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.

```
Q1 <- d %>% group_by(treatment) %>% summarise_at(tab2_vars,mean,na.rm=TRUE) %>% gather(key,value,-treatm

Q1 <- vars %>% inner_join(Q1,by="Variable")

kable(Q1,caption="Balance in the Data")
```

Table 2: Balance in the Data

Variable	Description	Late	Early
avgpoverty	Poverty in 1995	4.59	4.58
pobtot1994	Population in 1994/5	1851.61	2040.11
pobelegiblep	Population Eligible	0.85	0.89
villages	Number of Villages in Precinct	6.38	6.08
one_random	Randomly assigned villages=1	0.91	0.91
two_random	Randomly assigned villages=2	0.08	0.09
t1994	Turnout % in 1994	0.64	0.66
pri1994s	PRI vote share 1994	0.41	0.44
pan1994s	PAN vote share 1994	0.06	0.05
prd1994s	PRD vote share 1994	0.10	0.10

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.

```
library(broom)
Q2 <- vars %>% select(Variable)

Q2[,"Difference"] <- NA
Q2[,"p.value"] <- NA

for (outcome in tab2_vars){
  temp <- tidy(lm(as.formula(paste(outcome,"~treatment")),data=d))
  Q2[Q2$Variable==outcome,"Difference"] <- temp %>% filter(term=="treatment") %>% select(estimate)
  Q2[Q2$Variable==outcome,"p.value"] <- temp %>% filter(term=="treatment") %>% select(p.value)
}

Q2 <- Q1 %>% inner_join(Q2,by="Variable")

kable(Q2,caption="Balance in the Data",digits=2)
```

Table 3: Balance in the Data

Variable	Description	Late	Early	Difference	p.value
avgpoverty	Poverty in 1995	4.59	4.58	-0.02	0.73
pobtot1994	Population in 1994/5	1851.61	2040.11	188.50	0.77

Variable	Description	Late	Early	Difference	p.value
pobelegiblep	Population Eligible	0.85	0.89	0.04	0.15
villages	Number of Villages in Precinct	6.38	6.08	-0.29	0.44
one_random	Randomly assigned villages=1	0.91	0.91	0.00	0.92
two_random	Randomly assigned villages=2	0.08	0.09	0.01	0.83
t1994	Turnout % in 1994	0.64	0.66	0.01	0.67
pri1994s	PRI vote share 1994	0.41	0.44	0.03	0.31
pan1994s	PAN vote share 1994	0.06	0.05	-0.01	0.18
prd1994s	PRD vote share 1994	0.10	0.10	0.00	0.74

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?

Balance is a necessary condition for causal inference because systematic differences between the treatment and control groups on the covariates could be responsible for any diffences in the outcome variable and act as a confounder, preventing us from isolating the effect of treatment alone. However, balance is not a sufficient condition for causal inference. Even if we have balance on observed covariates, there may be imbalance on unobserved or unmeasurable covariates. For this reason it is also important to verify the *procedure* of treatment assignment could not have introduced any confounding. For example, if we have balance and we know that treatment assignment was randomized then we can be more confident in conducting 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?

After filtering for precincts that have only one treatment or control village in them, 90.7% of the data remains.

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. To keep things simple, ignore the robust standard errors and just use normal standard errors. Interpret the results of the regression.

```
library(xtable)
reg1 <- lm(t2000~treatment + avgpoverty + pobtot1994 + votos_totales1994 + pri1994 + pan1994 + prd1994 +

xtable(reg1,caption="Regression of turnout in 2000 on treatment with controls")
```

% latex table generated in R 3.3.3 by xtable 1.8-2 package % Thu Sep 28 10:58:16 2017

Precincts which received the Progresa treatment early experienced 5.3 percentage points higher turnout at the next election, controlling for covariates and fixed effects for the number of villages in the precinct.

