing treated and nontreated near c reveals a treatment effect

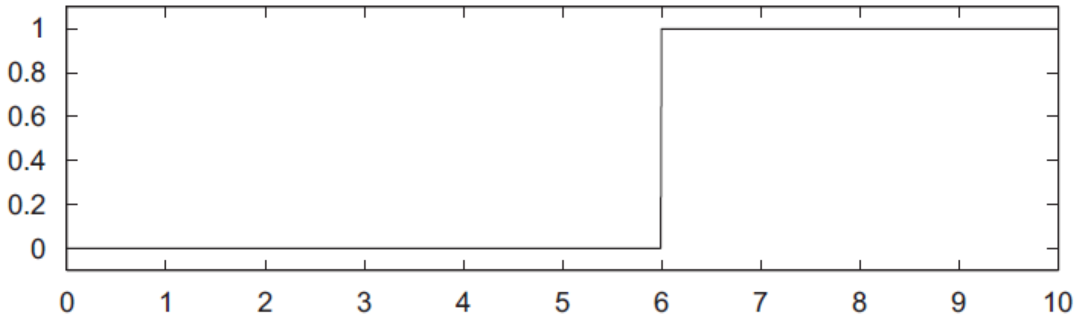


Fig. 1. Assignment probabilities (SRD).

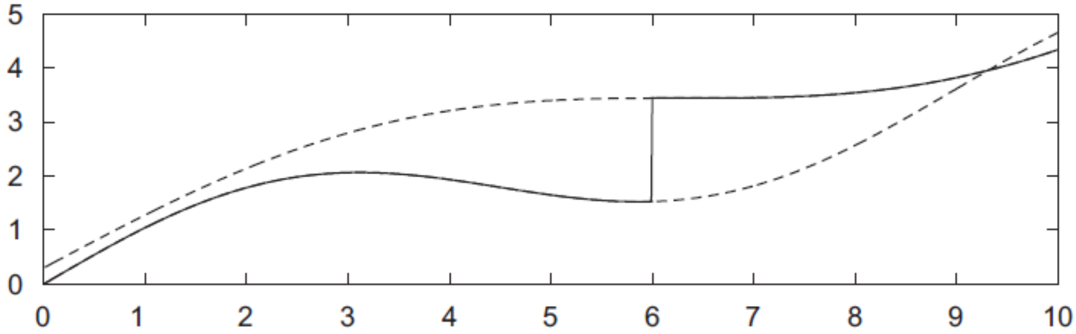


Fig. 2. Potential and observed outcome regression functions.

Sharp RD Identification

- Key assumption: potential outcomes are smooth at the threshold
- Formally:

$$\lim_{r \to c^+} E\left[Y_i(d)|R_i = r\right] = \lim_{r \to c^-} E\left[Y_i(d)|R_i = r\right], \; d \in \{0,1\}$$

- Potential outcome CEFs must be continuous at the threshold
- The population just below must not be discretely different from the population just above

Sharp RD Identification

- If this assumption holds we have

$$\lim_{r \to c^+} E\left[Y_i | R_i = r\right] - \lim_{r \to c^-} E\left[Y_i | R_i = r\right]$$

$$= \lim_{r \to c^+} E\left[Y_i(1) | R_i = r\right] - \lim_{r \to c^-} E\left[Y_i(0) | R_i = r\right]$$

$$= E\left[Y_i(1) | R_i = c\right] - E\left[Y_i(0) | R_i = c\right]$$

$$= E\left[Y_i(1) - Y_i(0) | R_i = c\right]$$

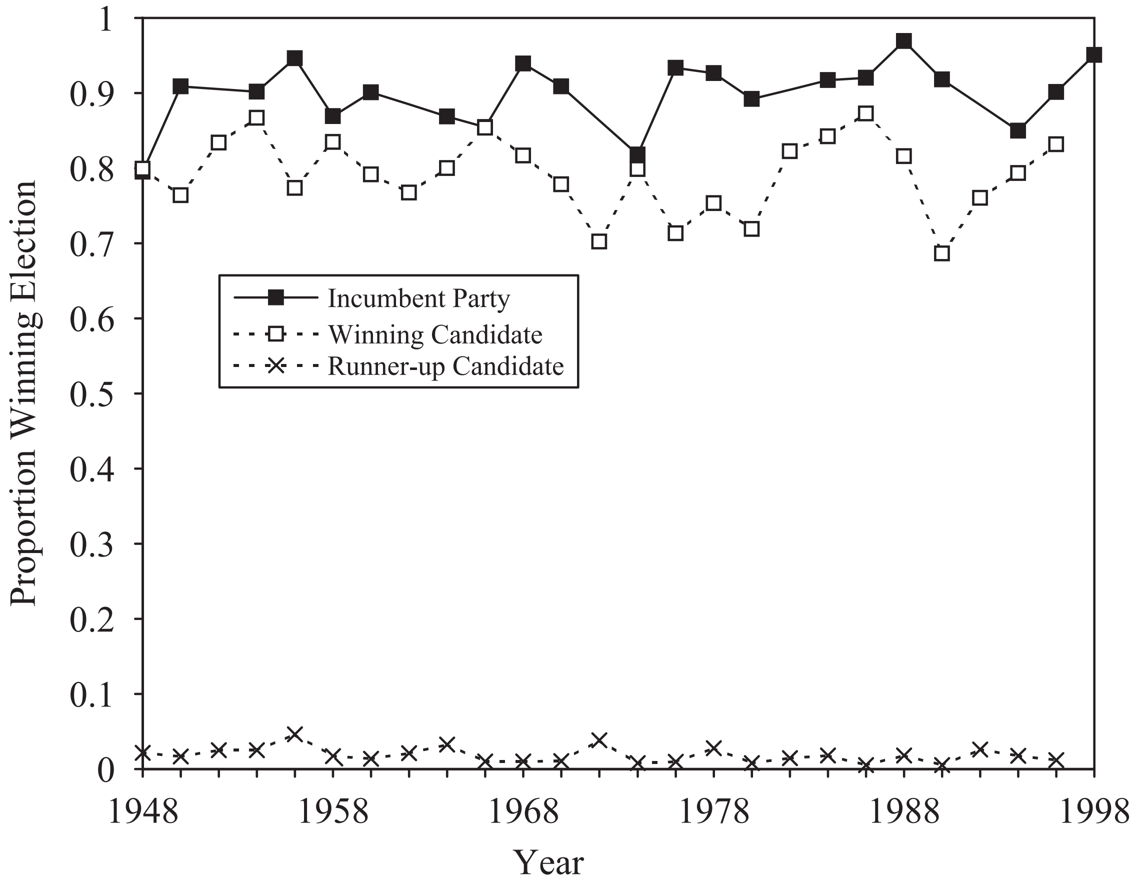
- When potential outcomes are smooth around the threshold, a comparison of individuals just above and just below yields the average treatment effect for those at the threshold
- Identification argument is nonparametric: we don't need to assume anything about the distribution of $Y_i(d)$ other than continuity of CEFs

