Chapter 5: Regression Discontinuity Designs

Pre-requisites

- Chapter 1: Intro to STATA
- Chapter 2: Review of Regression
- Chapter 3: Experiments
- Chapter 4: Problems with Experiments

Contents

1. Introduction

The previous chapters were devoted to experimental designs and its most common shortcomings. In this chapter, students will learn how to design impact evaluation studies when the treatment status is not a result of a random process as the one described in chapters 3 and 4. Instead, individuals are

considered beneficiaries of an intervention if a certain variable is above or below a determined threshold, and considered part of the control if they are not. These thresholds are common in social programs. For instance, fellowships programs are assigned according to test scores for which a minimum level if defined. Those applicants with scores slightly below this threshold value are not assigned a fellowship whereas those with scores slightly are granted one. The idea behind this approach is that those who were just below the cut-off (and did not receive the program) are a good counterfactual of those who scored just above the cut-off (and were assigned to the program). Since this approach exploits these discontinuous changes in treatment assignment related to an assignment variable (also known as "forcing" or "running" variable), it has received the name of **regression discontinuity design** (RD).

RD has become a popular approach in program evaluation. It is considered one of the most credible research designs, and the best approach when randomization is not feasible. It is transparent and easy to interpret. We will discuss the implementation of RD in the context of PROGRESA, exploiting the same datasets from previous chapters. In particular, we will take advantage of the poverty score used by the Mexican government to allocate the treatment within treated villages to discuss RD.

Specifically, we will cover the following issues:

- We analyze the visual analysis of RD. This is the most characteristic devise of RD because it helps to see the "jump" on the outcome at the discontinuity point of the assignment variable.
- We cover regression analysis for sharp and fuzzy regression discontinuity designs. In this chapter will be covering either parametric (or polynomial) and non-parametric (mainly, local regression) estimators.

At the end of this chapter we expect students to be able to:

- Recognize a situation when a regression discontinuity is feasible in programs where standard random assignment is not possible and be able organize the data appropriately to conduct an estimation.
- Draw the typical graphs of RD for visualizing the effect of treatment at the discontinuity point of the assignment variable.
- Implement the sharp and fuzzy regression discontinuity designs, either for parametric and non-parametric estimators.

2. Preliminaries

In this section, we will be using the dataset `PanelPROGRESA_Enrollment_97_99.dta` that we prepared for the previous chapter. This is a panel dataset for children aged six to sixteen. Our panel is comprised of a sample of individuals who were followed during three survey rounds for year 1997 (the baseline) and years 1998 and 1999 (the follow-up rounds). To begin our discussion about RD, let's open the dataset:

```
set more off
clear all
```

```
global   path="C:\Users\Stanislao\Dropbox\Teaching\1.   Current\Econometrics\4.
Handbook\2. Data\"

use"$path/PanelPROGRESA_Enrollment_97_99.dta", clear
```

Recall that the original dataset does not contain the information of treatment assignment for the baseline year. We replicate here the steps implemented in this previous chapter.

```
* Creating some auxiliary variables

destring year, replace

gen aux=D if year==1998
egen D_assig=mean(aux),by(villid)

gen aux1=pov_HH if year==1998
egen aux2=mean(aux1),by(hogid)
replace pov_HH=aux2 if year==1997
drop aux1 aux2
```

We refer the reader to the previous chapter for an explanation of these variables. We now inspect how the dataset is organized in the baseline year according to eligibility and treatment status:

```
* Distribution of the sample: eligibility and treatment status
tab D_assig pov_HH if year==1997
```

The result is the following:

```
. tab D_assig pov_HH if year==1997

           |        pov_HH
   D_assig |  Non poor       poor |     Total
-----------+----------------------+----------
         0 |     1,884      2,887 |     4,771
         1 |     2,935      4,905 |     7,840
-----------+----------------------+----------
     Total |     4,819      7,792 |    12,611
```

In the dataset, `D_assig` refers to the treatment status (which is given at a village level) and `pov_HH` refers to eligibility status (which is defined at household level). Because of randomization, we already know that the comparison of the outcome variable between treated and control groups gives the impact estimate (we should take care with noncompliance and spillover problems!). That is what we did in our former chapters. Now, using this dataset, we want to illustrate the use of RD.

For RD to provide a consistent estimate of the treatment effect, the treatment must be assigned following a rule that depends on an assignment or forcing variable. For example, scholarships (the treatment) could be assigned to those students that have a GPA (the assignment variable) above some score (the rule). Similarly, some kind of financial credit (the treatment) could be given to those small enterprises that have asset values (assignment variable) below some amount of money (rule). PROGRESA has some rules that can be applied since eligible households inside a beneficiary village were

selected into the program (the treatment) on a basis of a composed index of poverty (the assignment variable). Thus, in every region, those households whose index of poverty score `yycali` was below some cutoff (the specific cutoff varied by region) were selected into treatment and those above, were not (the rule). Given this rule, it is expected that the comparison of those treated and untreated individuals close to the threshold of the assignment variable is a feasible way to obtain an unbiased estimate of the true impact of the intervention.

However, in this comparison we must be careful with the selection of a valid control group given the fact that PROGRESA has externalities that affect non-beneficiaries of treated villages (as it was seen in the last chapter). This means that we cannot directly compare treated and untreated individuals from the same villages because, given the spillover effects, the impact size will be under estimated. It is more reasonable to compare treated individuals from treated villages with untreated individuals from untreated villages. That way our sample will be composed both by non-poor children from control villages and poor children from treated villages. In the dataset that we will use during this chapter, we can select this subsample with the following command line:

```
gen sampleRD = (pov_HH==1& D_assig==1)|(pov_HH==0 & D_assig==0)
```

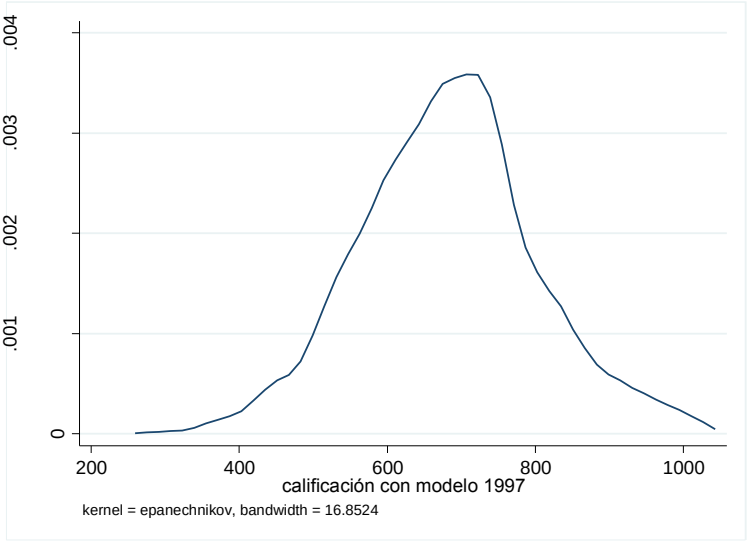
From now on, all the estimations will be conditioned to be part of the valid RD subsample.

