FLS 6415: Replication 3 - Natural Experiments

April 2020

To be submitted (code + answers) by midnight, Wednesday 8th April.

First read the paper by De La O (2013) on the class website.

The replication data is in the file *DelaO.csv*, and the important variables are described below. Each row of this dataset is one electoral precinct, some of which are considered treated because they received *Progresa*.

Table 1: Key Variables in De La O (2013)

Variable	Description
treatment	Whether the precinct received Progresa
numerotreated	Number of Treated Villages in Precinct
numerocontrol	Number of Control Villages in Precinct
avgpoverty	Poverty in 1995
pobtot1994	Population in 1994/5
pobelegiblep	Population Eligible
villages	Number of Villages in Precinct
t1994	Turnout % in 1994
pri1994s	PRI vote share 1994
pan1994s	PAN vote share 1994
prd1994s	PRD vote share 1994
votos_totales1994	Total Number of Votes in 1994
pri1994	Number of PRI votes in 1994
pan1994	Number of PAN votes in 1994
prd1994	Number of PRD votes in 1994
t2000	Turnout % in 2000
pri2000s	PRI Vote Share in 2000
pan2000s	PAN Vote Share in 2000
prd2000s	PRD Vote Share in 2000

1. First, what is treatment in this study? What is control? What is the outcome being measured?

2. To help assess the balance between treatment and control units, reproduce Table 2 in De La O (2013) (Don't worry about the standard errors in brackets in the 'Difference' column for now).

3. Is the balance shown in this table (Table 2 in De La O) a necessary condition for causal inference? Is it a sufficient condition for causal inference?

4. The main analysis in De La O is conducted on a subset of the full dataset. Filter the data so that only precincts that have either one treatment village (`numerotreated`) or one control village (`numerocontrol`) inside them are included in your new dataset. What percentage of the original precincts are included in the new dataset?

5. One of De La O's conclusions is that treatment (receiving Progresa) boosts turnout. Conduct a simple difference-in-means t-test on the filtered dataset from Q4 to assess this claim.

What is the estimated difference-in-means and how statistically significant is the result?

6. De La O's analysis of turnout is in the upper panel of Table 3, where she runs a regression, adding some controls. (We are going to focus on the 'ITT' estimates, we will talk about the 'IV' estimates next week). Replicate this turnout regression. The controls (listed under De La O's Table 3) are `avgpoverty,pobtot1994,votos_totales1994,pri1994,pan1994,prd1994` and there is a fixed effect for the *villages* variable. (Try to include the robust standard errors, but no problem if you cannot). Interpret the results.

7. Now run the same regression but exclude the number-of-villages fixed effects (keep the other controls). How does this change the comparisons we are making between treated and control villages? How do the results change?

8. Replicate all four columns of the upper panel of Table 3 in De La O (2013). Interpret the results.

9. Now let's look at some critiques of the paper. Normally, we measure turnout percentages and vote shares as being naturally bounded between 0 and 100% (or 0 and 1). Other numbers don't make sense. Use a boxplot or similar graph to assess the distribution of values on the four dependent variables. What do you find?

10. As a 'quick fix' replace all the unrealistic values above 100% (1) with `NA` for all the turnout percentage and vote share dependent variables. Re-run your regressions from question 8. Do your conclusions change? Why might this be?

11. Next, examine the control variable for population in 1994 (`pobtot1994`). Use a graph or other method to identify any extreme outliers. Extreme values of control variables are not a problem if they are balanced across treatement and control groups. But are they in this case? Identify whether the extreme outliers are in the control or treatment group.

12. Remove the extreme outliers you identified in Q9 from the dataset (the dataset before you removed the infeasible values of the dependent variables). Re-run your regressions. Do your conclusions change? Why might this be?

13. One more issue. The controls for the regressions you have conducted so far are the *absolute number* of votes for turnout, PRI, PAN and the PRD. But for the dependent variable, De La O is using the *percentage vote share of the population.* Arguably it might be more consistent to use the same measurement approach on both the left and right-hand sides of the regression. Try implementing the regressions using the controls `t1994, pri1994s, pan1994s, prd1994s` in place of `votos_totales1994, pri1994, pan1994, prd1994`. Ignore the other corrections you made in previous questions. Does this change your conclusions? Why might this be?