# FLS 6415: Class 5 Homework

*September 21, 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.

This week we analyze the data from De La O (2013), a randomized natural experiment. Each row of this dataset is one electoral precinct, some of which are considered treated because they received *Progresa*. The key variables in the dataset are defined below, and you can copy this code to help prepare your tables:

```
d <- read.dta("DeLaO_AJPS2013_rep_file.dta")
vars <- data.frame(names(d)[c(7, 8, 9, 5, 4, 6, 13, 10, 11, 20, 22, 24, 26,
    16, 15, 17, 18, 19, 21, 23, 25)], c("Whether the precinct received Progresa",
    "Number of Treated Villages in Precinct", "Number of Control Villages in Precinct",
    "Poverty in 1995", "Population in 1994/5", "Population Eligible", "Number of Villages in Precinct",
    "Randomly assigned villages=1", "Randomly assigned villages=2", "Turnout % in 1994",
    "PRI vote share 1994", "PAN vote share 1994", "PRD vote share 1994", "Total Number of Votes in 1994
    "Number of PRI votes in 1994", "Number of PAN votes in 1994", "Number of PRD votes in 1994",
    "Turnout % in 2000", "PRI Vote Share in 2000", "PAN Vote Share in 2000",
    "PRD Vote Share in 2000"))
colnames(vars) <- c("Variable", "Description")

tab2_vars <- names(d)[c(5, 4, 6, 13, 10, 11, 20, 22, 24, 26)]
tab3_outcomes <- c("t2000", "pri2000s", "pan2000s", "prd2000s")
```

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 |
| one_random | Randomly assigned villages=1 |
| two_random | Randomly assigned villages=2 |
| 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. To help assess the balance between treatment and control units, reproduce the 'Early' and 'Late'

columns of De La O's Table 2 (ignore the 'Difference' column for now). Hint: Use `summarise_at` to quickly summarise many columns. To turn a 'wide' data.frame into a 'long' data.frame with treatment and control columns you need to first `gather` all the variables except treatment, and then `spread` the treatment variable. To easily use the full variable names, try to `join` the data.frame you produce with the *vars* data.frame produced by the code above.

2. Use a `for` loop to apply a simple linear regression (no extra controls, normal standard errors) to assess the magnitude (coefficent) and significance (p-value) of the difference between treatment and control ('Early' and 'Late'). Create a data.frame as the output from this `for` loop and then use a `join` to combine it with the data.frame you produced in Q1 so you have a single table like De La O's Table 2.

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. Replicate the results of column (1) of the upper panel of Table 3 in De La O to estimate the effect of treatment on turnout. The controls (listed under De La O's Table 3) are *avg-poverty,pobtot1994,votos_totales1994,pri1994,pan1994,prd1994* and a fixed effect for the *villages* variable. To keep things simple, ignore the robust standard errors and just use normal standard errors. Interpret the results of the regression.

6. Replicate all columns of the upper panel of Table 3 in De La O (2013) by conducting the regression for all four dependent variables. To keep things simple, ignore the robust standard errors and just use normal standard errors. You should try to use a `for loop` to change the dependent variable for each regression, a `list` to store the results and `stargazer` to present the table.

7. When we calculate turnout percentages and vote shares, the data should be naturally bounded between 0 and 100% (or 0 and 1). Is that true of De La O's outcome data? Provide evidence to support your claim.

8. As a 'quick fix' replace all the values above 100% (1) with `NA` for all the turnout percentage and vote share dependent variables. Re-run your regressions from question 4. Do your conclusions change? Why might this be?

9. Examine the control variable for population in 1994 (*pobtot1994*). Identify any extreme outliers. Remember that the value of randomized treatment for causal inteface is in balancing the characteristics and potential outcomes across treatment and control groups. So extreme values are not a problem provided that they're 'balanced' and half of them are in the treatment group and the other half (if comparable) are in the control group. Is that the case for the extreme outliers in our data for population in 1994? Hint: Use `rank` and/or `arrange` to help you identify outliers.

10. Remove the extreme outliers you identified in Q9 from the dataset (start from the dataset after you applied the filter in Q4). Re-run your regressions. Do your conclusions change? Why might this be?

11. 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* so arguably it might be more natural to use the same measurement approach for the controls. Try implementing the regressions using the controls *t1994, pri1994s, pan1994s, prd1994s* in place of *votos_totales1994, pri1994, pan1994, prd1994*. Again, to focus just on this issue, start from the dataset after you applied the filter in Q4 - Don't adjust the control or dependent variables so that they are between 0 and 1, and don't remove the population 1994 outliers. Does this change your conclusions? Why might this be?