Now that we know what our usable subsample is, let's try to analyze the distribution of the assignment variable by using kernel density estimates. The code is the following:

```
* Inspecting the assignment variable (yycali)

kdensity yycali if sampleRD==1 & year==1997, graphregion(fcolor(white))
      title("")
```

We use in Figure 1 the `kdensity` command to draw a kernel density for the poverty index variable `yycali`. We see that the variable closely follows a normal distribution wit mean close to 700.

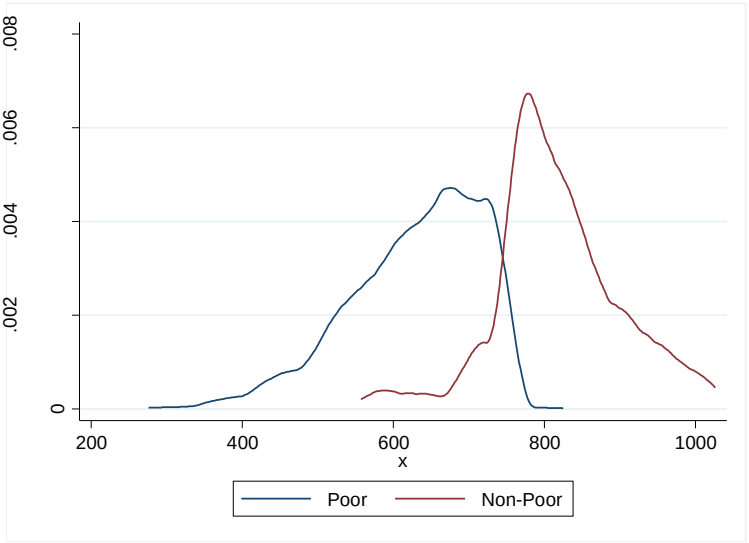
Figure 1: Kernel density for assignment variable



We can also draw this distribution for treatment and comparison groups. The code is the following:

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

Figure 2: Kernel density for assignment variable for treatment and comparison groups



The graphs show interesting features about the distribution of `yycali`. It has, apparently, an almost normal distribution with a modal and mean around 750 considering join data from treated and control groups. However, when we split the distributions for each group, it's clear that treated (or poor) households have values of `yycali` under 800, while non-treated ones have values higher than 600.

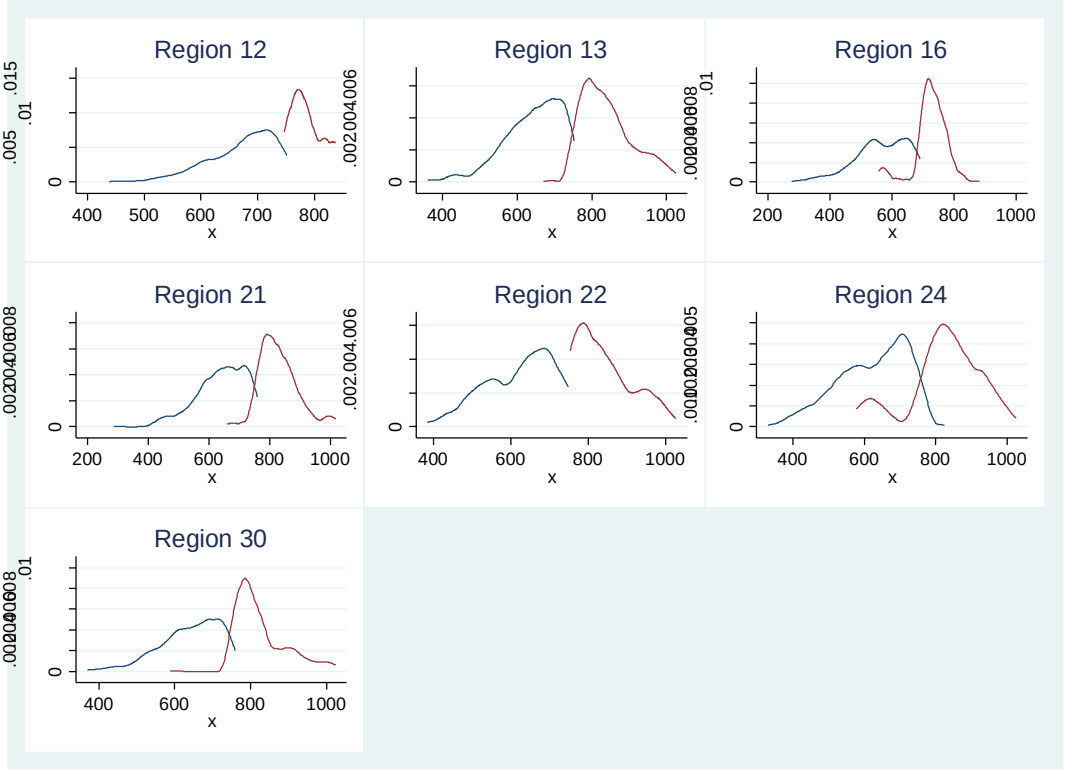
One reason why there appears to be a considerable range of overlap values between treated and untreated groups for this variable correspond to the fact that this threshold used by the Mexican government to discriminate between treated and untreated groups varied among regions. Specifically, the localities for which data were collected were grouped into seven broad geographical regions. For each region a separate discriminant analysis was performed to calculate `yycali` (the poverty threshold) which resulted in a situation where different regions have different threshold scores to determine whether a household is eligible. To observe this, the following command lines use a loop to compute the maximum values of `yycali` among beneficiary households in every region (`entidad`). As mentioned, there are seven thresholds for the seven regions. The code is below:

```
* Disentangle thresholds among entities

levelsof entidad, local(entidades)
foreach j of local entidades {
      summ yycali if year==1997 & pov_HH==1 & entidad==`j'
      scalar max_`j'=r(max)
```

```
      twoway (kdensity yycali if D_assig==1 & sampleRD==1 & year==1997 &
entidad==`j') (kdensity yycali if D_assig==0 & sampleRD==1 & year==1997 &
entidad==`j'), graphregion(fcolor(white)) ti(Region `j') ytitle("")
legend(off) saving("E:\Consultancy\IADB Course\1. Chapters\Chapter 5-
RDD\graph`j'.gph", replace)

}
scalar list
```

The graphs are presented below:

Figure 3: Kernel density for assignment variable for treatment and comparison groups for region



After obtaining the threshold in every region, we draw the empirical density estimate for treated and control in every region graphed each one individually in Figure 3. We see that, despite the fact that there are different cut off points among regions, the graph shows that there exist some regions where some untreated households have values of `yycali` less than their corresponding thresholds. This is particularly clear in Region 13, Region 16, Region 21, Region 24 and Region 30. In these regions, the treatment rules were not perfectly complied.

