MQE: Economic Inference from Data: Module 6: Regression Discontinuity

Claire Duquennois

6/9/2020

Regression discontinuity research designs

Introduced in other fields as far back as the 1960s.

Gain popularity in economics in the past 20 years or so as economists:

- increasingly focus on causal inference
- large administrative datasets became more widely available

When correctly applied, RD designs are very transparent in how they achieve causal identification which makes them very appealing.

Regression discontinuity research designs

RD designs leverage the researchers knowledge of a rule or policy that determines treatment.

Identification comes from how some rules are applied in a fairly arbitrary way.

This arbitrary application generates randomness we can exploit to estimate causal effects.

The Set Up

Suppose that we want to estimate the effect of some binary treatment D_i on an outcome Y_i. Using the potential outcomes framework:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

Suppose D_i is completely (or partially) determined by whether some predictor X_i lies above or below a certain threshold, c.

The Set Up

The "running variable" X_i need not be randomly assigned:

- X_i is related to Y_i
- absent treatment, this relationship is smooth (Y_i does not jump discontinuously as X_i changes).

$\Rightarrow$ any discontinuous change in Y_i as X_i crosses c can thus be interpreted as a causal effect of D_i.

Where to find RD setups?

Often from administrative situations in which units are assigned a program, treatment or award based on a numerical index being above or below a certain threshold.

Where to find RD setups?

Examples:

- A politician may be elected if and only if the differential between the vote share that she receives and the vote share that her opponent receives exceeds 0.
- A student may be assigned to summer school if and only if his performance on a combination of tests falls below a certain threshold.
- A toxic waste site may receive cleanup funds if and only if it's hazard rating falls above a certain level.

Why does RD work?

The idea:

- Units whose indices X lie directly below the threshold c are considered to be comparable to individuals or units whose indices X lie directly above the threshold c.
- We can estimate the treatment effect by taking a difference in mean outcomes for units directly above the threshold and units directly below the threshold.

RD Types

RD designs come in two flavors:

- Sharp:
  - the probability that $D = 1$ changes from 0 to 1 as the running variable crosses c.
  - no one with $X < c$ gets treated and everyone with $X \geq c$ gets treated
- Fuzzy:
  - the probability of treatment changes by some nonzero amount as the running variable crosses the threshold c,
  - the change in probability is less than 100 percentage points $\Rightarrow D_i$ is no longer a deterministic function of X_i.