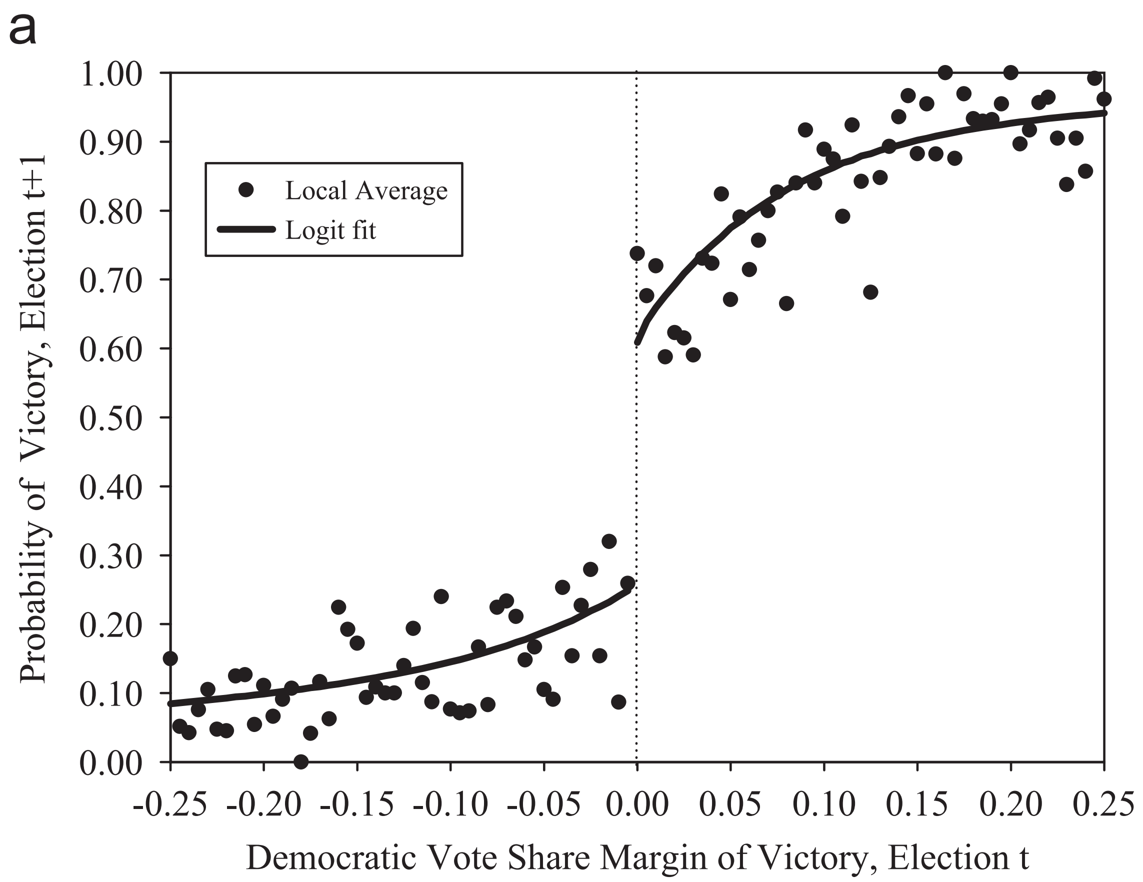
RD Interpretation

- RD as RCT
  - Core intuition: for those right around the threshold, things could have gone either way
  - Interpret RD as a local randomized trial among those sufficiently close to $R_i = c$
  - Explains why RD evidence can be especially compelling relative to other research designs – close to "gold standard" of RCT
- RD and CIA
  - In a sharp RD we have the conditional independence assumption (CIA) : $(Y_i(1), Y_i(0)) \perp\!\!\!\perp D_i | R_i$
  - CIA holds trivially because there is no variation in treatment conditional on the running variable
  - But there is also no common support
  - RD estimation is a local extrapolation outside the support of the data to predict the mean treated and untreated potential outcomes at $R_i = c$

Sharp RD Example: Lee (2008)

- Important phenomenon in politics: The incumbency advantage
- Candidates/parties who won the previous election are much more likely to win again
- By definition, incumbents are candidates who were successful last time. Some or all of incumbency advantage could be due to persistent unobservables
- How much of the incumbency advantage is causal?
- Lee (2008) uses an RD design to estimate the causal effect of winning US House elections





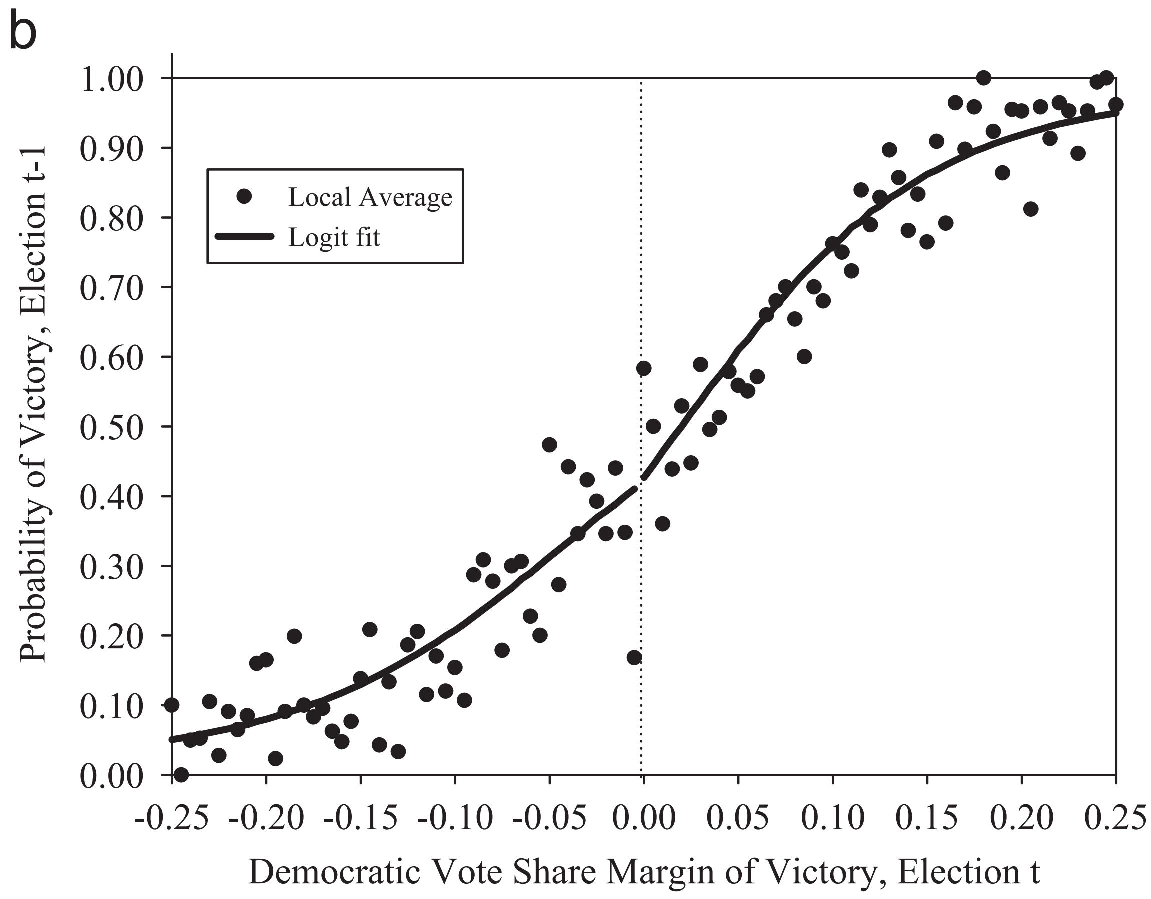
.

RD Diagnostics

- Fundamental RD identifying assumption: potential outcome distributions are smooth around the threshold
- $1\{R_i > c\}$ must be as good as randomly assigned in the neighborhood of $R_i = c$
- May be violated if individuals can exactly control the value of R_i and therefore location relative to the threshold
- Example: suppose savvy politicians manipulate close elections to ensure victory
- Identifying assumption is untestable, but some common diagnostics serve as a guide to its plausibility

RD Diagnostic I: Covariate Balance

- In the absence of sorting in the neighborhood of the threshold, we'd expect distributions of observed covariates to be smooth
- This motivates a check for whether there is a discontinuity in $E\left[X_i|R_i = r\right]$ at $R_i = c$ for covariates X_i
- Jumps in observables suggest sorting in the neighborhood of the threshold – people just above are different than people just below
- This is analogous to a balance check in a randomized trial



RD Diagnostic II: Bunching

- If individuals are strategically locating above or below the threshold, we'd expect "bunching" on whichever side of the discontinuity is preferable
- More generally, strategic manipulation may generate anomalies in the distribution of R_i around the threshold
- McCrary (2008) suggests looking for a discontinuity in the density of the running variable near $R_i = c$
- Bunching around the threshold compromises RD estimates of treatment effects, but may reflect a behavioral response of substantive interest
  - Urquiola and Verhoogen (2009): Sorting around class size caps in Chilean private schools