3. Setting the data for RD

To set the data for RD analysis in the context of PROGRESA, we have to make some adjustments. First, we need to express the assignment rule in terms of just a single cut-off point. Recall that each region has its own cut-off point, so we need to normalize the data in order to have a single cut-off for the whole sample. Conventionally, this is done by re-centering `yycali` around zero. In the more

common cases this implies only to subtract the value of the threshold from the assignment variable. In our case, since the rule has different thresholds depending on the region where the household is located, we have to do this by re-centering within each region. This is done using the following STATA code:

```
* Centering assignment variable

   gen double z = 0
   foreach j of local entidades {
      replace z = yycali - max_`j' if entidad==`j' & D_assig==1
      replace z = yycali - max_`j' if entidad==`j' & D_assig==0
   }
```

For each region, we normalize the assignment variable (`yycali`) using the maximum value for this variable in each region that were computed above. These values were computed with the following part of the code above:

```
summ yycali if year==1997 & pov_HH==1 & entidad==`j'

scalar max_`j'=r(max)
```

To see the results of the exercise we use the option `scalar list` is the following:

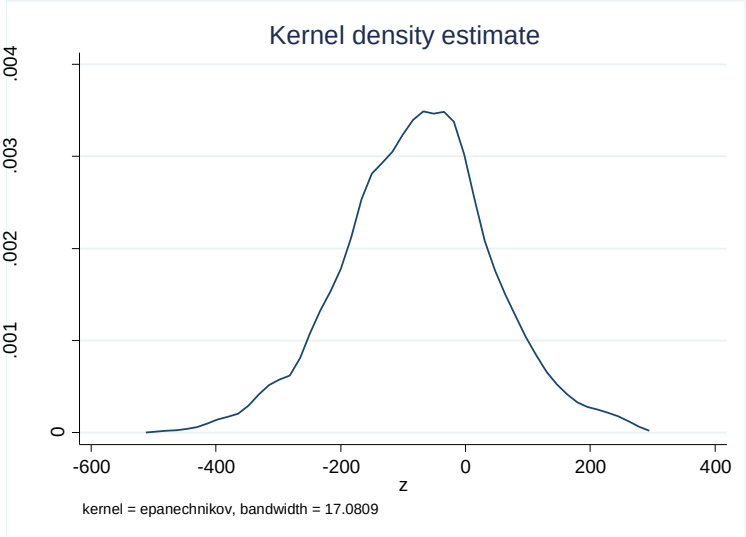
```
. scalar list
    max_30 =  759.35999
    max_24 =         825
    max_22 =         748
    max_21 =         759
    max_16 =         691
    max_13 =         753
    max_12 =         751
```

We have created a re-centered assignment variable called `z`. Its empirical distribution is draw using the following code:

```
kdensity z if sampleRD==1 & year==1997
```

The graph is the following:

Figure 4: Kernel density for the normalized assignment variable z



As we can see, this is quite similar to Figure 1 of `yycali` but translated to the left.

The next step consists in evaluating whether there is full compliance with treatment assignment. If the assignment rule were fully complied (all of those assigned to treatment or control status comply with their assignment), then a **sharp design** should be used. However, if there are a large proportion of non-compliers, then the **fuzzy design** is preferable. In order to analyze how large the non-complier group is, let's define a variable `out` as equal to 1 for those individuals that comply with the rule and zero if they have not. The code is the following:

```
* Settting Sample Range
keep if sampleRD==1

gen out = ~((z<=0 & pov_HH==1) | (z>0 & pov_HH==0))
```

The next box shows how large this group is:

```
. tab entidad out if year==1997

 clave del |          out
    estado |         0          1 |     Total
-----------+----------------------+----------
        12 |       449          2 |       451
        13 |     1,297          2 |     1,299
        16 |       909         24 |       933
        21 |     1,023         21 |     1,044
        22 |       417          0 |       417
        24 |       886        172 |     1,058
        30 |     1,553         34 |     1,587
-----------+----------------------+----------
     Total |     6,534        255 |     6,789
```

We find that 255 of our sample of 6,784 –about 4%-- fall into this group. Since the proportion of untreated individuals with values of `z` lower than zero (or `out=0`) is not too large, we could get rid of them. This is actually what Buddelmeyer and Skoufias (2004) do when analyzed the effectiveness of RD to replicates experimental estimates. However, just as an example, we will implement the fuzzy design as well, taking advantage of those individuals that don't comply with the rule. To make this feasible, we need to create a variable that represents the rule, i.e. a variable `E` that takes the value of one if the assignment variable is lower than zero:

```
gen E = z<=0

label var E "1 = z<0"
```