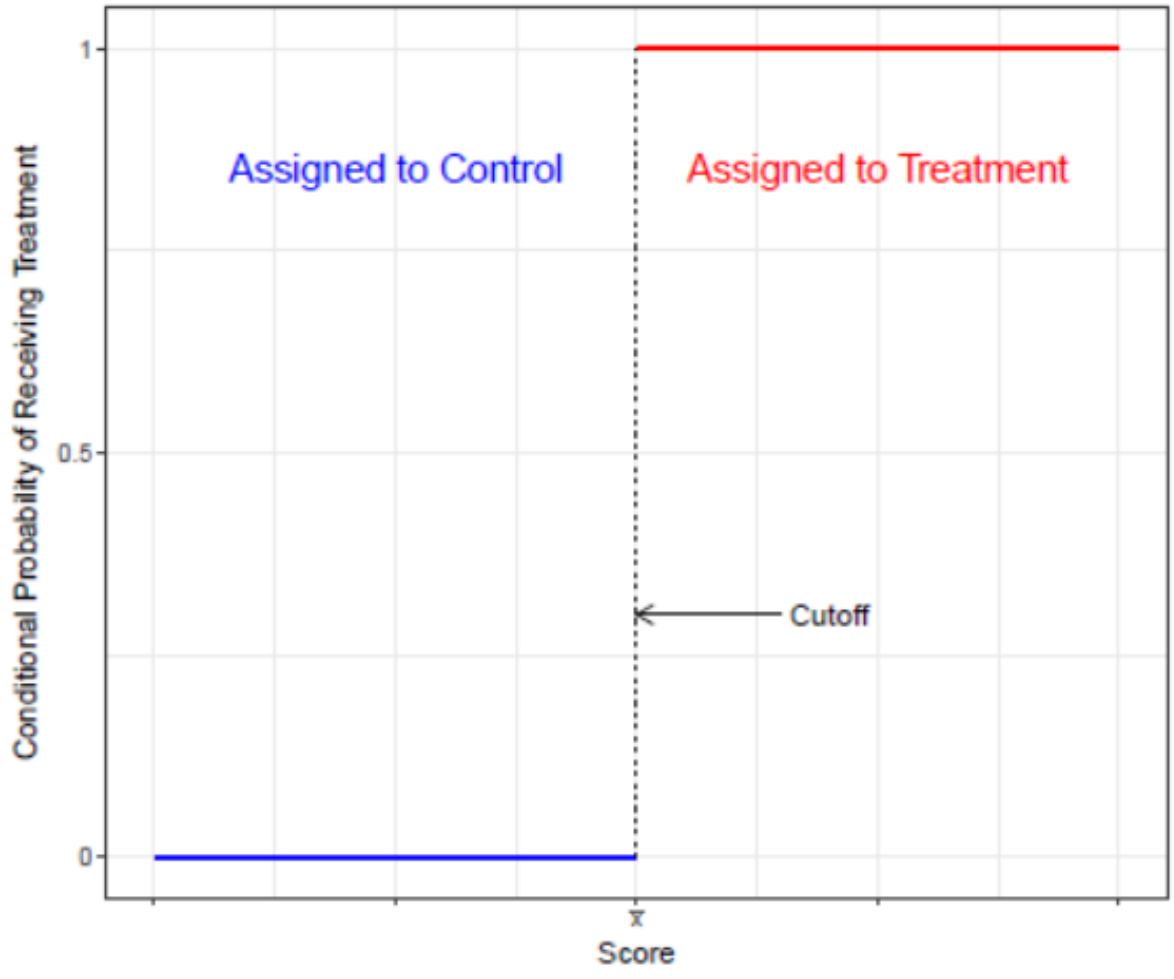
Sharp RD Set up:

The probability that $D = 1$ changes from 0 to 1 as the running variable crosses c.

- no one with $X < c$ gets treated
- everyone with $X \geq c$ gets treated
- $\Rightarrow D_i$ is a deterministic function of X_i : $D_i = 1$ if $(X_i \geq c)$.
- (or vice versa)

Sharp RD Set up:

Figure 2.1: Conditional Probability of Receiving Treatment in Sharp RD Design



Sharp RD Set up:

To estimate the causal effect of D_i on some outcome Y_i, we simply take the difference in mean outcome on either side of c:

$$\lim_{x \to c} E[Y_i|X_i = x] - \lim_{x \leftarrow c} E[Y_i|X_i = x] = \lim_{x \to c} E[Y_i(1)|X_i = x] - \lim_{x \leftarrow c} E[Y_i(0)|X_i = x]$$

This estimates τ_{SRD}, the causal effect for individuals with $X_i = c$:

$$\tau_{SRD} = E[Y_i(1) - Y_i(0)|X_i = c]$$

Sharp RD Assumption:

The continuity assumption:

$E[Y_i(0)|X_i = x]$ **and** $E[Y_i(1)|X_i = x]$ **are continuous in** x**.**

(Absent treatment, there would be no discontinuity in outcomes.)

Sharp RD Assumption:

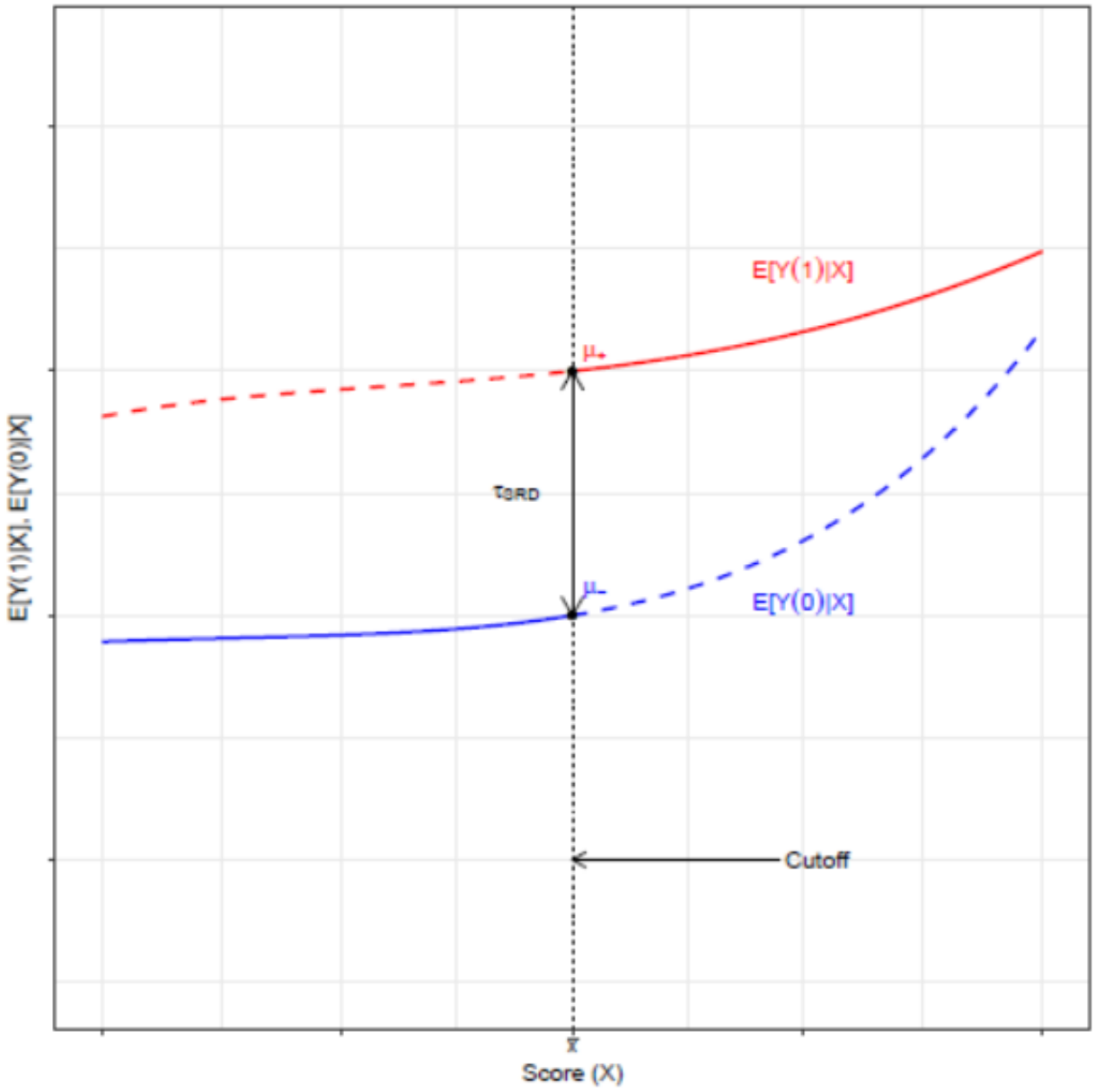
If the continuity assumption holds:

$$\tau_{SRD} = \lim_{x \to c} E[Y_i | X_i = x] - \lim_{x \leftarrow c} E[Y_i | X_i = x]$$

We can estimate τ_{SRD} as the difference between the two regression functions estimated in the **neighborhood** of c.

Sharp RD designs:

Figure 2.2: RD Treatment Effect in Sharp RD Design



Sharp RD Simulation

You are the superintendent of a large school district.

Last year you made participation in small reading groups **mandatory** for all students whose 3rd grade reading score was 75 points or less.

Your data includes the 3rd and 4th grade reading scores for all 5000 students in your school district.

How did these reading groups affected student performance on their 4th grade reading tests?

Sharp RD Simulation

```
set.seed(7000)

sharp<-rnorm(5000, mean=80, sd=5)
sharp<-as.data.frame(sharp)

names(sharp)<-c("read3")
sharp$error<-rnorm(5000, mean=0, sd=5)
sharp$pe3<-rnorm(5000, mean=90, sd=4)
sharp$height<-rnorm(5000, mean=130, sd=15)
```

Sharp RD Simulation

```
sharp$treated<-0
sharp$treated[sharp$read3<=75]<-1

tau=10
#the DGP
sharp$read4<-(-6)+0.8*sharp$read3+tau*sharp$treated+sharp$error

sharp<-sharp[sharp$read3<78 & sharp$read3>72,]
```

Fuzzy RD Set Up

Potentially more common than the sharp RD set ups.