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. The controls (listed under De La O's Table 3) are *avgpoverty,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. 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.

```
results <- vector("list",length(tab3_outcomes))
names(results) <- tab3_outcomes

for (i in tab3_outcomes){
  formula <-  as.formula(paste(i,"~treatment + avgpoverty + pobtot1994 + votos_totales1994 + pri1994 + p
  results[[i]] <- lm(formula,data=subset_d)
}
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.6192	0.1825	3.39	0.0008
treatment	0.0532	0.0315	1.68	0.0928
avgpoverty	0.0140	0.0367	0.38	0.7029
pobtot1994	-0.0000	0.0000	-5.08	0.0000
votos_totales1994	-0.0003	0.0004	-0.77	0.4444
pri1994	0.0004	0.0004	0.88	0.3806
pan1994	0.0004	0.0006	0.67	0.5011
prd1994	0.0002	0.0004	0.58	0.5649
factor(villages)2	-0.0580	0.0712	-0.81	0.4157
factor(villages)3	-0.0660	0.0748	-0.88	0.3782
factor(villages)4	-0.0293	0.0721	-0.41	0.6844
factor(villages)5	0.0264	0.0763	0.35	0.7295
factor(villages)6	-0.0508	0.0757	-0.67	0.5032
factor(villages)7	-0.0261	0.0815	-0.32	0.7491
factor(villages)8	-0.1345	0.0946	-1.42	0.1560
factor(villages)9	0.0815	0.0947	0.86	0.3895
factor(villages)10	-0.1336	0.1135	-1.18	0.2401
factor(villages)11	-0.1151	0.0991	-1.16	0.2459
factor(villages)12	-0.0315	0.1256	-0.25	0.8018
factor(villages)13	0.1492	0.1093	1.37	0.1730
factor(villages)14	-0.1159	0.0822	-1.41	0.1592

Table 4: Regression of turnout in 2000 on treatment with controls

```r
stargazer(results,style="ajps",keep=c("treatment"),intercept.bottom=TRUE,dep.var.labels=c("Turnout","PR
```

Table 5: Regressions of Outcomes on Treatment with Controls

	Turnout	**PRI**	**PAN**	**PRD**
treatment	0.053*	0.037**	0.007	0.002
	(0.032)	(0.017)	(0.013)	(0.013)
N	417	417	417	417
R-squared	0.116	0.288	0.197	0.318
Adj. R-squared	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***

$^{***}p < .01$; $^{**}p < .05$; $^{*}p < .1$

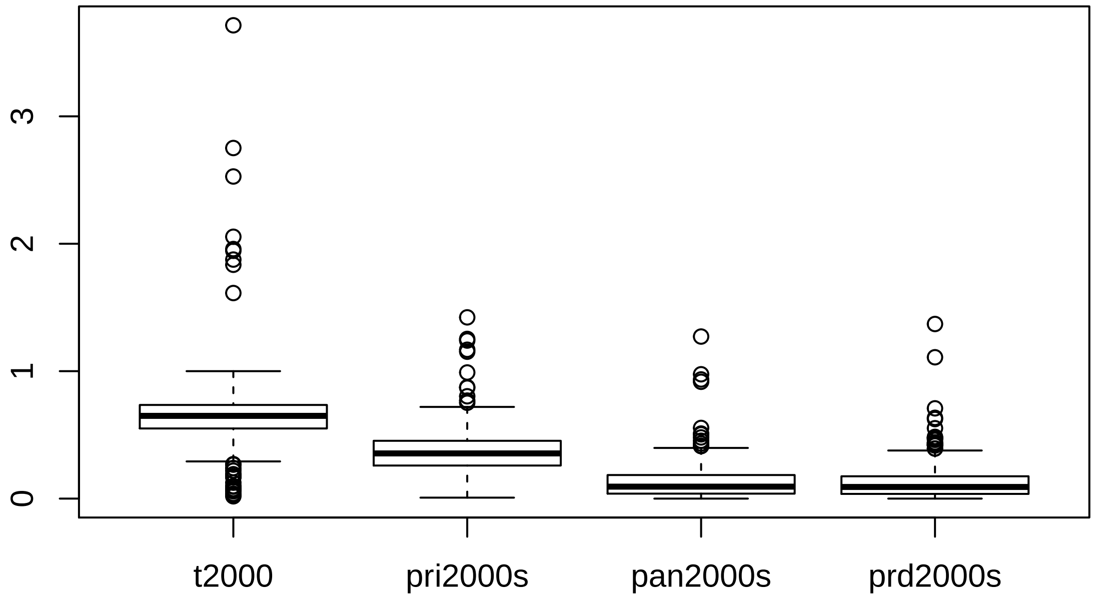
7. When we calculate turnout percentages and vote shares, the data should be naturally bounded between 0 and 100% (or 0 and 1). Is that true of De La O's outcome data? Provide evidence to support your claim.

