Bocconi University, Microeconometrics (cd. 20295)
Professor: Thomas Le Barbanchon and T.A.: Erick Baumgartner
Problem Set 3: Regression Discontinuity Design

Due: May 13^{th} by 11:59pm. Submit by email to `erick.baumgartner@unibocconi.it`.

Problem Set 3

This problem set is composed of two exercises, each exercise focusing on a different regression discontinuity design (RDD). In Exercise 1, we follow a standard RDD application, Meyersson (2014), to study the effect that Islamic political representation had on the educational attainment of women in Turkey during the late 1990s. In Exercise 2, we turn to a spatial RDD, Gonzalez (2021), to study the effect of cell phone coverage on electoral frauds.

Commands

Regression discontinuity designs are implementable with packages such as `rdrobust`, `rddensity` and `lpdensity`, among others, in both `R` and `Stata`. You should install these before proceeding.

Instructions

No need to produce a pdf file with your answers. Save all graphics and tables requested (e.g., as `pset_3_exercise_1_question_1_a.xlsx`) and a do-file summarizing all of your work in a zipped folder identifying your group and the problem set (e.g., as `pset_X_group_Y.zip`) to `erick.baumgartner@unibocconi.it`. In the sub-questions where you are asked to write, please add your answer as a comment in your do file.

Hint: Type * to add a comment in your do-file (e.g., `* this is a comment line *`).

Hint: You can choose your preferred way of preparing tables: **(1)** one option is to use the command `outsheet` to construct the tables of summary statistics and the command `outreg2` to construct the regression tables (you can use them to export results of summary statistics and regressions to an excel file); **(2)** another option is to save results using the command `eststo` and then export these directly to a .tex (latex) file using the command `esttab`. If using *R*, one option for exporting results is the `stargazer` package. Read carefully `help` for each command you choose and try different options, so as to have well-formatted tables.

Hint: `rdrobust` reports estimates from different estimation methods: **(1)** `Conventional`, **(2)** `Bias-corrected`, and **(3)** `Robust`. In this problem set, any `rdrobust` output should be reported with `Conventional` betas and standard errors. Nonetheless, note that in your own research it is recommended that you report `Conventional` betas and `Robust` standard errors.

Hint: Unless asked otherwise, use as default options for your `rdrobust` estimates:

```
kernel(triangular) p(1) bwselect(mserd)
```

Hint: Have in mind that some commands have different default procedures in *Stata* and *R*. Since we are not asking you to specify some of these procedures, it is normal that sometimes the results are not exaclty the same between the two languages.

Exercise 1

Short Summary of Discussion

When developing research with regression discontinuity designs, we ought to: (**S1**) perform diagnosis tests to motivate our RD identification assumptions; (**S2**) estimate our RD treatment effects; (**S3**) check whether our results are robust.

We will be asked to replicate these steps for a particular paper - Meyersson (2014).

Data

Meyersson (2014) studies how Islamic political representation affects the level of education attained by women, in Turkey. To do so, the author analyzes data from municipal Turkish elections held in 1994. In terms of each specific variable - Meyersson (2014) studies this effect while controlling for demographic data at the municipality level - this data goes from 1989 to 2000; the outcome of interest instead is referent to 2000 (in particular, the share of women with high school education at a given municipality in the year of 2000).

Questions

Before estimating our treatment effect, we ought to perform a set of diagnosis tests: (**T1**) show that a discontinuity in treatment exists at our cutoff; (**T2**) show that discontinuities in other covariates or pre-determined variables do not exist at our cutoff; (**T3**) test for the null hypothesis that the density of our running variable does not exhibit a discontinuity at the cutoff; (**T4**) show that discontinuities in our running variable and outcomes do not exist away from our cutoff.

In the following questions, we will be asked to execute each of these steps described above.

1. Use `pset_3`.

(a) Generate a RD Plot of `T` - Islamic mayor in 1994 - against `X` - Islamic Vote Margin in 1994 - when the Islamic party wins and lose an election.

Call the y-axis - `Treatment Variable`; call the x-axis - `Running variable`.

Is the current design a sharp or a fuzzy RD? Why?

(b) Create a macro named `covariates` containing the baseline variables: `hischshr1520m i89 vshr_islam1994 partycount lpop1994 merkezi merkezp subbuyuk buyuk`.

Create a table named `Table_1`, summarizing RD estimates for all baseline variables.

`Table_1` should have the following columns: **Label, MSE-Optimal Bandwidth, RD Estimator, p-value, and Effective Number of Observations**.

Hint: The *effective number of observations* column should be understood as the number of observations to the left of the cutoff plus the observations to the right of the cutoff given positive weight in the estimation.

(c) Generate a RD plot for each of the baseline variables on `covariates`.

Use `graph combine` to generate a **unique graphic** containing all 9 RD plots.

Title each RD subplot so that the reader is able to identify each subplot to the corresponding outcome. Save the **unique graphic** as `Graph_1`.

(d) Generate a graphic with histograms for the observations to the left and the observations to the right of our cutoff. Choose contrasting colors for the histograms on each side of our cutoff.

Use `rddensity` to generate a graphic of our running variable `X`'s estimated density.

