Replication Exercise 3 - Natural Experiments

Marina Merlo

21 de abril de 2019

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

Both treatment and control is to have been assigned to Progresa, but they differ on how long each group received the program: the treatment group received the benefits for 21 months and the control, for 6 months before the elections. The outcome being measured is the incumbent voting.

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

Variable	Early	Late	Difference
Poverty	4.576	4.593	-0.017
Population	2040.107	1851.610	188.496
Pop.Eligible	0.887	0.849	0.038
No.of.Villages	6.084	6.377	-0.292
Ran.Assigned.Villages1	0.906	0.909	-0.003
Ran.Assigned.Villages2	0.091	0.084	0.006
Turnout1994	0.658	0.643	0.015
PRI.voteshare94	0.435	0.410	0.025
PAN.voteshare94	0.051	0.062	-0.012
PRD.voteshare94	0.102	0.098	0.004

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?

The balance table can be considered a necessary condition for causal inference, but isn't sufficient. We would need an exhaustive list with the confounder variables, but this is impossible: there will always be latent or omitted variables that might explain our outcomes. More importantly, the balance table says nothing about the randomization and the treatment assignment mechanism in the experiment - usually, we need qualitative data (as briefly provided by De La O (2013) in page 5) to verify if the experiment happened "as-if" random. Knowing these aspects would be sufficient to assess if the treatment is independent of the potential outcomes.

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?

The new dataset has 420 observations, which represents 90.71% of the original precints.

5. One of De La O's conclusions is that treatment 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?

```
##
##  Welch Two Sample t-test
##
## data:  t2000 by treatment
## t = -1.6082, df = 376.12, p-value = 0.1086
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10144010  0.01016072
## sample estimates:
## mean in group 0 mean in group 1
##       0.6347740       0.6804137
```

The estimated difference-in-means in the 2000 turnout for treatment and control isn't statistically significant at a 95% confidence level. The estimated difference in turnout means is between -0.1014 and 0.01016.

6. De La O's analysis of turnout is in the upper panel of Table 3, where she runs a regression, adding some controls. Replicate this turnout regression. The controls (listed under De LaO'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.

	Dependent variable:
	t2000
treatment	0.053*
	(0.030)
Constant	0.619***
	(0.171)
Observations	417
R^2	0.116
Adjusted R^2	0.071
Residual Std. Error	0.298 (df = 396)
F Statistic	2.587*** (df = 20; 396)
Note:	*p<0.1; **p<0.05; ***p<0.01

The constant is different from the paper, but the rest of the coefficients are the same. The fixed effects in the village numbers per precint accounts for unobserved confounders related to it. The regression shows a positive and significant effect of the treatment in the turnout at 90% confidence. There's an average of 5.3% increase in the turnout in the precints with treated villages, but this effect vary from 8.3% to 2.3%.

7. Now run the same regression but exclude the village 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?

	Dependent variable:
	t2000
treatment	0.045*
	(0.028)
Constant	0.640***
	(0.147)
Observations	417
R^2	0.079
Adjusted R^2	0.063
Residual Std. Error	0.299 (df = 409)
F Statistic	4.978*** (df = 7; 409)
Note:	*p<0.1; **p<0.05; ***p<0.01

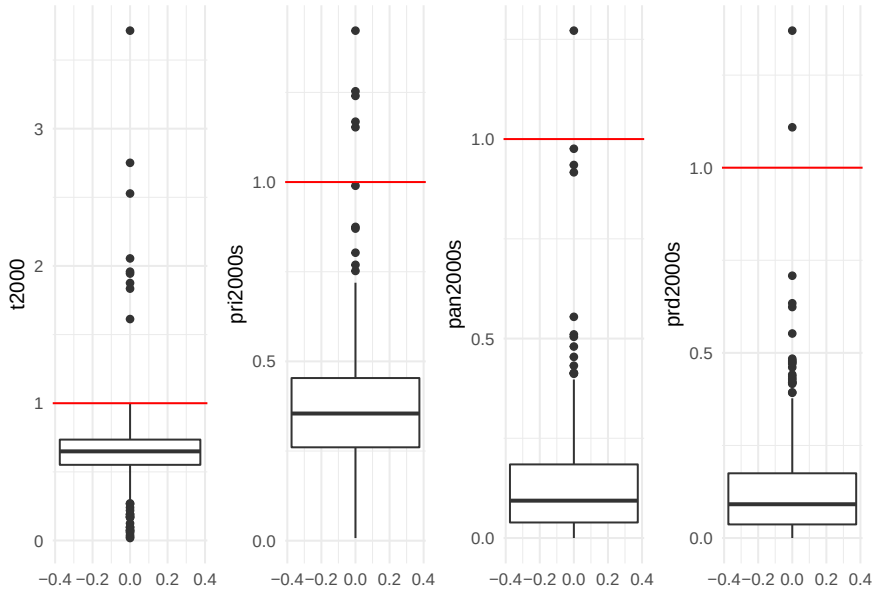
Without the villages fixed effects, we aren't taking into account possible unobserved difference in the precints due its number of treated villages. The treatment coefficient and standard errors are smaller, but still significant at 90% confidence. Now there's an average effect of 4.5% increase in the turnout in the precints with treated villages, varying from 1.7% to 7.3%.