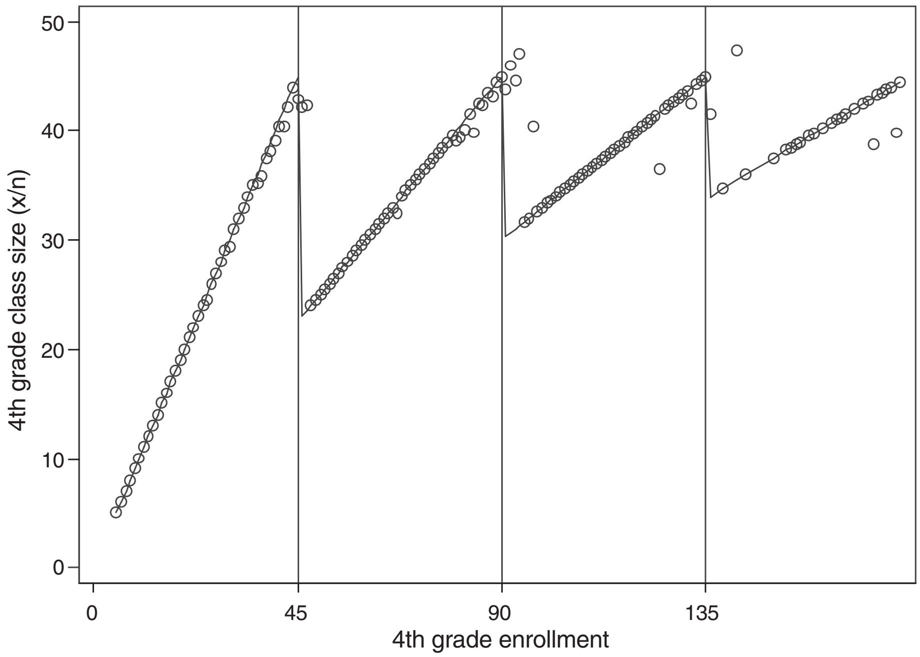
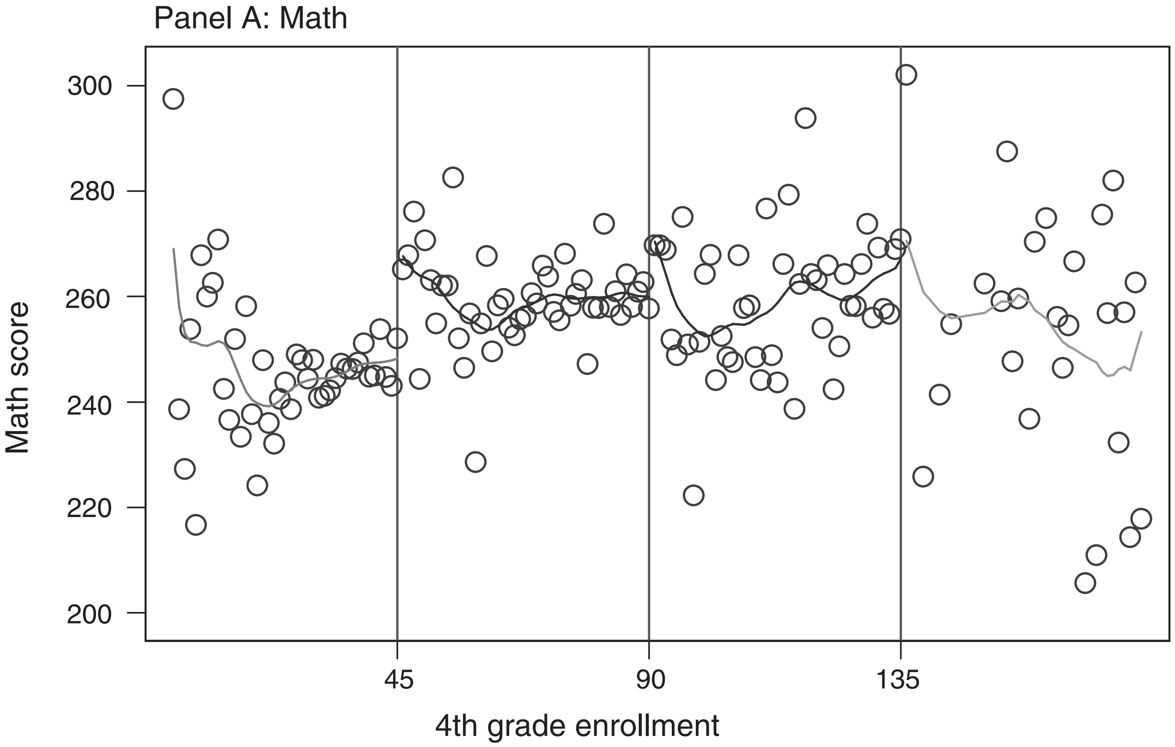
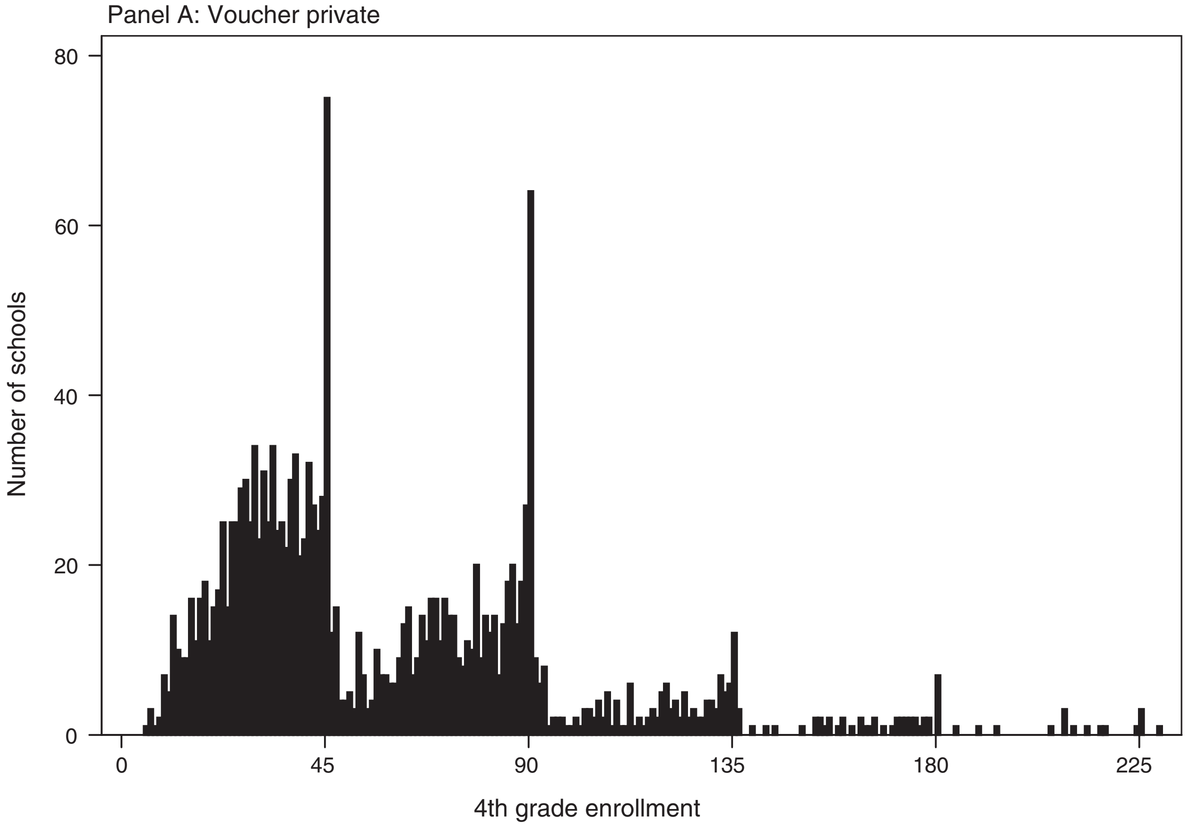
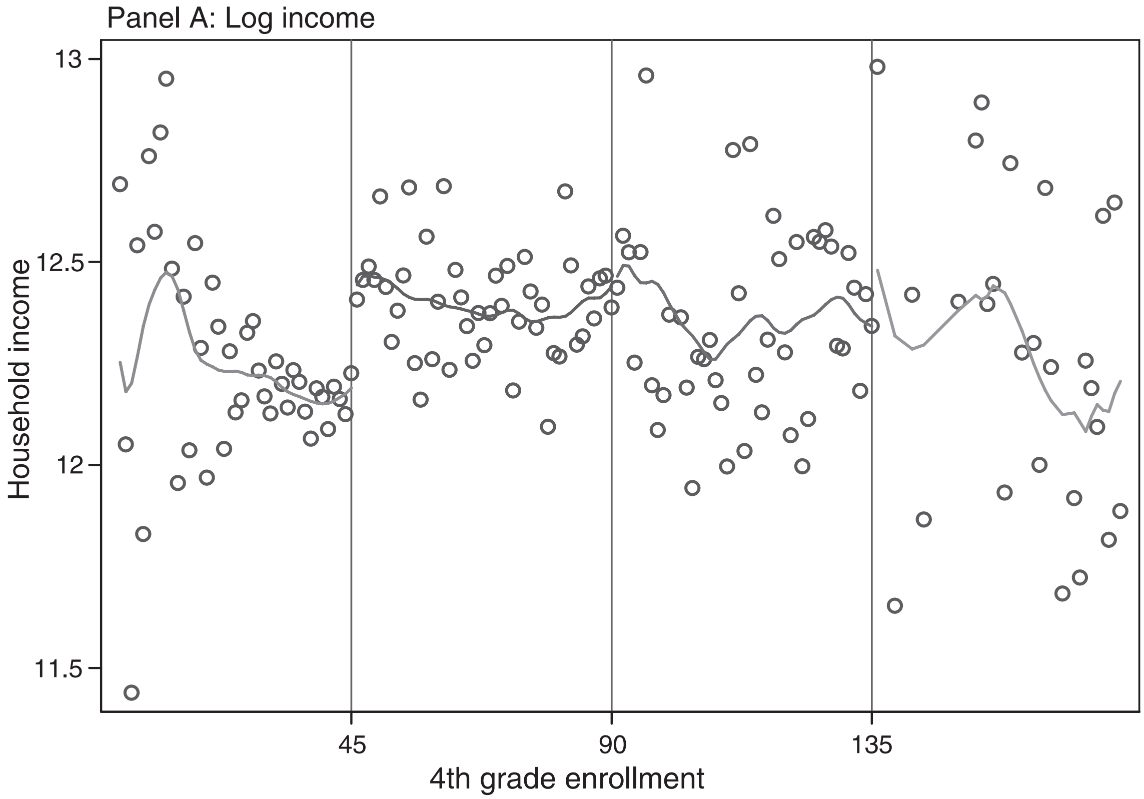


