Part E: Regression Discontinuity

E1: RDD Basics

Kirill Borusyak

ARE 213 Applied Econometrics

UC Berkeley, Fall 2024

E1 Outline

1. RDD idea and identification
2. Visualization, estimation, and inference
3. Falsification tests
4. A cautionary tale

Reading: Cattaneo, Idrobo, Titiunik ("Practical introduction: Foundations" 2019)

For code in Stata, R, and Python, see https://rdpackages.github.io/

Setting

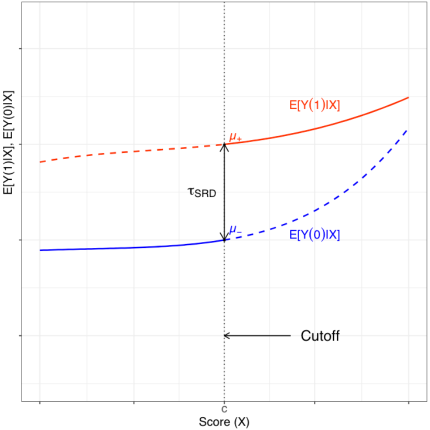
Sharp regression discontinuity design involves:

- Scalar continuous **score** (a.k.a. **running variable**, forcing variable) X_i
- Scalar **cutoff** c (with non-zero density of X_i on both sides)
- Binary treatment $D_i = \mathbf{1}\left[X_i \geq c\right]$
  - Fully determined by the score, no discretion
  - Rule is discontinuous in X_i at c
- **Continuity of** (expected) **potential outcomes** $\mathbb{E}\left[Y_i(0) \mid X_i\right], \mathbb{E}\left[Y_i(1) \mid X_i\right]$ at $X_i = c$
  1. No other determinant of Y_i jumps at $X_i = c$
  2. Score cannot be precisely (and endogenously) manipulated

Identification

Then $\lim_{x\downarrow c}\mathbb{E}\left[Y_i \mid X_i = x\right] - \lim_{x\uparrow c}\mathbb{E}\left[Y_i \mid X_i = x\right] = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right]$

- Discontinuity of regression $\mathbb{E}\left[Y_i \mid X_i\right]$ at c identifies a local causal parameter



(Cattaneo and Titiunik 2019, Figure 2)

Examples

- Effects of financial aid on college enrollment *(van der Klaauw 2002)*
  - Score $X_i = GPA_i + SAT_i$
- Effects of class size on educational achievement *(Angrist and Lavy 1999)*
  - $X_i =$ number of students in a school cohort
  - "Maimonides rule": max class size in Israel = 40; with $X_i = 41$ classes are small

Examples (2)

- Incumbency advantage in a two-party system *(Lee 2008)*
  - $Y_i =$ Democratic candidate elected to U.S. House in district i
  - $X_i =$ vote share of Democratic candidate in the previous election; $c = 0.5$
  - $D_i =$ Democrat is incumbent; $\mathbb{E}\left[Y_i(1) - Y_i(0)\right]$: incumbency advantage
- Effect of displayed Yelp rating on restaurant sales *(Anderson and Magruder 2012)*
  - $X_i =$ actual restaurant rating, e.g. 3.24 or 3.26
  - $D_i =$ displayed rating which is rounded to the nearest 0.5

Is RDD a special case of something?

- Is RDD like selection on observables, with X_i as a control variable?
- Is RDD like IV, with X_i as instrument?
- Is RDD like a RCT in the neighborhood of $X_i = c$?

Answer key

- Is RDD like selection on observables, with X_i as a control?

  No. By construction, there is no overlap: no value of X_i where both $D_i = 0$ and $D_i = 1$ are observed

- Is RDD like IV, with X_i as instrument?

  No. Exogeneity of X_i is not assumed, e.g. higher vote share in election $t - 1$ correlates with higher vote share in t

- Is RDD like a RCT in the neighborhood of $X_i = c$?

