Regression Discontinuity Empirics

Vinish Shrestha

October 4, 2021

Some Folklores About RD

- RD is invalid if individuals can manipulate the treatment assignment
- If individuals cannot precisely manipulate the treatment assignment, RD effects can be similar to localized randomization
- Baseline Test of RD has similar connotation to Baseline test of a randomized experiment.
- RD effects are also related to the literature of ATE, where RD effects are just the weighted ATE, and the weights are determined by how close the running variable is to the cutoff.
- Non-parametric RD is not a solution to the shortcoming of parametric RD design. NP RD should be viewed as complements rather than substitutes for non-parametric RD.
- Bandwidth selection is crucial in minimizing bias on RD estimate

What is RD?

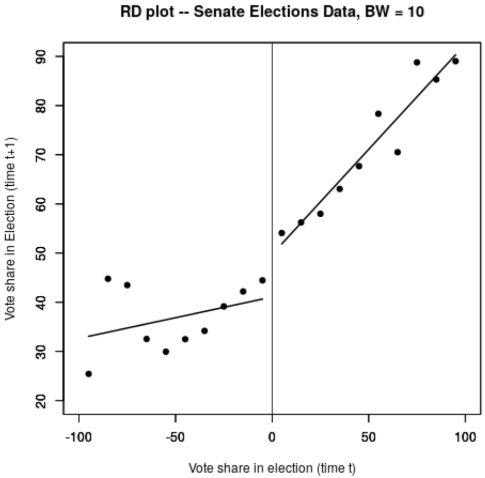


Figure:

From Lee (2008). Here, the bandwidth is 10.

Regression Discontinuity

- In Figure 1 the x-axis shows the vote share of the Demotracts in time (t) in a statewide election for a Senate seat (running variable), and y-axis pertains to time (t+1).
- If the margin is above 0, Democratic party wins against the strongest opponent.
- If one is interested in the effect of incumbent on vote share in coming election, one can compare cases right around the cutoff point–0.
- The idea is that these cases right around the cutoff point, to the left and right, will be quite comparable. We can compare situations when democrats barely won and lost, and at this margin, willing or losing can be dictated by luck or some other random events.

For RD Design to Work–Assumptions

1. All factors affecting the outcome variable $Y_i$should move smoothly with respect to X (running variable), without any discrete jump at point c (cutoff point).
   - This implies the balance test in RD setting analogous to the balance test done in RCT setting.
   - "continuity assumption"
2. Individuals should not be able to manipulate the treatment (in other words decide whether they want to have $c > 0$ or $c < 0$.
   - If individuals can precisely manipulate the treatment then individuals who benefit from the treatment will get the treatment and those who don't will not. This will create selection issue.

RD Framework

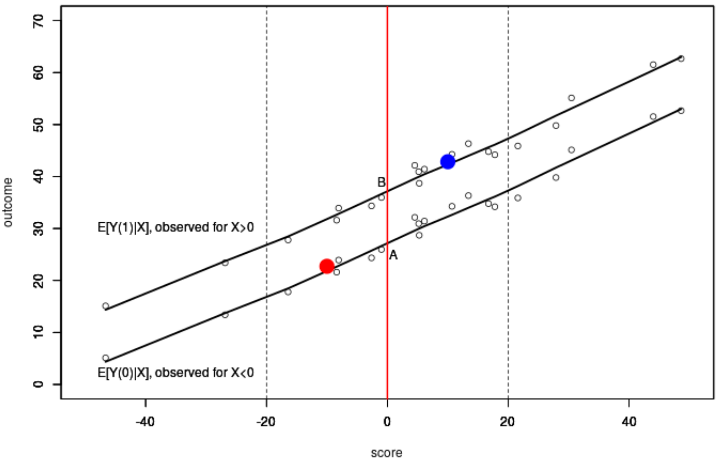
- For individual i, there exits two potential outcomes — $Y_i(1)$, person i is treated and $Y_i(0)$, person i is not treated.
- The treatment effect for person i is $Y_i(1) - Y_i(0)$. The ATE is $E[Y_i(1) - Y_i(0)]$.
- However, an individual i cannot be observed in two states at the same time.
- Consider Figure 2, RD estimate using the continuity of $E[Y_i(1)|X]$ and $E[Y_i(1)|X]$ is

