Regression Discontinuity

Han Zhang

Apr 20, 2021

Outline

Logistics

- Presentation next week
- 20 minutes each
- Additional scores if you ask questions/raise suggestions
- At least play with your dataset. You may have a hard time turning your regression on paper to codes in software

Recommended Readings

- MHE, Chapter 6 (too short)
- Cattaneo, Matias D., Nicolás Idrobo, and Rocío Titiunik. *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Cambridge University Press, 2019.
    - Freely available at the author's website: `https://sites.google.com/site/rdpackages/replication/cit-2019-cup`

Setup

- $0 < P(D = 1) < 1$
- Non-zero treatment probability means every unit has some probability to be treated or in the control group
  - Also called positivity, common support, overlap condition
- Regression/matching/fixed effects/DiD all assume this
- Basically, this assumption makes sure that we can find treatment and control units similar to each other, and use the counterfactual of control to predict that of the treated units
  - e.g., twin studies, DiD, matching, etc.

Why Non-zero treatment probability is important?

- If non-zero treatment probability does not hold, it means that some units can never be in treatment or in control groups
- IV: never-takers/always-takers
- How do we predict their counterfactual outcomes, then?
- IV estimates effect for compliers, not the other two groups.

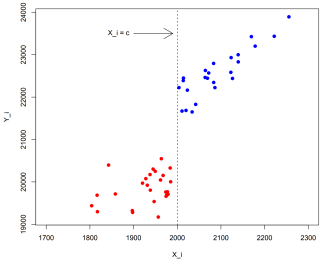
Setup

- Without positivity assumption, it's hard to do counterfactual predictions
- But in some special case, we can manage to work out something
- Example: Chinese College Entrance Exam (Gaokao)
  - Each university has a clear threshold
  - Students on or above the threshold can go to their dream school
  - Students below the threshold cannot
  - D: admitted by some college or not
    - D is fully determined by some other variable X (here, X is score in Gaokao)
  - Y: future earnings
- Example 2: voting share $>=$ 50% -$>$ win the election

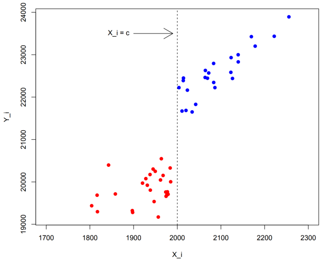
Sharp Regression Discontinuity

- Three core elements
- Running variable, or scores X
- Cutoff or threshold c
- Treatment assignment D, which is fully determined by X based on cutoff c

$$D_i = \begin{cases} 1 & \text{if } X_i \geq c \\ 0 & \text{if } X_i < c \end{cases} \tag{1}$$



RD Intuition



- Positivity assumption is broken: treated and control units do not have overlap on X
- But we can compare Y between points on and just below the cutoff
  - $X = c$ vs. $X = c - \epsilon$,

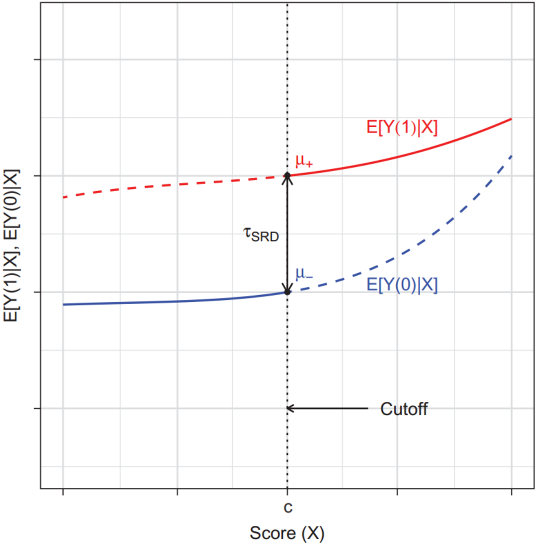
Sharp RD treatment effect

- If ϵ becomes infinitely small, we obtain the causal effect for $X = c$
- Sharp RD treatment effect

$$\tau_{\mathrm{SRD}} \equiv \mathbb{E}\left[Y_i^1 - Y_i^0 | X_i = c\right] = \lim_{x \downarrow c} \mathbb{E}\left[Y_i | X_i = x\right] - \lim_{x \uparrow c} \mathbb{E}\left[Y_i | X_i = x\right] \tag{2}$$

