Regression Discontinuity Design

Up close, we are all the same

Fernando Rios-Avila

Re-Cap: Potential outcome Model

In the ideal world, where we can see all possible outcomes and scenarios of your potential treatments, it will be very simple to estimate treatment effects:

$$\delta_i = Y_i(1) - Y_i(0)$$

This works because all observed and unobserved individual characteristics are kept fixed, except for the treatment Status.

$$y_i(D) = y_i(X, u, D)$$

So when comparing a person with himself (clones or parallel worlds), we know (or at least expect) that everything else is the same, and that differences between the two states are explained only by the treatment.

The Problem

We do not observe both **ALL** States at the same time. People will either be treated or untreated, not both.

So what can we do?

We need to find good counterfactuals!

This means finding people are very similar to the ones treated, so they can be used as the examples of the "what if" question.

But there is a problem. Even in the best scenarios, we can never be asure about how to control for unobservables...or can we?

- **You can always RCT** But it can be expensive
- **You can IV the problem** but its hard to justify
- **You can add FE**, but you have time varying errors

Then what?

- You could RDD the problem (if you have the right data!)

What is RDD?

RDD or Regression Discontinuity design is methodology that is known for its clean identification and with a relatively easy visualization tool to understand the identification, and solve the problem of unobserved distributions. (see here for a recent paper on how to make graphs on this).

In fact, the treatment Status has a very clear Rule!

Consider the following problem:

1. You want to study the role of college on earnings.
2. You have data on people who are applying to go to school. They all take some standardized tests. Their grade will determine if they get into College or not.
3. People with high skills will get a higher grades in the GRE, go to college, and probably get higher salaries.
4. But, there is a problem. How can you figure out if wages are due to College or skill?

Possible Solution

Say that we actually have access to the grades, which range from 100 to 200. And assume that you say, every one with grades higher than 170 will go to college.

Can you estimate the effect now?

- You can't compare people with more than 170 to those with less than 170. Because skill or ability will be different across groups.
- However, what if you compare individuals with 170-172 vs 167-169?
  - These individuals are so close togeter they problably have very similar characteristics as well!
  - you have a Localized randomization.

In this case, your analysis is those individuals just above the thresholds to those just below (counterfactual)

Unless you think grades near the threshold are as good as random, then you have a design to identify treatment effects!

RDD: How it works.

Selection and Index:

1. The first thing you need to see if you can use and RDD is to see if you have access to a variable that "ranks" units.

- This variable should be smooth and preferibly continuous.
  - age, distance from boarder, test score, poverty index

2. Assignment into treatment is a function of this index only, with a clear threshold for the index to have an impact the treatment.

- Those under the Threshold are not treated. Those above are.
- This is called a **Sharp RDD** design.

3. The threshold should be unique to the Treatment of interest (nothing else happens around that point)

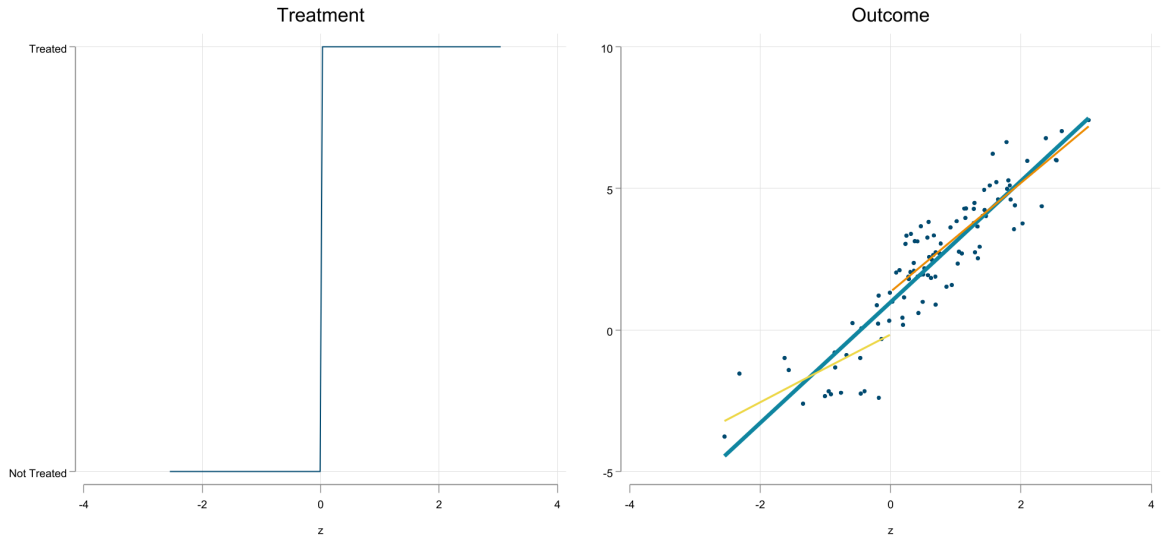
- In the College Case, we assume 170 triggers acceptance to School. But if it also triggers Scholarships??