  Yes. Continuity of potential outcomes implies their balance around $X_i = c$

  No. This only holds in an infinitesimal neighborhood. So we need to be careful with estimation

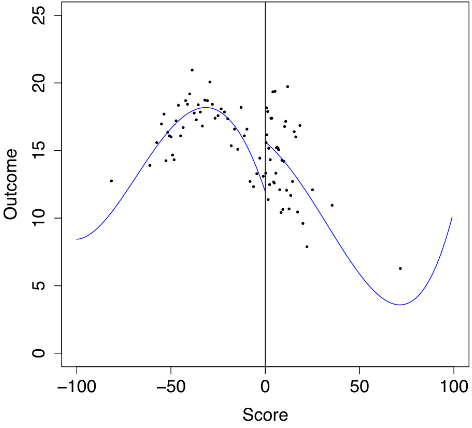
Checklist for sharp RD

- Visualization: “RD plot”
- Estimation and inference
- Falsification tests
  - Balance tests: RD plots and estimates for covariates and placebo outcomes
  - McCrary test for continuous density of X_i around the cutoff

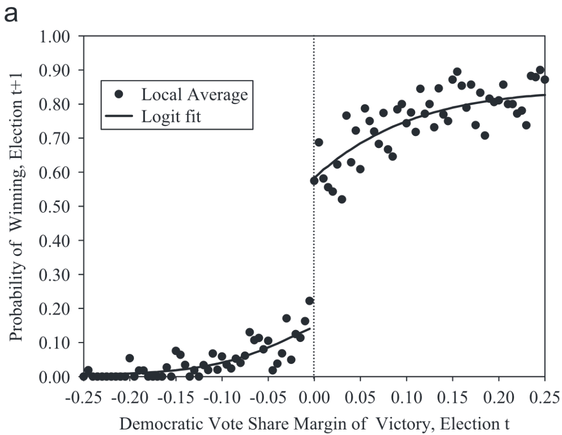
Outline

1. RDD idea and identification
2. Visualization, estimation, and inference
3. Falsification tests
4. A cautionary tale

RD plot



(Cattaneo and Titiunik 2019, Figure 11, Meyersson (2014) data)



(Lee 2008, Figure 2a)

RD plot: Details

Shows discontinuity in regression $\mathbb{E}[Y_i \mid X_i]$ in two ways:

- **Parametric fit**: shows the *global* shape and nonlinearity of the regression
  - Separately on the left and right of c, fitted values from

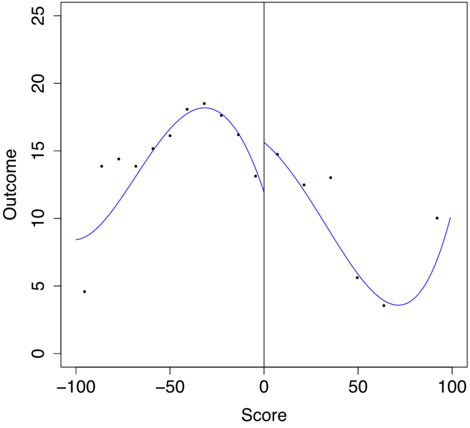
$$Y_i = \alpha_0 + \alpha_1(X_i - c) + \cdots + \alpha_p(X_i - c)^p + \text{error}$$

    e.g. with quartic polynomial ($p = 4$)

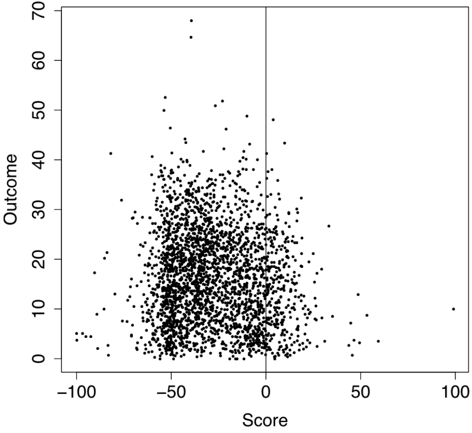
- **Binscatter**: a *local*/nonparametric estimator of the regression
  - Separately on each side, for some bins of X_i: average Y_i against bin's midpoint
  - Bins with similar numbers of observations (splitting by quantile) are more informative. But equal width is also common
  - Calonico, Cattaneo, Titiunik (2015) propose data-driven optimal number of bins...