- In contrast to ATT or ATE, Sharp RD identifies a local effect
- The key assumption is that $\mathbb{E}\left[Y_i^1 | X_i = c\right]$ and $\mathbb{E}\left[Y_i^0 | X_i = c\right]$ are continuous

Sharp RD

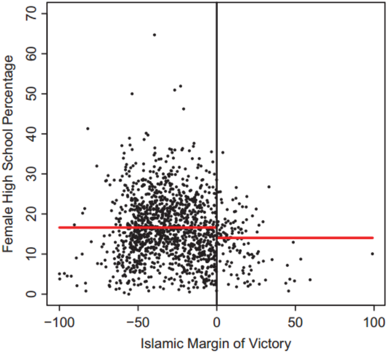


Example

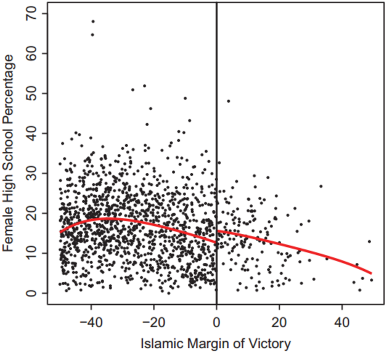
- Meyersson, "Islamic Rules and the Empowerment of the Poor and Pious", *Econometrica*, 2014
- Question: whether Islamic rules make female less likely to be educated?
- X: vote share margin; vote percentage of an Islamic mayor candidates - vote percentage of a secular candidate, in 1994 Turkish local elections
- c: 0 if Islamic candidate won
- Y: share of local women aged 15 to 20 in 2000 who had competed high school by 2000

RD plot: global vs local

- Globally: negative impact
- Locally: positive impact
- Since RD is about local effect, data far away from the cutoff are not useful

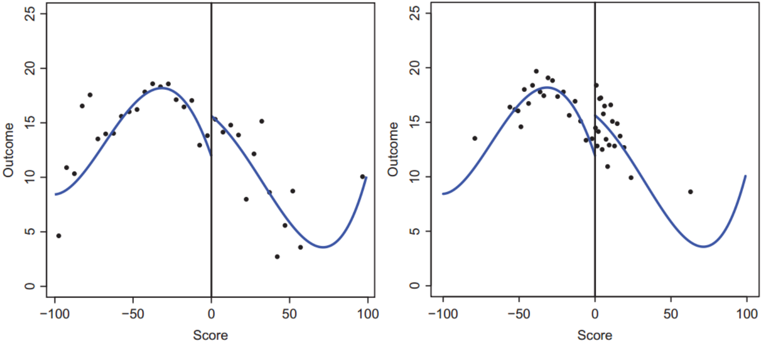


(a) Raw Comparison of Means



(b) Local Comparison of Means

RD plot: aggregated individual data



(a) 40 Evenly-Spaced Bins (b) 40 Quantile-Spaced Bins

Figure 7 RD Plots (Meyersson Data)

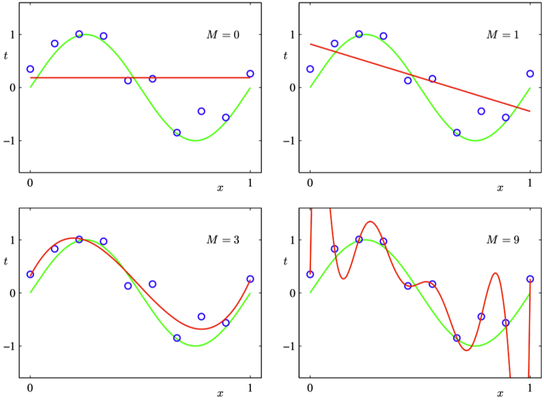
Estimating Sharp RD Treatment Effect

- Theory: calculate the vertical distance between those on the boundary and those just below the boundary
- Reality: if X is continuous, there are no (or sometimes in practice very few) observations around c
- Solution: use data in a small region $[c-h, c+h]$ and a prediction function to predict $E(Y^1|X=c)$ and $E(Y^0|X=c)$
- h is called bandwidth
- And the current default choice is local polynomial regression:

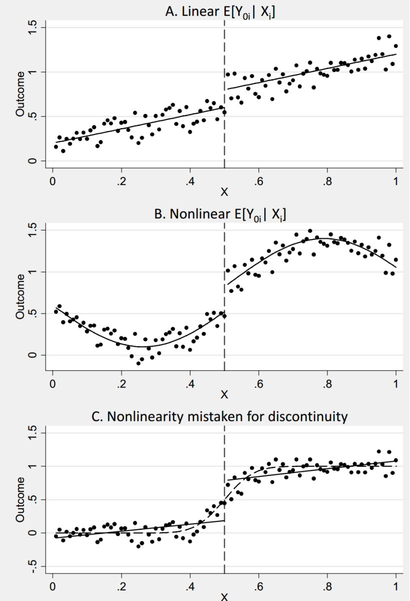
$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_p X^p$$

Polynomial example

- here M is p: order of polynomial

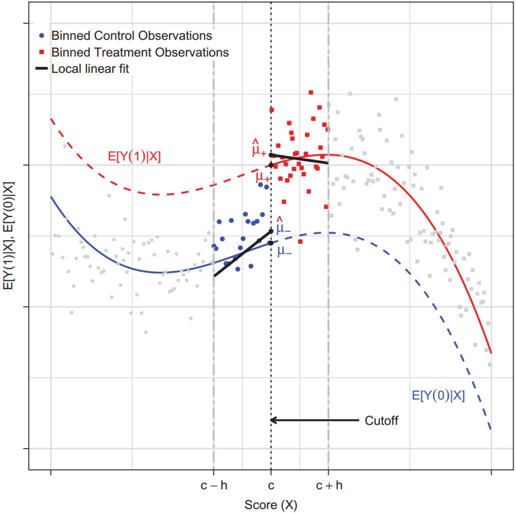


Extrapolation function matters



Using Local Polynomial Regression to predict

- here $p = 1$: linear prediction



Sharp RD procedures

1. Choose a polynomial order p and a kernel function $K()$.
   - Kernel essentially add weights to points according to their distances to c
   - p is typically 3 - 5
2. Choose a bandwidth h
3. For $X_i \in [c-h, c)$, Fit a weighted linear regression of Y on $X_i - c, (X_i - c)^2, , (X_i - c)^p$, use weights based on $K(\frac{X_i - c}{h})$. Estimate of $E(Y^0|X = c) = \hat{\mu}_-$

$$\hat{\mu}_- : \hat{Y}_i = \hat{\mu}_- + \hat{\mu}_{-,1}(X_i - c) + \hat{\mu}_{-,2}(X_i - c)^2 + \cdots + \hat{\mu}_{-,p}(X_i - c)^p$$

4. The same thing for $X_i \in [c, c+h]$; obtain estimate of $E(Y^1|X = c) = \hat{\mu}_+$
5. Sharp RD treatment effect is simply the difference in means:

$$\hat{\tau}_{\mathrm{SRD}} = \hat{\mu}_+ - \hat{\mu}_- \qquad (3)$$

Different kernels

- in practice, kernal choices are less sensitive

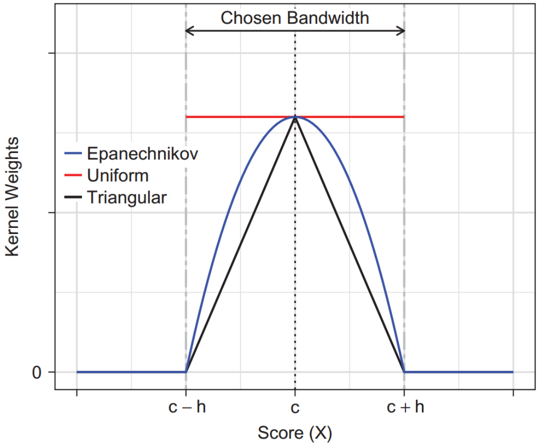
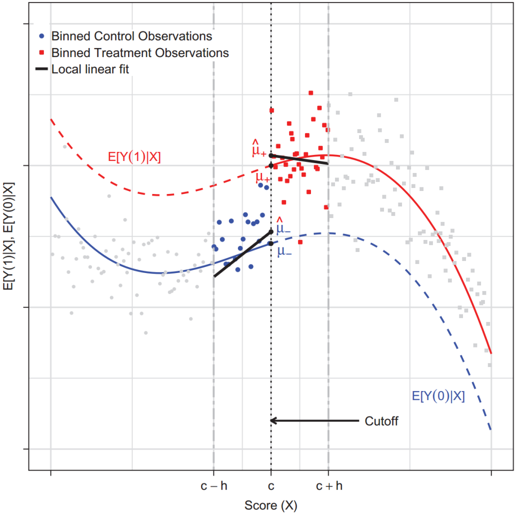


Figure 13 Different Kernel Weights for RD Estimation

Bandwidth matters



Bandwidth choices

- Smaller bandwidth implies that the choice of prediction function matters less
  - Extreme case: bandwidth = 0 and we just calculate difference in means
  - But we have fewer data to work with, leading to larger variance
- Larger bandwidth may lead to more biased results (if you chose the wrong prediction function)
  - But reduce the variances of estimates of $E(Y^0|X = c)$ and $E(Y^1|X = c)$
- Bias-variance trade-off again

Bandwidth Size

- A more promising approach is data-driven bandwidth selection
- First developed in Imbens and Kalyanaraman, 2012, *Review of Economic Studies*
- And then in Calonico, Cattaneo and Titiunik, 2014, *Econometrica*

Bandwidth selection

- Key idea: bias-variance trade-off
- Find a bandwidth that balances the prediction MSE

$$\mathsf{MSE}\left(\hat{\tau}_{\mathrm{SRD}}\right) = \mathsf{Bias}^2\left(\hat{\tau}_{\mathrm{SRD}}\right) + \mathsf{Variance}\left(\hat{\tau}_{\mathrm{SRD}}\right) \tag{4}$$

- With local polynomial regression, expected bias for $X \in [c-h, c)$ is (n_- is number of units in $[c-h, c)$

$$\begin{aligned}\mathsf{Bias}\left(\hat{\tau}_{\mathrm{SRD}}\right) &= \hat{E}(Y^0|X=c) - E(Y^0|X=c) \\ &= \frac{1}{n_-}\sum_i \left(\hat{\mu}_- + \hat{\mu}_{-,1}(X_i - c) + \cdots + \hat{\mu}_{-,p}(X_i - c)^p\right) - E(Y^0|X=c)\end{aligned} \tag{5}$$

- And bias for $X > c$ can be similarily written down
- Total bias is the sum of bias for $X < c$ and bias for $X > c$

Bandwidth selection

- With Taylor Expansion, we can also write $E(Y^0|X = X_i)$ as:

$$E(Y^0|X = X_i) = \sum_{p=0}^{\infty} \left(\frac{d^p E(Y^0|X)}{dX} (X_i - c)^p \right) \tag{6}$$

- So substititing the Taylor expansion into Equation (3), terms up to p-th order will be cancelled out
- And we also discard all terms from $p + 2$ (since $X_i - c$ is already small)
- And assume we have infinite amount of data (so called asymptotic analysis), and take the limits

$$\text{Bias}\left(\hat{\tau}_{\text{SRD}}\right) \approx \lim_{x_i \to c} E(Y^0|X = X_i) - E(Y^0|X = c) \approx h^{p+1} B, \tag{7}$$

- Where h is bandwidth and B is asymptotic bias (those cannot be removed by taking limits)

Bandwidth selection

- Using similar idea, variance can be roughly expressed as

$$\text{Variance}\left(\hat{\tau}_{\mathrm{SRD}}\right) = \frac{1}{nh} V, \tag{8}$$

