Center for Effective Global Action
University of California, Berkeley

Module 2.4: 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 STATA 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 since Module 2.2. 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 and process the data.
  - Open `PanelPROGRESA_Enrollment_97_99.dta`. 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.
  - Open the dataset and create some variables that we will need in our analysis. Please refer to the DO file to note these data processing changes. Basically, we assigned the household poverty status and enrollment in PROGRESA from 1998 to observations from 1997.
  - 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.

	pov_HH		
D_assig	Non poor	poor	Total
0	1,316	1,856	3,172
	41.49	58.51	100.00
1	1,993	3,160	5,153
	38.68	61.32	100.00
Total	3,309	5,016	8,325
	39.75	60.25	100.00

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 an 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, `twoway (kdensity yycali if pov_HH==1 & D_assig==1 & year==1997) (kdensity yycali if pov_HH==0 & D_assig==1 & year==1997), legend(lab(1 Poor) lab(2 Non-Poor)) graphregion(fcolor(white)) title("")` . 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`).

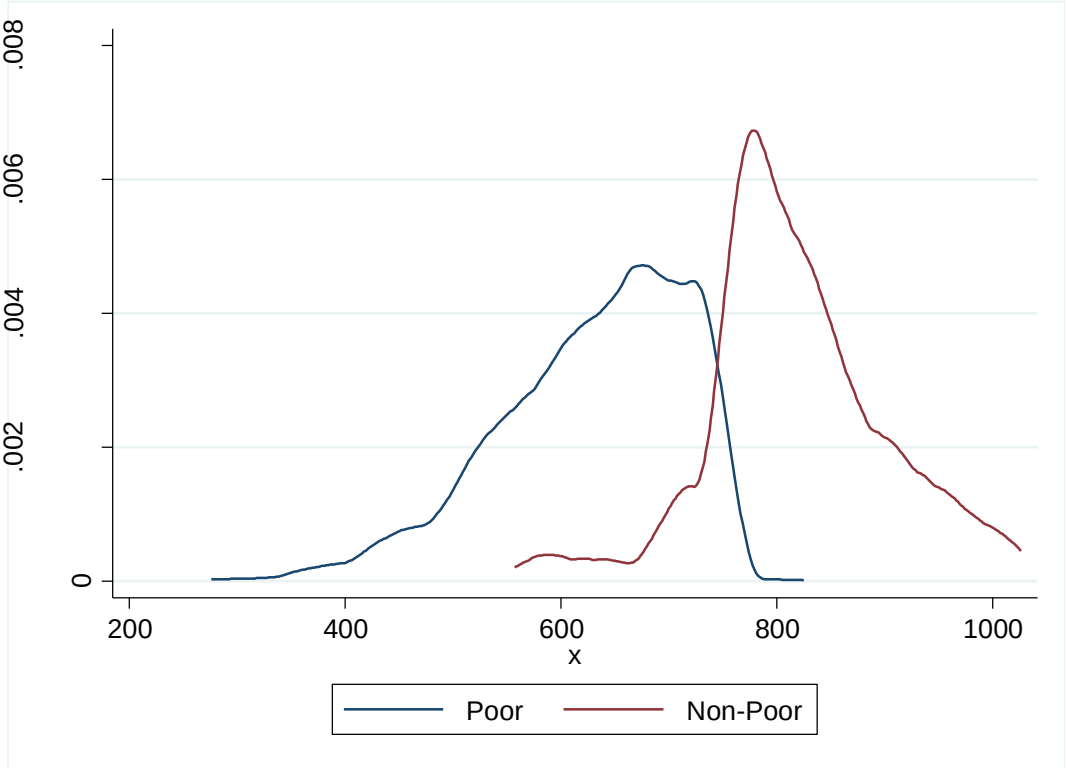

Figure 2. Distribution of forcing variable (poverty index) in treatment and control households

✓ 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 (`max_*`) for the regions in a loop:

```
gen maxcut = 0
levelsof entidad, local(entidades)
foreach j of local entidades {
      sum yycali if year==1997 & D_assig == 1 & pov_HH==1 &
      entidad==`j'
         replace maxcut=r(max) if entidad==`j'

}
```

This provides a good example of the usefulness of loops in STATA code; see the help files for the 'forvalues' and 'foreach' commands for more information.