$$
\begin{aligned}
B - A &= lim_{\epsilon \downarrow 0} E[Y_i|X_i = c + \epsilon] - lim_{\epsilon \uparrow 0} E[Y_i|X_i = c + \epsilon] \\
&= E[Y_i(1)|X = c] - E[Y_i(0)|X = c]
\end{aligned}
$$

- RD estimate at point c (if feasibly estimated), is the just the average treatment effect. Those denied treatment right around the cut off serves as a valid counterfactual.

Conceptual Figure

Figure:



This figure will be used to provide intution throughout.

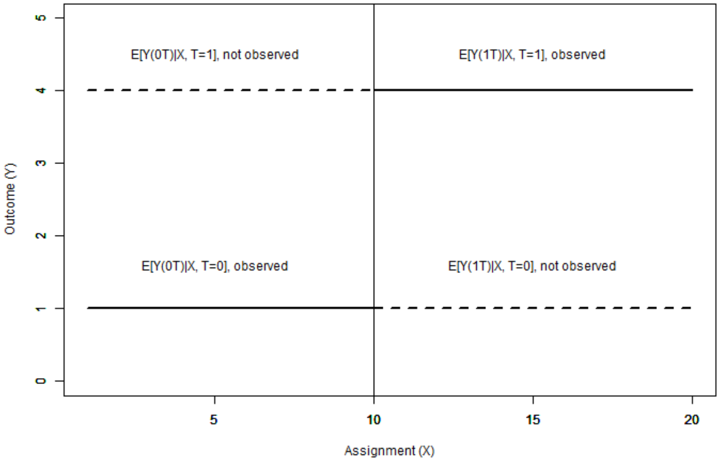
RD and Ignorability or Unconfoundedness and Overlap

1. Ignorability: $E[Y_T(1)|T=1,X]=E[Y_C(1)|T=1,X]$ Effects would be the same to the treated group had the control group hypothetically received the treatment.
   - ignorability is trivally satisfied
   - Treatment follows a random distribution conditional on X

2. Overlap: This assumption is violated, as it is not possible to observe units with both $T=1$ and $T=0$.
   - need continuity assumption to compensate fo the lack of overlap assumption.

Viewing RD as a Randomized Experiment?

Figure:



RD in relation to Randomized Experiment and Vice-Versa.

RD and Randomized Experiment

- Note that

$$ATE = E[Y(1T)|X, T = 1] - E[Y(1T)|X, T = 0]$$

$$= E[Y(0T)|X, T = 1] - E[Y(0T)|X, T = 0];$$

$$= E[Y(1T)|X, T = 1] - E[Y(0T)|X, T = 0]$$

- As the assignment variable is completely flat in the case of a randomized experiment, the curves (analogous to the RD curves) are flat.
- The curves being flat, trivally statisfies the continuity argument, given by the dotted lines. Hence, continuity directly comes from randomized experiment when related to the RD desgin.

RD and Randomized Experiment

- In Figure 3, Assignment into the Treatment group is given by X. Say, from a lottery draw, people above the 10 are given the job program treatment and those less below aren't. The outcome variable is time taken to find a job.
  - This is a classic case of a randomized experiment.
- Let's bring in some systematic assignment that contaminates the treatment assignment. Say, people are compensated for being unlucky by the size of monetary benefits; i.e. monetary benefits is proportional to X. In this case, the curve $E[Y(0T)|X, T = 0]$, would be upward sloping. This resembles the typical RD design.
  - In this case, comparison of means won't show the treatment effects
- However, look closeby to the cutoff point, people would receive similar monetary benefits that flows continuously over the threshhold.
- This creates local randomized experiments around the cutoff.

