Regression Discontinuity Design II

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- In sharp RDD, the treatment is determined by the value of the running variable.
- We can only identify the causal effect for units at the cutoff.
- We assume the continuity of the expected outcome across the cutoff and rely on kernel regression for estimation.
- It is necessary to evaluate the assumption of continuity with placebo outcomes, placebo cutoffs, and the McCrary test.
- Bandwidth is selected to minimize the MSE of the intercept estimates.
- Bias correction ensures that the estimate converges to a normal distribution.

Fuzzy RDD

- Today we are going to discuss several variants of the sharp RDD.
- The first variant is the fuzzy RDD, in which D_i is affected by Z_i in the following way:

$$D_i = \begin{cases} D_i(1) \text{ if } \mathbf{1}\{Z_i \geq 0\} \\ D_i(0) \text{ if } \mathbf{1}\{Z_i < 0\} \end{cases}$$

- $D_i(1)$ may not be 1 and $D_i(0)$ may not be 0.
- In other words, we now have non-compliance in the ideal experiment.
- $\mathbf{1}\{Z_i \geq 0\}$ is an instrument of D_i.

Fuzzy RDD

- $\mathbf{1}\{Z_i \geq 0\}$ should satisfy all the requirements for an instrument.
- Then, we can identify the treatment effect on the compliers when $Z_i = 0$:

$$\tau_{FRD} = \frac{E[Y_i(1) - Y_i(0)|Z_i = 0]}{E[D_i(1) - D_i(0)|Z_i = 0]}.$$

- Naturally, we can estimate the quantity with

$$\hat{\tau}_{FRD} = \frac{\hat{\mu}_{Y+} - \hat{\mu}_{Y-}}{\hat{\mu}_{D+} - \hat{\mu}_{D-}}.$$

  where all the four intercepts are estimated via local regression as in the previous lecture.

Fuzzy RDD

- $\hat{\tau}_{FRD}$ is a Wald estimator and approximately a linear combination of two sharp RDD estimates (one for Y and the other for D).
- Bias correction is also necessary for the estimate to be asymptotically normal.
- It must be conducted for both sharp RDD estimators.
- Calonico, Cattaneo, and Titiunik (2014) show that

$$\frac{\hat{\tau}^{bc}_{FRD} - \tau_{FRD}}{\sqrt{Var[\hat{\tau}^{bc}_{FRD}]}} \to N(0,1)$$

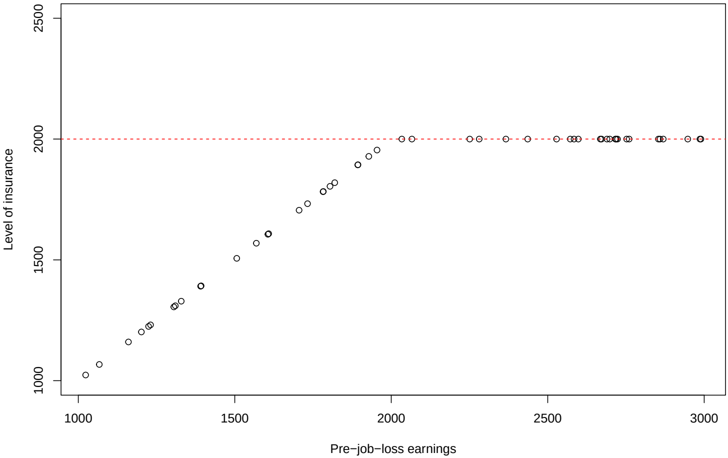
under the same condition: $N \min\{b^5, h^5\} \max\{b^2, h^2\} \to 0$.

Kink

- In a kink design, the treatment's level may not vary across the cutoff, but its rate of change might.
- For example, Y_i is the duration of unemployment, D_i is the level of unemployment insurance, and Z_i is one's pre-job-loss earnings.
- We want to estimate how Y_i responses to the change of D_i.
- Clearly, the correlation between Y_i and D_i is inevitably driven by confounders.

Kink

- Suppose the level of unemployment insurance equals to one's pre-job-loss earnings but there is a cap: $D_i = \min\{Z_i, 2000\}$:



Kink

- We can exploit this kink in the treatment to identify the response of the outcome to the treatment.
- Notice that the treatment is a deterministic and continuous function of running variable, $D = d(z)$.
- Suppose there is also a kink in the outcome across the cutoff, then the quantity of interest is

$$\tau_{SKRD} = \frac{\mu_{+}^{(1)} - \mu_{-}^{(1)}}{d_{+}^{(1)} - d_{-}^{(1)}}$$

where $d_{+}^{(1)} = \lim_{z \to 0_{+}} \frac{d(d(z))}{dz}$ ($D_{-}^{(1)}$ is similarly defined).