Binscatter vs. scatterplot

Few bins $\Longrightarrow$ doesn't trace $\mathbb{E}\left[Y_i \mid X_i\right]$ (bias); many bins (e.g. scatterplot) $\Longrightarrow$ noisy



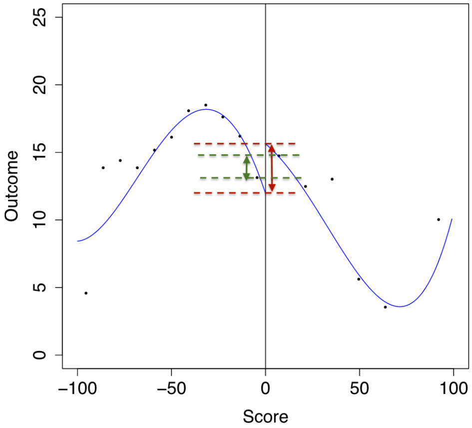
(Figure 8, integrated MSE-optimal number of bins)



(Figure 5: Scatterplot)

Estimation

RD plots yield two estimates of the causal effect $\tau = \mathbb{E}\left[Y_i(1) - Y_i(0) \mid X_i = c\right]$:



(Figure 8 again)

Estimation (2)

Are these estimators good?

- Global extrapolation using higher-order polynomials has bad properties at the border *(Gelman and Imbens 2019)*:
  - Noisy and highly sensitive to the order of the polynomial
- Difference in outcome means between the nearest bin on the right vs. on the left?
  - This "local constant regression" is too biased
  - Instead, use **local polynomial regression**, e.g. **local linear** ($p = 1$)

Local linear regression

- Estimate $\lim_{x \downarrow c} \mathbb{E}\left[Y_i \mid X_i = x\right]$ by $\hat{\alpha}_+$ from

$$(\hat{\alpha}_+, \hat{\beta}_+) = \arg\min_{\alpha_+, \beta_+} \sum_{i:\ c \leq X_i \leq c + h_+} \left(Y_i - \alpha_+ - \beta_+(X_i - c)\right)^2 \kappa\left(\frac{X_i - c}{h_+}\right)$$

where $h_+ > 0$ is some **bandwidth** and $\kappa(\cdot)$ is a kernel function, e.g.

  - Uniform kernel: $\kappa(x) = \mathbf{1}\left[|x| \leq 1\right]$ (uses all obs. in the neighborhood)
  - Triangular kernel: $\kappa(x) = \max\left\{1 - |x|, 0\right\}$ (weights obs. closer to c more)

Local linear regression (2)

- Estimate $\lim_{x \uparrow c} \mathbb{E}\left[Y_i \mid X_i = x\right]$ by $\hat{\alpha}_-$ from

$$(\hat{\alpha}_-, \hat{\beta}_-) = \arg\min_{\alpha_-, \beta_-} \sum_{i:\ c - h_- \leq X_i < c} \left(Y_i - \alpha_- - \beta_-(X_i - c)\right)^2 \kappa\left(\frac{X_i - c}{h_-}\right)$$

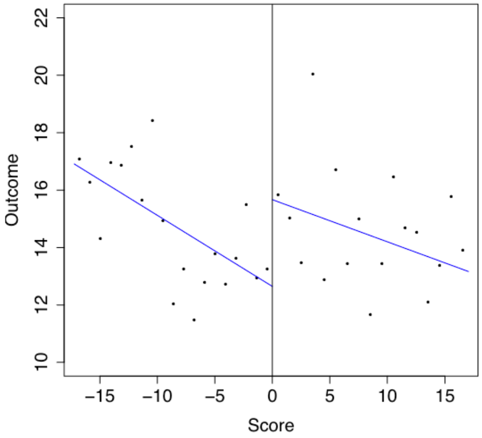
  (with $h_- = h_+$ or $h_- \neq h_+$)

- Estimate τ by $\hat{\tau} = \hat{\alpha}_+ - \hat{\alpha}_-$
  - Can implement by a single regression (on $X_i \in [c - h_-, c + h_+]$ and with kernel weights):

$$Y_i = \tau D_i + \gamma_0 + \gamma_1(X_i - c) + \gamma_2(X_i - c)D_i + \text{error}$$

    (but don't treat it like a true model, and be careful with SE!)