Identification

- Technically, it is important to have the following assumptions to have a valid RD design

1. Individuals should not be able to precisely control the assignment.
   - It is ok if there is imprecise control, due to some stochastic error term.
2. Baseline characteristics need to be balanced.
   - smoothly moving over the cut-off point
   - no discrete jumps around the cut-off

Identification: Precise vs Imprecise Control

- Let's take an example: Say a policy was implemented such that households within a boundary receive an unconditional cash-transfer, whereas those outside of the boundary receive nothing. Say, this cash trasfer program was implemented in 11/1/2019.
    - Imprecise control: It is true that people can relocate and if the benefits are high enough, people will relocate to live in an area with the program. However, while analyzing the short term impact, if we rely on survey that was conducted in 2020, RD design will be feasible. Yes, people will move, but to a certain degree the actuality of moving is determined by some level of randomness (given the short amount of time). Hence, people below the cut off and above the cut off will still be comparable.
    - Precise control: But evaluating the long term impact is problematic, as people who highly benefit from the program will choose to move and may relocate within but close to the boundary where the program was implemented. This creates a selection problem, and comparing households around the boundary will not be appropriate.

Identification: Precise vs Imprecise Control

- Let's take another example: Say a student receives a scholarship if they score above 50 out of 100. We want to evaluate the effects of scholarship on future outcomes, such as college attendance and wage. There are two types of student group: 1) that cares about the scholarship; and 2) those who are not that concerened about the scholarship.
    - Imprecise control: The students can study with an aim of getting the scholarship, but they cannot effectively influence their score. So people right above and below the cut off will be comparable as people right below and above the cut off will have had the score due to random elements such as luck.
    - Precise control: Now, say we inform the group suggesting that 50 percent of questions are quite simple and they can get them correct if they revise their work. This will create systematic differences in terms of who will revise and who won't. The group that cares about the scholarship will go over their work, and they can precisely control their score. Here, we are looking at two types below and above the cut off, who are fundamentally different. People of type 1) can precisely self select themselves in winning the scholarship.

Thinking of RD as a Randomized Experiment from Non-random Selection

- Take

$$Y = D\tau + W\delta_1 + U; \qquad D = 1[X \geq c]; \qquad X = W\delta_2 + V.$$

- Assume that assignment variable X only comes in through U. $U = \delta_3 X + U'$. W is a vector of all predetermined characteristics that determines Y or/and X. The individual heterogeneity for the outcome variable comes from W and U.

RD as a Randomized Experiment from Nonrandom Selection

1. If individuals can pick exactly where to live, we an model X as a degenerate distribution.
2. If individuals have precise enough control, those individuals who benefit from the program, will make sure that they are above the cut off. For them, $V > c - \delta_2 W$. This creates a truncated density on X for people who will be benefitted by the program.
3. If individuals have imprecise control, randomness in V will determine X. In this case, density of people who would have benefitted from the program will be smooth over the cut off.
   - If individuals cannot precisely pick V, then conditional on W and U, the density of V (and hence X) will be continuous.
   - This gives: $P(W = w, U = u|X) = f(X|W, U) * \frac{P(W,U)}{f(x)}$. If $f(X|W, U)$ is continuous, distribution of W, U will be continuous. All observed or unobserved factors that determine the outcome will be continuous on X. The reason why individuals are just above and below the threshold is sheerly because of randomness (conditional on W). Local Randomization Around the Cut Off!

Identification of Treatment Effect

- See

$$lim_{\epsilon \downarrow 0} E[Y_i|X_i = c + \epsilon] - lim_{\epsilon \uparrow 0} E[Y_i|X_i = c + \epsilon]$$

$$= \tau + lim_{\epsilon \downarrow 0} E[W\delta_1 + U|X = c + \epsilon] - lim_{\epsilon \uparrow 0} E[W\delta_1 + U|X = c + \epsilon]$$

$$= \tau + lim_{\epsilon \downarrow 0} \sum_{w,u} [W\delta_1 + U] * P(W = w, U = u|X = c + \epsilon)$$
$$- lim_{\epsilon \uparrow 0} \sum_{w,u} [W\delta_1 + U] * P(W = w, U = u|X = c + \epsilon)$$

$$= \tau$$

- Since the denisty of W, U are continuous, we get τ as the treatment effect.