Figure 5. Fourth Grade Enrollment and Class Size in Urban Private Voucher Schools, 2002

Notes: Based on administrative data for 2002. The solid line describes the relationship between enrollment and class size that would exist if the class size rule (equation (30) in the text) were applied mechanically. The circles plot actual enrollment cell means of fourth grade class size. Only data for schools with fourth grade enrollments below 180 are plotted; this excludes less than 2 percent of all schools.


Panel A: Math
Math score
300
280
260
240
220
200
45
90
135
4th grade enrollment


Panel A: Voucher private
Number of schools
80
60
40
20
0
0
45
90
135
180
225
4th grade enrollment



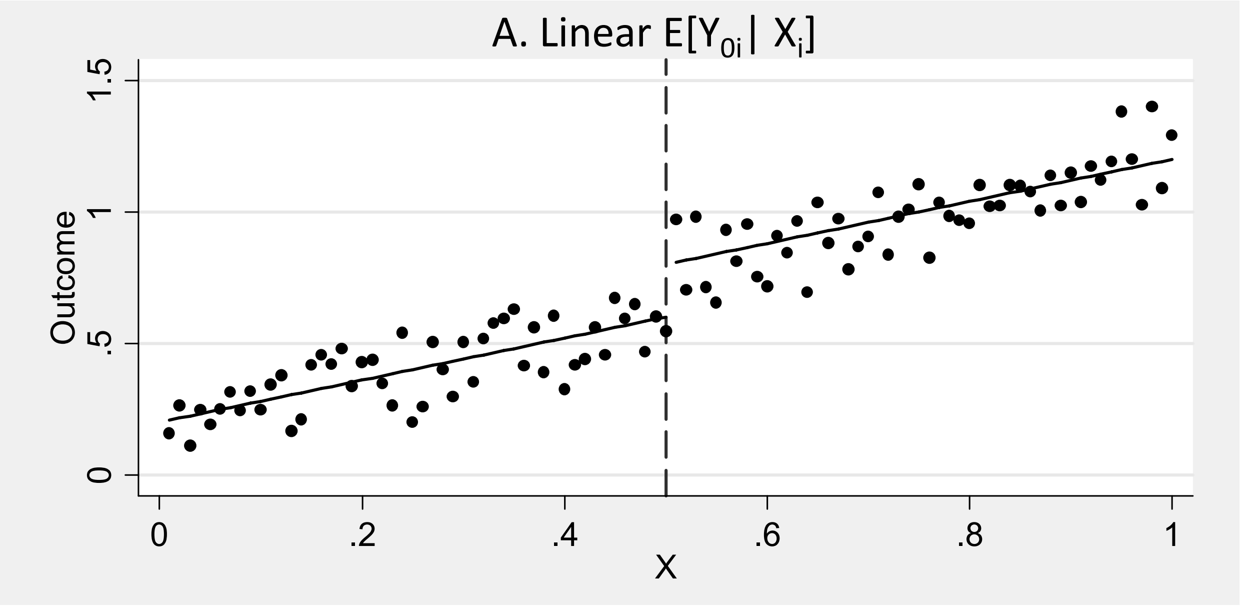
RD Estimation

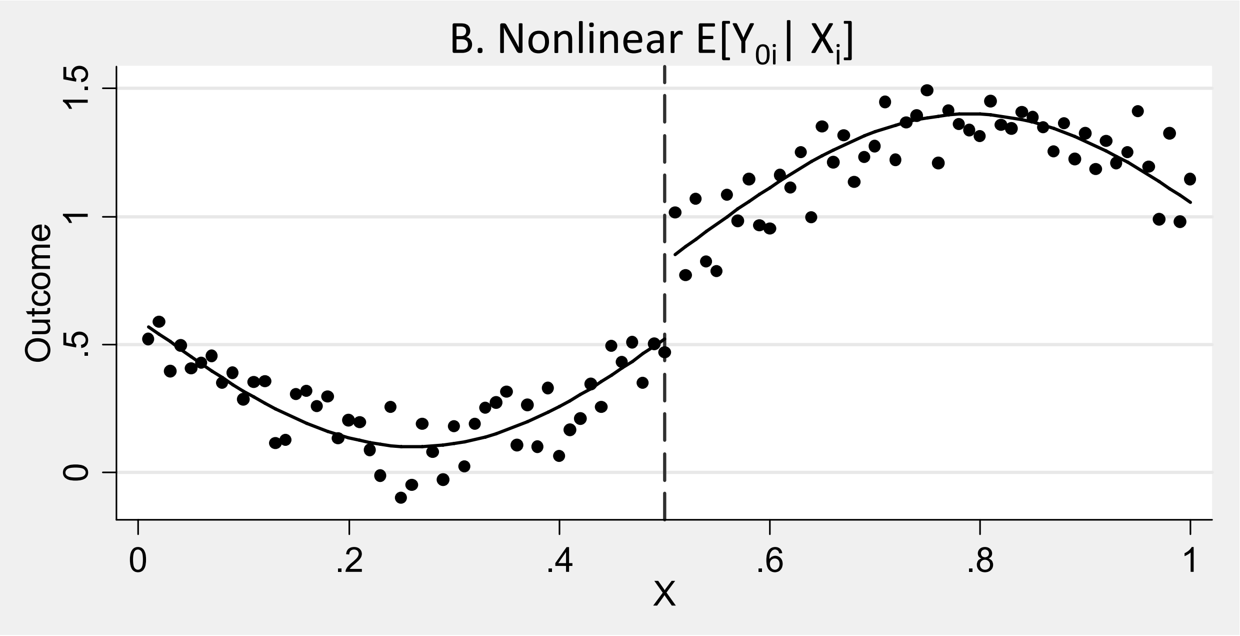
- Implementing a sharp RD requires estimating the right- and left-hand limits