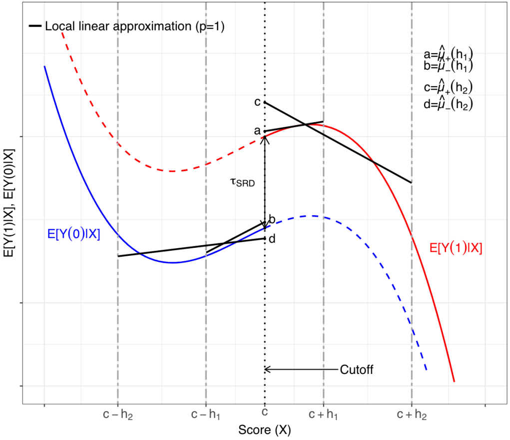
Local linear regression (3)



(Figure 15. Note more narrow range of scores)

Bandwidth choice

- Bandwidth choice is much more important the choice of kernel:



(Figure 14)

Optimal bandwidth choice

- For local polynomial of order p,

$$\text{bias}(\hat{\tau}) \approx B \cdot h^{p+1}, \qquad \mathrm{Var}\left[\hat{\tau}\right] \approx \frac{V}{Nh}$$

    - B is determined by the curvatures $d^{p+1}\mathbb{E}\left[Y \mid X = x\right]/dx^{p+1}\mid_{x=c}$ on each side
    - V is determined by the variance of Y_i and density of X_i
- Thus, $MSE \approx (Bh^{p+1})^2 + \frac{V}{Nh}$ is minimized at $h^* = \left(\frac{2(p+1)B^2}{V}N\right)^{-1/(2p+3)}$
    - E.g. $\propto N^{-1/5}$ when $p = 1$
    - $h^* \uparrow$ when bias is smaller: $|B| \downarrow$, $p \uparrow$
    - $h^* \downarrow$ when variance is smaller: $V \downarrow$, $N \uparrow$
- If we can estimate V (easy) and B, we can compute h^*
    - Calonico, Cattaneo, Titiunik (2014): to estimate B, run local polynomial estimation with order $q \geq p + 1$ (with a larger "pilot" bandwidth)

Inference and bias correction

- Problem: h^* minimizes MSE by trading off bias^2 and variance
  - At h^*, bias and SE of the same order $\implies$ conventional conf. intervals are wrong!

$$\sqrt{Nh^*}\left(\hat{\tau}-\tau\right)\xrightarrow{d}\mathcal{N}\left(B^*,V^*\right),\qquad B^*\neq 0$$

- One solution: "undersmoothing"
  - Use bandwidth h much smaller than h^*. Then inference is fine (bias $\ll$ SE)
  - But unclear how to choose h, and would yield a higher-MSE estimator
- Another solution: "robust bias correction" *(Calonico et al. 2014,* `rdrobust`*)*
  - We already estimated the bias $\implies$ let's subtract it from $\hat{\tau}$
  - Adjust SE for noise in bias estimation $\implies$ "Robust bias-corrected conf. interval"
  - Why not debias $\hat{\tau}$, too? Higher MSE because of bias estimation

Nearest neighbor inference

- Problem 2: recall the local linear estimating equation

$$Y_i = \tau D_i + \gamma_0 + \gamma_1 (X_i - c) + \gamma_2 (X_i - c) D_i + \text{error}$$

- Even if the bias doesn't arise (undersmoothing), SEs are upward biased
  - The error includes nonlinearity of $\mathbb{E}\left[Y(d) \mid X\right]$ $(d = 0, 1)$