- 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 STATA code is,
  ```
  gen z = yycali - 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 STATA code used to create this graph is,
  ```
  twoway (kdensity z if pov_HH==1 & D_assig==1 & year==1997)(kdensity z if pov_HH==0 & D_assig==1 & year==1997), legend(lab(1 Poor) lab(2 Non-Poor)) graphregion(fcolor(white)) title("")
  ```

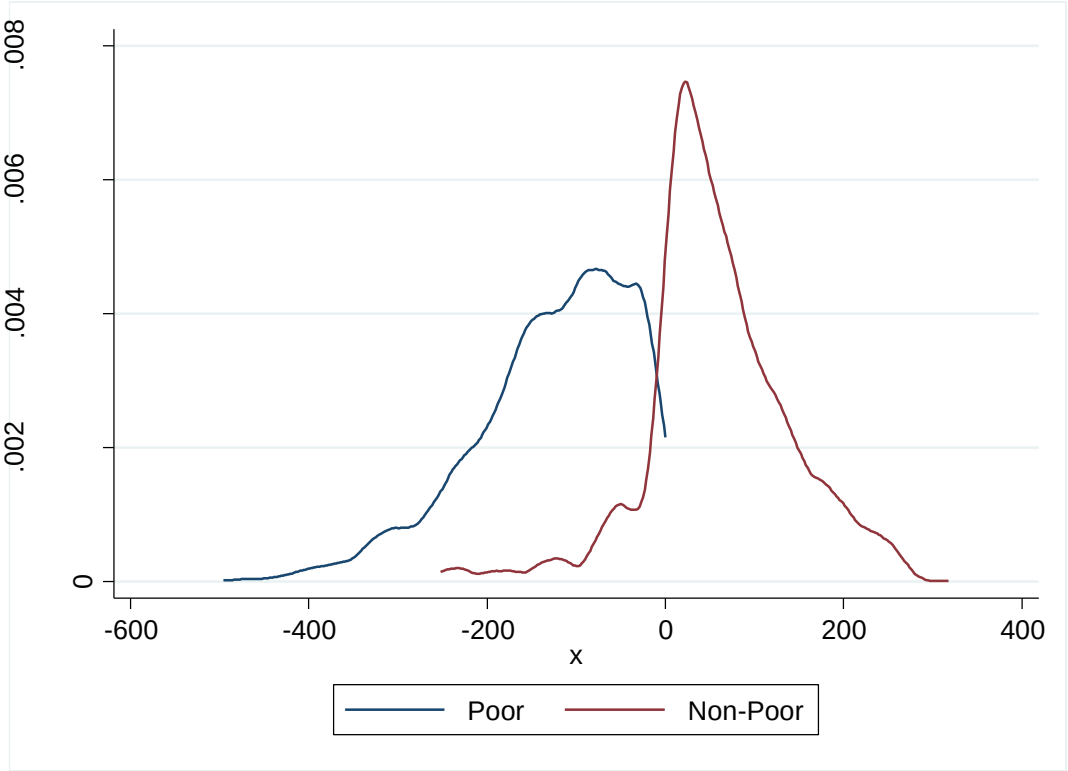

Figure 3. Distribution of cantered cut-off values for the two comparison groups

- We create a dummy assignment variable to flag those households who are eligible versus those who are not: `gen E = z<=0`

✓ 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).

  Within this subsample, we calculate 60 bins of equal size (30 on each side around the discontinuity point) of z and take means for the variable `enroll` inside every bin. Then, we plot the mean of every bin of `z` against the mean of `enroll`. The STATA code is as follows:

```
gen sampleRD = D_assig==1 & ( (pov_HH == 1 & z>=-200 & z<=0)
| (pov_HH == 0 & z>0 & z<=200) ) & year == 1999

xtile hl = z if pov_HH==1 & sampleRD == 1, n(30)
xtile hu = z if pov_HH==0 & sampleRD == 1, n(30)

gen hd = -hl if pov_HH==1
replace hd = hu if pov_HH==0

egen meanZ = mean(z), by(hd)
egen meanEnroll = mean(enroll), by(hd)
egen meanpov_HH = mean(pov_HH), by(hd)

gen meanZ2 = meanZ ^ 2

reg meanEnroll meanZ meanZ2 if meanpov_HH==1
predict yhat1 if e(sample)

reg meanEnroll meanZ meanZ2 if meanpov_HH==0
predict yhat0 if e(sample)

sort meanZ

twoway   (scatter meanEnroll meanZ if meanpov_HH ==1) (line
yhat1 meanZ if meanpov_HH ==1) || ///
   (scatter meanEnroll meanZ if meanpov_HH ==0) (line yhat0
meanZ if meanpov_HH ==0), ///
   ylabel(0          1)          xline(0)          legend(off)
graphregion(fcolor(white))
```

While the above graphical analysis is beyond the basic STATA skill level required in this course and we will not test you on this code, we also hope that you will try to understand the concepts and STATA commands presented here.

- 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).

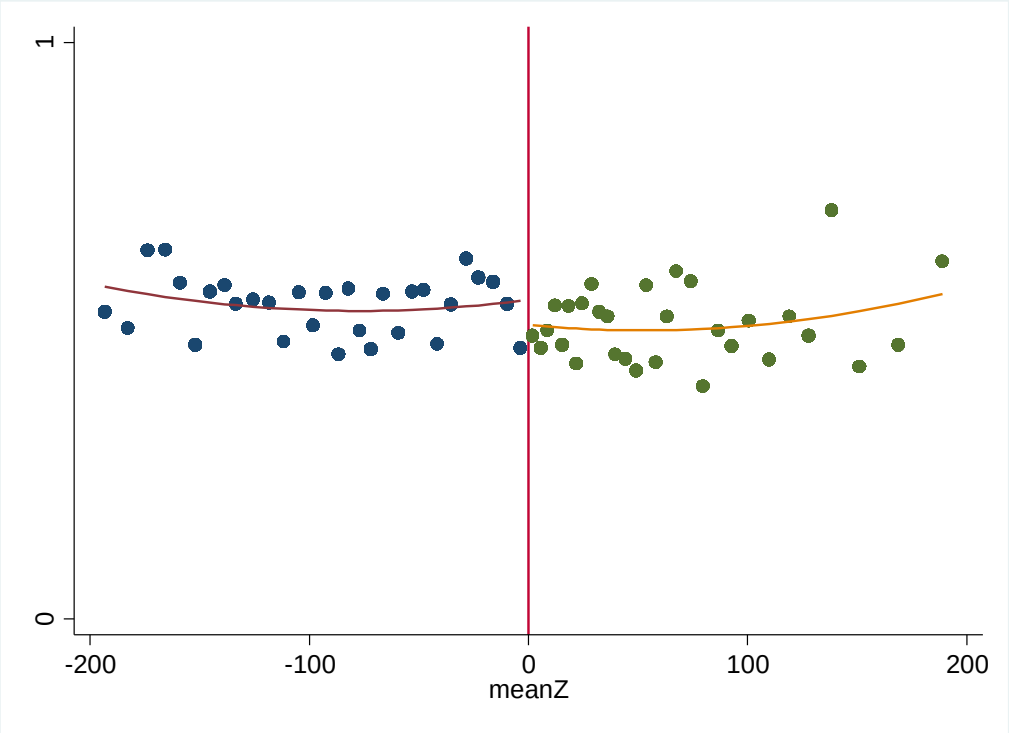

Figure 4. RDD graphical analysis: comparing the enrollment effect on eligible and non-eligible households around the cut-off value of the forcing variable

✓ RDD comparing eligible individuals in treatment group with ineligible individuals in the control group

- Above, we compared the households from the treatment group at the cut-off to evaluate the causal effect of the treatment. However, PROGRESA may have had local positive externalities, in which the ineligible households would also have benefitted from the program. Therefore, comparison within the treatment villages alone may underestimate the causal effect of the treatment. However, we could instead use the households from the control villages who were below the cut-off point (and thus doubly-ineligible to receive PROGRESA). This is reasonable because the assignment of villages

to PROGRESA is random and we can assume that the confounders are balanced at the village level.

