1 Sharp vs. Fuzzy Regression Discontinuity

Regression discontinuity 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 treatment is completely determined by crossing the threshold, then the RD is called *sharp.* Then, if the variable falls on one side of the cutoff, the person is assigned to the *treatment group.* If the value of the running variable falls on the other side, the person is assigned to the *control group.*
- If the treatment is not completely determined by crossing the threshold, then crossing the threshold can be used as an instrument for the treatment in what is called a *fuzzy* RD.
- Note that the treatment group can be defined by either $W_i > w$ (e.g., a test score cutoff for scholarship eligibility) or by $W_i < w$ (e.g., babies below a certain birthweight receiving extra medical care).

Which of the following are sharp, and which are fuzzy?

- The cutoff implicit in majority rule elections to test for the existence of an incumbency advantage in elections (as in Lee 2008, discussed in Section 3)
- Class size caps to study the effect of class size on achievement
- Income-based cutoffs for eligibility to measure the effect of social programs
- Test-score-based cutoffs for mandating summer school
- Medicare health insurance at age 65

Note one important takeaway from this exercise is that the sharp/fuzzy distinction often requires knowing something about the context you are studying!

2 Fuzzy Regression Discontinuity Design

Suppose a running variable W_i with threshold w_0. We construct an instrument Z_i such that:

$$Z_i = \begin{cases} 0 & \text{if } W_i < w_0 \\ 1 & \text{if } W_i \geq w_0 \end{cases}$$

Note that this is the same as the way we constructed the treatment variable X_i when the threshold perfectly determined treatment.

We'll construct our first stage, second stage, and reduced form equations as usual but use the regression-discontinuity-style control function for our controls:

$$\begin{aligned} X_i &= \pi_0 + \pi_1 Z_i + f\left(W_i - w_0\right) + \nu_i && \text{(First Stage)} \\ Y_i &= \alpha_0 + \alpha_1 Z_i + f\left(W_i - w_0\right) + \xi_i && \text{(Reduced Form)} \\ Y_i &= \beta_0 + \beta_1 \hat{X}_i + f\left(W_i - w_0\right) + u_i && \text{(Second Stage)} \end{aligned}$$

The control function $f(W_i - w_0)$ must be the same in all three equations.

- Why do I center the running variable? That is, why do I use $f(W_i - w_0)$ instead of $f(W_i)$?
- Interpret β_1. Does it have the same interpretation as in a sharp RD?
- How would we calculate β_1 using coefficients from the first stage and reduced form regressions?

3 2SLS Assumptions Recast as RD Assumptions

The identifying assumptions for fuzzy RD are similar to those for IV regressions more generally.

- For an IV design to be internally valid the instrument must satisfy __________ and __________.

In the context of fuzzy RD, we state these assumptions slightly differently, because we're looking at them right around the threshold.

$$\begin{aligned} \lim_{\Delta \to 0^+} \mathbb{E}\left[X_i | W_i = w_0 + \Delta\right] &\neq \lim_{\Delta \to 0^+} \mathbb{E}\left[X_i | W_i = w_0 - \Delta\right] && \text{(Relevance)} \\ \lim_{\Delta \to 0^+} \mathbb{E}\left[u_i | W_i = w_0 + \Delta\right] &= \lim_{\Delta \to 0^+} \mathbb{E}\left[u_i | W_i = w_0 - \Delta\right] && \text{(Exogeneity)} \end{aligned}$$

To put these in words, the relevance condition says that we **do** see a jump in the treatment at the running variable threshold. The exogeneity condition says that there is **no jump** in $u_i = Y_i(0) - E[Y_i(0)]$ at the running variable threshold.

There are two additional important components of the fuzzy RD:

Control function: Since we want to use only the variation in the outcome right around the running variable threshold, we need to do a good job of controlling for the broader relationship between the running variable and the outcome. Typically, we allow the polynomial to be different on each side of the cutoff by including the interaction of each term of the polynomial with the indicator for the threshold. This is called an *asymmetric control function.* When the terms of the polynomial are not interacted with the indicator for the threshold, the control function is the same on both sides of the threshold.

Bandwidth: Given that we're most interested in variation right around the cutoff, we don't want observations far away from the cutoff to have excessive influence on the fit of the control function. The robustness check here is to change the *bandwidth*, that is, the range of running variable values over which we estimate. When we do this, we zoom in on the area of interest, but we do so at the price of reduced sample size.

- Write out a quadratic control function with running variable W_i and threshold indicator Z_i.