In both graphics, plot a vertical line to signal our cutoff. Save a graphic named `Graph_2` containing the histogram plot and the estimated density plot **side-by-side**.

(e) Use `rddensity` to test if a discontinuity in our running variable `X`'s density does not exist in our cutoff.

What are we able to conclude from such test?

Is it favorable or against the validity of our RD design?

(f) Test if alternative discontinuities do not exist in the following alternative cutoffs:

-10, -5, 5, 10.

Did we found any evidence in favor of the absence of alternative discontinuities?

After validating our RD design, we can estimate our treatment's effect on our outcomes and check for the robustness of our results. That is what we will do in the following questions.

(g) Generate a RD Plot of `Y` - Share Women aged 15-20 with High School Education - against `X` - Islamic Vote Margin in 1994 - when the Islamic party wins and loose an election.

Use 40 Evenly-Spaced Bins.

Call the y-axis - `Outcome`; call the x-axis - `Running Variable`.

(h) Use `rdrobust` to estimate the effect of `T` - Islamic mayor in 1994 - on `Y` - Share Women aged 15-20 with High School Education.

Use a linear polynomial.

Try both an uniform and triangular kernel.

Does electing a mayor from an Islamic party has a significant effect on the educational attainment of women? Do results differ significantly for different kernel choices?

Use a triangular kernel for these next items.

(i) Estimate the effect of `T` on `Y` but using a **global** approach.

Do not choose any bandwidth.

Use a polynomial of order 4.

Run a regular linear regression instead of `rdrobust`.

(j) Estimate the effect of `T` on `Y` but using a **local** approach by restricting our sample to a window within an optimal bandwidth that we should have obtained with `rdrobust` (`mserd` bandwidth).

Run a regular linear regression.

Use a linear polynomial.

Do we get the exact same result as in item **(h)**? If not, explain why.

Hint: In the `rdrobust` post-estimate, save our optimal bandwidth in a local using:

```
local opt_i = e(h_l)
```

(k) Save item **(h)**'s bandwidth as a scalar named `opt_i`.

Re-estimate item **(h)**'s RD using as alternative bandwidths:

`0.5*opt_i`, `0.75*opt_i`, `1.25*opt_i`, and `1.5*opt_i`.

Plot each five RD point estimates, including that from item **(h)**, with their respective confidence intervals in a graphic named `Graph_3`. What can we say about the robustness of our results with respect to bandwidth choice?

Exercise 2

Short Summary of Discussion

Spatial RDDs are multidimensional RDDs that can be mapped onto a one-dimensional design. This mapping can be done by summarizing geographical coordinates (i.e., latitude and longitude) in a distance metric. Read Gonzalez (2021) to become acquainted with a spatial RD design.

Gonzalez (2021) is a paper that studies the effect of cell phone coverage on electoral frauds. It compares voting centers that are in different sides of cell phone coverage boundaries, in Afghanistan. This comparison is restricted to centers that are geographically close to one another, for causal purposes.

Data

We will partially replicate Gonzalez (2021). We will focus on his one-dimensional results related to how cell phone coverage affects electoral frauds (Table 2, page. 18).

To do so, we will use (a) polling-center specific data from the Afghan Electoral Complaints Commission (ECC) on fraud, (b) geographical coordinates for each polling center (also provenient from the ECC) and (c) geo-data on 2G coverage in Afghanistan (we will not explicitly use this dataset, we will use a metric computed by Gonzalez (2021) that measures the distance between each polling center and the closest point in which there is 2G coverage).

Questions

Assume Gonzalez (2021) did not have the exact longitude of each voting center in his sample, only a proxy. Instead, latitude was correctly measured. Endowed with the latitude and the proxy for longitude of each polling center, Gonzalez (2021) went on and measured the distance between each polling center "location" and the closest point with 2G coverage. In addition, Gonzalez (2021) has a coverage indicator for each polling center that has been collected by ECC officials.

Both variables can be found in `fraud_pcenter_final`. The distance between the polling centers and their closest points with 2G coverage is titled "`_dist`"; the cell phone coverage indicator is titled "`cov`".

(a) Plot the treatment variable used at Gonzalez (2021) as a function of this new running variable. In addition, compute the RD estimate for a regression where you model the same treatment variable as a function of the new running variable.

Is the current design a sharp or a fuzzy RD? Which assumptions must hold in order for the one-dimensional RD estimates of Gonzalez (2021) to be valid?

(b) Point out in which setting does having a proxy for longitude does not require you to change RD design (relative to Gonzalez, 2021). **Hint:** Read the "*Additional Results*" section of Gonzalez (2021) and reflect on which type of cell phone coverage boundary would deliver you this result.

(c) Use `fraud_pcenter_final` to partially replicate Columns 1, 3 and 5 of Table 2 under this new RD setting (present only point estimates). Interpret your new estimates. **Hint**: use `Table_onedim_results.do` and review your RDD slides.

References

Gonzalez, R. M. (2021). Cell Phone Access and Election Fraud: Evidence from a Spatial Regression Discontinuity Design in Afghanistan. *American Economic Journal: Applied Economics*, 13(2):1–51.

Meyersson, E. (2014). Islamic Rule and the Empowerment of the Poor and Pious. *Econometrica*, 82(1):229–269.