Economics 50/HKS SUP 135 Gregory A. Bruich, Ph.D.
Spring 2023 Harvard University

Lab 5 Regression Discontinuity Design (RDD)

1. Introduction to Regression Discontinuity Design (RDD)

Regression discontinuity design (RDD) is a powerful tool for causal inference that, when done well, provides evidence that is as good as that from a randomized experiment. The basic set up is as follows.

Treatment is assigned on the basis of a variable W_i which is called the running variable crossing a threshold $W_i = w$. If the value of running variable falls on one side of the cutoff, $W_i > w$, the person is assigned to the "treatment" group. If the value of assignment variable falls on the other side, $W_i < w$, the person is assigned to the "control" group. If treatment is completely determined by crossing the threshold, then the RDD is called sharp. If the treatment is not completely determined by crossing the threshold, then it is called a fuzzy RDD.

The *identification assumption* needed for regression discontinuity design, is that all other determinants of the outcome evolve smoothly across the threshold. If that is the case, then observations just above and just below the threshold are very similar, like treatment and control groups in a randomized experiment.

2. Validating a RDD

Graphical analyses are extremely valuable for RDD. To validate a RDD, we perform three graphical analyses.

First, we examine the binscatter plots of the main treatment and outcome variables of interest at the threshold value. We check for clear visual evidence of discontinuities at the threshold. We also check that there is no evidence of other large jumps at other values of the running variable.

Second, we check whether the treatment is the only characteristic that changes sharply at the threshold. If observable characteristics are similar on either side of the threshold, this gives us some confidence that *unobservable* characteristics are also similar on either side of the threshold.

Third, we check for manipulation of the running variable W_i by plotting histograms of the running variable. If we see a spike on one side of the threshold, then this may suggest that individuals are able to control which side of the threshold they are on. The issue here is that comparing individuals on either side of the threshold may then confound the impact of the treatment with characteristics of the types of individuals who are informed enough about the threshold to locate on one side or the other.

3. Estimation via linear regression

A very simple regression would be as follows:

$$Y_i = \beta_0 + \ \beta_{RD} T_i + \beta_2 (W_i - w) + \ \beta_3 \, T_i \times (W_i - w) + u_i$$

where Y_i is an outcome variable, T_i is an indicator variable for one side of the threshold:

$$T_i = \begin{cases} 1 \text{ if } W_i - w \geq 0 \\ 0 \text{ if } W_i - w < 0 \end{cases}$$

The variable $W_i - w$ is the running variable W_i minus the threshold w. The third term $T_i \times (W_i - w)$ is the product between the indicator and the running variable. In the regression discontinuity design terminology this regression would be described as being estimated using a *linear control function*.

The coefficient β_{RD} measures the jump in the predicted value at the threshold. To see this, notice that for observations with W_i less than w who therefore have $T_i = 0$, the predicted value from the regression is:

$$\hat{Y}_i = \beta_0 + \beta_2 (W_i - w)$$

For observations with W_i greater than w who therefore have $T_i = 1$, the predicted value from the regression is:

$$\hat{Y}_i = \beta_0 + \beta_{RD} + \beta_2 (W_i - w) + \ \beta_3 \ (W_i - w)$$

Taking the difference and evaluating it at the threshold $W_i = w$, shows why the coefficient on T is the jump at the threshold:

$$\Delta \hat{Y}_i = \{\beta_0 + \beta_{RD} + \beta_2 (W_i - w) + \beta_3 (W_i - w)\} - \{\beta_0 + \beta_2 (W_i - w)\}$$

$$\Rightarrow \Delta \hat{Y}_i = \beta_{RD} + \beta_3 (W_i - w)$$

$$\Rightarrow \Delta \hat{Y}_i = \beta_{RD}$$

4. Bandwidth

Because regression discontinuity design is based on comparing observations just above and just below the threshold $W_i = w$, it is common for the graphs and regressions described above to be restricted to only observations in a narrow range around the threshold. In the regression discontinuity design terminology, the size of the window around the threshold is called the *bandwidth*.