Kink

- Card et al. (2015) assume a non-separable outcome model

$$Y = y(D, Z, U),$$

  where U represents unobservable confounders.
- Card et al. (2015) show that τ_{SKRD} equals to the "local average response" defined in Altonji and Matzkin (2005):

$$\tau_{SKRD} = \int_u \frac{\partial y(d, z, u)}{\partial d}|_{d=d(0)} \frac{f_{Z|U=u}(0)}{f_Z(0)} f_U(u) du.$$

- It equals to the average marginal effect if U and Z are independent.

Kink

- The denominator is not at random and can be directly obtained.
- The numerator is the difference between two slopes and they can be similarly estimated via local regression.
- The asymptotic distribution of the numerator can be derived based on the theory in Calonico, Cattaneo, and Titiunik (2014):

$$\frac{\hat{\tau}^{bc}_{SKRD} - \tau_{SKRD}}{\sqrt{Var[\hat{\tau}^{bc}_{SKRD}]}} \rightarrow N(0,1)$$

  as long as $N \min\{b_N^7, h_N^7\} \max\{b_N^2, h_N^2\} \rightarrow 0$.
- In theory, we can estimate higher order derivatives but their substantive meanings are unclear.
- We also have the fuzzy kink design where there is non-compliance in the treatment.

Multidimensional RDD

- So far, we have assumed that the running variable Z_i is uni-dimensional.
- But in practice, Z_i could be the location of unit i on a geography, which is decided by both its latitude and longitude.
- L. J. Keele and Titiunik (2015) show that the identification assumption is that $\mu(\cdot)$ is continuous along both dimensions across the cutoff (border).
- First, we select a series of locations along the border.
- Next, we apply the local regression estimator to generate one estimate for each location.
- Each unit is weighted by its distance to this location.
- Finally, we aggregate the estimates by re-weighting them with the density at the corresponding location.
- Another option is to use the minimum distance to the border as a uni-dimensional running variable.
- This approach 1. loses information and 2. prevents us from seeing the heterogeneity in treatment effects.

Multidimensional RDD

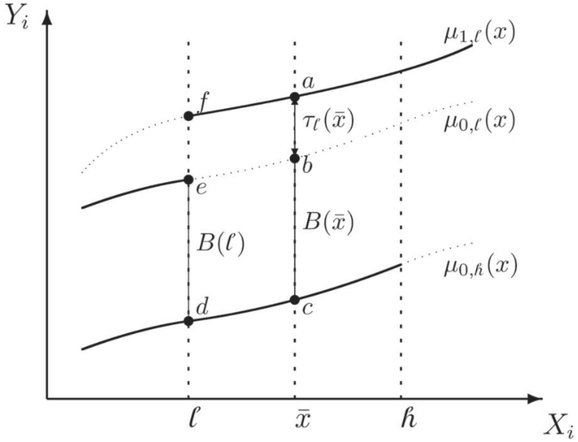
- It is usually harder to clarify what the treatment captures when Z_i is multi-dimensional (Dell, Lane, and Querubin 2018).
- Many things change together across the border of an administrative unit.
- Sorting (immigration) and autocorrelation are more likely to occur in the spatial setting.
- L. Keele, Titiunik, and Zubizarreta (2015) propose that researchers first match units on both covariates and their geographic locations and then generate RDD estimates on the matched sample.
- The validity of the method relies on the perspective of local randomization (see below).

RDD with multiple cutoffs

- When there are multiple cutoffs in the sample, researchers usually normalize the data such that all cutoffs equal to zero and run a pooled RDD.
- Cattaneo et al. (2016) show that the obtained estimate is a weighted average of the effect at each cutoff.
- The weights are proportional to the conditional density at each cutoff point $f(c|C = c)$, where $f(Z|C)$ is the density of the running variable when the cutoff is C.
- The estimate has various interpretations under different assumptions.
- Researchers can also estimate the effect around each cutoff and then take their average.

RDD with multiple cutoffs

- Cattaneo et al. (2020) further argue that the existence of multiple cutoffs helps us extrapolate results from RDD.



RDD with multiple cutoffs

- It requires an assumption that is similar to "parallel trends".
- Suppose there are two groups H and L with cutoffs h and l ($l < h$), respectively.
- Only the outcomes of units in L change across l ($\overline{ef}$).
- In addition, we assume that without the treatment, the conditional expectation of the outcome increases with z at the same rate across the two group: $\overline{ed} = \overline{bc}$.
- Then, we can estimate the treatment effect at $\bar{x}$ as

$$\hat{\tau}(\bar{x}) = \overline{ab} = \overline{ac} - \overline{bc} = \overline{ac} - \overline{de}.$$

RDD with a discrete running variable

- The classic theory developed in Calonico, Cattaneo, and Titiunik (2014) runs into difficulties when the running variable is discrete.
- By definition, there cannot be more observations around the cutoff point as N increases.
- Kolesár and Rothe (2018) provide a finite-sample confidence interval in this case.
- The CI does not rely on asymptotics and holds for any fixed N.
- The intuition is to bound the curvature of $\mu(z)$ and consider the worst scenario.
- The RDD estimator is also a weighting estimator.
- We can directly search for weights that optimize the bias-variance tradeoff.
- It can be implemented with the package *RDHonest*.