D_i is no longer a deterministic function of X_i.

The probability of treatment changes by some nonzero amount as the running variable crosses the threshold c.

The change in probability is less than 100 percentage points.

$$0 < \lim_{x \to c} P(D_i = 1 | X_i = x) - \lim_{x \leftarrow c} P(D_i = 1 | X_i = x) < 1$$

Fuzzy RD Set Up

There are now two causal effects to be estimated:

- the effect of crossing the threshold on the probability of treatment (which is 1 in the sharp RD)
- the effect of crossing the threshold on the outcome.

Formally, the fuzzy RD estimator is

$$\tau_{FRD} = \frac{\lim_{x \to c} E[Y_i | X_i = x] - \lim_{x \leftarrow c} E[Y_i | X_i = x]}{\lim_{x \to c} E[D_i | X_i = x] - \lim_{x \leftarrow c} E[D_i | X_i = x]}.$$

What should this remind you of?

Fuzzy RD IV

RD analog of an IV estimator:

- the instrument is an indicator for whether X_i lies directly above c.
- similar to how the IV estimator allowed us to recover the LATE estimate in an RCT.

Fuzzy RD IV

In a fuzzy RD:

- the ITT group (say who have $X_i \geq c$ for example)
- the control group ($X_i < c$).

Because we are in fuzzy land:

- if you are in the ITT group does not necessarily mean you get treated (there are never takers)
- some in the control get treated (always takers).
- but there is a group of observations, the compliers, whose treatment status changes if they go from control to ITT (ie if they were moved across the threshold).

Fuzzy RD IV

If I just compare the outcomes of those that are above and below the threshold ($\lim_{x \to c} E[Y_i|X_i = x] - \lim_{x \leftarrow c} E[Y_i|X_i = x]$) the effect is "diluted". **Why?**

- for many observations, crossing the threshold has no effect (since they are always or never takers).

To get the LATE, scale treatment estimates by the change in probability of being treated ($\lim_{x \to c} E[D_i|X_i = x] - \lim_{x \leftarrow c} E[D_i|X_i = x]$).

$\Rightarrow$ the fuzzy RD design measures the average treatment effect for RD compliers at the threshold (the LATE),

$$\tau_{FRD} = E[Y_i(1) - Y_i(0)|\text{unit } i \text{ is a complier and } X_i = c].$$

Fuzzy RD Simulation:

You are the superintendent of a large school district.

Last year you **strongly encouraged** students to participate in small reading groups if their 3rd grade reading score fell below 75 points.

Your data includes the 3rd and 4th grade reading scores for all 5000 students in your school district.

How did these reading groups affected student performance on their 4th grade reading tests?

Fuzzy RD Simulation:

Fuzzy RD Simulation:

RD and External Validity

Notice the conditioning statements in our treatment estimators.

- With fuzzy RD's, we are estimating effects on compliers
- We focus on observations that are in the neighborhood of the threshold. **Why?**
  - implicitly we are assuming that an observation falling just above or just below the threshold is effectively random
  - $\Rightarrow$ we can measure a LATE that is valid around c
  - observations far below and far above the threshold likely differ from each other in observable and unobservable ways.

RD and External Validity

Because of these conditions, RD estimates are:

- inherently localized since the effects are estimated for a sub population where their X_i is in the neighborhood of c.
- have a relatively high degree of internal validity,
- it is important to think about their external validity:
  - treatment effects could be quite different for observation where X_i is quite different from c.

RD Graphs

Key to a good RD project is the graphical analysis (statistical results really take a back seat).

A discontinuous change in treatment and outcomes (if $\tau \neq 0$) should be visible as X_i crosses the threshold c.

Failure to show visually perceptible breaks at c challenges the credibility of the approach (regardless of the regression results).

A break that is visually perceptible will almost surely be statistically significant.

RD Graphs

A simple scatterplot of your data is unlikely to reveal the patterns you wish to illustrate.

The key RD graphs are histogram-type plot that presents the average value of:

- the outcome
- treatment status
- covariates

at evenly spaced values of the running variable.