4. Perhaps the Most important: The score cannot be manipulated
   - Only then we have true local randomization.

5. You want the potential outcomes to be smooth functions of Z. (so we do not mix treatment effects with Nonsmooth changes in outcomes)

Sharp RDD

```
clear
set scheme white2
color_style bay
set seed 1
qui:set obs 100
gen e=rnormal()
gen z=runiform()+e
gen t = z>0
gen y = 1 + z + e + rnormal() + (z>0)
two line t z, sort title("Treatment") ylabel(0 "Not Treated" 1 "Treated") name(m1, replace) y
graph export resources\rdd1.png,  width(1000) height(1000) replace
two scatter y z, sort title("Outcome") pstyle(p1) || lfit y z, lw(1) pstyle(p2) ///
    || lfit y z if z<0, lw(0.5) || lfit y z if z>0, lw(0.5) , legend(off) name(m2, replace)
graph export resources\rdd2.png,  width(1000) height(1000)  replace
```



How it works : p2

Recall that in an RCT (or under randomization) treatment effects are estimated by comparing those treated and those not treated.

$$E(y|D=1) - E(y|D=0)$$

Under SRDD, you can also think about the same experiment, except that we would need to compare individuals AT the theshold.

$$\lim_{z \downarrow c} E(y|Z = z) - \lim_{z \uparrow c} E(y|Z = z)$$
$$E(y(1)|Z = c) - E(y(0)|Z = c)$$

- In this case, the overlapping assumption is violated. So we need to attemp obtaining effects for groups AT the limit when $Z = c$.

Estimation

The most simple way to proceed is to estimate the model using a parametric approach (OLS)

$$y = a_0 + \delta D_{z>c} + f(z - c) + e$$

The idea here is to identify a "jump" in the outcome (treatment effect) at the point where z crosses the threshold.

But to identify the jump only, we also need to model the trend observe before and after that threshold ($f(z - c)$), which can be modelled as flexible as possible. (this include interactions with the jump)

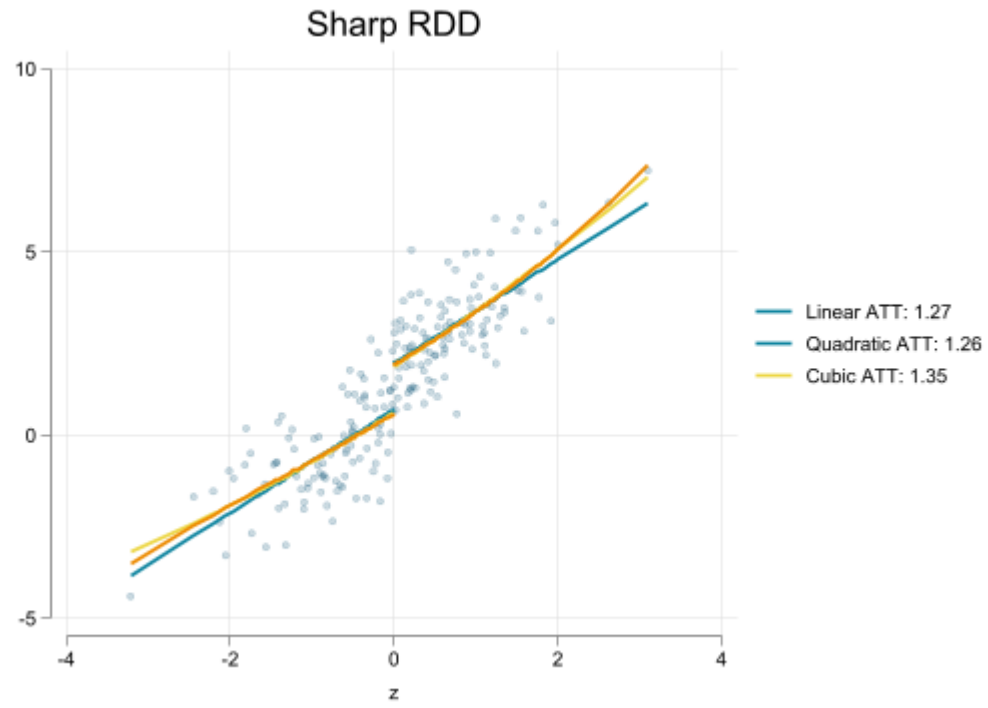
Alternatively, we could use smaller bandwidths (nonparametric)