- One solution: estimate errors $\varepsilon_i = Y_i - \mathbb{E}\left[Y_i \mid X_i\right]$ from variation of Y among J neighbors, nearest in terms of the score and on the same side:

$$\hat{\varepsilon}_i = \sqrt{\frac{J}{J+1}} \left(Y_i - \bar{Y}_{\text{neighbors}(i)}\right)$$

rdrobust with default options

```
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 2314         315
Eff. Number of Obs.             529         266
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  17.239      17.239
BW bias (b)                  28.575      28.575
rho (h/b)                     0.603       0.603

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     3.020     1.427     2.116     0.034     [0.223 , 5.817]
        Robust         -         -     1.776     0.076    [-0.309 , 6.276]
=============================================================================
```

(Cattaneo and Titiunik 2019, Code snippet 21)

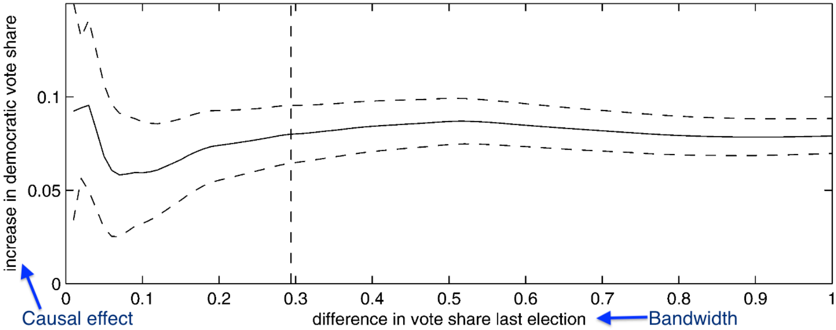
If you insist to report bias-corrected estimate

```
Call: rdrobust

Number of Obs.                 2629
BW type                       mserd
Kernel                   Triangular
VCE method                       NN

Number of Obs.                 2314         315
Eff. Number of Obs.             529         266
Order est. (p)                    1           1
Order bias  (p)                   2           2
BW est. (h)                  17.239      17.239
BW bias (b)                  28.575      28.575
rho (h/b)                     0.603       0.603

=============================================================================
        Method     Coef. Std. Err.         z     P>|z|      [ 95% C.I. ]
=============================================================================
  Conventional     3.020     1.427     2.116     0.034     [0.223 , 5.817]
Bias-Corrected     2.983     1.427     2.090     0.037     [0.186 , 5.780]
        Robust     2.983     1.680     1.776     0.076    [-0.309 , 6.276]
=============================================================================
```

(Cattaneo and Titiunik 2019, Code snippet 22)

Robustness to bandwidth choice

While we know the optimal bandwidth, checking sensitivity to this choice is also useful:



(Imbens and Kalyanaraman 2012, Figure 3)

Outline

1. RDD idea and identification
2. Visualization, estimation, and inference
3. Falsification tests
4. A cautionary tale

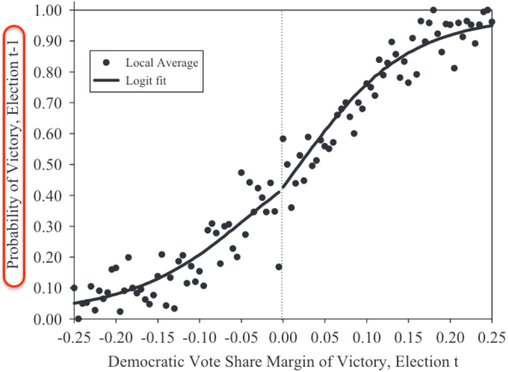
Falsification tests

- The assumption of continuity of potential outcomes is not testable
- But a slightly stronger assumption — "that X_i is influenced partially by random chance" *(Lee 2008)* — has two testable implications:

  1. Balance: distribution of predetermined variables W_i (lagged covariates or outcomes) should be continuous at the cutoff
  2. Density of X_i should be continuous at the cutoff

- *Note*: contextual knowledge, e.g. on how easy it is to manipulate X_i, is still indispensible