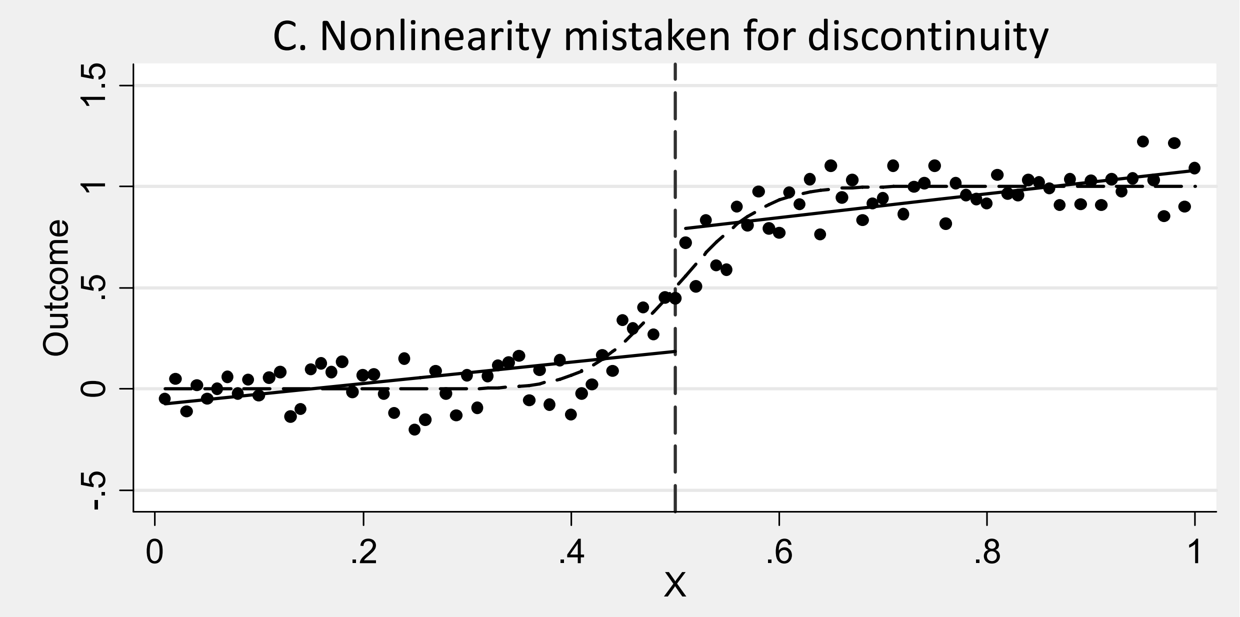
$$\lim_{r \to c^+} E\left[Y_i | R_i = r\right]$$

$$\lim_{r \to c^-} E\left[Y_i | R_i = r\right]$$

- Since extrapolation is central to the RD design, the functional form used to approximate $E\left[Y_i | R_i\right]$ really matters – it's not enough to just run OLS and rely on approximation theorems
- An insufficiently flexible specification of the CEF runs the risk of mistaking nonlinearity for treatment effect
- But an overly flexible specification reduces precision and runs the risk of overfitting


A. Linear $E[Y_{0i}| X_i]$
Outcome
0
.5
1
1.5
X
0
.2
.4
.6
.8
1


B. Nonlinear $E[Y_{0i}| X_i]$
Outcome
0
.5
1
1.5
X
0
.2
.4
.6
.8
1



RD Estimation

- Two general approaches to estimation:
  - Global parametric
  - Local nonparametric

Global Parametric Estimation

- The global approach uses parametric functions to approximate $E\left[Y_i|R_i\right]$, typically polynomials
- OLS regression:

$$Y_i = \alpha + \beta 1\left\{R_i > c\right\} + \sum_{k=1}^{K} \gamma_{0k} 1\left\{R_i \leq c\right\}\left(R_i - c\right)^k$$

$$+ \sum_{k=1}^{K} \gamma_{1k} 1\left\{R_i > c\right\}\left(R_i - c\right)^k + \epsilon_i$$

- This specification uses a Kth order polynomial with coefficients that differ on each side of the threshold
- Think of this as fitting $E\left[Y_i(0)|R_i\right]$ and $E\left[Y_i(1)|R_i\right]$ with two separate polynomials
- The parameter β measures the jump at the threshold

Local Linear Regression

- More modern approach: use nonparametric techniques to approximate the left- and right-hand limits
- Local linear regression:

$$\left(\hat{\alpha}_0, \hat{\delta}_0\right) = \arg\min_{\alpha_0, \delta_0} \sum_i 1\left\{R_i \leq c\right\} K\left(\frac{R_i - c}{h}\right)\left[Y_i - \alpha_0 - \delta_0\left(R_i - c\right)\right]^2$$

- Here $K(\cdot)$ is a symmetric kernel function maximized at 0, and h is a bandwidth that vanishes to zero asymptotically

Local Linear Regression

$$\left(\hat{\alpha}_0, \hat{\delta}_0\right) = \arg\min_{\alpha_0, \delta_0} \sum_i 1\left\{R_i \leq c\right\} K\left(\frac{R_i - c}{h}\right) \left[Y_i - \alpha_0 - \delta_0 \left(R_i - c\right)\right]^2$$

- This is weighted least squares using observations to the left of the threshold, and weighting by proximity to the threshold
- The bandwidth h determines how quickly weight falls off away from the discontinuity
- The linear term eliminates "boundary bias" exhibited by local constant estimators (Fan and Gijbels, 1992)

Local Linear Regression

$$\left(\hat{\alpha}_0, \hat{\delta}_0\right) = \arg\min_{\alpha_0, \delta_0} \sum_i 1\left\{R_i \leq c\right\} K\left(\frac{R_i - c}{h}\right) \left[Y_i - \alpha_0 - \delta_0 \left(R_i - c\right)\right]^2$$

- $\hat{\alpha}_0$ is an estimate of $\lim_{r \to c^-} E\left[Y_i | R_i = r\right]$
- Run similar regression for the sample with $R_i > c$ to get $\hat{\alpha}_1$:

$$\left(\hat{\alpha}_1, \hat{\delta}_1\right) = \arg\min_{\alpha_1, \delta_1} \sum_i 1\left\{R_i > c\right\} K\left(\frac{R_i - c}{h}\right) \left[Y_i - \alpha_1 - \delta_1 \left(R_i - c\right)\right]^2$$

- Treatment effect estimate is $\hat{\alpha}_1 - \hat{\alpha}_0$

RD Estimation: Global or Local?

- When should we use a global polynomial approach vs. a nonparametric local approach?
- There is no real conceptual distinction. We need to choose a bandwidth, a kernel, and a polynomial order
- Global estimators use an infinite bandwidth, a uniform kernel, and a relatively high-order polynomial
- Local estimators use a smaller bandwidth, a kernel that places more weight near the threshold, and a lower-order polynomial (often linear)

Global vs. Local

- Some researchers dislike global estimators because these estimators can rely heavily on data far from the discontinuity (Gelman and Imbens 2019, "Why high-order polynomials should not be used in regression discontinuity designs")
- There can be no theorem on which approach works better in general. It depends on the data generating process
- If the outcome CEF is well-approximated by a polynomial, then using points far from the discontinuity to estimate the polynomial is legitimate and increases precision. If not, the global approach may perform poorly
- Not much is known about choosing the optimal order of a global polynomial for estimating the causal effect of interest; in contrast, there is a large literature on optimal bandwidths for local estimators of treatment effects
- Most researchers have gravitated towards local approaches

Kernels and Bandwidths

- Suppose we want to run a local linear regression. What kernel and bandwidth should we choose?
- The **edge** or **triangular** kernel is a popular choice:

$$K(u) = 1\{|u| \leq 1\} \times (1 - |u|)$$

- The edge kernel has optimality properties in boundary estimation problems (Cheng et al., 1997)
- Also intuitively appealing: it generates the weighting function

$$K\left(\frac{R_i - c}{h}\right) = 1\{|R_i - c| \leq h\} \times \left(1 - \frac{|R_i - c|}{h}\right)$$

- The bandwidth h can then be interpreted as cutoff distance beyond which data are not used, and weights fall linearly from 1 to 0 in remaining sample

Optimal Bandwidth

- Recent literature focuses on optimal choice of bandwidth h
- Bias/variance tradeoff: Smaller bandwidth reduces bias from using points away from the boundary, but also reduces precision
- Intuitively, if there is not a lot of curvature in the CEF of Y_i given R_i, the bias from using points away from the boundary to estimate a regression slope will be small
- Imbens and Kalyanaraman (IK, 2012) use an asymptotic approximation to the mean squared error of the RD estimator and derive the MSE-minimizing bandwidth
- The optimal bandwidth depends on the curvature of the CEF near the discontinuity – IK propose to use plug-in estimators of parameters governing curvature

Robust Confidence Intervals

- The IK bandwidth minimizes MSE and is therefore well-suited to estimation
- Calonico, Cattaneo and Titiunik (CCT, 2014) show that it is poorly suited for inference, however: the IK bandwidth leaves an asymptotically non-negligible bias term in the estimate, so naive inference can lead to misleading confidence intervals
- CCT advocate using a second, smaller bandwidth that removes this bias term when constructing confidence intervals
- The IK bandwidth and CCT confidence intervals are automated in the *rdrobust* stata package

Fuzzy RD

- Sometimes treatment is generated by a discontinuous assignment rule that isn't deterministic
- Suppose that

$$\lim_{r \to c^-} Pr\left[D_i = 1 | R_i = r\right] < \lim_{r \to c^+} Pr\left[D_i = 1 | R_i = r\right]$$

- The probability of treatment jumps at $R_i = c$, but not necessarily from zero to one
- This is a **fuzzy RD** scenario because treatment is only partly determined by the threshold
- Example (Carneiro and Ginja, 2014): An income threshold determines eligibility for a government program, but not every eligible household participates