- where V is asymptotic variance
- Overall, we want our choice of h to minimize prediction MSE:

$$\begin{aligned}\text{MSE}\left(\hat{\tau}_{\mathrm{SRD}}\right) &= \text{Bias}^2\left(\hat{\tau}_{\mathrm{SRD}}\right) + \text{Variance}\left(\hat{\tau}_{\mathrm{SRD}}\right) \\ &\approx h^{2(p+1)} B^2 + \frac{1}{nh} V\end{aligned}$$

- Solve the above and the estimate of h is :

$$h_{\mathrm{MSE}} = \left(\frac{V}{2(p+1)B^2}\right)^{1/(2p+3)} n^{-1/(2p+3)} \tag{9}$$

Intuition of bandwidth

- Say if $p = 1$,

$$h_{\mathrm{MSE}} = \left(\frac{V}{2B^2}\right)^{1/5} n^{-1/5} \tag{10}$$

- It's on a scale of $n^{1/5}$; $n = 1000$ leads to 3.98
- $p = 3$ leads to a scale of $n^{1/9} = 2.15$
- And if the asymptotic bias B increasese, it suggests that you should narrow your region, hence leading to smaller h
- If the asymptotic bias V increases, we have relied on so few data points, hence leading to larger h

Some extensions on bandwidth selection

- Bandwidth chosen from the above can be quite small; so people often add regularization term to force it bigger
- There are alternative ways to do it: cross-validations
- Cross-validation bandwidth selection ideas:
- Use data in $[c - h - c1, c - h]$ (training data), c_1 is again a small number
- Predict data in $[c - h, c)$ (test data)
- And the best choice of h should make this prediction MSE the smallest

Cross-validation example

- Lee and Lemieux, 2010, *Journal of Economic Literature*

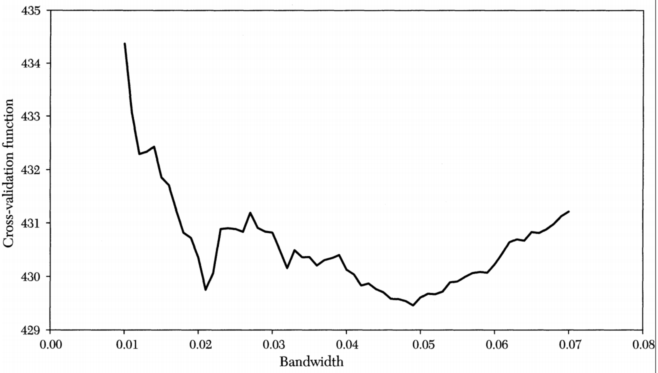


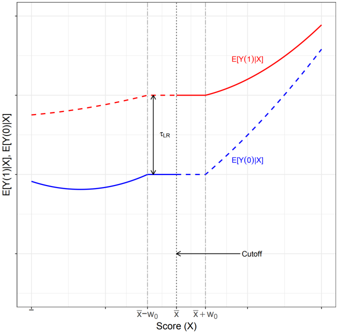
Figure 12. Cross-Validation Function for Choosing the Bandwidth in a RD Graph: Winning the Next Election

RD assumption Clarification

- The key assumption is that $\mathbb{E}\left[Y_i^1 | X_i = c\right]$ and $\mathbb{E}\left[Y_i^0 | X_i = c\right]$ are continuous
- You may see a misunderstanding of RD as a local randomness assumption:
  - with a small boundary $[c - h, c + h]$,
- The local randomness assumption is not necessary! It is more demanding than the continuity assumption
- E.g., in voting example, it's hard to argue that districts in which parties has a narrow win share is due to randomness.