RD Graphs

Gernerating these plots requires choosing two key parameters:

- the binwidth, h,
- the number of bins shown to the left and right of the threshold value, K_0 and K_1.

Once these choices are made, construct:

- K_0 evenly spaced bins of width h below the threshold value
- K_1 evenly spaced bins of width h above the threshold value
- (avoid having any bin crossing the threshold value c).

RD Treatment status graph

RD papers often include a graph that plots treatment by the running variable.

We expect to see a visually perceptible discontinuity in the probability of treatment as X crosses the threshold.

After constructing the bins described above, calculate $\bar{D}_k$, the average treatment level in the bin

$$\bar{D}_k = \frac{1}{N_k} \sum_{i=1}^{N} D_i * 1(b_k < X_i \leq b_{k+1}).$$

Plot these values against the midpoint of each of the bins.

Sharp RD: Treatment status graph

In a sharp RD, $\bar{D}_k$ should be either 0 or 1.

Sharp RD: Treatment status graph Simulation

```
#I will break up the data into 60 bins (30 above and 30 below the threshold)
cuts<-c(72,72.1,72.2,72.3,72.4,72.5,72.6,72.7,72.8,72.9,73,
        73.1,73.2,73.3,73.4,73.5,73.6,73.7,73.8,73.9,74,
        74.1,74.2,74.3,74.4,74.5,74.6,74.7,74.8,74.9,75,
        75.1,75.2,75.3,75.4,75.5,75.6,75.7,75.8,75.9,76,
        76.1,76.2,76.3,76.4,76.5,76.6,76.7,76.8,76.9,77,
        77.1,77.2,77.3,77.4,77.5,77.6,77.7,77.8,77.9,78)
midpoints<-cuts[2:61]-0.05

sharp$bins <- cut(sharp$read3,
                  breaks=cuts,
                  include.lowest=TRUE,
                  right=FALSE,
                  labels=midpoints)


sharp_mean<-sharp %>%
    group_by(bins) %>%
    dplyr::summarize(outbinmean = mean(read4, na.rm=TRUE),
                     treatbinmean=mean(treated, na.rm=TRUE),
                     pebinmean=mean(pe3, na.rm=TRUE),
                     heightbinmean=mean(height, na.rm=TRUE), numb=n())
```