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

	Dependent variable:			
	t2000	pri2000s	pan2000s	prd2000s
	(1)	(2)	(3)	(4)
treatment	0.053*	0.037**	0.007	0.002
	(0.030)	(0.015)	(0.012)	(0.014)
Constant	0.619***	0.357***	0.135*	0.137*
	(0.171)	(0.083)	(0.072)	(0.075)
Observations	417	417	417	417
R^2	0.116	0.288	0.197	0.318
Adjusted R^2	0.071	0.252	0.157	0.284
Residual Std. Error (df = 396)	0.298	0.158	0.126	0.122
F Statistic (df = 20; 396)	2.587***	7.998***	4.862***	9.244***
Note:				*p<0.1; **p<0.05; ***p<0.01

The positive effect of the treatment only happens for the turnout rates and the votes for the PRI party, the incumbent, over the total population.

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 graphic to assess the distribution of values on the four dependent variables. What do you find?



There are turnout percentages and vote shares for all the parties above 100% (all the dots above the red line at each boxplot), meaning that our estimates in the previous regressions were made with illogical values and probably are biased.

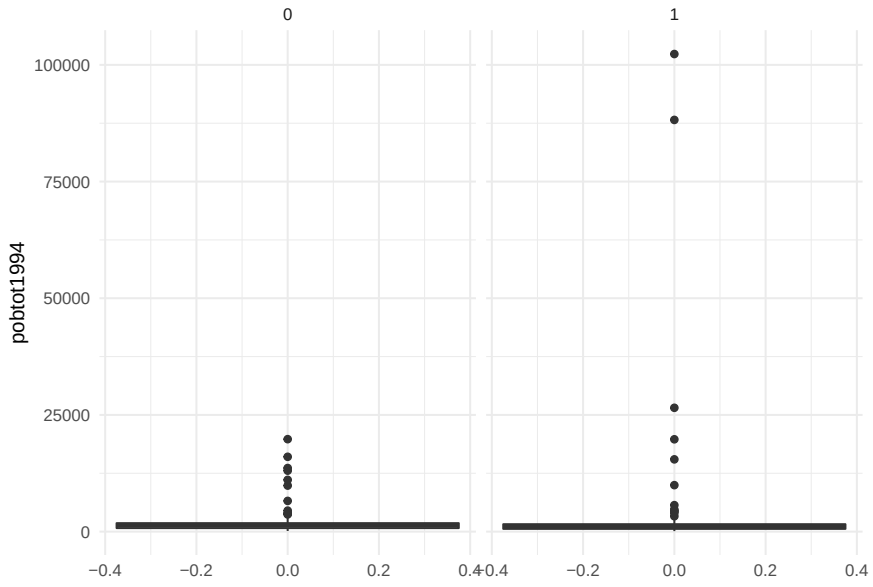
8. As a 'quick fix' replace all the unrealistic values above 100% (1) with NA for all the turnout percentage and vote share dependent variables. Re-run your regressions from question 8. Do your conclusions change? Why might this be?

	Dependent variable:			
	t2000	pri2000s	pan2000s	prd2000s
	(1)	(2)	(3)	(4)
treatment	0.016	0.017	−0.006	−0.002
	(0.017)	(0.012)	(0.009)	(0.011)
Constant	0.783***	0.446***	0.171***	0.166***
	(0.089)	(0.067)	(0.054)	(0.056)
Observations	408	408	408	408
R^2	0.233	0.388	0.291	0.386
Adjusted R^2	0.194	0.356	0.255	0.354
Residual Std. Error (df = 387)	0.162	0.122	0.087	0.097
F Statistic (df = 20; 387)	5.884***	12.268***	7.961***	12.149***
Note:				*p<0.1; **p<0.05; ***p<0.01

There were 9 observations with percentage values above 100%. When we run the same regressions with this filtered dataset, the treatment loses its significance for all the dependent variables. Probably the earlier

results were driven by those higher outliers.

9. 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.



The outliers are unbalanced between control and treatment groups - in the latter, there are three observations with populations above 25.000

10. 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?

	Dependent variable:			
	t2000	pri2000s	pan2000s	prd2000s
	(1)	(2)	(3)	(4)
treatment	0.032	0.028*	0.001	−0.004
	(0.027)	(0.015)	(0.011)	(0.014)
Constant	0.636***	0.364***	0.140**	0.142*
	(0.162)	(0.081)	(0.069)	(0.074)
Observations	414	414	414	414
R^2	0.206	0.320	0.246	0.359
Adjusted R^2	0.166	0.285	0.208	0.326
Residual Std. Error (df = 393)	0.279	0.153	0.122	0.118
F Statistic (df = 20; 393)	5.098***	9.235***	6.407***	10.989***
Note:				*p<0.1; **p<0.05; ***p<0.01

Without these extreme values in the population size, the treatment is statistically significant just for the vote share of PRI party. The significance for the turnout is likely to vanish because the outliers in the population are directly related to the turnout variable, biasing its estimation.

11. 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 LaO 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?

	Dependent variable:			
	t2000	pri2000s	pan2000s	prd2000s
	(1)	(2)	(3)	(4)
treatment	0.009	0.013	−0.005	−0.005
	(0.014)	(0.011)	(0.009)	(0.010)
Constant	−0.042	0.007	−0.046	0.026
	(0.132)	(0.073)	(0.065)	(0.051)
Observations	417	417	417	417
R^2	0.710	0.604	0.505	0.666
Adjusted R^2	0.696	0.584	0.480	0.649
Residual Std. Error (df = 396)	0.171	0.118	0.099	0.085
F Statistic (df = 20; 396)	48.535***	30.162***	20.236***	39.508***
Note:				*p<0.1; **p<0.05; ***p<0.01

Using the same measurement for the control variables also vanishes all the statistical significance of the treatment for all the dependent variables. The huge difference in the magnitudes between dependent (going from 0 to 1) and the independent (going as high as 1319 in votos_totales1994, for example) variables probably drove the correlation between the variables to be higher than they actually are. [not sure about this though]