Generability of RD

- Let's introduce the concept of heterogeneous treatment effect
  - This is when the treatment varies with each individual
  - $\tau(w,u)$ in equation $Y = \tau(w,u)D + W\delta_1 + U$

$$lim_{\epsilon \downarrow 0} E[Y|X = c + \epsilon] - lim_{\epsilon \uparrow} E[Y|X = c + \epsilon]$$

$$= \sum \tau(w,u) * P(W = w, U = u|X = c)$$

$$= \sum \tau(w,u) * \frac{f(c|W = w, U = u)}{f(c)} * P(W = w, U = u)$$

Generability of RD

- Where $\frac{f(c|W=w,U=u)}{f(c)}$ are the weights. If we do not have this ratio term, the RD effect is just similar to the ATE.
- These weights are directly proportional to value of $X = x$ being close to the threshold.
  - However, in an applied setting we only get 1 observation for an indvidual. Hence, we know nothing about the ex-ante distribution of X for every individual.
- If weights are strongly placed for observations close to the cut off then RD effects are very similar to the ATE
- If everyone gets similar weights then RD effects can vary from the ATE.

Generability of RD: What remains to be found?

- To speak to the comparison of RD estimate and the overall population effect, we need the density of the assignment variable X (say score) at the individual
    - But we may have only one observation of the score, so cannot have the density at the individual level.
- If someone scoring 90 will have very low chance of scoring 50, then singal to noise ratio will be very high. And the RD effects will just be dominated by those around the cut off point (score of 50).
    - However, if the signal to noise ratio is high, then RD will pertain to a larger group of subpopulation, as those scoring 90 will also have significant probability of scoring 50. (reliability)
- How can information regarding reliability be used together with the RD gap, to understand the relationship between RD effects and the ATE of the overall population?

Estimation and Inference

1. Graphical Presentation: Bandwidth selection is a major part of RD design.

- Construct bins of $(b_k, b_{k+1}]$ both on the left and to the right of the cut off
  - let K be the total number of bins. $K = K_0 + K_1$, where $K_0$is the numer of bins to the left and $K_1$is the number of bins to the right.
  - $b_k = c - (K_0 - k + 1) * h$
- Look at the average value of the outcome within each bin
  - $\bar{Y}_k = \frac{1}{N_k} \sum_{i=1}^{N} Y_i * 1(b_k < X \leq b_{k+1})$, where $N_k = \sum_{i=1}^{N} 1(b_k < X \leq b_{k+1})$
- Bins represent nonparametric estimates of the regression function
  - mean of Yin a bin, for nonparameric Kernel estimator, is evaluated at the midpoint of the bin using a rectangular kernel.
  - Graphical potrayal can present a guide to choose functional form for regression models.
- How to choose the bandwidth, or the bin width?
  1. if the bin width is too small, noisy estimates

Cross Validation Approach of Choosing Bandwidth

1. Leave one out process
   - for the left: determine the point X_i s.t. $X_i < c$. Now, exclude X_i and only focus on set $X = X_i - h \leq X < X_i$.
   - for the right: determine the point X_i s.t. $X_i > c$. Now, exclude X_i and only focus on set $X = X_i < X \leq X_i + h$.
2. Run regression on the set of observations, excluding X_i. Use the regression estimates to get the predicted value of X_i.
3. Repeat 1 and 2 for each and every $X_i$and get predicted value of Y for every X.
4. Then calcuate the criterion function $CV_Y(h) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$
5. Pick a bandwidth h that minimizes the criterion function in 4.