RDD in time

- In practice, people have been using RDD with time as the running variable (RDiT).
- Hausman and Rapson (2018) reviewed common issues with this method.
- They suggest that researchers should first account for the influence of temporal shocks such as weather or seasonality.
- Then, the classical estimator is applied to the residuals.
- More justifications are needed for this method.
- Time is always a discrete running variable.
- Which time interval should we use?
- What if the effect manifests gradually over time?
- What should we do when there are multiple treated units?

RDD as a local randomization

- We have discussed that the classic RDD can be seen as a simple experiment when $Z = 0$.
- Yet the conventional analysis is not a design-based approach as we do not try to model the assignment process.
- A recent perspective treats RDD as an experiment that is implemented where $Z \in [-h, h]$.
- If we know h, classic approaches (regression, weighting, matching, etc.) can be applied to estimate the treatment effect.
- Cattaneo, Frandsen, and Titiunik (2015) suggest that we should find h by balancing all the covariates.

RDD as a local randomization

- Eckles et al. (2020) extend this idea and argue that Z should be seen as a noisy measure of the true confounder U.
- For example, U_i is one's true capability and Z_i is her test score.
- We assume the measurement error is normally distributed

$$Z_i | U_i \sim \mathcal{N}(U_i, \nu^2)$$

and conditionally independent to the potential outcomes

$$\{Y_i(1), Y_i(0)\} \perp Z_i | U_i$$

- The assumptions imply that

$$\{Y_i(1), Y_i(0)\} \perp D_i | U_i$$

and the only issue is that U_i is unobservable.

RDD as a local randomization

- Eckles et al. (2020) claim that it is possible to estimate the causal effect using the noisy measure Z_i.
- When the confounders are observable, all we need to do is to balance the confounders.
- Now the confounder is unobservable, hence we should balance a functional f of its noisy measure Z_i.
- Z_i contains information of U_i and we know their relationship.
- If $f(Z_i)$ is balanced and f is properly chosen, we expect U_i to be balanced as well.
- They analyzed an example where U_i is the true level of CD4 count for HIV carriers and Z_i is the measured level.

References I

Altonji, Joseph G, and Rosa L Matzkin. 2005. "Cross Section and Panel Data Estimators for Nonseparable Models with Endogenous Regressors." *Econometrica* 73 (4): 1053–1102.

Calonico, Sebastian, Matias D Cattaneo, and Rocio Titiunik. 2014. "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs." *Econometrica* 82 (6): 2295–2326.

Card, David, David S Lee, Zhuan Pei, and Andrea Weber. 2015. "Inference on Causal Effects in a Generalized Regression Kink Design." *Econometrica* 83 (6): 2453–83.

Cattaneo, Matias D, Brigham R Frandsen, and Rocio Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the US Senate." *Journal of Causal Inference* 3 (1): 1–24.

References II

Cattaneo, Matias D, Luke Keele, Rocio Titiunik, and Gonzalo Vazquez-Bare. 2020. "Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs." *Journal of the American Statistical Association*, 1–12.

Cattaneo, Matias D, Rocio Titiunik, Gonzalo Vazquez-Bare, and Luke Keele. 2016. "Interpreting Regression Discontinuity Designs with Multiple Cutoffs." *The Journal of Politics* 78 (4): 1229–48.

Dell, Melissa, Nathan Lane, and Pablo Querubin. 2018. "The Historical State, Local Collective Action, and Economic Development in Vietnam." *Econometrica* 86 (6): 2083–2121.

Eckles, Dean, Nikolaos Ignatiadis, Stefan Wager, and Han Wu. 2020. "Noise-Induced Randomization in Regression Discontinuity Designs." *arXiv Preprint arXiv:2004.09458*.

Hausman, Catherine, and David S Rapson. 2018. "Regression Discontinuity in Time: Considerations for Empirical Applications." *Annual Review of Resource Economics* 10: 533–52.

References III

Keele, Luke J, and Rocio Titiunik. 2015. "Geographic Boundaries as Regression Discontinuities." *Political Analysis* 23 (1): 127–55.

Keele, Luke, Rocio Titiunik, and José R Zubizarreta. 2015. "Enhancing a Geographic Regression Discontinuity Design Through Matching to Estimate the Effect of Ballot Initiatives on Voter Turnout." *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, 223–39.

Kolesár, Michal, and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable." *American Economic Review* 108 (8): 2277–2304.