Module 7: Regression Discontinuity Design

Contents

List of Figures

1. INTRODUCTION

In the previous modules, we have studied counterfactuals, the exchangeability of the treatment and control groups, and how randomization minimizes selection bias. We have applied t-test and OLS regression analysis to determine the causal effects of randomized experiments. We have also reviewed some of the problems in conducting such randomized experiments. Now, we will discuss how to analyze causal effects when randomization is not possible using a quasi-experimental method.

In this module we discuss Regression Discontinuity Designs (RDD). This is a particularly useful tool to use when there is a cut-off criterion used to identify the target or eligible beneficiaries of an intervention. RDD exploits the fact that the eligible beneficiaries just above the cut-off are highly similar to those ineligible just below the cut-off. The degree of dissimilarity between these two groups will increase as we move away from the cut-off. However, the groups just above and below this "administratively-" decided cut–off will be highly similar, and the "selection bias" may be minimal. For example, fellowships may be awarded according to a cut-off in test scores: say, the 95th percentile. Would those scoring between the 95th and 96th percentiles be different than those between the 94th and 95th percentile? The difference is only because of a somewhat-arbitrary administrative criterion, which is established as a rule or convenience for decision making. The confounders can be expected to be well-balanced between people or groups just above and below such eligibility criteria. Therefore, those who were just below the cut-off (and did not receive the treatment) are a good counterfactual of those who scored just above the cut-off (and were assigned the treatment). Since this design exploits these *discontinuous* changes in a treatment assignment variable (also known as a "forcing" or "running" variable), we call it a regression discontinuity design. It is considered one of the most robust non-experimental evaluation designs when it is feasible to implement. The learning objectives of this module are:

- ✓ Identify interventions or program designs where RDD is applicable
- ✓ Learn how to visualize data from RDD studies
- ✓ Understand the difference between sharp and fuzzy designs and their basic application.

RDD can be complicated to analyze, and we recognize that more developed skills in econometrics and R are necessary. However, the purpose of this module is more to inform than to build your capacity to actually analyze an RDD design. However, adequate information will be provided in case you want further learn about RDD on your own.

2. VISUALIZATION AND GRAPHICAL ANALYSIS

Let's work with the PROGRESA panel data we have been using in previous modeules. The following steps help us process the data, understand its structure and how the program was assigned and adopted by households, and then graphically visualize and analyze the data.

- ✓ Open the data.
  - Open `PROGRESA_RD_Mod7.csv`. This is a panel dataset for children aged six to sixteen years. The panel consists of households and individuals from selected villages who were tracked annually from 1997 to 1999.
  - Figure 1 describes the program assignment and eligibility criteria. Households who were "poor" according to a government classification were eligible to receive the cash transfer under the PROGRESA. In the treatment group, about 53% households were eligible for the program. In control group, 51% household could have been eligible.

	PROGRESA_Data_1997$pov_HH		
PROGRESA_Data_1997$D_assign	0	1	Row Total
0	1316	1856	3172
	0.415	0.585	0.381
1	1993	3160	5153
	0.387	0.613	0.619
Column Total	3309	5016	8325

Figure 1. Distribution of hypothetical household eligibility across treatment and control groups

- ✓ Exploring the forcing variable
  - For RD to provide a consistent estimate of the treatment effect, the treatment must be assigned following a rule that depends on a forcing variable as discussed in the introduction. In PROGRESA, households from the treatment villages were eligible on the basis of a poverty index (`yycali`, the assignment variable). Those households where `yycali` was below a cutoff value were eligible for PROGRESA and offered the program benefits in the treatment villages.
  - We plot the distribution of `yycali` poverty index score to visualize the treatment and control households (note, we are now restricting the sample and don't estimate ITT but

ATET) as per the following code.

```
PROGRESA_Data_1997$pov_HH<- as.factor(PROGRESA_Data_1997$pov_HH)
ggplot(subset(PROGRESA_Data_1997,D_assign == 1 & pov_HH %in% c(0, 1)), aes(x=yycali)) + geom_density(aes(fill=pov_HH), alpha=0.25)
```

Figure 2 shows that the poverty index averaged 800 for the treated households and 700 for the ineligible control households.

