FLS 6415: Class 3 Homework

August 29, 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.

First, some very quick questions to encourage you to explore the app that simulates potential outcomes and randomized experiments. Use the app at https://poliong.shinyapps.io/causation_app/ to answer the following questions:

1. Set the number of units to at least 20. Try generating random units a few times to understand the patterns. Comparing the 'Actual' Average Treatment Effect with the 'Observed' ATE, which treatment assignment mechanism appears to provide an unbiased estimate? Which provides an over-estimate, and which an under-estimate?

The Random treatment assignment mechanism provides an unbiased estimate. The self-selection mechanism provides an over-estimate. The 'By covariates' method provides an under-estimate to the actual ATE.

2. Using the table of potential outcomes (and, if helpful, the 'data' tab), explain why the 'self-selection' treatment assignment mechanism does not accurately estimate the actual ATE.

With self-selection, the treated units tend to have an unusually high y_1, the potential outcome if treated, compared to the control units. This is because the units that might benefit the most (i.e. have a high y_0) choose to acquire treatment. since the observed ATE is calculated from the y_1 for treated units versus the y_0 for control units, the observed values therefore provide a biased estimate of the actual ATE which is based on the y_1 and y_0 for *all* units.

3. On the 'Analysis' tab, why does the comparison of means estimate exactly the same ATE as the regression?

The simple regression of the outcome on a binary treatment (with no other controls) estimates exactly the difference in the average outcome measure between the two treatment categories. But that is simply the difference in means between the two groups. So we can use either measure to calculate the ATE.

4. Set a high number of units and random treatment assignment. On the 'Repeated Experiments' tab, describe how the bias of the observed ATE compared to the actual ATE changes as the number of repeated experiments increases from 1 to 100.

When we only conduct a single experiment there is always a chance that the potential outcomes will not be particularly well balanced. So the actual ATE and the observed ATE can vary considerably. As we gradually increase the number of experiments, the distribution of the average treatment effects centres on the actual treatment effects and becomes normally distributed. On average, therefore, over many repeated experiments, random assignment reliably recovers the actual ATE. But in a single experiment (and obviously depending on the sample size) there remains some risk that our estimate is inaccurate. Note that even after 100 repeated experiment the distribution of the observed ATE is still more dispersed - and therefore less precisely estimated - than the distribution of the actual ATE. This is because there is always lower precision (and a risk of imbalance) from our experimental method that only uses half of the real data (the observed half of the potential outcomes), while the actual ATE uses the full dataset.

The questions below focus on replicating the results of the analysis in Olken (2010). While you should try and conduct the same analysis and see if you reach the same conclusion as Olken, it is NOT necessary that you get exactly the same numerical result. Try, but if you can't that's okay. Similarly, don't worry about copying the same formatting of the tables - as long as the results are clear. The aim is to focus on the understanding, skills and coding knowledge needed to apply the same type of analysis.

A dictionary of variable names and description for each table is provided at the end of this document, along with code you can copy and paste into your own analysis to save you time typing.

1. Open the file *clean_table2_village_data.dta*. Implement the balance test that Olken conducts on pages 247-248 to compare the village population in treatment and control villages.

```
library(foreign)
library(sandwich)
library(lmtest)
library(tidyverse)
library(knitr)

vill <- read.dta("clean_table2_village_data.dta")

bal_reg <- lm(desa_pop ~ elect + wave,data=vill)
bal_reg <- coeftest(bal_reg,vcovHC(bal_reg,"HC1"))[,c("Estimate","Pr(>|t|)")] #With Robust SEs
kable(bal_reg,caption="Balance Test for Village Population")
```

Table 1: Balance Test for Village Population

	Estimate	Pr(>|t|)
(Intercept)	2.8128416	0.0000518
elect	-0.2945508	0.6249111
wave	-1.1989937	0.0400029

2. Conduct a simpler balance test on village population by using a t-test to assess the difference in means for village population between treatment and control villages. How does the result differ from the regression in Q1? What explains the difference?

```
t_test_popn <- t.test(vill[vill[,"elect"]==0,"desa_pop"],vill[vill[,"elect"]==1,"desa_pop"])
```