```
share_vars <- c("t2000","pri2000s","pan2000s","prd2000s")
kable(subset_d %>% summarise_at(share_vars,max,na.rm=TRUE),caption="Maximum values for each variable")
```

Table 6: Maximum values for each variable

t2000	pri2000s	pan2000s	prd2000s
3.714286	1.422222	1.271795	1.37013

```
boxplot(subset_d[,share_vars])
```



As the table and the boxplot illustrate, there are numerous values for all variables which are above the maximum feasible value of 1.

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?

```
d_corrected <- subset_d
d_corrected <- d_corrected %>% mutate(t2000=ifelse(t2000>1,NA,t2000)) %>%
  mutate(pri2000s=ifelse(pri2000s>1,NA,pri2000s)) %>%
  mutate(pan2000s=ifelse(pan2000s>1,NA,pan2000s)) %>%
  mutate(prd2000s=ifelse(prd2000s>1,NA,prd2000s))
```

```r
results <- vector("list",length(tab3_outcomes))
names(results) <- tab3_outcomes

for (i in tab3_outcomes){
  formula <-  as.formula(paste(i,"~treatment + avgpoverty + pobtot1994 + votos_totales1994 + pri1994 + 
  results[[i]] <- lm(formula,data=d_corrected)
}

stargazer(results,style="ajps",keep=c("treatment"),intercept.bottom=TRUE,dep.var.labels=c("Turnout","PR
```

Table 7: Regressions of Outcomes on Treatment with Controls with Recoded Data

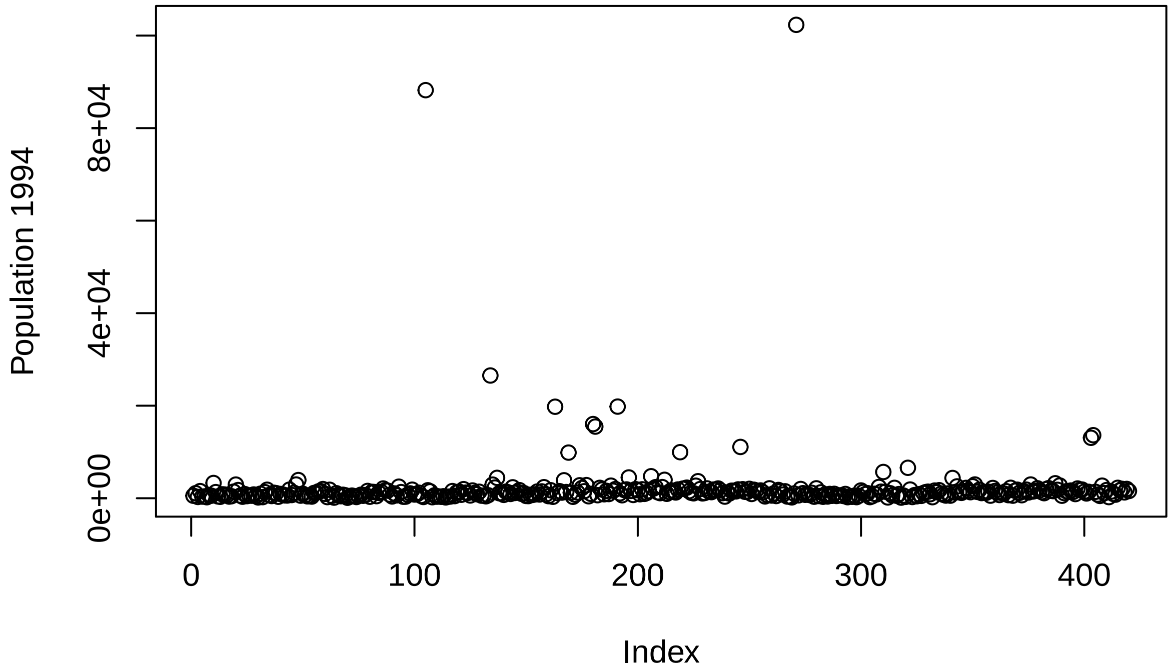
	Turnout	**PRI**	**PAN**	**PRD**
treatment	0.016	0.018	0.004	0.003
	(0.017)	(0.014)	(0.012)	(0.011)
N	408	412	416	415
R-squared	0.233	0.353	0.221	0.360
Adj. R-squared	0.194	0.320	0.182	0.328
Residual Std. Error	0.162 (df = 387)	0.129 (df = 391)	0.113 (df = 395)	0.099 (df = 39
F Statistic	5.884*** (df = 20; 387)	10.669*** (df = 20; 391)	5.619*** (df = 20; 395)	11.088*** (df = 20

***p < .01; **p < .05; *p < .1

The coefficient on treatment is no longer significant for both turnout and the PRI's vote share and the estimates are much lower. This may be because the extreme values exaggerate the turnout and vote share for the treatment units more than for control units, biasing the coefficient estimate upwards.

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.