Placebo RD plots and estimates

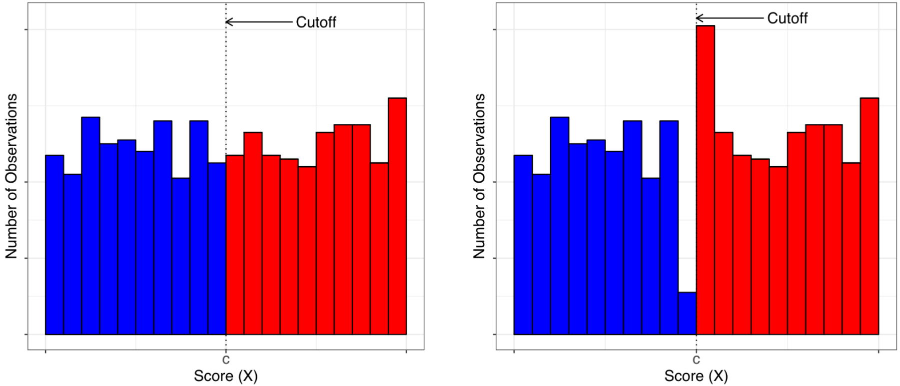


Dependent variable	(1) Vote share $t+1$	(2) Vote share $t+1$	(3) Vote share $t+1$	(4) Vote share $t+1$	(5) Vote share $t+1$	(6) Res. vote share $t+1$	(7) 1st dif. vote share, $t+1$	(8) Vote share $t-1$
Victory, election t	0.077 (0.011)	0.078 (0.011)	0.077 (0.011)	0.077 (0.011)	0.078 (0.011)	0.081 (0.014)	0.079 (0.013)	−0.002 (0.011)

(Lee 2008, Figure 5b and Table 2)

Discontinuity of density ("bunching") test

McCrary (2008): Discontinuity of density of X_i around $X_i = c$ suggests manipulation

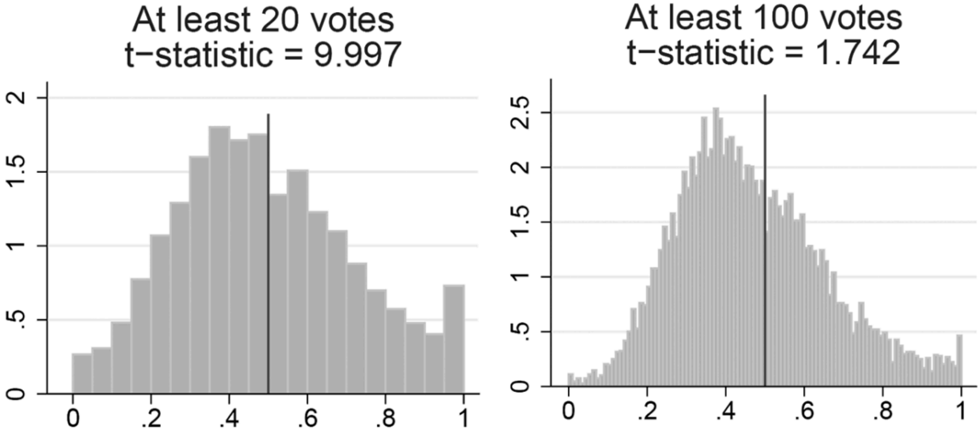


(a) No Sorting (b) Sorting

(Illustration from Cattaneo and Titiunik 2019, Fig. 18)

Discontinuity of density in practice

A real issue in the literature on the effect of unionization, using close elections RDD:

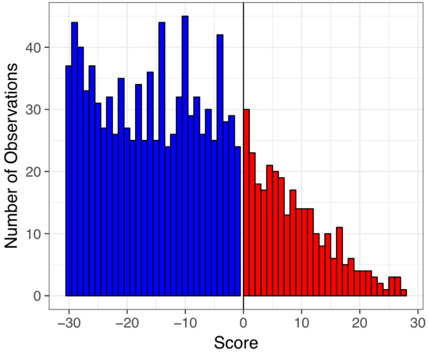


(Frandsen 2021, Fig. 1)