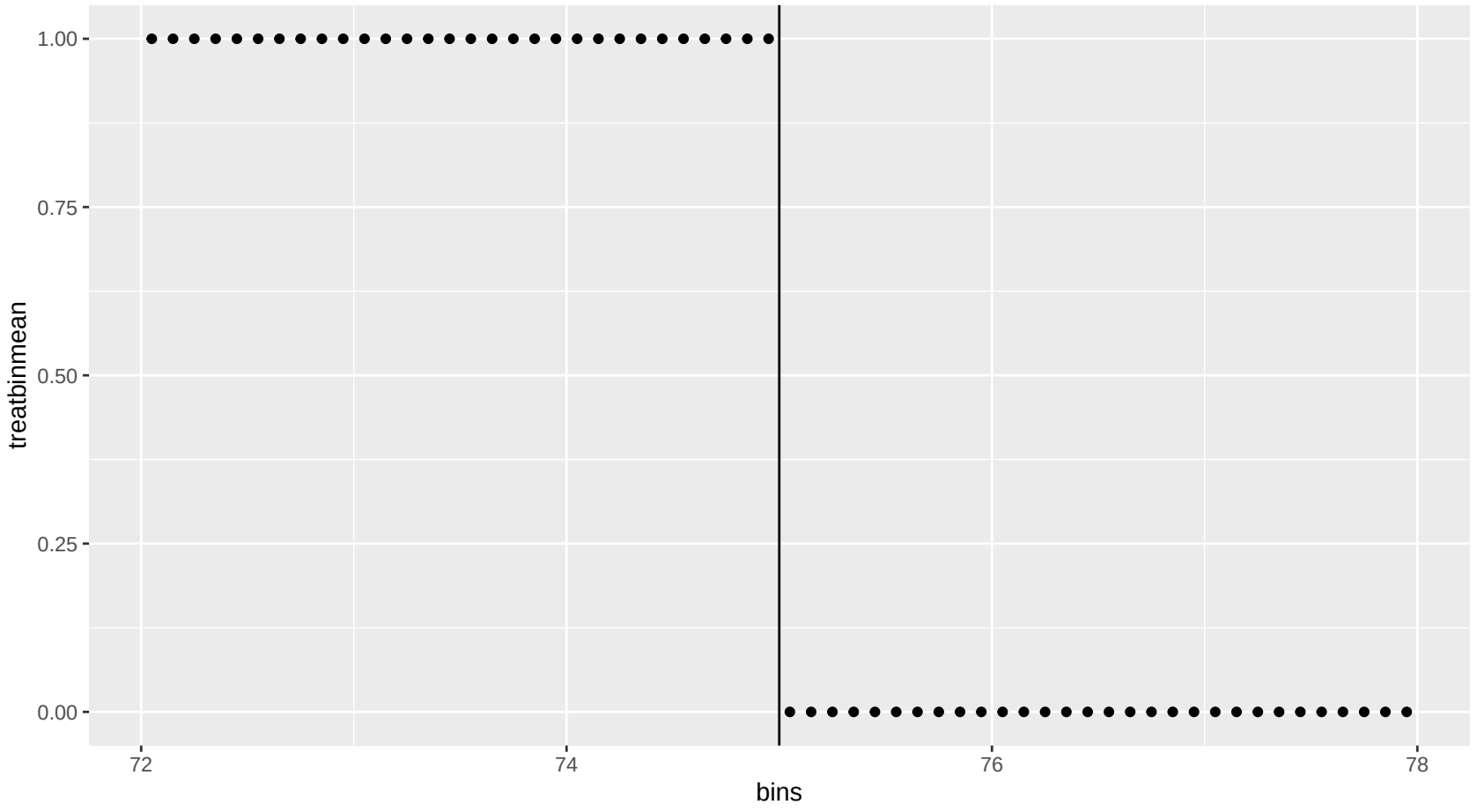
```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
sharp_mean$bins<-as.numeric(as.character(sharp_mean$bins))

plot1shp<-ggplot(sharp_mean, aes(x=bins, y=treatbinmean))+
         geom_point()+
         geom_vline(xintercept = 75)
```

Sharp RD: Treatment status graph Simulation

```
plot1shp
```



Fuzzy RD: Treatment status graph

In a fuzzy RD, $\bar{D}_k$ can take on many possible values.

This plot should show that there is a visual discontinuity in the probability of getting treated at the threshold c.

This graph is equivalent to the first stage in an IV analysis: It shows that we have found a tool that generates some random variation we can leverage to estimate unbiased treatment effects.

Fuzzy RD: Treatment status graph Simulation

```
#I will break up the data into 60 bins (30 above and 30 below the threshold)

cuts<-c(72,72.1,72.2,72.3,72.4,72.5,72.6,72.7,72.8,72.9,73,
        73.1,73.2,73.3,73.4,73.5,73.6,73.7,73.8,73.9,74,
        74.1,74.2,74.3,74.4,74.5,74.6,74.7,74.8,74.9,75,
        75.1,75.2,75.3,75.4,75.5,75.6,75.7,75.8,75.9,76,
        76.1,76.2,76.3,76.4,76.5,76.6,76.7,76.8,76.9,77,
        77.1,77.2,77.3,77.4,77.5,77.6,77.7,77.8,77.9,78)
midpoints<-cuts[2:61]-0.05

fuzzy$bins <- cut(fuzzy$read3,
                  breaks=cuts,
                  include.lowest=TRUE,
                  right=FALSE,
                  labels=midpoints)


fuzzy_mean<-fuzzy %>%
    group_by(bins) %>%
    dplyr::summarize(outbinmean = mean(read4, na.rm=TRUE),
                     treatbinmean=mean(treated, na.rm=TRUE),
                     pebinmean=mean(pe3, na.rm=TRUE),
                     heightbinmean=mean(height, na.rm=TRUE), numb=n())
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
fuzzy_mean$bins<-as.numeric(as.character(fuzzy_mean$bins))

plot1fuz<-ggplot(fuzzy_mean, aes(x=bins, y=treatbinmean))+
         geom_point()+
         geom_vline(xintercept = 75)
```

Fuzzy RD: Treatment status graph Simulation

```
plot1fuz
```