- We will be comparing poor individuals from the treatment villages with non-poor individuals from the control villages, and that there could be "selection bias" in these two groups. A quasi-experimental RDD allows us to minimize this bias by restricting the comparison to a small region around the cut-off point. Again, we are doing so just because we don't want to underestimate the effects due to positive externalities.

- We repeat the graphical analysis above in STATA as follows. Figure 5 is the output of this graphical analysis. We find that the impacts around the discontinuity point are larger and clearer than those in Figure 4 as expected.

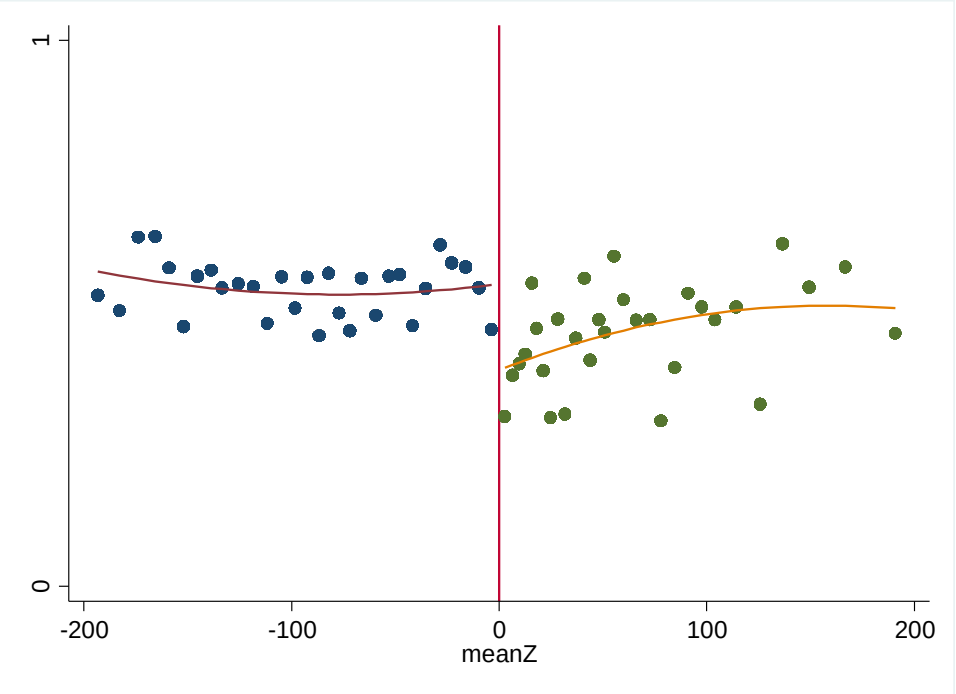

Figure 5. RDD graphical analysis – comparing enrolment effects on eligible individuals from treatment villages with that on ineligible individuals from the control villages around the cut-off value of the forcing variable

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 STATA demonstration below.

- ✓ We will compare the conditional outcome in enrollment rates between the participating eligible households and the ineligible households from the control group. We restrict the observations appropriately as shown in the DO file. We also create the variable `z2` by squaring `z`. Note how we are creating a flag variable to restrict the observations to the subset that we want to analyze: `replace sampleRD = ( (D_assig==1 & pov_HH == 1 & z>=-200 & z<=0) | (D_assig==0 & pov_HH == 0 & z>0 & z<=200) )`

✓ Check for the baseline balance

- We use two alternative specifications of $f(Z_i)$:,

```
reg enroll D_assig z if sampleRD==1 & year==1997,
vce(cluster villid)
eststo r1_97

reg enroll D_assig z z2 if sampleRD==1 & year==1997,
vce(cluster villid)
eststo r2_97
```

- We find that the coefficient for D_assig is not significant in any model, indicating a good baseline balance in the outcome.

✓ Estimate the ATE at the endline

- We estimate the impacts in year 1999 by specifying the same regression models as above:

```
reg enroll D_assig z if sampleRD==1 & year==1999,
vce(cluster villid)
eststo r3_99

reg enroll D_assig z z2 if sampleRD==1 & year==1999,
vce(cluster villid)
eststo r4_99
```

✓ Export the stored regression results. This is not required for RDD, but we are demonstrating how regression analysis results can be exported for your reference.