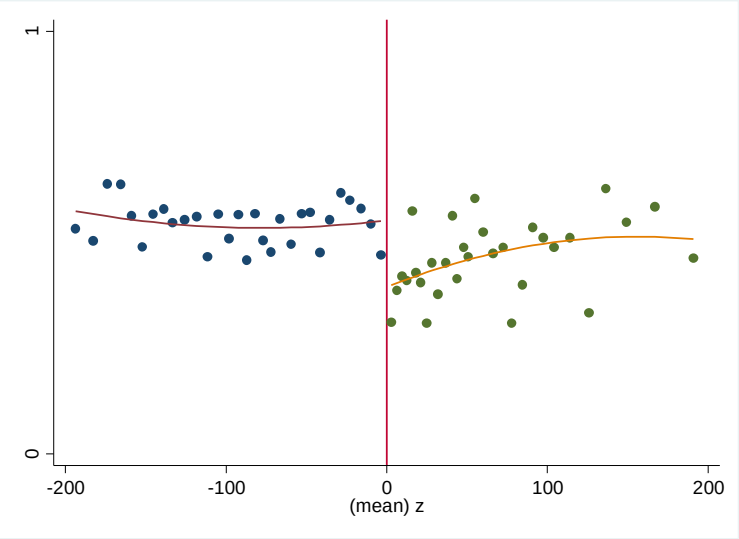
4. Graphical Analysis for RD

One advantage of RD over other techniques is the credibility of its graphical analysis. Particularly, we are interested in evaluating the validity of the assumptions behind the RD design. For a treatment effect exists, it should be the case that there is a jump in the outcome variable at the discontinuity point (the cutoff point between poor and non-poor). To create such graph, we select only observations whose value of the assignment variable, `z`, is between -200 and 200. Although this is not necessary because comparisons are done only around zero, it help us to facilitate the display of observations close enough to be similar. Then, we restrict the sample to observations from year 1999 to include only the individuals that would fit in a sharp design. Within this subsample, we calculated 60 bins of equal size (30 in both sides around the discontinuity point) of z and took means for the variable `enroll` inside every bin. Then, we plot the mean of every bin of z against the mean of enroll. The code is below:

```
preserve
keep if z>=-200 & z<=200
keep if out==0
keep if year==1999
xtile hl = z if D_assig==1, n(30)
xtile hu = z if D_assig==0, n(30)
gen hd = -hl if D_assig==1
replace hd = hu if D_assig==0

collapse (mean) z enroll D_assig, by(hd)
gen z2 = z^2

reg enroll z z2  if D_assig==1
predict yhat0 if e(sample)

reg enroll z z2  if D_assig==0
predict yhat1 if e(sample)

sort z

tw  (scatter enroll z if D_assig==1) (line yhat0 z if D_assig==1) || ///
      (scatter enroll z if D_assig==0) (line yhat1 z if D_assig==0), ///
      ylabel(0 1) xline(0) legend(off) graphregion(fcolor(white))
restore
```

The graph is the following:

Figure 5: Discontinuous impact around the cut-off



The resulting graph shows a clear discontinuity change of enrollment around zero, the re-centered value of the cutoff. If PROGRESA has an impact on the enrollment rate of the beneficiaries, this jump would be clearly seen at the discontinuity point of z as it is the case in our example. Later we will test this hypothesis more formally in the context of regression, but the graph is illustrative enough to support the idea that enrollment discontinuously change around the (normalized) cut-off. The term "impact around the discontinuity point" is not rhetoric. It explains one key characteristic of the RD design: RD captures only local impacts on those individuals that are close to z=0. In other words those just above and just below the cutoff point.

5. Sharp RD

Let now estimate the impact of PROGRESA on enrollment more formally. The general formula of the Wald RD estimator is the following:

$$\tau_{RD} = \frac{\lim_{z \to c^+} E(y_i|z_i) - \lim_{z \to c^-} E(y_i|z_i)}{\lim_{z \to c^+} E(D_i|z_i) - \lim_{z \to c^-} E(D_i|z_i)} \qquad (1)$$

Where y_i is the outcome variable (enrollment rate), D_i the treatment variable, z_i the assignment variable and c the cutoff point. When the treatment status changes deterministically (the rule is complied sharply), then the denominator of the formula is one and the RD design is called **sharp RD**. Conversely, when the treatment status changes only probabilistically with the rule (there are some non-compliers), the denominator of the formula is different from 1 (and zero) and the RD design is called **fuzzy RD**.

In the present case, even though there are some observations that don't comply with the rule within each region, the proportion is small enough that it is feasible to use a sharp design. Given this design, the parametric estimator of (1) is the following:

$$y_i = \alpha + \beta D_i + f(z_i) + u_i$$

Here $f(z_i)$ is a function of the re-centered assignment variable that controls for differences between treated and control individuals outside the discontinuity point, u_i is the error term and α, β are parameters. This last parameter β is the one of interest because it recovers the impact of PROGRESA under this RD design.

It is important to model correctly $f(z_i)$ to guarantee the consistency of β, but this function is not observed. However, the former graph seems to suggest that this function is almost linear. Consequently, we will run a base regression with this specification. Additionally, to analyze the sensitivity of the estimates to changes in this function, we will run a second regression with a quadratic form of $f(z_i)$. The STATA code is presented below:

```
gen z2 = z^2

* SHARP RD
*---------

* 1997
reg enroll D_assig z if out==0 & year==1997, vce(cluster villid)
estimates store r1_97

reg enroll D_assig z z2 if out==0 & year==1997, vce(cluster villid)
estimates store r2_97

* 1999
reg enroll D_assig z if out==0 & year==1999, vce(cluster villid)
estimates store r3_99

reg enroll D_assig z z2 if out==0 & year==1999, vce(cluster villid)
estimates store r4_99

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

After creating the second order of z, we have run four regressions by OLS, the first two for year 1997 (prior to PROGRESA) and the second two for year 1999 (two year after the beginning of PROGRESA). In all the regressions we use clustered standard errors and we then stack all the regressions in one table using the user-written command `xml_tab`.

The results of the four estimations are shown in the table I.

Table I: Sharp RD for Enrollment

	r1_97	r2_97	r3_99	r4_99
	coef/se	coef/se	coef/se	coef/se
D_assig	0.030	0.048	0.100***	0.117***
	(0.033)	(0.035)	(0.038)	(0.040)

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	0.639***	0.625***	0.458***	0.444***
	(0.023)	(0.025)	(0.026)	(0.029)
Number of observations	6,490	6,490	5,262	5,262
R2	0.000	0.001	0.005	0.005
note: .01 - ***; .05 - **; .1 - *;				

Clearly, the results are quite reasonable: in the baseline period (year 1997), the difference between treated and untreated groups was not statistically significant (columns 1 and 2), either under a linear or quadratic specification of the function $f(z_i)$. After two years of the implementation of PROGRESA, the difference has become positive around 0.10 and significant at 1%. Assuming the empirical model is correctly specified, this parameter reflects the impact of the treatment.

6. Fuzzy RD

In the case of imperfect compliance, we can implement a fuzzy RD design. As it was done for the case of the previous chapter, we can follow an instrumental variable approach (IV). We can then estimate the Wald ratio (1) by using the rule as an instrument of the treatment status around the discontinuity point. The parametric procedure is similar to IV and is implemented in two states:

First Stage: $D_i = \alpha_0 + \alpha_1 E + f(z_i) + e_i$
Second Stage: $y_i = \alpha_0 + \alpha_1 \widehat{D}_i + f(z_i) + u_i$

The STATA command for this estimation, ivregress, is the same that was used for IV in the former chapter. The next box shows the STATA commands for the estimation using data from 1997 and 1999, using both linear and quadratic functions of $f(z_i)$:

```
* FUZZY RD
*---------

* 1997
ivregress 2sls enroll (D_assig=E) z if year==1997, vce(cluster villid)
estimates store r1_97

ivregress 2sls enroll (D_assig=E) z z2 if year==1997, vce(cluster villid)
estimates store r2_97

* 1999
ivregress 2sls enroll (D_assig=E) z if year==1999, vce(cluster villid)
estimates store r3_99

ivregress 2sls enroll (D_assig=E) z z2 if year==1999, vce(cluster villid)
estimates store r4_99

xml_tab r1_97 r2_97 r3_99 r4_99, replace save("RD_TableII.xml") ///
      title("Table II: Fuzzy RD for Enrollment") below stats(N r2)
```

Results are reported in the following table:

Table II: Fuzzy RD for Enrollment

	r1_97	r2_97	r3_99	r4_99
	coef/se	coef/se	coef/se	coef/se
D_assig	0.053	0.076**	0.114***	0.136***
	(0.035)	(0.037)	(0.040)	(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	0.633***	0.616***	0.453***	0.437***
	(0.023)	(0.025)	(0.026)	(0.029)
Number of observations	6,743	6,743	5,476	5,476
R2	.	.	0.005	0.005

note: .01 - ***; .05 - **; .1 - *;

Table 2 results again show that the difference in year 1997 was not statistically significant (at least under the linear specification), but it was significant for 1999. Again, assuming the model is correctly specified, this parameter captures the causal effect of PROGRESA on the enrollment rate among compliers located close to the discontinuity point.

7. Specification and Robustness Checks

7.1. Sensitivity to functional form assumptions

An essential exercise when using RD for impact evaluation (and with every econometric exercise) is to verify the sensitivity of the estimations to variations in the basic specification. As we said above, even when the graphical analysis can shed some light about the form of $f(z_i)$, we can not be completely sure. It is necessary to test if results are sensitive to the inclusion of higher order polynomials. This is done in the following box for year 1999 and only for sharp design (as an exercise, you should consider replicating the estimations for the fuzzy design). The code is the following:

```
* HIGHER ORDER POLINOMIALS
gen z3 = z^3
gen z4 = z^4
gen z5 = z^5

reg enroll D_assig z z2 z3 if out==0 & year==1999, vce(cluster villid)
estimates store r1

reg enroll D_assig z z2 z3 z4 if out==0 & year==1999, vce(cluster villid)
estimates store r2

reg enroll D_assig z z2 z3 z4 z5 if out==0 & year==1999, vce(cluster villid)
estimates store r3

xml_tab r1 r2 r3, replace save("RD_TableIII.xml") ///
```

```
      title("Table III: Sensitivity of functional form assumptions") below
stats(N r2)
```

The resulting table is below:

Table III: Sensitivity of functional form assumptions

	r1	r2	r3
	coef/se	coef/se	coef/se
D_assig	0.105**	0.121***	0.125***
	(0.042)	(0.046)	(0.045)
z	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
z2	0.000	0.000	0.000***
	(0.000)	(0.000)	(0.000)
z3	0.000	-0.000	0.000
	(0.000)	(0.000)	(0.000)
z4		-0.000	-0.000***
		(0.000)	(0.000)
z5			-0.000**
			(0.000)
_cons	0.444***	0.435***	0.419***
	(0.029)	(0.032)	(0.032)
Number of observations	5,262	5,262	5,262
R2	0.005	0.005	0.007

note: .01 - ***; .05 - **; .1 - *;

As table 3 shows, the estimated parameter of D changes a little bit, but not in a significant manner. In the basic specification (column 3 and 4 of table I), this parameter was 0.10 and 0.117 for linear and quadratic specification, respectively, and now it has a size from 0.105 and 0.125. Thus, the impact estimated is quite robust to the form of $f(z_i)$. This is good news since it means that our results are not driven by miss-specification problems.

7.2. Adding control variables

A second useful robustness exercise is to try to analyze how the parameter changes when we add control variables. It may be the case that our results are driven by children, household, or region level omitted variables. To test this possibility, we run three more regressions. The STATA code is below:

```
* ADDING COVARIATES
reg enroll D_assig z z2 age sex lang if out==0 & year==1999, vce(cluster
villid)
estimates store r1

reg enroll D_assig z z2 age sex lang ageHH sexHH eduHH if out==0 &
year==1999, vce(cluster villid)
estimates store r2

reg enroll D_assig z z2 age sex lang ageHH sexHH eduHH i.entidad if out==0 &
year==1999, vce(cluster villid)
```

```
estimates store r3

xml_tab r1 r2 r3, replace save("RD_TableIV.xml") ///
      title("Table IV: Sensitivity to additional covariates") below stats(N
r2)
```

The results are presented in the next (edited) table:

Table IV: Sensitivity to additional covariates

	r1 coef/se	r2 coef/se	r3 coef/se
D_assig	0.087**	0.083**	0.091***
	(0.036)	(0.035)	(0.034)
z	0.000**	0.000*	0.000**
	(0.000)	(0.000)	(0.000)
z2	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)
age	-0.130***	-0.127***	-0.127***
	(0.004)	(0.004)	(0.004)
Gender: 1= Male, 0=Female	0.013	0.013	0.015
	(0.013)	(0.014)	(0.013)
Language: 1= Indigenous, 0=Non Indigenous	0.143***	0.134***	0.099***
	(0.028)	(0.026)	(0.026)
age of HH		0.001*	0.001
		(0.001)	(0.001)
sex of HH		-0.017	-0.022
		(0.029)	(0.029)
years of education of HH		0.000***	0.000***
		(0.000)	(0.000)
Entidad fixed effects	No	No	Yes
_cons	2.410***	2.245***	2.265***
	(0.069)	(0.087)	(0.099)
Number of observations	5,249	5,195	5,195
R2	0.194	0.206	0.219

note: .01 - ***; .05 - **; .1 - *;

First, we have add some variables defined at the child level, namely age, sex and a dummy for indigenous people. Results of the estimation after controlling for these variables are presented in the table 4, column 1. Compared to the base estimation (column 4 of table 1), the estimated parameter now is a little smaller (0.087), but still significant. We then add household variables such as the age, sex and education level of the household head. Result shows that the parameter is practically the same and, again, statistically significant (column 2).

Even when the impact parameter does not change when we control for personal or household characteristics, it's possible that changes could be driven entirely by differences among regions. We

therefore must control for fixed effects at region levels. The results are shown in column 3. Once more the parameter changes just a little bit, but now it is even more significant. Again, these results suggest that specification problems are not the ones that explain the size and direction of the impact estimated.

7.3. Evaluating the discontinuity: placebo discontinuity points

A third exercise to evaluate the validity of the estimated parameters is to determine whether results change if we move the discontinuity point at which we evaluate the impact. If the impact of the treatment is truly occurring around the discontinuity point, creating placebo discontinuity points and then evaluating their impact should not give significant results. We do this in the following box for a three order polynomial. The STATA code is below:

```
* PLACEBO EXPERIMENTS
* Testing different cut offs on the left
foreach k of numlist 30 60 90 {
      gen zf = z + `k'
      gen zf_2 = zf^2
      gen zf_3 = zf^3
      gen Ef = zf < 0
      reg enroll Ef zf zf_2 zf_3 if out==0 & year==1999, vce(cluster villid)
      estimates store SHL`k'
      drop Ef zf zf_2 zf_3
}

* Testing different cut offs on the right
foreach k of numlist 30 60 90 {
      gen zf = z - `k'
      gen zf_2 = zf^2
      gen zf_3 = zf^3
      gen Ef = zf < 0
      reg enroll Ef zf zf_2 zf_3 if out==0 & year==1999, vce(cluster villid)
      estimates store SHR`k'
      drop Ef zf zf_2 zf_3
}

xml_tab SHL30 SHL60 SHL90 SHR30 SHR60 SHR90, replace save("RD_TableV.xml")
///
      title("Table V: Placebo discontinuity points") below stats(N r2)
```

The results of the experiment are reported in the next table:

Table V: Placebo discontinuity points

	SHL30 coef/se	SHL60 coef/se	SHL90 coef/se	SHR30 coef/se	SHR60 coef/se	SHR90 coef/se
Ef	0.014	-0.013	0.011	0.002	-0.009	-0.074
	(0.032)	(0.031)	(0.030)	(0.039)	(0.040)	(0.049)
zf	-0.000	-0.000*	-0.000	-0.000	-0.000	-0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
zf_2	0.000	-0.000	-0.000	0.000	0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
zf_3	0.000*	0.000*	0.000	0.000**	0.000**	0.000

	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
_cons	0.505***	0.529***	0.527***	0.492***	0.492***	0.531***
	(0.022)	(0.022)	(0.022)	(0.031)	(0.031)	(0.036)
Number of observations	5,262	5,262	5,262	5,262	5,262	5,262
R2	0.003	0.003	0.003	0.003	0.003	0.003
note: .01 - ***; .05 - **; .1 - *;						

First, taking advantage of a loop, we move the re-centered discontinuity point in 30, 60 and 90 point increments to the left, generate a new treatment variable `Tf` and estimate the impact of the intervention at each new point. Then, using a similar loop, we move it to the right in the same magnitudes, generate the same variable and run the corresponding regressions. Results are stacked in table V.

According to our results, none of these false experiments gives significant estimates. This is a good result, because it not only says that the parameter estimated at the real cutoff is the impact of PROGRESA, but also reflects that the outcome variable is continuous outside zero.

8. Testing the Validity of RD

RD design requires assumptions that must be validated:

Assumption 1: $E[y_{0i}|z_i]$ and $E[y_{1i}|z_i]$ must be continuous at $z_i = 0$, the discontinuity point.

However, since we do not observe the potential outcome of treatment and control groups at $z_i = 0$, we must try to test this assumption indirectly. Given the fact that we have baseline data, a first way to do this is to test if there is any difference in the outcome variable. We did so in table I and found that in 1997 the estimated impact parameter was statistically indistinguishable from zero under either linear or quadratic specification of $f(z_i)$.

8.1. Pre-treatment balance in covariates

If we do not have baseline data, a useful exercise would be to verify whether some covariates that explain the outcome variable but are not influenced by the treatment are continuous at the discontinuity point. If assumption 1 holds, then these covariates should be continuous at the discontinuity point. We do so in the next box:

```
* BALANCING TEST (JUST BASELINE COVARIATES AND A FUZZY RDD)
replace edu=edu/100
replace eduHH=eduHH/100
foreach var of varlist sex lang edu ageHH sexHH eduHH {
      reg `var' D_assig z z2 z3 if out==0 & year==1997, vce(cluster villid)
      estimates store BT`var'
}
xml_tab BTsex BTlang BTedu BTageHH BTsexHH BTeduHH, replace
save("RD_TableVI.xml") ///
      title("Table VI: Pre-treatment balance") below stats(N r2)
```

The results are the following:

Table VI: Pre-treatment balance

	BTsex	BTlang	BTedu	BTageHH	BTsexHH	BTeduHH
	coef/se	coef/se	coef/se	coef/se	coef/se	coef/se
D_assig	0.061**	0.038	0.068	-1.546**	-0.001	-0.040
	(0.025)	(0.061)	(0.051)	(0.787)	(0.017)	(0.191)
z	0.000	-0.001***	0.001***	0.018***	-0.000	0.000
	(0.000)	(0.000)	(0.000)	(0.004)	(0.000)	(0.001)
z2	-0.000**	-0.000	-0.000	-0.000***	0.000**	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
z3	-0.000	0.000**	-0.000***	-0.000***	0.000	0.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
_cons	0.485***	0.218***	6.080***	48.518***	0.913***	2.571***
	(0.017)	(0.039)	(0.035)	(0.546)	(0.011)	(0.124)
Number of observations	6,532	6,518	6,534	6,526	6,532	6,474
R2	0.003	0.029	0.011	0.044	0.002	0.004

note: .01 - ***; .05 - **; .1 - *;

It seems that speaking an indigenous language, education and the sex and education years of the household head are similar among treated and control groups around the discontinuity point. However, there exists some discontinuity in the sex of the children and the age of the household head. This is evidence of unbalance in some characteristics since it could mean that, at $z = 0$, we are comparing individuals that differ in some variables that could partially explain the outcome. However, given the fact that we have just tested the discontinuity of six variables, the differences could be driven by sample variability, so these results cannot be invalidated immediately. In any case, we can control for these differences by incorporating them into the econometric specification. Since we have already discussed this issue above, we don't further discuss it here.

8.2. McCrary test

A more technical way to test assumption 1 is through the **McCrary test**. According to McCrary (2008), when the assignment rule is public knowledge, individuals could manipulate certain characteristics, so they meet the rule criteria in order to obtain the benefits. If this were a systematic behavior, potential outcomes would differ between treatment and control and our assumptions could be invalidated. The test proposed by McCrary verifies if the density of the assignment variable changes discontinuously at $z = 0$. If individuals were strategically adjusting their behavior or characteristics to meet the rule, then we would find people concentrated just to the left of z=0 (i.e. an adjustment done just enough to be eligible).To implement this test in STATA, we can use the ado file `DCdensity`, available on Justin McCrary's website[1]. Download the file and save it wherever you want. To run the program you only have to call it before writing the syntax of the command. This is shown in the next box, where the ado file is saved in the folder "E:\Consultancy\IADB Course\1. Chapters\Chapter 5-RDD".