Cross Validation Approach of Choosing Bandwidth

objective function: $h_{CV}^{opt} = arg\ min\ CV_Y(h)$

- Since we are concerned about action in the threshold, only consider observations close to the threshold.
  - from median of observations of the left of the cut off to the median on the right.
- Note that the specification used to predict the values of Y_i is given in the Estimation section (next slide).
- The CV approach can be carried out for both (left and right side) at once, or right and the left side separately.

Cross Validation Approach of Choosing Bandwidth

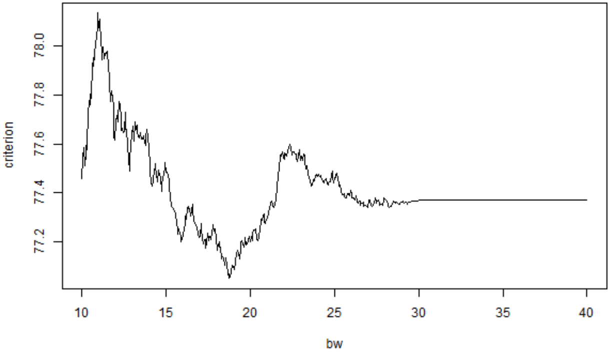


Figure:

This figure plots the values from CV criterion function when the sample is restricted between margin of -30 and 30 percent of vote share. Local linear regression is used within the allocated bandwidth, with a rectangular Kernel. In this case, selecting a bandwidth greater than 30 will produce exactly similar CV criterion value as the bandwidth of 30. How do we deal with this?

Estimation

- Regression on the left hand of the cut off $Y = \alpha_l + \beta_l(X - c) + \epsilon$, where $c - h \leq X < c$.
- Regression on the right hand $Y = \alpha_r + \beta_l(X - c) + \epsilon$, where $c < X \leq c + h$.
- Combining: $Y = \alpha_l + \alpha_r D + \beta_l(X - c) + (\beta_r - \beta_l)D(X - c) + \epsilon$, for $c - h \leq X < c + h$
- Question is: how to choose the bandwidth h?
  - We have discussed the CV approach above.
  - But one should not rely on just one option. Other alternatives: 1) Visual inspection, 2) ROT
- Note that bandwidth selection severely depends on the number of observations
  - as N increases it is possible to use smaller and smaller bandwidth.

Testing for the bandwidth size

1. Say you have decided on using K' of bins. The first test is to check the fittness of regression model that uses K' bin compared to $K'/2$ bins using F-test. $H_o : K'$ bins provide a better model fit, H_1: $K'/2$ provides a better model fit. If the $F - stat$ rejects the null, then the bandwidth you are currently using is oversmoothing the data and should use smaller bandwidth.
2. If bins are narrow enough, there should not be any systematic relationship between X and Ywithin each bin. Run $Y = \alpha + \sum_{b=1}^{K'} \gamma_b 1(X_b - c) + \epsilon$. Then conduct a joint test for the coefficients on $\gamma_{b,}$ *for* $b = \{1, 2, ..., K'\}$. If the $F - stat$ is significant, bins are too large – use smaller bins and conduct this exercise again.

Parametric or Non-Parametric Regression?

$Y = \alpha + D\tau + X\beta + \epsilon$ where Y is linear on X. But no real reason to believe that such is the case.

Typically estimate $Y = \alpha + D\tau + f(X - c) + \epsilon$

- Some choices of parametric: polynomials of X.
  - This choice provides a global estimates of RD and disregards the essense that focuse should be provided around the cut off.
  - Such a problem leads to non-parametric approach
- Non parametric approach:
  - Kernel regression performs poorly around the boundary. This can be explained by using Figure 2. The red and blue dots represent the means at the specific bins to the left and right of the cut off, respectively. But the treatment effect based on the means, $Y(blue) - Y(red)$ over states the actual treatment effect as the lines are upward sloping. Hence, this approach of non-parametric regression based on rectangular kernel fails.
  - Run a local linear regression: Run a standard regression on the bin then predict the values of the outcome variable. Use the predicted value of Y at the corner/boundary to estimate the RD effects.