```
xml_tab r1_97 r2_97 r3_99 r4_99, replace
save("[PATH]/RD_TableI.xml") title("Table I: Sharp RD for
Enrollment") below stats(N r2)
```

Again, some of these commands exceed the necessary knowledge of STATA required for this course, but we use this code for our own ease of demonstrating the results and with a hope that you may become more conversant with these commands over time.

✓ Figure 6 displays the results of the regression analyses conducted above. We find that PROGRESA increased enrollment rate by between 9 and 11 percentage points at the discontinuity. We also find that the functional form of $f(Z_i)$ has only a minor effect on the treatment size.

Table I: Sharp RD for Enrollment				
	r1_97 coef/se	r2_97 coef/se	r3_99 coef/se	r4_99 coef/se
D_assig	0.026	0.036	0.096**	0.119***
	(0.036)	(0.037)	(0.041)	(0.043)
z	0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)
z2		0.000		0.000**
		(0.000)		(0.000)
_cons	0.638***	0.627***	0.457***	0.429***
	(0.024)	(0.026)	(0.028)	(0.031)
Number of observations	5,472	5,472	4,457	4,457
R2	0.000	0.001	0.006	0.007
note: .01 - ***; .05 - **; .1 - *;				

Figure 6. 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 STATA code is provided below.

- ✓ Restrict the sample to households as above, but now allow for some non-compliance. We change the values of the flag variable to restrict the sample used in analysis: `replace sampleRD = ( (D_assig==1 & pov_HH == 1 ) | (D_assig==0 & pov_HH == 0) ) & (z>= -200 & z <=200)`

- ✓ The baseline-balance check is the same as in the case of Sharp RDD
  - `ivregress 2sls enroll (D_assig=E) z if sampleRD == 1 & year==1997, vce(cluster villid)`
    `eststo r1_97`

    `ivregress 2sls enroll (D_assig=E) z z2 if sampleRD == 1 & year==1997, vce(cluster villid)`
    `eststo r2_97`

  - We find that the baseline difference in outcomes is significant, suggesting that our estimate of the treatment effect will be biased. There exist methods to deal with this bias, but they are beyond the scope of this module.

✓ Estimate the average treatment effect in year 1999:

- ```
  ivregress 2sls enroll (D_assig=E) z if sampleRD == 1 & year==1999, vce(cluster villid)
  eststo r3_99

  ivregress 2sls enroll (D_assig=E) z z2 if sampleRD == 1 & year==1999, vce(cluster villid)
  eststo r4_99
  ```

- Figure 7 presents the results of Fuzzy RD analysis. Again, assuming the model is correctly specified, the coefficient on `D_assig` captures the causal effect of PROGRESA on the enrolment rate among compliers located close to the discontinuity point.

Table II: Fuzzy RD for Enrollment

	r1_97 coef/se	r2_97 coef/se	r3_99 coef/se	r4_99 coef/se
D_assig	0.060	0.069*	0.123***	0.148***
	(0.039)	(0.039)	(0.045)	(0.047)
z	0.000	0.000*	0.000	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)
z2		0.000		0.000*
		(0.000)		(0.000)
_cons	0.627***	0.619***	0.447***	0.422***
	(0.024)	(0.027)	(0.028)	(0.032)
Number of observations	5,695	5,695	4,644	4,644
R2	.	.	0.006	0.006
note: .01 - ***; .05 - **; .1 - *;				

Figure 7. Fuzzy RD regression analysis results

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.