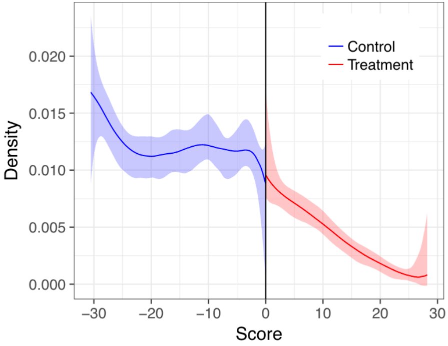
Convenient implementation

Cattaneo, Jansson, Ma (2020), `rddensity` package:

- Density is the slope of CDF, which is easy to estimate $\Longrightarrow$ estimate from a local polynomial approximation to the CDF



(a) Histogram



(b) Estimated Density

(Cattaneo and Titiunik 2019, Fig. 19)

Outline

1 RDD idea and identification

2 Visualization, estimation, and inference

3 Falsification tests

4 A cautionary tale

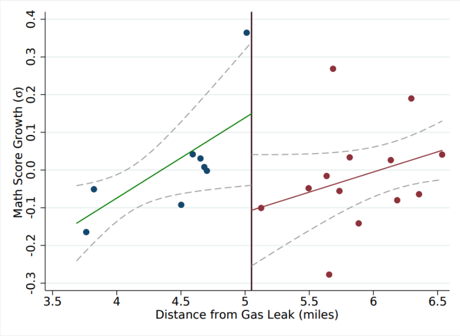
A cautionary tale

- We have discussed an *algorithm* for estimating causal effects in RDDs
  - Plots, estimators, inference methods, tests
- But blindly following the algorithm is not enough to get to the truth
  - Illustration following Andrew Gelman's blog posts in 2019 and 2020 in the RDD context but the lesson is broader

Gilraine (2020): Effect of air filters on student achievement

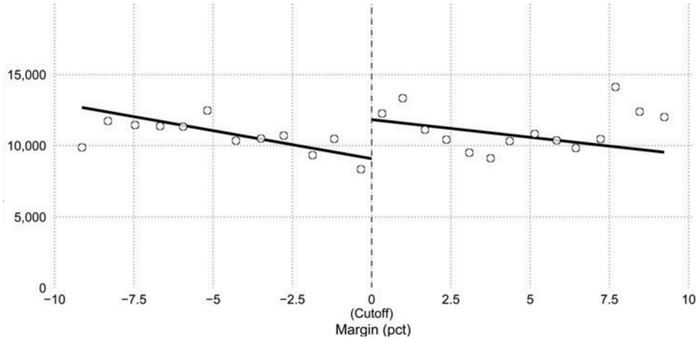
- Filters installed in all schools within 5 miles of a big gas leak in Los Angeles
- *"Once the distance to the gas leak exceeds five miles we see a substantial drop in test score growth in both math and English. This provides clear and convincing evidence that air filters substantially raised test scores."*

(a) Math Score Growth



Barfort et al. (2021)

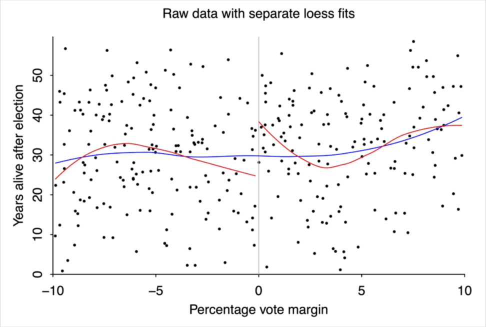
- Study the effects of a candidate winning gubernatiorial election on life expectancy!
- Significant local linear regression estimates $\hat{\tau} \approx 2000$–3000 days
  - Report placebo outcomes, robustness to bandwidth and polynomial order, etc.



Lessons from Barfort et al. (2021)?

- Gelman tries to replicate it from scratch and finds it difficult
  - Many choices during data cleaning not be captured by robustness checks
  - *"The garden of forking paths"*
- The effect magnitude is entirely implausible
  - But how should we use our priors?
- Raw data is noisy
  - Different models can fit them in different ways. *"No smoking gun"*
  - But should we just give up?

Lessons from Barfort et al. (2021)?



(Gelman's reanalysis of the raw data)