```r
plot(d_corrected$pobtot1994,ylab="Population 1994")
```



```
outlier_status <- d_corrected %>% mutate(rank_pobtot1994=rank(-pobtot1994)) %>% filter(rank_pobtot1994<

kable(outlier_status,caption="Treatment Status of Extreme Outliers")
```

Table 8: Treatment Status of Extreme Outliers

Population_1994	Rank_Population_1994	Treatment
102322	1	1
88205	2	1

The plot identifies two extreme outliers far from the rest of the data. As the table shows, these two extreme values are both treatment values, which means neither can have a suitable counterfactual that is a control unit.

10. Remove the extreme outliers you identified in Q7 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?

```
data_Q10 <- d_corrected %>% mutate(rank_pobtot1994=rank(-pobtot1994)) %>% filter(rank_pobtot1994>2)

results <- vector("list",length(tab3_outcomes))
names(results) <- tab3_outcomes

for (i in tab3_outcomes){
  formula <-  as.formula(paste(i,"~treatment + avgpoverty + pobtot1994 + votos_totales1994 + pri1994 + 
  results[[i]] <- lm(formula,data=data_Q10)
}

stargazer(results,style="ajps",keep=c("treatment"),intercept.bottom=TRUE,dep.var.labels=c("Turnout","PR
```

Table 9: Regressions of Outcomes on Treatment with Controls with Population Outliers Removed

	Turnout	**PRI**	**PAN**	**PRD**
treatment	0.005	0.013	−0.0005	−0.001
	(0.015)	(0.013)	(0.012)	(0.010)
N	406	410	414	413
R-squared	0.384	0.383	0.269	0.399
Adj. R-squared	0.352	0.351	0.231	0.369
Residual Std. Error	0.142 (df = 385)	0.124 (df = 389)	0.110 (df = 393)	0.096 (df = 3
F Statistic	11.997*** (df = 20; 385)	12.061*** (df = 20; 389)	7.216*** (df = 20; 393)	13.027*** (df = 2

***p < .01; **p < .05; *p < .1

The coefficients are again smaller and no longer significant. This is likely because the outliers on high population were both treated and had unusually high levels of turnout and PRI vote share (perhaps because urban areas are confounded and systematically differnet in these ways). So their inclusion in the original regression biases the coefficients upwards.

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?

```
results <- vector("list",length(tab3_outcomes))
names(results) <- tab3_outcomes

for (i in tab3_outcomes){
  formula <-  as.formula(paste(i,"~treatment + avgpoverty + pobtot1994 + t1994 + pri1994s + pan1994s + 
  results[[i]] <- lm(formula,data=subset_d)
}

stargazer(results,style="ajps",keep=c("treatment"),intercept.bottom=TRUE,dep.var.labels=c("Turnout","PR
```

The Table shows the coefficients are again much lower, and no longer significant. This may be because the controls in absolute numbers controlled more for the size of the precinct than for the relative strength of the parties. So when using the relative measure of support for each party we are controlling for potential confounders and imbalance that reduces the treatment effect. Moreover, the interpretation of the regression

Table 10: Regressions of Outcomes on Treatment with Controls with Population Outliers Removed

	Turnout	**PRI**	**PAN**	**PRD**
treatment	0.009	0.013	−0.005	−0.005
	(0.018)	(0.013)	(0.010)	(0.009)
N	417	417	417	417
R-squared	0.710	0.604	0.505	0.666
Adj. R-squared	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***

***p < .01; **p < .05; *p < .1

is much more natural with the same format for vote share on the left hand side and the right hand side, because we can interpret the outcome now as the change in vote share between 1994 and 2000.