- We find an overlap of the distributions because the Mexican government had different cut-offs for the poverty index in different regions (region variable: `entidad`).

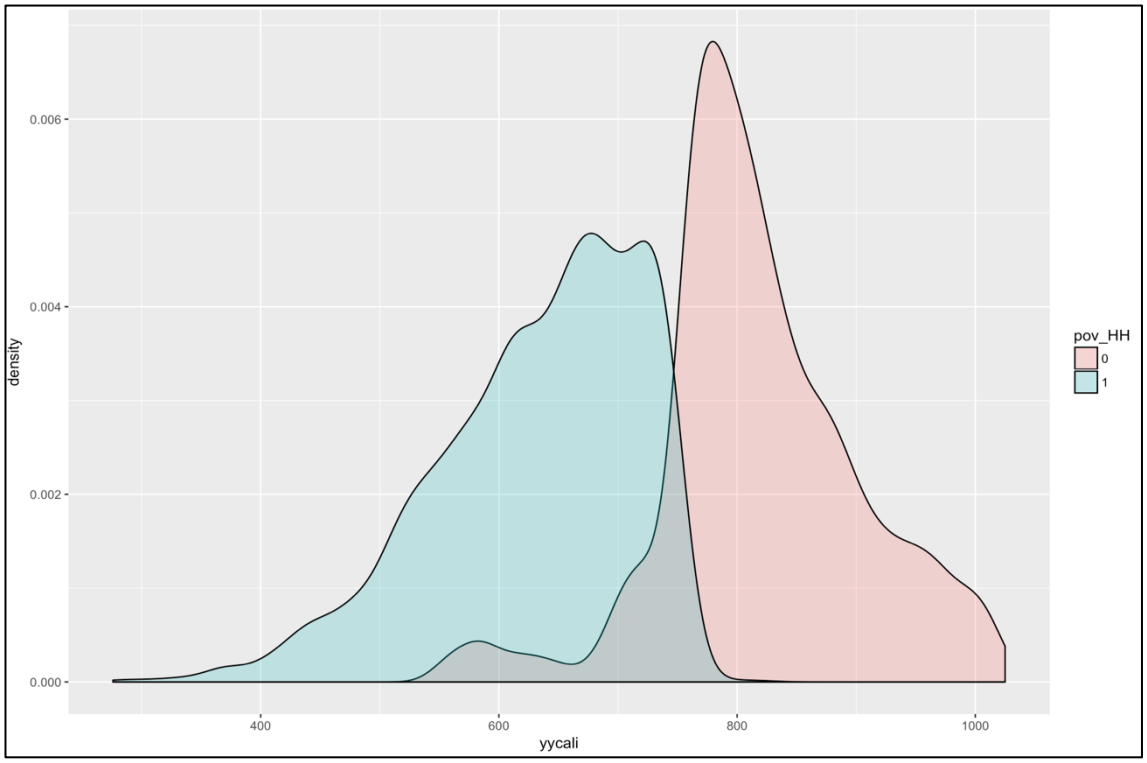

Figure 2. Distribution of forcing variable (poverty index) in treatment and control households

✓ Processing the data for RDD analysis

- Let's assume that the maximum `yycali` value for poor households from the treatment villages in each region is the cut-off value for that region. We can create cut- off values (`maxcut`) for the regions in a loop:

```
# aggregating and finding maximum per region
entidads_aggr <- aggregate(yycali ~ entidad, subset(PROGRESA_Data_1997, D_assign == 1 & pov_HH == 1), max)

colnames(entidads_aggr) <- c('entidad', 'maxcut')

# One - to - many merge to original dataset
PROGRESA_Data_1997 <- merge(PROGRESA_Data_1997, entidads_aggr, by = 'entidad')
```

- Next, we need to "normalize" the forcing variable value so that there is a single cut-off point in all regions. We will simply subtract the cut-off value determined in step (a) above from the `yycali` value in each region such that the cut-off value of the new variable is centered around 0.