4 Example: Health Insurance and Healthcare Utilization

There are a number of papers that exploit sharp changes in eligibility for health insurance at specific age thresholds. The age thresholds are as follows:

- Age 65, where Medicare health insurance eligibility changes sharply (Card, Dobkin, and Maestas 2009; Appendix to Dobkin, Finkelstein, Kluender, and Notowidigdo 2016)
- Age 19, where those not in school "age out" of being on their parents' health insurance plans, before the ACA mandated that young adults could stay on their parents' plans to age 26 (Anderson, Dobkin, and Gross 2012)
- Age 23, which insurers have used as a cutoff after which students are no longer eligible for their parents' health insurance, before the ACA mandated that young adults could stay on their parents' plans to age 26 (Anderson, Dobkin, and Gross 2014)
- Age 21, where Medicaid health insurance eligibility changes sharply pre-ACA vs. post-ACA (Duggan, Gupta, and Jackson 2017)

Example 1: Health Insurance and Healthcare Utilization at 65
Card, Dobkin, and Maestas (2008) exploit the age eligibility threshold for the U.S. universal health insurance program for the elderly, Medicare, to estimate the effect of health insurance on health

care utilization. They use data from the 1992-2003 National Health Insurance Survey, which is a repeated cross-section survey collected annually. We'll use some of their data to assess the effect of health insurance on one of their outcomes, an indicator equal to whether or not an individual spent at least one night in a hospital in the past year.

- How can we visualize the first stage regression?

- How can we visualize the reduced form regression?

- What are the components of the exogeneity assumption? How can we test each of them?

- In this example, what does a bandwidth of 2 years mean? How would we modify our code to use this bandwidth?

McCrary (2008) test
One important test of exogeneity involves looking at the density of the running variable. In particular, we might be concerned if we see too many or too few people on one side of the threshold.

- Why would bunching above or below the cutoff suggest that exogeneity isn't satisfied?

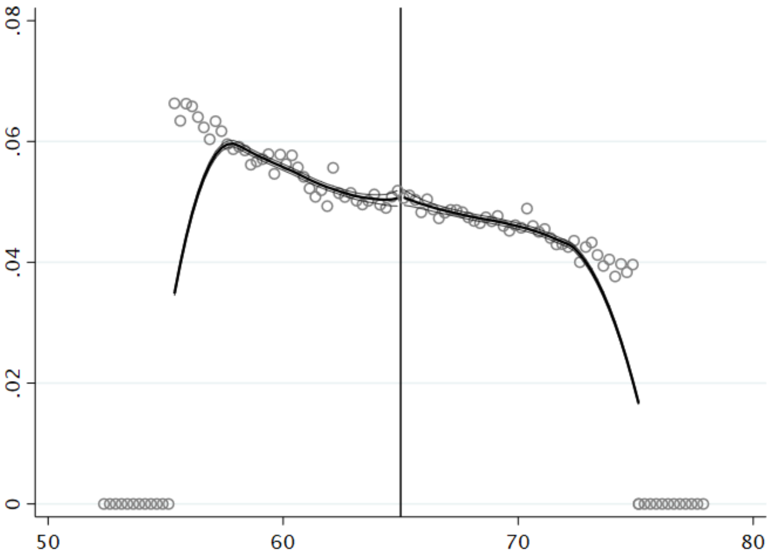
While we may not be particularly worried about people manipulating their age in our health insurance example, we'll conduct a McCrary test to show how it's done. We'll use a Stata .ado command called `DCdensity`. This is a user-written command that lives outside of Stata. Be sure to have it in your working directory so that Stata can execute it.

```
DCdensity age4, breakpoint(65) generate(Xj Yj r0 fhat se_fhat) b(.25)

// b(.25) specifies bin width; necessary here
// because the data aren't really continuous

Using default bandwidth calculation, bandwidth = 2.82818905

Discontinuity estimate (log difference in height): .003959117
                                                   (.020372898)
```



- What do we conclude from the results of this test?

Example 2: Health Insurance and Healthcare Utilization at Ages 19 and 23
Anderson, Dobkin, and Gross (2012, 2014) study the effect of health insurance on healthcare utilization at ages 19 and 23. These examples nicely illustrate the LATE vs. ATE distinction.

At age 19, those who are not still in school are no longer eligible to be covered as a dependent on their parents' insurance plans (but the ACA changed this). Therefore, turning 19 increases the probability of becoming uninsured discontinuously. Use $Z_i = 1\{\text{Age}_i \geq 19\}$ as an instrument for health insurance coverage. There is a similar discontinuity at age 23 driven by students losing eligibility for being on their parents' plan.

The first stage estimates changes in insurance coverage at 19 using a regression of the form:

$$\text{Insurance}_i = \pi_0 + \pi_1 Z_i + f(\text{Age}_i - 19) + v_i$$

where $f(\cdot)$ is a polynomial (usually interacted with Z_i so it can be different on either side of age 19). The reduced form simply replaces Insurance_i with a variable measuring use of medical services. The 2SLS estimate combines these two estimates to yield an estimate of the causal effect of insurance coverage on use of services.

- This fuzzy RD estimates a local average treatment effect (LATE). In your judgment, is this LATE greater than, less than, or equal to the average treatment effect? Explain. What other caveats would you put on interpreting these results?