14.661: Recitation 10 Notes

Chris Walters

November 19, 2010

Regression Discontinuity Design

The regression discontinuity (or "RD") design is a quasi-experimental identification strategy that is used when treatment is assigned as a function of some observable variable (often called the "running" variable). We will go through some of the theory of RD, though the basic idea is quite intuitive and the most convincing RD evidence can be presented in simple graphs.

RD Intuition

In an RD design, we want to know the effect of a treatment T_i on an outcome variable Y_i. In the simplest RD case (called "sharp" RD), the treatment is assigned to all observations beyond some threshold in "running variable" x_i:

$$T_i = 1\{x_i \geq c\}$$

Examples of sharp RDs include:

- Union status (DiNardo and Lee)
- Political incumbency (Lee)

What is T and what is x in each of these examples?

The basic idea of RD is very simple: We look for a discontinuity in the relationship between Y_i and x_i near the threshold c, and call this the treatment effect. This is easiest to think about in a graph. Intuitively, we compare observations with x_i just below c and those just above c.

RD Theory

Let's use the potential outcome notation we introduced earlier in the course. Let Y_{0i} and Y_{1i} be random variables that tell us i's outcome as a function of the treatment T_i. Then

$$\begin{aligned} Y_i &= Y_{1i}T_i + Y_{0i}(1 - T_i) \\ &= Y_{0i} + (Y_{1i} - Y_{0i})T_i \\ &= Y_{0i} + \beta_i T_i \end{aligned}$$

Now let's make the following assumption:

$$E[Y_{0i}|x_i] \text{ is continuous at } x_i = c$$

What does this get us? Let's think about following the RD intuition and comparing observed outcomes for individuals on each side of the threshold:

$$\lim_{x_i \to c^-} E\left[Y_i | x_i\right] = \lim_{x_i \to c^-} E\left[Y_{0i} + \beta_i T_i | x_i\right]$$

$$= \lim_{x_i \to c^-} E\left[Y_{0i} + \beta_i 1\left\{x_i > c\right\} | x_i\right]$$

$$= \lim_{x_i \to c} E\left[Y_{0i} | x_i\right]$$

Similarly

$$\lim_{x_i \to c^+} E\left[Y_i | x_i\right] = \lim_{x_i \to c^+} E\left[Y_{0i} + \beta_i 1\left\{x_i > c\right\} | x_i\right]$$

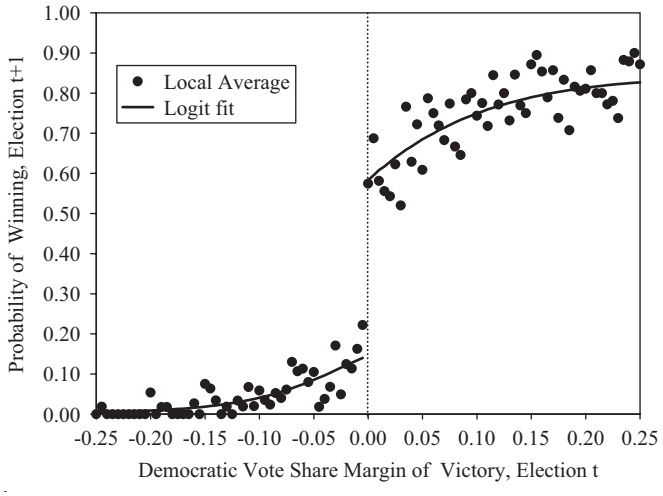
$$= \lim_{x_i \to c^+} E\left[Y_{0i} + \beta_i | x_i\right]$$

$$= E\left[\beta_i | x_i = c\right] + \lim_{x_i \to c} E\left[Y_{0i} | x_i\right]$$

so

$$\lim_{x_i \to c^+} E\left[Y_i | x_i\right] - \lim_{x_i \to c^-} E\left[Y_i | x_i\right] = E\left[\beta_i | x_i = c\right]$$

That is, comparing limits on each side of the threshold should give us a treatment effect of interest. Using this method, Lee (2008) shows that the "incumbency effect" in congressional elections is about 40%. The running variable here is the Democratic vote share in the previous election cycle; at 50%, the Democrats go from being losers to winners. The outcome is winning today. The incumbency effect appears to be on the order of 35 percent.



Computation of RD

From the above discussion, it's pretty obvious that what matters for RD is getting consistent estimates of the limits on the right and left sides of the threshold. There are a few ways we can do this.

Method 1: Series Regression

The most straightforward way of computing these limits is to estimate a flexible relationship between Y_i and x_i, allowing for a discontinuity at $x_i = c$. That is, we can run the regression

$$Y_i = \alpha + \beta T_i + \sum_{k=1}^{K} (x_i - c)^k + \sum_{k=1}^{K} T_i \cdot (x_i - c)^k + \epsilon_i$$

This allows for flexible k-dimensional polynomial fits on each side of the threshold, and the coefficient β measures the "break" at $x_i = c$. One advantage of this method is that the standard errors are easy to get – we are just running OLS! A disadvantage is that it's not clear how many polynomial terms we should include, and it's often hard to decide (though there are principled ways to do this).