The R code is,

```
PROGRESA_Data_1997$z <- PROGRESA_Data_1997$yycali - PROGRESA_Data_1997$maxcut
```

- Figure 3 plots the distribution of the new centered forcing variable `z` by the two groups of households we compare. We find that 0 is now the cutoff point. The R code used to create this graph is,

```
ggplot(subset(PROGRESA_Data_1997,D_assign == 1 & pov_HH %in% c(0,
    1)), aes(x=z)) + geom_density(aes(fill=pov_HH), alpha=0.25)
```

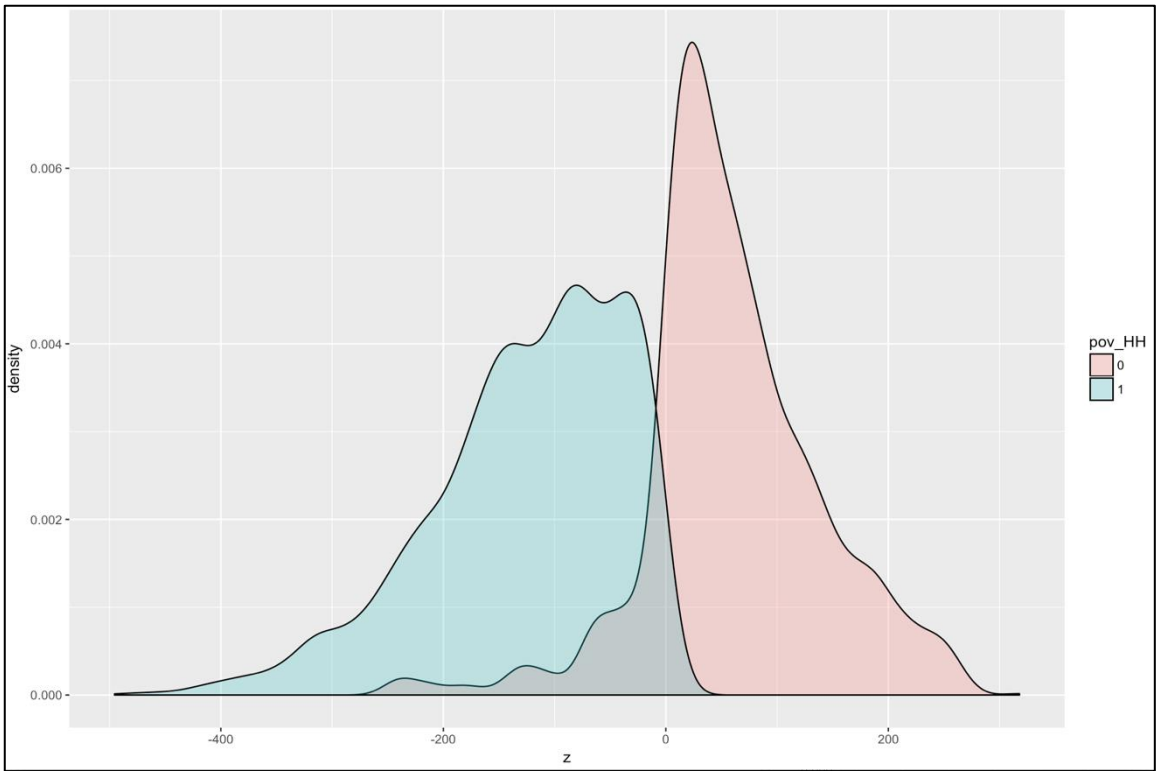

Figure 3. Distribution of cantered cut-off values for the two comparison groups

- We create a dummy assignment variable to flag those households who are eligible versus those who are not:

```
PROGRESA_Data_1997$E <- ifelse(PROGRESA_Data_1997$z <= 0, 1, 0)
```

For the subsequent sections, we will be using the code from Lab 7's jupyter notebook.

Graphical analysis to visualize RDD based impacts

- We had discussed that in RDD, the comparison should be done as close to the cut-off point as possible. However, the narrower the range, the smaller is the sample available for analysis, implying a trade-off between sample size and theoretical consistency. Let's constrain the analysis to only those households where the centered poverty index `z` is +/- 200 (arbitrarily chosen for sake of demonstration).