Example

```
 qui: {
clear
set seed 1
set obs 200
gen e=rnormal()
gen z=runiform()+e
sum z
replace z=(z-r(mean))/r(sd)
gen t = z>0
gen y = 1 + 0.5*z + e + rnormal() + (z>0)

qui:reg y t z
predict yh1
local b1:display %3.2f _b[t]
```

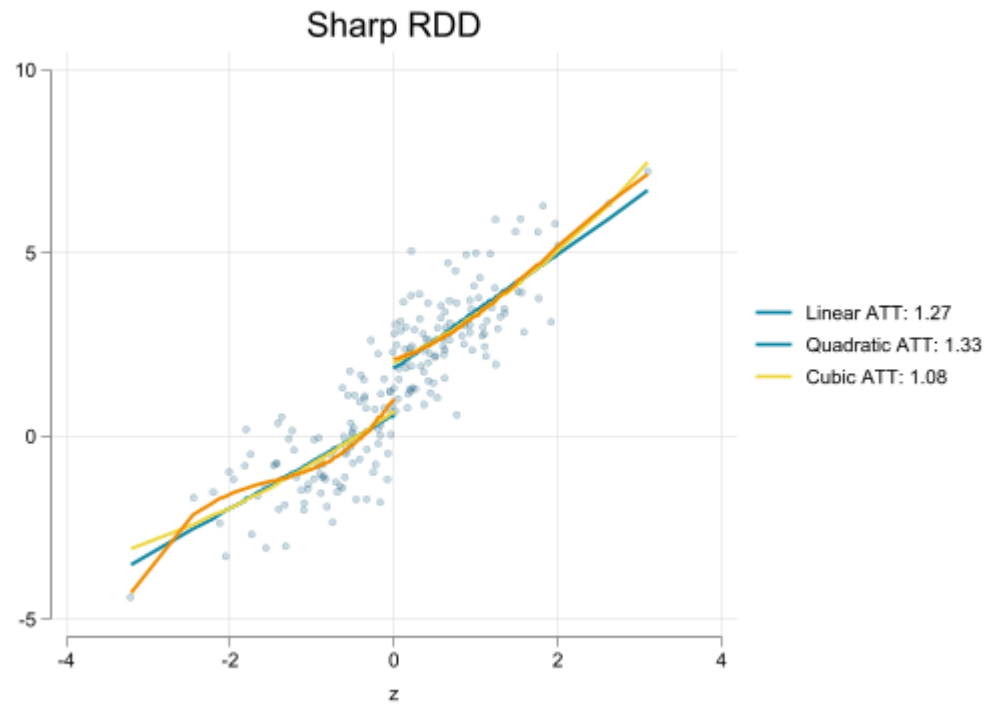
```
qui:reg y t c.z##c.z
predict yh2
local b2:display %3.2f _b[t]
qui:reg y t c.z##c.z##c.z
predict yh3
local b3:display %3.2f _b[t]
sort z
}
two (scatter y z, sort title("Sharp RDD") pstyle(p1) color(%20)) ///
    (line yh1 z if z<0, pstyle(p2) lw(0.5)) (line yh1 z if z>0, pstyle(p2) lw(0.5)) ///
    (line yh2 z if z<0, pstyle(p3) lw(0.5)) (line yh2 z if z>0, pstyle(p3) lw(0.5)) ///
    (line yh3 z if z<0, pstyle(p4) lw(0.5)) (line yh3 z if z>0, pstyle(p4) lw(0.5)) , ///
    legend(order(2 "Linear ATT: `b1'" 3 "Quadratic ATT: `b2'" 4 "Cubic ATT: `b3'")) name(m1,
```



Example

```
qui: {
qui:reg y t c.z#t
```

```
predict yh11
local b1:display %3.2f _b[t]
qui:reg y t (c.z##c.z)#t
predict yh21
local b2:display %3.2f _b[t]
qui:reg y t (c.z##c.z##c.z)#t
predict yh31
local b3:display %3.2f _b[t]
}
two (scatter y z, sort title("Sharp RDD") pstyle(p1) color(%20)) ///
    (line yh11 z if z<0, pstyle(p2) lw(0.5)) (line yh11 z if z>0, pstyle(p2) lw(0.5)) ///
    (line yh21 z if z<0, pstyle(p3) lw(0.5)) (line yh21 z if z>0, pstyle(p3) lw(0.5)) ///
    (line yh31 z if z<0, pstyle(p4) lw(0.5)) (line yh31 z if z>0, pstyle(p4) lw(0.5)) , ///
    legend(order(2 "Linear ATT: `b1'" 3 "Quadratic ATT: `b2'" 4 "Cubic ATT: `b3'")) name(m2,
```



Fuzzy RD: Imperfect compliance

While the Idea Scenario happens when there is perfect compliance (above the threshold you are treated), this doesnt happen all the time.