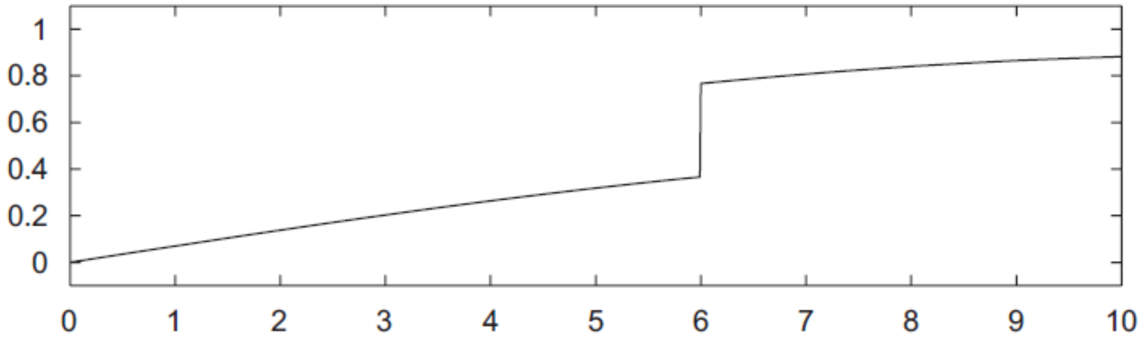


Fig. 3. Assignment probabilities (FRD).

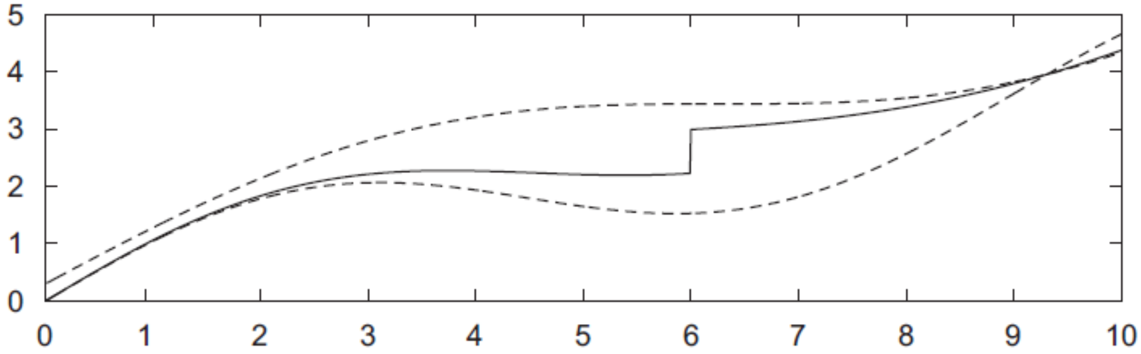


Fig. 4. Potential and observed outcome regression (FRD).

Fuzzy RD Assumptions

- As before, assume the distributions of $Y_i(1)$ and $Y_i(0)$ are smooth around the threshold
- Let $D_i(1)$ and $D_i(0)$ denote potential treatment statuses for individual i if s/he were located above and below the threshold. Assume these are also smooth across the threshold, and

$$D_i(1) \geq D_i(0)\ \forall i$$

- Crossing the threshold weakly increases the likelihood of treatment for everyone

Fuzzy RD

- Under these assumptions, we have

$$\frac{\lim_{r \to c^+} E\left[Y_i|R_i = r\right] - \lim_{r \to c^-} E\left[Y_i|R_i = r\right]}{\lim_{r \to c^+} E\left[D_i|R_i = r\right] - \lim_{r \to c^-} E\left[D_i|R_i = r\right]}$$

$$= E\left[Y_i(1) - Y_i(0)|D_i(1) > D_i(0), R_i = c\right]$$

- The numerator on the left is the jump in outcomes at the threshold, as in a sharp RD
- The denominator is the change in the probability of treatment at the threshold
- The ratio of the jump in the outcome CEF to the jump in the treatment probability identifies an average treatment effect for individuals who switch treatment status at the threshold
- Sound familiar?

Fuzzy RD is IV

- Fuzzy RD is IV using a threshold indicator $Z_i = 1\left\{R_i > c\right\}$ as an instrument for treatment in the neighborhood of the threshold
- Think of Fuzzy RD as a local randomized trial with non-compliance

Fuzzy RD and LATE

- This IV interpretation implies that fuzzy RD estimates are local in two senses
  - First, they are local to the threshold, $R_i = c$
    - Also applies to sharp RD estimates
  - Second, they apply only to compliers at the threshold, rather than everyone with $R_i = c$
    - This is the “local” in LATE

Fuzzy RD Implementation

- As with sharp RD, we can implement fuzzy RD with a global parametric or local nonparametric approach
- Global polynomial 2SLS:

$$D_i = \lambda + \pi 1\{R_i > c\} + \sum_{k=1}^{K} \theta_{0k} 1\{R_i \leq c\}(R_i - c)^k$$

$$+ \sum_{k=1}^{K} \theta_{1k} 1\{R_i > c\}(R_i - c)^k + \eta_i$$

$$Y_i = \alpha + \beta \hat{D}_i + \sum_{k=1}^{K} \gamma_{0k} 1\{R_i \leq c\}(R_i - c)^k$$

$$+ \sum_{k=1}^{K} \gamma_{1k} 1\{R_i > c\}(R_i - c)^k + \epsilon_i$$

- Excluded instrument is $1\{R_i > c\}$
- Alternatively, we can estimate each of the four limits in the Wald ratio by local linear regressions of Y_i and D_i on R_i
- IK and CCT provide optimal bandwidths and robust confidence intervals for fuzzy RD, also automated in *rdrobust*

Fuzzy RD Example: Clark and Martorell (2014)

- Clark and Martorell (2014) use an RD design to estimate the causal effect of high school graduation on earnings
- Two views on the causal effect of schooling on earnings:
  - Human capital: Schooling raises productivity
  - Signaling: Schooling reveals ability but has no productive value
- OLS returns to education are especially large for grade 12
- How much of this "sheepskin effect" reflects signaling?

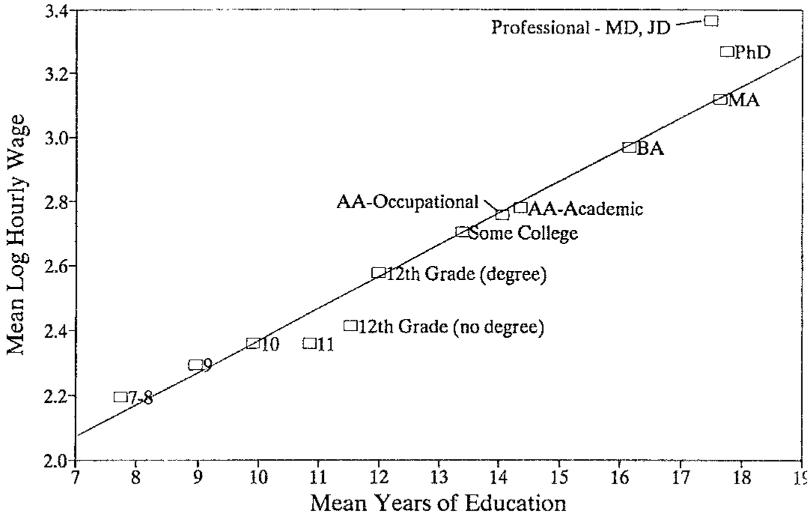


Fig. 2. Relationship between mean log hourly wages and completed education, men aged 40–45 in 1994–1996 Current Population Survey. Mean education by degree category estimated from February 1990 CPS.