- We restrict the sample to observations from 1999 and include only the individuals that

would fit in a sharp design (which will be explained in the next section). This filtered dataset is PROGRESA_RD.csv, which we will be using for the rest of the analysis.

- We limit the data close to the cutoff, within 200 points of the poverty_index cutoff z.

```
sampleRD <- subset(PROGRESA_RD,  ( PROGRESA_RD$z>=-200 &
PROGRESA_RD$z<=200) )
```

- Within this subsample, we then calculate 10 bins of z

```
sampleRD<-sampleRD[order(sampleRD$z),]
sampleRD$bin<-cut_interval(sampleRD$z, length=10, labels =
   FALSE)
sampleRD$bin<- -210+(sampleRD$bin*10)
   summary(sampleRD$bin)
```

- Next, we are going to focus on the variable of child enrollment in school as our outcome. We want to compare enrollment rates for those just below and above the cutoff, so we calculate the mean enrollment values for each of the bins we created above and plot these means against the forcing variable (relative poverty index). We will also fit regression lines for below and above the cutoff.

```
#Make enroll_child factor variable and limit datset to only relevant
   variables
 sampleRD$enroll_child<-ifelse(sampleRD$enroll_child == "si" ,1,0)
 v<-c("bin", "z", "enroll_child")
 sampleRD2<-sampleRD[v]

 #collapse dataset to get bin means
 aggdata     <-aggregate(sampleRD2,     by=list(sampleRD2$bin),
FUN=mean, na.rm=TRUE)

 #create separate variables for below and above cutoff
 aggdata$below<-aggdata$enroll_child
 aggdata$below[aggdata$bin>=0]<-NA
 aggdata$above<-aggdata$enroll_child
 aggdata$above[aggdata$bin<0]<-NA

 #Plot bin averages
 L <- aggdata %>% ggplot(aes(bin, enroll_child))
 P <- geom_point(color='blue')
 ## add fit lines for above/below cutoff
 L  <-  L  +  geom_smooth(color='red',method='lm',  formula=y~x,
aes(aggdata$bin, aggdata$above))
 L  <-  L  +  geom_smooth(color='green',method='lm',  formula=y~x,
aes(aggdata$bin, aggdata$below))

 #Display Plot
   L+P + xlab("X") +  ylab("Child Enrollment")
```

- Figure 4 compares the impact around the cut-off point between eligible (poor) and ineligible (non-poor) households from the treatment villages on child enrollment. We do find a discrete shift in the beneficiary enrollment rate at the discontinuity point. Later, we will test the statistical significance of this shift using regression analysis. It is important to

remember that RDD estimates the conditional impact around the discontinuity point (the local impact on those individuals that are close to z=0, which is not necessarily generalizable to the broader population).

