Methods III: Replication Exercise 1

March, 2020

The data for Gerber, Green and Larimer (GGL, 2008) is available on the course website. You should first read through GGL 2008 quickly to understand the context of the field experiment.

Download the data from the course website and answer the following questions in either R or Stata. You should submit (i) your code, and (ii) a document or PDF containing your answers to jonnyphillips@gmail.com by midnight on Wednesday 25th March. If you get stuck, please feel free to email me, or in the worst case skip that question and continue to the next.

For the replication, don't worry about copying the specific details or formatting of the tables - as long as the results are clear. If you are using R, I encourage you to use R markdown to make it easy to combine your code and answers.

Here is a table of variable names and descriptions from the dataset:

Variable	Description
hh_id	Household Identifier
hh_size	Household Size
cluster	Block
treatment	Treatment status
g2000	Voted in 2000 General Election
g2002	Voted in 2002 General Election
p2000	Voted in 2000 Primary Election
p2002	Voted in 2002 Primary Election
p2004	Voted in 2004 Primary Election
sex	Sex
age	Age
voted	Voted in 2006 Primary Election

1. What hypotheses are GGL testing? Where did they get these hypotheses from?

2. What are the treatment and control conditions? What is the outcome variable?

3. What is the unit of analysis? What is the level at which treatment is applied?

4. Did randomization work? Let's reproduce Table 1 of GGL to conduct balance tests between the treatment and control groups on pre-treatment covariates. Note that GGL evaluate balance at the *household* level so you first need to aggregate the individual data to the household level by finding the household mean on each of the variables we want to assess balance for. Then calculate the mean across household separately for the control and treatment groups. What do we learn from the results?

Variable	Control	Civic Duty	Hawthorne	Self	Neighbors
hh_size	1.9124	1.9108	1.9100	1.9109	1.9100
g2002	0.8344	0.8361	0.8357	0.8352	0.8351
g2000	0.8663	0.8654	0.8668	0.8625	0.8653
p2004	0.4166	0.4155	0.4188	0.4209	0.4227
p2002	0.4087	0.4099	0.4117	0.4104	0.4060
p2000	0.2649	0.2664	0.2632	0.2629	0.2631
sex	0.5023	0.5027	0.5032	0.5014	0.5046
age	51.3140	51.1790	51.2041	51.2442	51.3423

5. GGL don't bother to do a t-test for the difference-in-means, but let's do it ourselves. Conduct a t-test for the difference in mean household age between the Control and 'Neighbors' conditions. Interpret the result.

```
##
##  Welch Two Sample t-test
##
## data:  age by treatment
## t = -0.28124, df = 28381, p-value = 0.7785
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.2258195  0.1691470
## sample estimates:
##   mean in group Control mean in group Neighbors
##              51.31400                51.34233
```

6. Now let's look at the results of the experiment. Perform a simple difference-in-means t-test for voter turnout between the Control and 'Neighbors' groups in the individual data. Interpret the result.

```
##
##  Welch Two Sample t-test
##
## data:  voted by treatment
## t = -30.207, df = 52613, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.08658577 -0.07603405
## sample estimates:
##   mean in group Control mean in group Neighbors
##             0.2966383               0.3779482
```

7. Now run an OLS regression to understand the effect of each treatment on voter turnout, to replicate column (a) of Table 3. (If you prefer you can use treatment as a factor variable, not a series of dummies like the authors use. For this question do not adjust the standard errors). Interpret the results.

9. Repeat your regression but this time with standard errors clustered to the household level. What difference does this make? Why do we do this?

Table 2:

	Dependent variable:
	voted
treatmentCivic Duty	0.018***
	(0.003)
treatmentHawthorne	0.026***
	(0.003)
treatmentSelf	0.049***
	(0.003)
treatmentNeighbors	0.081***
	(0.003)
Constant	0.297***
	(0.001)
Observations	344,084
Note:	*p<0.1; **p<0.05; ***p<0.01

term	estimate	std.error	p.value
(Intercept)	0.3145	0.0024	0.0000
treatmentControl	-0.0179	0.0026	0.0000
treatmentHawthorne	0.0078	0.0034	0.0201
treatmentNeighbors	0.0634	0.0034	0.0000
treatmentSelf	0.0306	0.0034	0.0000

10. Next, we want to add block-level fixed effects to our model to reproduce column (b) of Table 3. However, there are many (10,000) blocks so your computer probably doesn't have enough memory to run this regression directly. An equivalent methodology is to remove the between-group variation in the treatment variable manually and then run the same regression as in Q7. To do this for the 'Neighbors' treatment:

1. Create a dummy variable for individuals that received the 'Neighbors' treatment,
2. Remove the other treatments from your dataset (so you are left with just 'Neighbors' and 'Control' units),
3. For each 'cluster' group calculate the mean value of the binary 'Neighbors' treatment variable,
4. Subtract the cluster mean from the individual values of the Neighbors treatment variable.
5. Run the same regression as in Q8 but using the cluster-mean-centered treatment variable you just created as the explanatory variable.

How do the results change? How does this change the comparisons we are making in the regression?

11. Add covariates (g2000,g2002,p2000,p2002,p2004) to your model from Q10 to reproduce column 3 of Table 3. How do the results change when we add covariates?

12. In place of an OLS regression, use a logit regression model to run the same model as in Q8. How would you interpret the results?

13. Predict the mean first difference (the mean change) of the probability of voting when moving from the 'Control' to 'Neighbors' treatment category using your logit model from

Table 3:

	Dependent variable:
	voted
Neighbors	0.082^{***} (0.003)
Constant	0.306^{***} (0.001)
Observations	229,444
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4:

	Dependent variable:
	voted
Neighbors	0.082^{***} (0.003)
g2000	-0.005^{*} (0.003)
g2002	0.099^{***} (0.003)
p2000	0.099^{***} (0.002)
p2002	0.134^{***} (0.002)
p2004	0.155^{***} (0.002)
Constant	0.090^{***} (0.003)
Observations	229,444
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 5:

	Dependent variable:
	voted
treatmentCivic Duty	0.084***
	(0.012)
treatmentHawthorne	0.120***
	(0.012)
treatmentSelf	0.223***
	(0.012)
treatmentNeighbors	0.365***
	(0.012)
Constant	−0.863***
	(0.005)
Observations	344,084
Note:	*p<0.1; **p<0.05; ***p<0.01

Q12. ***Hint: Use Zelig in R and Clarify in Stata.***

```
## [1] 0.0838
```

14. How does the data processing the authors conduct on p.36-37 (under 'Study Population') affect your interpretation of the conclusions?

15. How generalizable are the findings of this study to other elections? To the same set of elections in 2010? To neighbouring Indiana in the same year? To elections in Brazil?