Method 2: Nonparametric Regression

There are now somewhat more sophisticated methods for nonparametrically computing conditional expectations of functions. The simplest of these is a kernel regression:

$$\widehat{E}\left[Y_i | x_i = a\right] = \frac{\sum Y_i \cdot K\left(\frac{x_i - a}{h}\right)}{\sum K\left(\frac{x_i - a}{h}\right)}$$

Here, K is some (usually symmetric) kernel weighting function, and h is a bandwidth. The larger is h, the more weight is given to observations that are far from a, the point of interest. To estimate the limits for RD, we could run kernel regressions at $x_i = c$ using only data from one side or the other of the boundary.

However, it is now well understood that kernel regressions are badly biased at the boundary of the data. Since all of the action for RD is at the boundary, this is bad news. Fortunately, an alternative called "local linear regression" is much better. Here, we run

$$\begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = \arg\min_{(\alpha,\beta)'} \sum K\left(\frac{x_i - a}{h}\right) \left(Y_i - \alpha - \beta(x_i - a)\right)^2$$

and

$$\widehat{E}\left[Y_i | x_i = a\right] = \hat{\alpha}$$

In some ways this method is more elegant than series regression. However, there is an obvious problem: Bandwidth selection. Imbens and Kalyanaraman (2009) derive the optimal bandwidth for minimizing MSE. Standard errors for this estimator also require a little care.

Generalization: "Fuzzy RD"

Now let's consider a more general version of the model. Let's write

$$Y_i = \alpha_i + \beta_i T_i$$

T_i can now be continuous, and treatment is no longer a deterministic function of the threshold. Instead, the average value of the treatment changes discretely at $x_i = c$ (if T_i is a dummy, the probability changes). Let's assume

1. $\lim_{x_i \to c^-} E\left[T_i | x_i\right] \neq \lim_{x_i \to c^+} E\left[T_i | x_i\right]$
2. $E\left[\alpha_i | x_i\right]$ is continuous at $x_i = c$
3. $E\left[\beta_i | x_i\right]$ is continuous at $x_i = c$
4. Given x_i, β_i is independent of T_i near c

Now, let's think about comparing limits again:

$$\lim_{x_i \to c^-} E[Y_i|x_i] = \lim_{x_i \to c^-} E[\alpha_i + \beta_i T_i|x_i]$$

$$= \lim_{x_i \to c} E[\alpha_i|x_i] + \lim_{x_i \to c^-} E[\beta_i T_i|x_i]$$

$$= \lim_{x_i \to c} E[\alpha_i|x_i] + \lim_{x_i \to c^-} E[T_i|x_i] \cdot \lim_{x_i \to c^-} E[\beta_i|x_i]$$

$$= \lim_{x_i \to c} E[\alpha_i|x_i] + \lim_{x_i \to c^-} E[T_i|x_i] \cdot E[\beta_i|x_i = c]$$

Similarly,

$$\lim_{x_i \to c^+} E[Y_i|x_i] = \lim_{x_i \to c} E[\alpha_i|x_i] + \lim_{x_i \to c^+} E[T_i|x_i] \cdot E[\beta_i|x_i = c]$$

so

$$\lim_{x_i \to c^+} E[Y_i|x_i] - \lim_{x_i \to c^-} E[Y_i|x_i] = E[\beta_i|x_i = c] \cdot \left(\lim_{x_i \to c^+} E[T_i|x_i] - \lim_{x_i \to c^-} E[T_i|x_i] \right)$$

and

$$\frac{\lim_{x_i \to c^+} E[Y_i|x_i] - \lim_{x_i \to c^-} E[Y_i|x_i]}{\lim_{x_i \to c^+} E[T_i|x_i] - \lim_{x_i \to c^-} E[T_i|x_i]} = E[\beta_i|x_i = c]$$

What does this remind you of? This is IV! In fact, fuzzy RD is just IV using the threshold as an instrument for treatment, controlling flexibly for continuous functions of x_i. We can literally run this as IV with series regression, where we instrument for T_i using a dummy for crossing the threshold. Alternatively, we could do nonparametric regressions to approximate all 4 limits in this formula. One famous example of fuzzy RD is Angrist's paper on Maimonides' rule for class size.

One other note on RD: Like with IV, the assumption that nothing else changes at the threshold is not testable. However, it is possible to make some arguments. In particular, one can check whether the values of relevant covariates shift near the threshold. If there is sorting going on, this may detect it.

14.661 Labor Economics I

Fall 2010