The difference in means reported by the t-test is 0.517 and the p-value is 0.418

3. Reproduce the full balance table (Table 2 in Olken) using the files *clean_table2_village_data.dta* and *clean_table2_hh_data.dta*. (You can make separate tables for the village and respondent characteristics. Try to report the summary tables, but if you can't just report each regression separately. In your output there's no need to include the mean of the meeting group, the standard errors in brackets or the number of observations column; just the coefficient and p-value is fine).

```
library(miceadds)

tab2_vill$coef <- NA
tab2_vill$p.value <- NA

for (out in t2_vars_vill){
  bal_reg <- lm(as.formula(paste(out," ~ elect + wave")),data=vill)
  bal_reg <- coeftest(bal_reg,vcovHC(bal_reg,"HC1"))[,c("Estimate","Pr(>|t|)")] #With Robust SEs - to b
  tab2_vill[tab2_vill$Variable==out,"coef"] <- round(bal_reg["elect","Estimate"],3)
  tab2_vill[tab2_vill$Variable==out,"p.value"] <- round(bal_reg["elect","Pr(>|t|)"],3)
}


hh <- read.dta("clean_table2_hh_data.dta")
```

```
tab2_indi$coef <- NA
tab2_indi$p.value <- NA

for (out in t2_vars_indi){
  bal_reg <- lm.cluster(as.formula(paste(out," ~ elect + wave")),data=hh,cluster="desaid")
  tab2_indi[tab2_indi$Variable==out,"coef"] <- round(summary(bal_reg)["elect","Estimate"],3)
  tab2_indi[tab2_indi$Variable==out,"p.value"] <- round(summary(bal_reg)["elect","Pr(>|t|)"],3)
}
```

```
kable(tab2_vill,caption="Balance Tests for Village Characteristics")
```

Table 2: Balance Tests for Village Characteristics

Variable	Description	coef	p.value
desa_pop	Village population (1,000 inhabitants	-0.295	0.625
ag_wage	Agricultural wage (1,000 Rupiah)	-1.061	0.466
asphalt	Percent village roads that are asphalt	-0.042	0.507
dusun_count	Number of hamlets per village	-0.633	0.142
church_mosque	Number of churches and mosques per village	-0.220	0.698
dist_capital	Distance to subdistrict capital (km)	3.548	0.109
v_ethnic_frag	Village ethnic fragmentation	-0.075	0.190
v_rel_frag	Village religious fragmentation	0.011	0.827
vh_age	Village head age	2.368	0.443
educ	Village head years of education	-1.409	0.081
v_numvhcandidates	Number of village head candidates in last village head election	0.304	0.432
v_morethan1candidate	More than one candidate in last village head election	0.089	0.449
v_sharevhvoted	Share of population that voted in last village head election	-0.004	0.910
v_vhvictorymargin	Village head's margin of victory in last election (if challenger)	-0.011	0.870
v_numaparat	Number of village government executive branch members	-0.616	0.386
v_sharedusunsinaparat	Share of hamlets represented in village executive branch	0.043	0.442
bpd_numpeople	Number of people in village parliament	-0.976	0.249
v_sharedusunsinbpd	Share of hamlets represented in village parliament	0.054	0.339
bpd_nummeetings	Number of village parliament meetings in last year	-1.853	0.041
bpd_districtsystem	Villiage parliament district system (1 = district, 0 = at large)	0.081	0.587
ppk_numproj	Number of previous KDP projects	-0.239	0.455

```
kable(tab2_indi, caption="Balance Tests for Respondent Characteristics")
```

Table 3: Balance Tests for Respondent Characteristics

Variable	Description	coef	p.value
zlhhexpcap	Survey respondent predicted log per capita expenditure	0.034	0.599
educ	Survey respondent years education	-0.519	0.399
female	Survey respondent is female	0.025	0.286
age	Survey respondent age	1.896	0.265
farmer_buruh	Survey respondent is farmer	-0.052	0.538

4. Interpret the balance table. What does this table tell us about (i) the randomization assignment process, and (ii) the prospects for causal inference?

The balance table identifies only one significant difference at the 0.05 level, which for 26 variables is about

what we would expect by chance (remember a threshold of 0.05 means 1 in 20 variables will be significant on average by chance). This suggests that (i) randomization was successful and was not interefered with, and (ii) it is likely to be feasible to conduct a reliable causal inference analysis with a simple difference in means assessment. However, we might still want to include a control for the one imbalanced variable (number of village parliament meetings) to ensure this is not confounding the results.

5. Open the file *clean_table7_10_data.dta*. The following variables are currently scaled so that they can take on integer values between 1 and 4: sat_prop, benefit, fair, aspire, kdp_satw2. Transform these variables so that they are on a scale between 0 and 1.

```
d <- read.dta("clean_table7_10_data.dta")

d <- d %>% mutate(sat_prop=(sat_prop-1)/3)
d <- d %>% mutate(benefit=(benefit-1)/3)
d <- d %>% mutate(fair=(fair-1)/3)
d <- d %>% mutate(aspire=(aspire-1)/3)
d <- d %>% mutate(kdp_satw2=(kdp_satw2-1)/3)
```