```
* MCCRARY DENSITY TEST
```

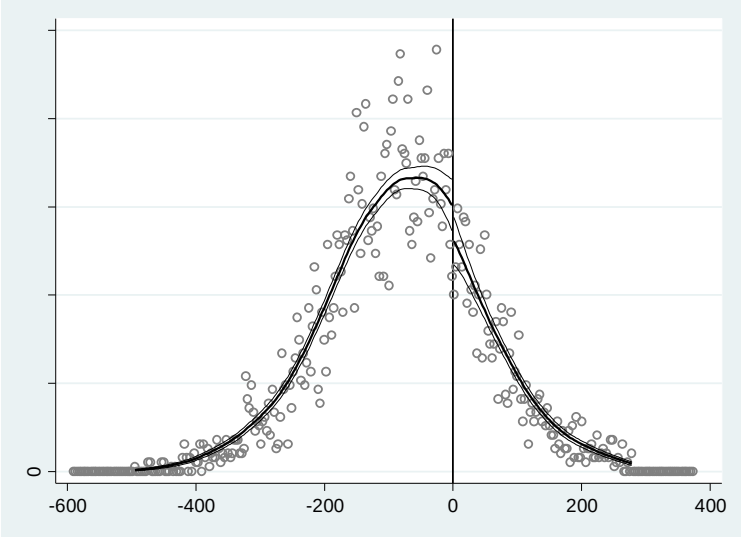
[1] http://emlab.berkeley.edu/~jmccrary/

```
DCdensity z if out==0 & year==1997, breakpoint(0) generate(Xj Yj r0 fhat
se_fhat)
drop Yj Xj r0 fhat se_fhat
```

The output and resulting graph are below:

```
. * MCCRARY DENSITY TEST
. DCdensity z if out==0 & year==1997, breakpoint(0) generate(Xj Yj r0 fhat
se_fhat)
Using default bin size calculation, bin size = 2.97428769
Using default bandwidth calculation, bandwidth = 94.0358965

Discontinuity estimate (log difference in height): -.132753566
                                                   (.074569725)
Performing LLR smoothing.
261 iterations will be performed
.........................
```

Figure 6: McCrary density test



The exercise was done for a sharp design and only for the year 2007. According to our results, there is a small discontinuity in the density of z around 0 which is statistically significant. If the cut-off threshold in every region was known publicly prior to the beginning of the baseline period, then this evidence could mean that some people lied about their characteristics to become eligible beneficiaries. Again, even with this evidence we cannot invalidate the results we have obtained so far. It merely makes us aware of the possible existence of some non-observed variables that could partially influence the size of the impact.

9. Non-parametric RD

9.1. Basics

So far, the exercises were done following a parametric framework for the estimation of the equation (1). However, the non-parametric estimation has limitations because its convergence does not rely on a specification assumption. In STATA, there are many user-written commands to implement a nonparametric estimation of the RD design. Here we will use the `rd` command written by Austin Nichols. This is a flexible command that uses one order local linear regressions to estimate the limits of formula (1) with different options of parameters and kernel functions (see the help file of this command for more information).

The next box shows how to implement a RD estimation for a sharp non-parametric design. As we see, after typing `rd` we have to tell STATA what the outcome and the assignment variables are, followed by the usual restrictions. In this case, we are running the estimations only for the sharp sample (out=0) and year 2009. The options needed are the threshold or discontinuity point (z0) and the bandwidth for the nonparametric estimations (`bwidth`). Also it is necessary to define the `mbw`, which allows us to specify many bandwidths in percentage terms with regard to `bwidth` (when we give a value of 100 to this parameter, we are asking STATA to consider just one bandwidth with the value we gave). The STATA code is the following:

```
* SHARP: Triangle Kernel and bandwidth=3

rd enroll z if out==0 & year==1999, z0(0) bwidth(3) mbw(100)
```

The output is the following:

```
. * SHARP: Triangle Kernel and bandwidth=3
. rd enroll z if out==0 & year==1999, z0(0) bwidth(3) mbw(100)
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is z
 Treatment variable X_T unspecified
 Outcome variable y is enroll_child

(13928 missing values generated)
(13928 missing values generated)
(13928 missing values generated)
Estimating for bandwidth 3
------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwald |    .207781   .3120609     0.67   0.506    -.4038472    .8194092
------------------------------------------------------------------------------
```

According to our results, using a sharp design with triangle kernel function and a bandwidth of 3, the impact estimated is around 0.2, but it is not statistically different from zero. Curiously, this parameter is different from the one in table I. Probably, given the huge data requirement, the non-significance is driven mainly by the low power of the estimator.

The fuzzy design can also be implemented in a non-parametric fashion. Here, before defining the assignment variable, we have to tell STATA which is the treatment variable. The STATA code is the following:

```
* FUZZY: Triangle Kernel and bandwidth=3

rd enroll D_assig z if year==1999, z0(0) bwidth(3) mbw(100)
```

The output is the following:

```
. * FUZZY: Triangle Kernel and bandwidth=3
. rd enroll D_assig z if year==1999, z0(0) bwidth(3) mbw(100)
Three variables specified; jump in treatment
at Z=0 will be estimated. Local Wald Estimate
is the ratio of jump in outcome to jump in treatment.

