FLS 6415: Class 7 Homework

October 5, 2017

Remember to answer all the questions in R markdown and produce a PDF. Email your completed homework (R markdown file and PDF) to jonnyphillips@gmail.com by midnight the night before class. Remember to refer to the example code from this week and the last couple of weeks for coding guidance.

To analyze a regression discontinuity we will conduct the same analysis as Titiunik (2011). However, we will not use her pre-prepared dataset - we will start the analysis by constructing our own. As with all analysis, 90% of the work is in preparing the dataset.

1. Titiunik measures the effect of being an incumbent mayor for the 2000-2004 period on the vote share of the incumbent in 2004. We can download this in a clean simple format from cepesp.io. Choose "Eleicoes por Cargo" and we want prefeito data at the municipal level for *parties* in the 2000 elections first. Then 'Consultar' and 'Adicionar Colunas' to add the COD_MUN_IBGE variable. Then export this dataset to 'CSV'. Then make the same selection for 2004 and download this as a separate CSV. Finally, load the data in R. Details for the description of each variable can be found on cepesp.io (see adicionar colunas).

2. First, prepare the 2000 dataset:
a. Filter only for the first round (primeiro turno), calculate the percentage vote share for each party in each municipal contest and calculate which position the party came in the municipal election (their `rank`).
b. Next, we need to calculate the winning margin of each party as defined on page 9 of Titiunik (2011). One way to do this is to `arrange` the data by municipality (*COD_MUN_IBGE*) and by the *rank* of vote share that you just created (so the winning party is at the top of each municipality in your data.frame). Then calculate new columns for the highest and second-highest vote share in each municipality. *Hint:* Use `nth` within `mutate` to calculate the second-highest vote share.
c. Then create another column for *win_margin* using `ifelse` or `case_when` to calculate the gap between each party and the second-placed party (if they win), or between each party and the first-placed party (if they did not win). Check your *win_margin* values have the correct sign.
d. Finally, add an 'Incumbent' variable to code for treatment, i.e. whether the party actually won the election and an 'Electorate' variable to measure the total number of votes in each municipality (we'll use this later).

3. Now prepare the 2004 data. Filter for the first round (primeiro turno), and calculate our outcome measure: the vote share of each party in each municipal contest.

4. We will first conduct the analysis for the PMDB. Filter the 2000 dataset so it only contains the voting results data for the PMDB. Then each row in the 2000 dataset needs to also have a value for the outcome in 2004 - we need to merge the 2004 dataset into the 2000 data using the keys (merging variables) for municipality and party. How many rows are now in your dataset?

5. If we did not know about regression discontinuity, the observational regression we might run is of the outcome (vote share in 2004) on treatment (becoming an incumbent in 2000). Run and interpret this regression. Give specific examples of reasons why the estimated causal effect may be biased.

6. Before we implement the regression discontinuity methodology, we can check for balance. We don't have many variables in our dataset, but we can at least check if the size of the electorate (a good proxy for population) is the same for municipalities where the PMDB just lost compared to where they just won. First, assess balance for the 'Electorate' variable you created in Q2 within a 5% bandwidth either side of the cutoff of winning margin=0

for treatment and control. Then compare this difference with the balance of 'Electorate' on Incumbency in the full dataset. Interpret the results.

7. Before we implement the regression discontinuity methodology, we can check for sorting of units to either side of the threshold. Implement the McCrary density test and interpret both the graphical and statistical results. *Hint:* Use the `DCdensity` function with option verbose=TRUE.

8. Before we run the analysis, let's construct a regression discontinuity plot to visually inspect the causal effect of incumbency at the threshold. Create two charts, one using linear (`lm`) and one using smooth local regression (`loess`) trend lines. Interpret what each chart tells you about the size and direction of the causal effect. *Hint:* Use two layers of geom_smooth to create separate 'regression' lines for each half of the data (above and below winning margin=0).

9. This graph is 'messy' because there are too many points on it to see clearly the pattern. To create a clearer regression discontinuity chart, first we need to lump (aggregate) observations with similar values of the running variable (win margin) together - so round each value to the nearest 0.05. Next, for each of these rounded win margins, calculate the average value of the outcome (vote share in 2004). Then plot this average outcome on the y-axis against the rounded values of the win margin on the x-axis. Finally, add the same two non-linear regression lines you used in Q8 - but these lines should use the full original data, not the rounded data you plotted as points. *Hint:* To round to the nearest 0.1 we would use `round(x,1)`; to round to the nearest 0.05 we can use `round(x/0.05)*0.05`

10. For the first version of the regression discontinuity analysis, implement a simple difference-in-means test, comparing the average vote share received by the PMDB in 2004 for all contests in 2000 where the PMDB just won (treatment) or lost (control) by less than 3%. I.e. keep only observations with a win margin of +/-3%. Interpret these results and compare to the observational regression in Q5.

11. For the second version of the regression discontinuity analysis, use *all* the data and run the linear regression of the outcome (vote share in 2004) on treatment (Incumbency) and the running variable (win margin). Interpret this regression and compare it to your results in Q10.

12. The regression in Q11 assumes that there is a linear relationship between the winning margin in 2000 and the party's vote share in 2004. However, the chart in Q8 suggests the relationship might be non-linear. To get an accurate measure of the causal effect we need to accurately match the shape of the relationship between the running variable and the toucome. We can do this by adding quadratic (`^2`) and cubic (`^3`) terms of win margin as controls in the regression. Implement this non-linear regression discontinuity and interpret the results.

13. For the third version of the regression discontinuity analysis, we apply the regression method only to data within a small window (bandwidth) of the threshold. Subset the data to +/-5% of the threshold and apply the same regression as in Q11 to this smaller dataset. Interpret the results and compare to your results in questions 10,11 and 12.

14. An alternative way to implement the regression discontinuity analysis in Q13 is to use the package `RDestimate`. Report and interpret the results for each of the three automatically-selected bandwidths. *Hint:* You can access the output from the `RDestimate` object using `$est` for the coefficient on treatment and `$p` for the p-value.

15. Use a `for` loop to implement the regression discontinuity method in Q11 (linear regression on all the data) for each of the three parties: the PMDB, PSDB and PFL. Summarise the results in a single table and compare the results between parties.

16. The Mayor of a small municipality calls you for political advice. He wants to know what vote share his party (the PMDB) is likely to receive in the next election. He is very confident because at the last election he won easily with a winning margin of 30%. Based on the evidence

you have recorded above from the regression discontinuities, how would you advise the Mayor about his likely performance in the next election?