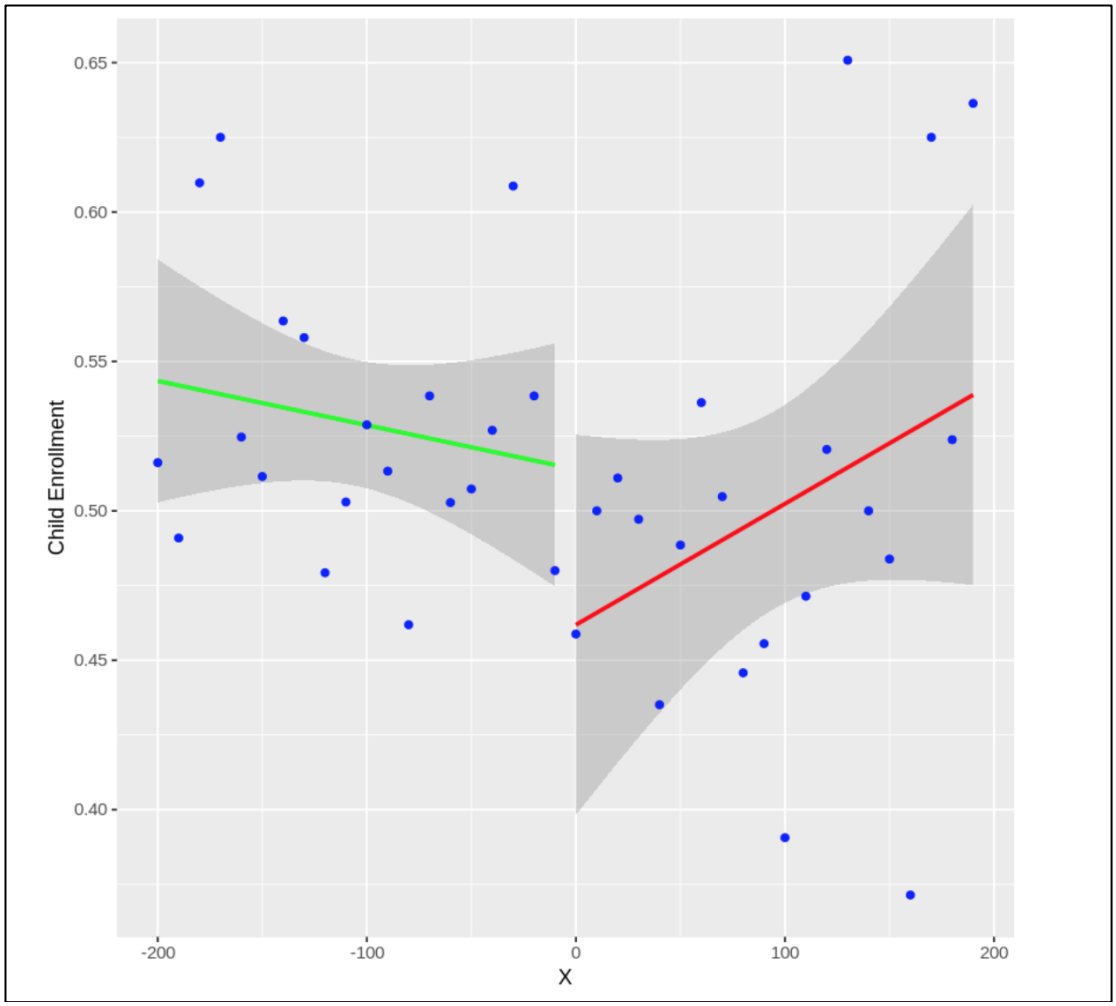

Figure 4. RDD graphical analysis: comparing the enrollment effect on eligible and non-eligible households around the cut-off value of the forcing variable

3. REGRESSION ANALYSIS IN RDD

In RDD, the treatment or the intervention (T) is a deterministic function of the forcing variable (Z), such that:

$$T_i = 1 \quad if \quad Z_i \geq C$$

where T_i is the treatment or intervention indicator variable for an individual i, Z_i is the value of assignment variable for that individual and C is the critical value above which the individual received treatment. We usually estimate the average causal effect of the treatment at the discontinuity point as a difference in conditional outcomes (Y):

$$\tau_{SRD} = \lim_{Z_i \to C^+} E[Y_i|Z_i] - \lim_{Z_i \to C^-} E[Y_i|Z_i]$$

This kind of effect estimation is called ***Sharp*** RDD. The Sharp RDD requires that the treatment goes from 0 to 1 at the cut-off value of the forcing variable. However, in ***Fuzzy*** RDD, the probability of receiving the treatment can change on a continuous scale from 0 to 1. This situation is more commonly encountered in real life evaluations because of non-compliance. For example, when the threshold value for eligibility or the program benefits are not broad enough for all eligible households/individuals to participate in the program, several of them will not! This results in the probability of participation among the eligible households being less than 1.

Similarly, the ineligible may somehow circumvent the official threshold and participate in the program, so their participation probability may be non-zero. We account for such continuous probability of participating in the treatment ($E[T_i|Z_i]$) in Fuzzy RDD to estimate the impacts as:

$$\tau_{FRD} = \frac{\lim_{Z_i \to C^+} E[Y_i|Z_i] - \lim_{Z_i \to C^-} E[Y_i|Z_i]}{\lim_{Z_i \to C^+} E[T_i|Z_i] - \lim_{Z_i \to C^-} E[T_i|Z_i]}$$

Note, sharp RDD average treatment effects are a special case of fuzzy RDD effect when the denominator is 1.