Clark and Martorell (2014)

- CM use the fact that students in Texas must pass exams before graduating high school
- Testing starts in 10th grade and students can try multiple times, but eventually face a "last chance" exam at the end of 12th grade
- Students who just barely fail vs. barely pass should have similar human capital, but differ in educational credentials
- RD therefore plausibly identifies the signaling value of a diploma
- There is some "slippage" even with last-chance exams – so the RD is fuzzy

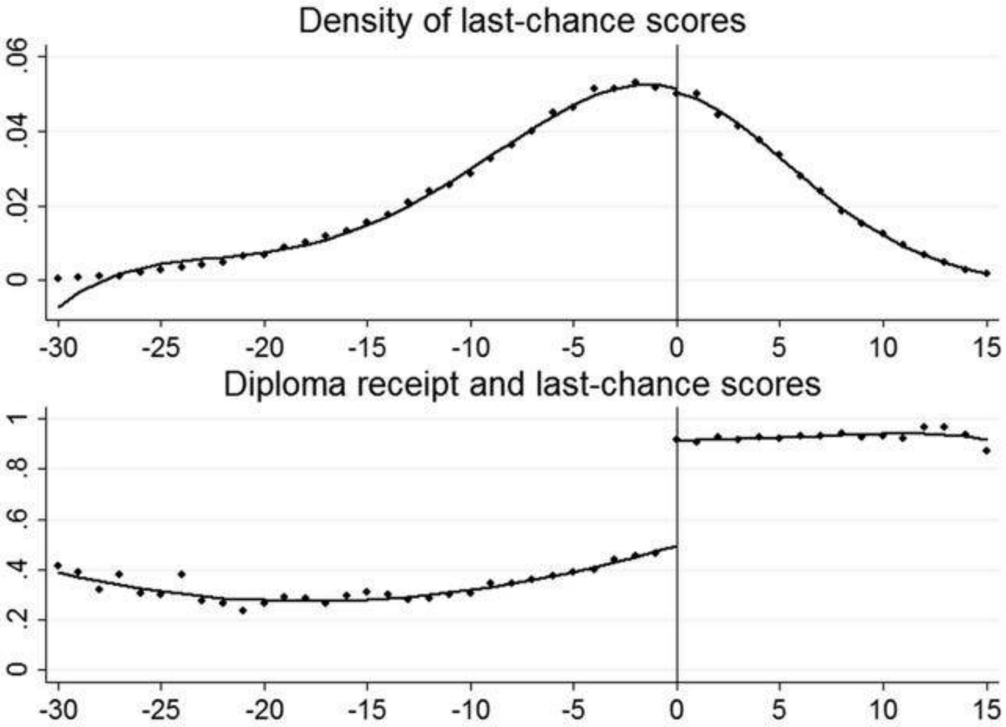
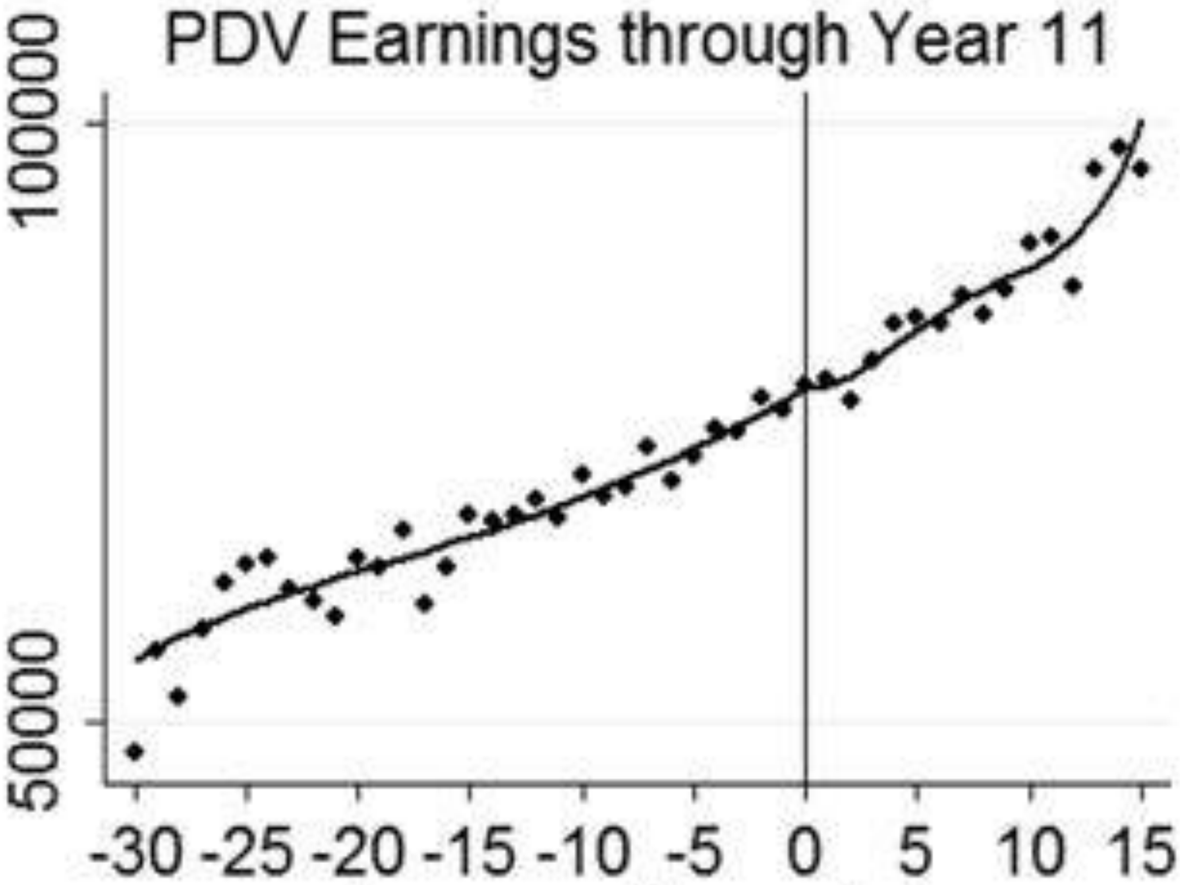


Fig. 1.—Last-chance exam scores and diploma receipt. The graphs are based on the last-chance sample. See table 1 and the text. Dots are test score cell means. The scores on the x-axis are the minimum of the section scores (recentered to be zero at the passing cutoff) that are taken in the last-chance exam. Lines are fourth-order polynomials fitted separately on either side of the passing threshold.

TABLE 2
Impact of Passing the Last-Chance Exam on the Probability of Earning a Diploma

Receive High School Diploma	(1)	(2)	(3)	(4)	(5)
By end of summer after 12th grade (sample mean = .363)	.545 (.007)	.484 (.009)	.481 (.012)	.475 (.016)	.486 (.009)
Within 1 year of last-chance exam (sample mean = .452)	.480 (.007)	.420 (.009)	.425 (.012)	.424 (.016)	.422 (.009)
Within 2 years of last-chance exam (sample mean = .465)	.472 (.007)	.415 (.009)	.419 (.012)	.417 (.016)	.417 (.009)
Within 3 years of last-chance exam (sample mean = .468)	.468 (.007)	.412 (.009)	.416 (.012)	.414 (.016)	.414 (.009)
Baseline covariates?	No	No	No	No	Yes
Degree of test score polynomial	1	2	3	4	2

Note.—The table is based on last-chance samples (see table 1 and the text). "Degree of test score polynomial" refers to the test score polynomials controlled for in these regressions (all interacted with a dummy for passing the exam). Column 5 presents estimates based on models that also control for covariates (see note to table 1). Robust standard errors are in parentheses. There are 37,571 observations in each panel.



Regression Kink Design

- Recent extension of RD: The **regression kink design** (RKD; Card et al., 2015)
- Instead of exploiting a discontinuity in the CEF of the treatment variable, the regression kink design exploits a kink in the CEF of a continuous treatment (i.e. a discontinuity in the first derivative)
- A corresponding kink in the distribution of the outcome variable suggests the presence of a treatment effect

Regression Kink Design

- Suppose the treatment of interest is a deterministic function of the running variable:

$$S_i = b\left(R_i\right)$$

- Here $b(\cdot)$ is a continuous function with a kink at c
- Example (Card et al., 2015): Unemployment benefit is a kinked function of past earnings
- Let $f_i(s)$ denote i's potential outcome as a function of the treatment. The observed outcome is

$$Y_i = f_i(S_i)$$

Regression Kink Design

- Then under mild regularity conditions:

$$\frac{\lim\limits_{r\to c^+}\frac{dE[Y_i|R_i=r]}{dr} - \lim\limits_{r\to c^-}\frac{dE[Y_i|R_i=r]}{dr}}{\lim\limits_{r\to c^+} b'(r) - \lim\limits_{r\to c^-} b'(r)} = E\left[f_i'(S_i)|R_i = c\right]$$

- The ratio of the discontinuity in the outcome derivative to the discontinuity in the treatment derivative identifies the average marginal effect of treatment for individuals at the threshold
- As before, the key assumption is that potential outcomes are smooth around the threshold – any kink in the outcome CEF must be due to the treatment
- Diagnostics: Look for kinks in covariate distributions, or bunching in the density of R_i
- As with RD, we can generalize RKD to a "fuzzy" scenario where the treatment is not a deterministic function of R_i, but $E\left[S_i|R_i\right]$ is kinked at $R_i = c$
- Can be implemented via local polynomial regression with the analogue of the IK bandwidth and CCT robust CI, automated in *rdrobust*