6. Open the file *clean_table7_10_data.dta* and get as close as possible to replicating column (1) of Table 7 using OLS. Use the transformed versions of the outcome variables you generated in Q5. The control variables are phase2, male, age, npers, c_servant, enterp, hwife, laborer, not-in_lf, other, priv_comp, student, teacher, trader, hexpcap. (Note there is an error that means the coefficient values for 'benefit' and 'use' you calculate will not be the same as in Olken's table (see Q9)).

```
tab7$coef <- NA
tab7$p.value <- NA

for (out in t7_vars){
  reg1 <- lm.cluster(as.formula(paste(out," ~ elect + phase2 + male + age + npers + c_servant + enterp +
  tab7[tab7$Variable==out,"coef"] <- round(summary(reg1)["elect","Estimate"],3)
  tab7[tab7$Variable==out,"p.value"] <- round(summary(reg1)["elect","Pr(>|t|)"],3)
}
```

```
kable(tab7,caption="Replicating the first column of Table 7 of Olken (2010)")
```

Table 4: Replicating the first column of Table 7 of Olken (2010)

Variable	Description	coef	p.value
sat_prop	Was the project chosen in accordance with your wishes?	0.059	0.018
benefit	Will the proposal benefit you personally?	0.120	0.006
use	Will you use the project?	0.075	0.087
fair	Was the chosen proposal fair?	0.060	0.028
aspire	Is the chosen proposal in accordance with the people's aspirations?	0.050	0.041
kdp_satw2	Are you satisfied with the KDP?	0.103	0.018
jobapproval_pres	Job approval of president of Indonesia	0.032	0.186
jobapproval_vh	Job approval of village head	-0.023	0.667

7. Replicate column (1) of Table 7 again but without including the control variables (but keep variable phase2). Does this suggest any different conclusions about the average treatment effect compared to the regression with controls? Why?

```
tab7$coef_no_controls <- NA
tab7$p.value_no_controls <- NA
```

```
for (out in t7_vars){
  reg1 <- lm.cluster(as.formula(paste(out," ~ elect + phase2")),data=d,cluster="desaid")
  tab7[tab7$Variable==out,"coef_no_controls"] <- round(summary(reg1)["elect","Estimate"],3)
  tab7[tab7$Variable==out,"p.value_no_controls"] <- round(summary(reg1)["elect","Pr(>|t|)"],3)
}

tab7[,"Description"] <- NULL
```

```
kable(tab7,caption="Replicating the first column of Table 7 of Olken (2010) without control variables")
```

Table 5: Replicating the first column of Table 7 of Olken (2010) without control variables

Variable	coef	p.value	coef_no_controls	p.value_no_controls
sat_prop	0.059	0.018	0.059	0.030
benefit	0.120	0.006	0.116	0.009
use	0.075	0.087	0.062	0.159
fair	0.060	0.028	0.062	0.022
aspire	0.050	0.041	0.051	0.033
kdp_satw2	0.103	0.018	0.110	0.007
jobapproval_pres	0.032	0.186	0.032	0.198
jobapproval_vh	-0.023	0.667	-0.026	0.634

Without controls, the coefficient estimates are broadly comparable in direction and magnitude. The p-values are slightly different but generally do not change the conclusions about statistical significance. This illustrates that the inclusion of covariates in randomized experiments is generally unnecessary, but can play a minor role in addressing concerns of residual imbalance and helping to increase the precision (reduce the p-value) of estimates.

8. For the variable 'use' ("Will you use the project?"), conduct a t-test for the differences in means between the treatment and control groups. Report the difference in means and the p-value of the t-test. Does this suggest any different conclusions about the average treatment effect compared to the regression without controls? Why?

```
tab7$t.test_dif_means <- NA
tab7$t.test_p.value <- NA

for (out in t7_vars){
  t_obj <- t.test(d[d$elect==0,out],d[d$elect==1,out])
  tab7[tab7$Variable==out,"t.test_dif_means"] <- t_obj$estimate[2]-t_obj$estimate[1]
  tab7[tab7$Variable==out,"t.test_p.value"] <- t_obj$p.value
}
tab7_use <- tab7[tab7$Variable=="use",]
```

The t-test for the difference in means for the variable "Will you use the project?" between plebiscite and meeting conditions is 0.086 and the p-value is 0.012. Compared to the regression without controls this coefficient estimate is slightly larger and the p-value is slightly smaller. The difference is likely to be explained by the presence of the remaining control variable for 'phase2' in the regression which means the regression does not reproduce the simple difference in means.

9. (Challenging) Investigate why the coefficient values for 'benefit' and 'use' in Q7 are different to those in Olken's Table 7. It may help to refer to the stata do file *100305tables.do*.

From the do file, Olken makes some special adjustments for these two variables.

```
replace benefit = . if benefit > 4

replace use = . if benefit > 4
```

The first adjustment seems to adjust for miscoding of the data to make sure everything is on a 1 to 4 scale. That's fine, but we haven't made this adjustment in our code.

The second adjustment contains a serious error - presumably it should read `replace use = . if use > 4`, but he copied and pasted the `benefit` line and forgot to change one variable. This affects the reported values.