3.1 Regression Analysis for Sharp RDD

We can estimate an average treatment effect as follows if the compliance with treatment protocol is perfect; that is, all eligible individuals who are assigned to the treatment (because they were above a cut-off) actually participate in it, and those who are not assigned or ineligible do not participate:

$$Y_i = \beta_0 + \beta_1 D_i + f(Z_i) + \varepsilon_i$$

where Y_i is the outcome for individual i and β_1 is the average treatment effect. To control for differences between the treatment and control individuals away as their distance from the discontinuity point, we control for a function of the forcing variable (Z_i) as $f(Z_i)$ in the estimation. In reality we cannot observe $f(Z_i)$, but Figure 4 and 5 suggest a somewhat linear relationship. However, it is always good practice to evaluate robustness of our effects to different specifications of $f(Z_i)$. For

example, we use the functions $f(Z_i) = Z_i$ and $f(Z_i) = Z_i + Z_i^2$ in the R demonstration below.

- ✓ If we assume that everyone below the cutoff received treatment and everyone above did not receive treatment, we can run a regression using an indicator for "below the cutoff" as the independent variable, that estimates the treatment effect on the dependent variable. We also control for the forcing variable (X, the poverty index) and should try doing this linearly and with other functional forms (e.g. quadratic) Again, we want to make sure we limit the data to be relatively close to the cutoff (we've already limited to +-200).

```
#Cluster at the village level and inlude a linear control for z
    #pov_HH=1 is the indicator for being below the cutoff
    model <- lm.cluster(data = sampleRD, formula= enroll_child ~
pov_HH+z, cluster=sampleRD$villid)
    summary(model)
    #next, try a quadratic control for z
    sampleRD$z2<-sampleRD$z^2
    model2 <- lm.cluster(data = sampleRD, formula= enroll_child ~
pov_HH+z+z2, cluster=sampleRD$villid)
    summary(model2)
```

- ✓ Figure 6 displays the results of the regression analyses conducted above.

```
R^2 = 0.00119

                 Estimate   Std. Error    t value      Pr(>|t|)
(Intercept)  4.943732e-01 0.0195547719 25.2814633 5.108385e-141
pov_HH       2.953866e-02 0.0281280015  1.0501512  2.936486e-01
z           -3.816885e-05 0.0001534685 -0.2487079  8.035867e-01
R^2 = 0.00178

                Estimate   Std. Error    t value      Pr(>|t|)
(Intercept) 4.789523e-01 2.252099e-02 21.2669323 2.298280e-100
pov_HH      3.885764e-02 2.869374e-02  1.3542201  1.756662e-01
z           5.639947e-05 1.648254e-04  0.3421771  7.322176e-01
z2          1.285038e-06 8.800953e-07  1.4601123  1.442592e-01
```

Figure 5. Sharp RD regression analysis results

3.2 Regression Analysis for Fuzzy RDD

In the case of imperfect compliance, we can implement a fuzzy RDD. We use instrumental variable / 2 stage least square (IV/2SLS) method to estimate the effects in Fuzzy RDD as follows,