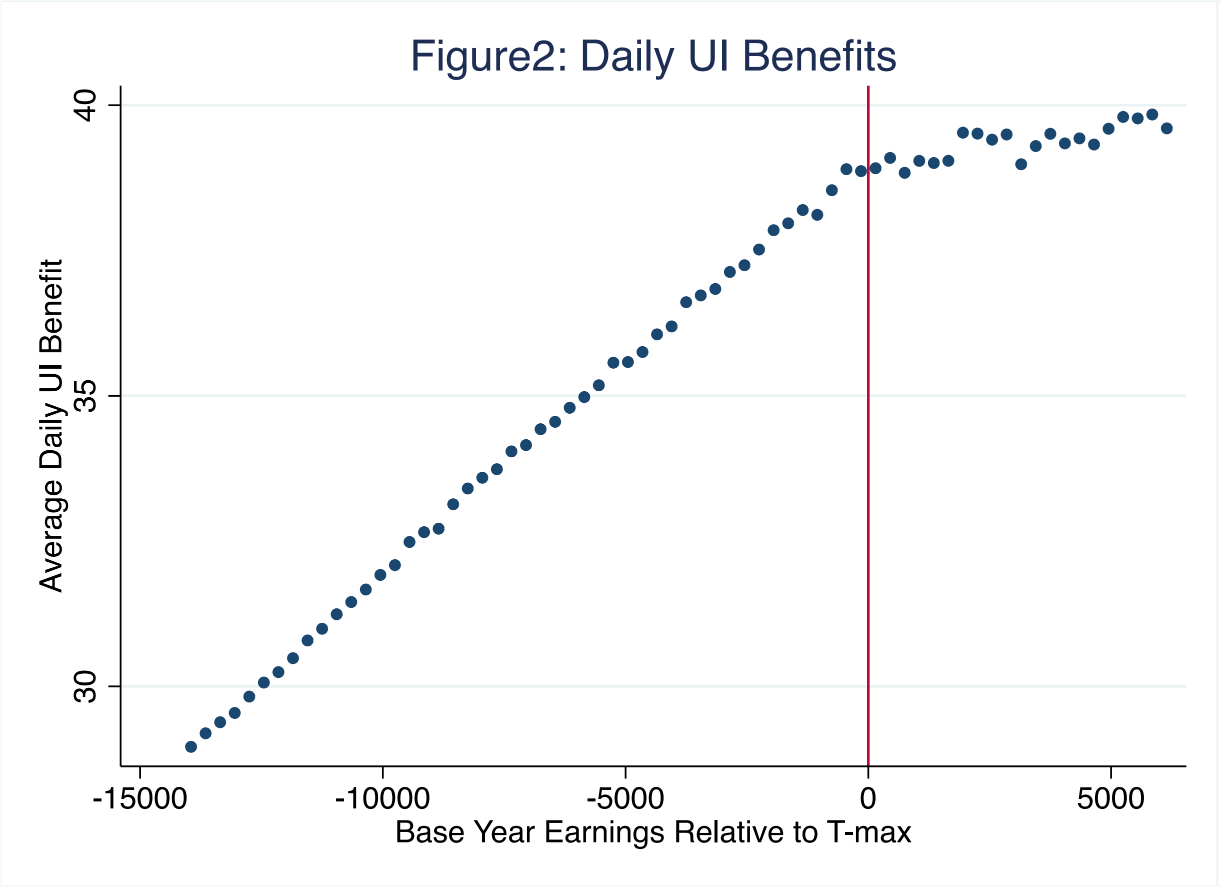
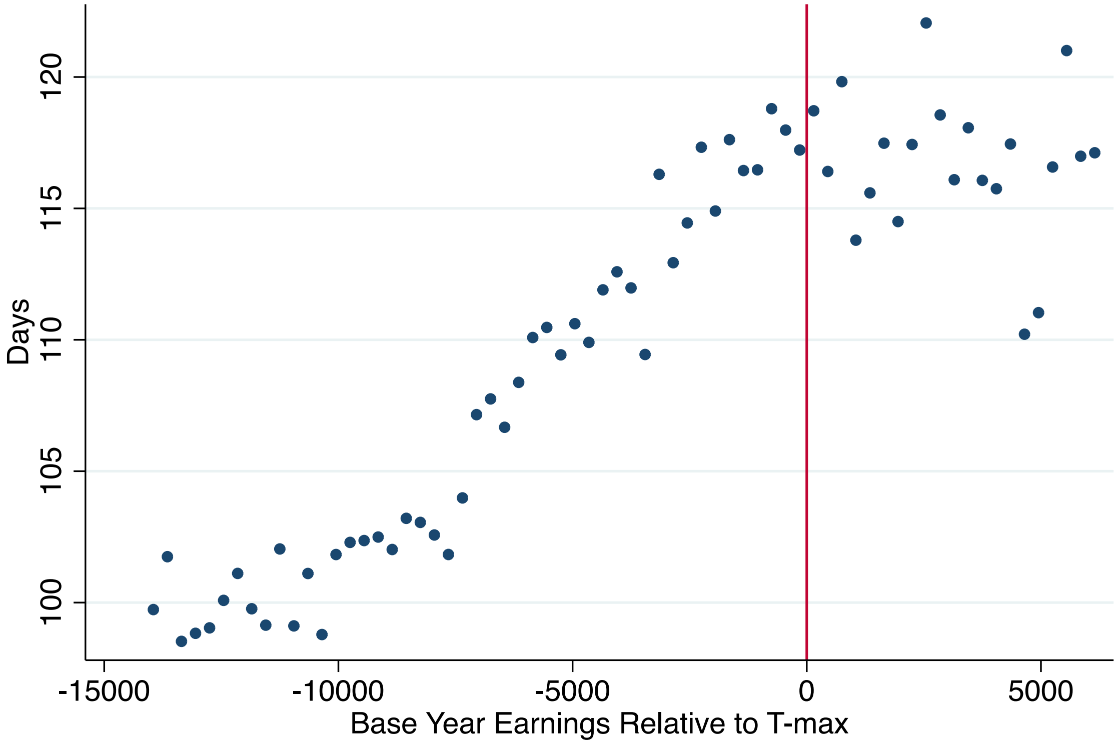


Figure2: Daily UI Benefits

Figure 3: Unemployment Duration



References

- Card, D., Lee, D., Pei, Z., and Weber, A. (2015). "Inference on causal effects in a generalized regression kink design." *Econometrica* 83(6).
- Calonico, S., Cattaneo, M., and Titiunik, R. (2014). "Robust nonparametric confidence intervals for regression-discontinuity designs." *Econometrica* 82(6).
- Carneiro, P., and Ginja, R. (2014). "Long-term impacts of compensatory preschool and health and behavior: evidence from Head Start." *American Economic Journal: Economic Policy* 6(4).
- Cheng, M., Fan, J., and Marron, J. (1997). "On automatic boundary corrections." *Annals of Statistics* 25(4).
- Clark, D., and Martorell, P. (2014). "The signaling value of a high school diploma." *Journal of Political Economy* 122(2).
- Fan, J., and Gijbels, I. (1992). "Variable bandwidth and local linear regression smoothers." *Annals of Statistics* 20(4).
- Gelman, A., and Imbens, G. (2019). "Why high-order polynomials should not be used in regression discontinuity designs." *Journal of Business and Economic Statistics* 37(3).

References

- Imbens, G., and Lemieux, T. (2008). "Regression discontinuity designs: a guide to practice." *Journal of Econometrics* 142(2).
- Imbens, G., and Kalyanaraman, K. (2012). "Optimal bandwidth choice for the regression discontinuity estimator." *Review of Economic Studies* 79(3).
- Lee, D. (2008). "Randomized experiments from non-random selection in U.S. House elections." *Journal of Econometrics* 142.
- Lee, D., and Lemieux, T. (2010). "Regression discontinuity designs in economics." *Journal of Economic Literature* 48(2).
- McCrary, J. (2008). "Manipulation of the running variable in the regression discontinuity design: a density test." *Journal of Econometrics* 142(2).
- Thistlethwaite, D., and Campbell, D. (1960). "Regression-discontinuity analysis: An alternative to the ex post facto experiment." *Journal of Educational Psychology* 51(6).
- Urquiola, M., and Verhoogen, E. (2009). "Class-size caps, sorting, and the regression discontinuity design." *American Economic Review* 99(1).