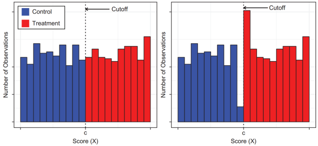
Local randomness assumption in figures

- Local randomness assumption essentially means that the counterfactual outcome is independent of the treatment within $[c - h, c + h]$
- And independence implies straight lines of $E(Y^1|X = c)$ and $E(Y^1|X = c)$
- This is a stronger assumption than local continuity



Credibility of RD assumptions

- Density test is popularized by McCrary, 2008, *Journal of Econometrics*
- There may be selection: people realized the cutoff and select themselves into treatment
- Density test: even people try, they have limited power and the number of observations below/above cutoff should be continuous
- Density test: plot X against the number of observations
    - note that RD plot is X against Y



(a) No Sorting (b) Sorting

Credibility of RD assumptions

- We can again use placebo test to falsify RD assumptions
- X determines D, but it should not determine other controls

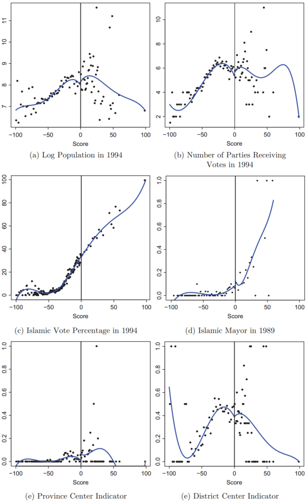


Figure 16 RD Plots for Predetermined Covariates (Meyersson Application)

RD as a linear regression

- Some older literature write RD as a linear regression (e.g., MHE)

$$
\begin{aligned}
Y_i = \alpha + \tau_{\mathrm{SRD}} D_i + \mu_{-,1}(X_i - c) + \cdots + \mu_{-,p}(X_i - c)^p \\
+ \mu_{+,1} D_i (X_i - c) + \cdots + \mu_{+,p} D_i (X_i - c)^p
\end{aligned}
\tag{11}
$$

- Note the interaction between D and distances to cutoff for treated users
- Regression estimator provides the equivalent point estimate to the difference-in-mean estimator
- But the standard error estimates from raw regression are usually wrong; they have not considered the variances in the bandwidth selection process
- Modern RD packages ('rdrobust' in R and Stata) generally does not use regression under the hood.

Covariates in RD

- We have discussed how to estimate RD by comparing means at the cutoff
- Note that no covariates have been used so far
- Like randomized experiments, adding controls reduces standard errors, but should not change point estimate too much
- Cattaneo et al., 2019, p. 71:

Analogously to the case of randomized experiments, the generally valid justification for including covariates in RD analysis is the potential for efficiency gains, not the promise to fix implausible identification assumptions.

Covariates in RD

- If you insist to add covariates, you can fit a regression like the below:

$$
\begin{aligned}
Y_i = \alpha + \tau_{\mathrm{SRD}} D_i + \mu_{-,1}(X_i - c) + \cdots + \mu_{-,p}(X_i - c)^p \\
+ \mu_{+,1} T_i (X_i - c) + \cdots + \mu_{+,p} T_i (X_i - c)^p + \mathbf{Z}_i' \gamma
\end{aligned}
\tag{12}
$$

- Coefficient on D_i returns estimates for τ_{SRD}
- Z_i are additional covariates
- Z_i should have been balanced (use placebo test in the previous slides)
- And if Z_i is balanced, it can be proven that τ_{SRD} estimated with and without Z should be similar

Fuzzy RD

- In Sharp RD, running variable X and cutoff c fully determines treatment assignment D
- In Fuzzy RD, running variable X and cutoff c probabilistically determines D
- But X does not directly influence Y
- Sounds familiar?
- Fuzzy RD is equivalent to IV
  - Running variable X becomes the instrument
  - The only practical difference is that you need to select bandwidth

Fuzzy RD

- And the causal effect under can again be obtained from Wald estimator:

$$\tau_{FRD} = \frac{\hat{E}(Y^1|X=c) - \hat{E}(Y^0|X=c)}{\hat{E}(D^1|X=c) - \hat{E}(D^0|X=c)} \tag{13}$$

- Numerator: effect of treatment assignment X on outcome Y, at the cutoff c (ITT)
- Denominator: effect of treatment assignment X on treatment take-up D, also at the cutoff
- τ_{FRD} again estimates local treatment effect for compliers (those who actually follows the assignment of the running variable)