 Assignment variable Z is z
 Treatment variable X_T is D_assig
 Outcome variable y is enroll_child

(13714 missing values generated)
(13714 missing values generated)
(13714 missing values generated)
Estimating for bandwidth 3
------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       numer |   .0788286   .2381351     0.33   0.741    -.3879076    .5455648
       denom |  -.1868498   .2240979    -0.83   0.404    -.6260736    .2523741
       lwald |  -.4218823   1.162068    -0.36   0.717    -2.699495     1.85573
------------------------------------------------------------------------------
```

The estimated parameter in this case is around -0.41, again not different from zero. Again, this could reflect the low power of the estimator in our sample and not the real absence of impact of PROGRESA.

9.2. Bandwidth choice

The selection of the bandwidth is the most important part of the exercise when using a non-parametric estimation. As a general rule, the higher the bandwidth, the bigger the power of the tests; However, bigger power can also lead to larger biases in the estimations. Contrarily, a smaller bandwidth leads to smaller biases in the estimation, but also smaller power. Therefore, given these tradeoffs, it is important to try to select an "optimal" bandwidth. Imbens and Kalyanaraman's (IK) optimal bandwidth is one useful option (see Imbens and Kalyaranaman, 2012)

In the `rd` package, the optimal bandwidth is obtained by default. It means, if we do not give STATA any value of the bandwidth, it will automatically compute the IK's optimal bandwidth. In the context of our exercise, this is shown in the following box.

```
. * Imbens & Kalyaraman Optimal bandwith
. rd enroll z if out==0 & year==1999, mbw(100) z0(0)
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is z
 Treatment variable X_T unspecified
```

```
 Outcome variable y is enroll_child

(13928 missing values generated)
(13928 missing values generated)
(13928 missing values generated)
Estimating for bandwidth 3.762761094839582
------------------------------------------------------------------------------
enroll_child |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwald |   .2381414   .2619179     0.91   0.363    -.2752083     .751491
------------------------------------------------------------------------------
```

As it can be seen, the optimal estimated bandwidth is 3.76, a little bigger than the one we have been using above.

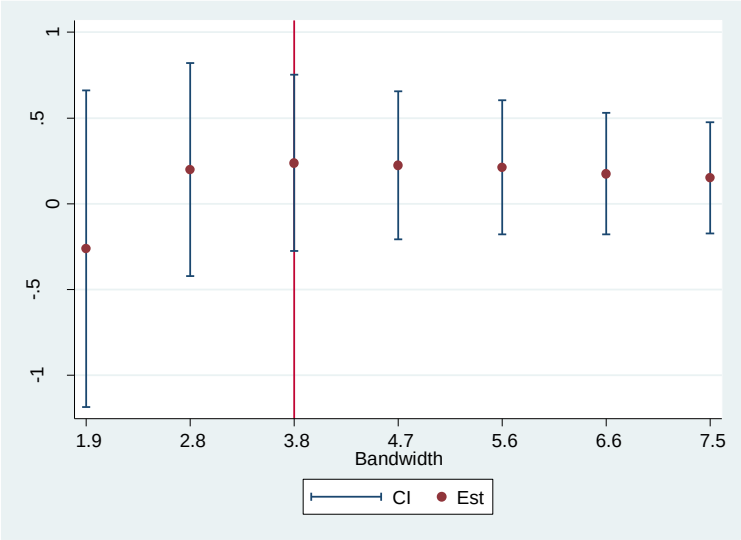
Imbens and Kalyanaraman (2012) offer a solution to the bandwidth selection problem. However, it is useful to verify how sensitive the estimations are to different bandwidth values. The `rd` command allows us to do this and present the results in a nice and very illustrative graph. The next box shows these results. Now, we have permitted the option `mbw` to vary from 50 to 200 with increments of 25. Using this option, we are asking STATA to compute many estimates with bandwidths that go from 50% of the IK's bandwidth to 200%. Then, with the option `bdep` and `ox` we are asking STATA to plot those results in a graph.

```
. * Sensitivity of impact effects to changes in bandwith size (graph)
. rd enroll z if out==0 & year==1999, z0(0) mbw(50(25)200) bdep ox
Two variables specified; treatment is
assumed to jump from zero to one at Z=0.

 Assignment variable Z is z
 Treatment variable X_T unspecified
 Outcome variable y is enroll_child

(13928 missing values generated)
(13928 missing values generated)
(13928 missing values generated)
Estimating for bandwidth 3.762761094839582
Estimating for bandwidth 1.881380547419791
Estimating for bandwidth 2.822070821129687
Estimating for bandwidth 4.703451368549477
Estimating for bandwidth 5.644141642259373
Estimating for bandwidth 6.584831915969269
Estimating for bandwidth 7.525522189679164
  ------------------------------------------------------------------------------
  enroll_child |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
  -------------+----------------------------------------------------------------
         lwald |   .2381414   .2619179     0.91   0.363    -.2752083     .751491
       lwald50 |  -.2624644   .4705027    -0.56   0.577    -1.184633     .659704
       lwald75 |   .1992289   .3170899     0.63   0.530    -.4222559    .8207136
      lwald125 |   .2245478   .2202439     1.02   0.308    -.2071223    .6562179
      lwald150 |   .2119926    .199099     1.06   0.287    -.1782344    .6022195
      lwald175 |   .1752539   .1807862     0.97   0.332    -.1790806    .5295884
      lwald200 |   .1507803    .166045     0.91   0.364    -.1746619    .4762226
  ------------------------------------------------------------------------------
```

Results are quite illustrative. With a bandwidth of 1.9, the impact estimated parameter is negative in 0.26. With one of 2.82, the parameter is 0.19. From this value, all the estimated parameters are stable around 0.15-0.22.

The resulting graph is the following:

Figure 7: Treatment estimates for different bandwidth choices



10. Final Remarks

We have started in this chapter the coverage of non-experimental designs. RD is considered the closest "cousin" of randomized designs. In fact, DiNardo and Lee (2011) suggest that RD is just a form of experimental design and as such RD designs can be analyzed as experiments. Give its proximity to experimental designs, RD is increasingly recognized as one of the most credible research design when a standard experiment is not feasible.

In this chapter, we have covered the basic of RD in terms of implementation. Our treatment of the topic has been biased to parametric versions of RD and some of the most common specification checks used by applied evaluators. Although we also covered the basics of non-parametric RD, the technical details of this version of RD demand a basic command of nonparametric econometrics which is beyond the goal of this chapter. Those interested in getting a depth treatment of this topic can consult the readings for this section.

11. Further Readings

Imbens, Guido and Thomas Lemieux (2008). "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142, 615-635.

Lee, David and Thomas Lemieux (2010). "Regression Discontinuity Design in Economics," *Journal of Economic Literature*, 48(2), 281-355.

Imbens, Guido and Karthik Kalyanaraman (2012). "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79, 933-959.

McCrary, Justin (2008). "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test," *Journal of Econometrics*, 142(2), 698-714.

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.