First Stage: $T_i = \beta_0 + \beta_1 E_i + f(Z_i) + \varepsilon_i$

Second Stage: $Y_i = \alpha_0 + \alpha_1 \hat{T}_i + f(Z_i) + u_i$

where E_i is the dummy variable we created earlier to note whether a household was eligible for participation on basis of Z_i. Demonstration of the R code is provided below.

```
# Estimate the average treatment effect in year 1999,
including a linear control for z
iv_model <- ivreg(enroll_child ~  pov_HH + z, data = sampleRD)
summary(iv_model)
# Next try a quadratic control
iv_model_2 <- ivreg(enroll_child ~ pov_HH + z + z2 , data =
```

```
sampleRD)
summary(iv_model_2)
```

```
Call:
ivreg(formula = enroll_child ~ pov_HH + z, data = sampleRD)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5315 -0.5252  0.4691  0.4745  0.5132

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.944e-01  1.441e-02  34.299   <2e-16 ***
pov_HH       2.954e-02  2.524e-02   1.170    0.242
z           -3.817e-05  1.298e-04  -0.294    0.769
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4996 on 5259 degrees of freedom
Multiple R-Squared: 0.001188,   Adjusted R-squared: 0.0008081
Wald test: 3.127 on 2 and 5259 DF,  p-value: 0.04391


Call:
ivreg(formula = enroll_child ~ pov_HH + z + z2, data = sampleRD)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5579 -0.5178  0.4495  0.4820  0.5210

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.790e-01  1.686e-02  28.413   <2e-16 ***
pov_HH      3.886e-02  2.578e-02   1.507   0.1319
z           5.640e-05  1.404e-04   0.402   0.6880
z2          1.285e-06  7.288e-07   1.763   0.0779 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4995 on 5258 degrees of freedom
Multiple R-Squared: 0.001778,   Adjusted R-squared: 0.001209
Wald test: 3.122 on 3 and 5258 DF,  p-value: 0.02488
```

Figure 6. Fuzzy RD regression analysis results

Exercise 7.1

(a) Repeat what we've done above (Graph and Sharp RD regressions) for at least two separate versions, varying the choice for the values of Z we include in the sample (band size-- something other than +-200) in one version, and the bin size in another (something other than 10). Interpret the results of the graphs and regression. What effect do the selected bin and band size have on the results?

(b) In the above analysis, we compare household within treatment villages. Describe at least one reason we might be concerned that a comparison of households below and above the cutoff to identify the treatment effect of PROGRESA might lead to a biased result.

4. SPECIFICATION AND ROBUSTNESS CHECKS

- ✓ **Sensitivity to functional form assumptions:** We have seen that changes in our assumptions about the functional form of $f(Z_i)$ can change the estimated magnitude of the treatment effect. We can evaluate the robustness of our results by estimating treatment effects for a wide variety of $f(Z_i)$ specifications. Above, we tried two different functional forms; we can add higher-order terms for Z_i and evaluate their effect.

- ✓ **Effect of socioeconomic and other factors on the treatment effect:** RDD is a quasi- experimental design, and it remains possible that some confounders (measured or unmeasured) will remain unbalanced at the baseline. For example, in the fuzzy RDD above, we found some evidence of imbalance in the outcome baseline itself. We can thus add individual-
, household- and village-specific control variables to the regression model in order to assess the robustness of the estimated causal effect.

- ✓ **Effect of the choice of the discontinuity criteria:** There may not be well-defined eligibility criteria, so we may have to determine the cut-off value for the assignment variable based on the data (as we did above). In such case, it is a good idea to estimate the effect at a few other discontinuity points to assess the robustness of the treatment effect. An alternative is the use is of "placebo" discontinuity points to build confidence that the detected association is truly causal. If the impact of the treatment is actually occurring around the discontinuity point, then we should not see differential treatment effects at "placebo" discontinuity points.

5. BILIOGRAPHY/FURTHER READINGS

More detailed information about RDD is available at:

1. Gertler, Paul J., Sebastian Martinez, Patrick Premand, Laura B. Rawlings, and Christel MJ Vermeersch. "Impact evaluation in practice." World Bank Publications, 2011.
2. Imbens, Guido and Thomas Lemieux (2008). "Regression Discontinuity Designs: A Guide to Practice," Journal of Econometrics, 142, 615-635.
3. Lee, David and Thomas Lemieux (2010). "Regression Discontinuity Design in Economics," Journal of Economic Literature, 48(2), 281-355.
4. Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," Review of Economic Studies, 79, 933-959.
5. McCrary, Justin (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," Journal of Econometrics, 142(2), 698-714.
6. Cook, Thomas and Vivian Wong (2008). "Empirical Test for the Validity of the Regression Discontinuity Design," Annals of Economics and Statistics, 91/92, 127-150.