Problem Set #3

Andrew Walters

July 11, 2018

```r
# load packages
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 3.4.2
```

```r
library(foreign)
```

```
## Warning: package 'foreign' was built under R version 3.4.4
```

```r
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```r
library(sandwich)
library(multiwayvcov)
library(stargazer)
```

```
##
## Please cite as:

## Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

## R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

0 Write Functions You're going to be doing a few things a *number* of times – calculating robust standard errors, calculating clustered standard errors, and then calculating the confidence intervals that are built off these standard errors.

After you've worked through a few of these questions, I suspect you will see places to write a function that will do this work for you. Include those functions here, if you write them.

1 Replicate Results Skim [Broockman and Green's](http://link.springer.com/article/10.1007/s11109-013-9239-z) paper on the effects of Facebook ads and download an anonymized version of the data for Facebook users only.

```r
d <- read.csv("./data/broockman_green_anon_pooled_fb_users_only.csv");
colnames(d)
```

```
## [1] "studyno"            "treat_ad"             "cluster"
## [4] "name_recall"        "positive_impression"
```

a. Using regression without clustered standard errors (that is, ignoring the clustered assignment), compute a confidence interval for the effect of the ad on candidate name recognition in Study 1 only (the dependent

```r
study1_1a <- d[d$studyno==1,]
lm_1a <- lm(name_recall ~ treat_ad, study1_1a)
lm_1a$robust_se <- sqrt(diag(vcovHC(lm_1a)))
for(ii in 1:length(lm_1a$coefficients)){
  print(sprintf("The 95%% confidence interval (robust) for %s is between %.3f and %.3f",
    names(lm_1a$coefficients[ii]),
    lm_1a$coefficients[ii]-lm_1a$robust_se[ii]*1.96,
    lm_1a$coefficients[ii]+lm_1a$robust_se[ii]*1.96
    ))
}

## [1] "The 95% confidence interval (robust) for (Intercept) is between 0.150 and 0.215"
## [1] "The 95% confidence interval (robust) for treat_ad is between -0.051 and 0.032"

stargazer(lm_1a,se=list(lm_1a$robust_se), header=F)
```

Table 1:

	Dependent variable:
	name_recall
treat_ad	−0.010
	(0.021)
Constant	0.182***
	(0.016)
Observations	1,364
R^2	0.0002
Adjusted R^2	−0.001
Residual Std. Error	0.382 (df = 1362)
F Statistic	0.217 (df = 1; 1362)
Note:	*p<0.1; **p<0.05; ***p<0.01

b. What are the clusters in Broockman and Green's study? Why might taking clustering into account increase the standard errors?

The clusters are demographic groups that can be uniquely identified both by facebook ad targeting and public records (voter rolls, campaign contributions). Specifically that means that each cluster has a gender, age, and geographic region (ie 24-year-old men in San Francisco). A clustered experiment will increase the standard error becuause the sharp null hypothesis states that treated and untreated potentail outcomes are identical. For a clustered experiment this means we must consider every potential assignment of treatment and cluster. Therefore the standard error for a clustered experiment is multiplied by a $\sqrt{\frac{N}{k}}$ term (N is the number of participants, k is the number of clusters) which will be greater than 1, increasing the standard error.

c. Now repeat part (a), but taking clustering into account. That is, compute a confidence interval for the effect of the ad on candidate name recognition in Study 1, but now correctly accounting for the clustered nature of the treatment assignment. If you're not familiar with how to calculate these clustered and robust estimates,

there is a demo worksheet that is available in our course repository: '/code/week5clusterAndRobust.Rmd'.

```
lm_1c <- lm_1a
lm_1c$robust_se <- sqrt(diag(cluster.vcov(lm_1c, ~ cluster)))
lm <- lm_1c
for(ii in 1:length(lm$coefficients)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for (Intercept) is between 0.146 and 0.219"
## [1] "The 95% confidence interval (cluster-robust) for treat_ad is between -0.056 and 0.037"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 2:

	Dependent variable:
	name_recall
treat_ad	−0.010
	(0.024)
Constant	0.182***
	(0.018)
Observations	1,364
R^2	0.0002
Adjusted R^2	−0.001
Residual Std. Error	0.382 (df = 1362)
F Statistic	0.217 (df = 1; 1362)
Note:	*p<0.1; **p<0.05; ***p<0.01

d. Repeat part (c), but now for Study 2 only.

```
study2_1d <- d[d$studyno==2,]
lm_1d <- lm(name_recall ~ treat_ad, study2_1d)
lm_1d$robust_se <- sqrt(diag(cluster.vcov(lm_1d, ~ cluster)))
for(ii in 1:length(lm_1d$coefficients)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm_1d$coefficients[ii]),
    lm_1d$coefficients[ii]-lm_1d$robust_se[ii]*1.96,
    lm_1d$coefficients[ii]+lm_1d$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for (Intercept) is between 0.570 and 0.641"
## [1] "The 95% confidence interval (cluster-robust) for treat_ad is between -0.072 and 0.067"
```

```
stargazer(lm_1d,se=list(lm_1d$robust_se), header=F)
```

<div align="center">

Table 3:

	Dependent variable:
	name_recall
treat_ad	−0.003
	(0.036)
Constant	0.606***
	(0.018)
Observations	1,337
R^2	0.00001
Adjusted R^2	−0.001
Residual Std. Error	0.489 (df = 1335)
F Statistic	0.008 (df = 1; 1335)
Note:	*p<0.1; **p<0.05; ***p<0.01

</div>

e. Repeat part (c), but using the entire sample from both studies. Do not take into account which study the data is from (more on this in a moment), but just pool the data and run one omnibus regression. What is the treatment effect estimate and associated p-value?

```
print("Assuming you still want clusters to be considered accross both studies")
```

```
## [1] "Assuming you still want clusters to be considered accross both studies"
```

```
combinedStudies_1e <- rbind(study1_1a,study2_1d)
lm_1e <- lm(name_recall ~ treat_ad, combinedStudies_1e)
lm <- lm_1e
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ cluster)))
for(ii in 1:length(lm$coefficients)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for (Intercept) is between 0.418 and 0.491"
## [1] "The 95% confidence interval (cluster-robust) for treat_ad is between -0.207 and -0.103"
```

```
print(sprintf("The ATE of treatemnt x is %.3f",lm$coefficients[2]))
```

```
## [1] "The ATE of treatemnt x is -0.155"
```

```
print(sprintf("The p-value is %.3f",summary(lm)$coefficients[2,4]))
```

```
## [1] "The p-value is 0.000"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 4:

	Dependent variable:
	name_recall
treat_ad	−0.155***
	(0.027)
Constant	0.454***
	(0.019)
Observations	2,701
R^2	0.025
Adjusted R^2	0.024
Residual Std. Error	0.482 (df = 2699)
F Statistic	68.313*** (df = 1; 2699)
Note:	*p<0.1; **p<0.05; ***p<0.01

f. Now, repeat part (e) but include a dummy variable (a 0/1 binary variable) for whether the data are from Study 1 or Study 2. What is the treatment effect estimate and associated p-value?

```
print("Assuming you still want clusters to be considered accross both studies")
```

```
## [1] "Assuming you still want clusters to be considered accross both studies"
```

```
study1_1a$isStudy1 <- 1
study2_1d$isStudy1 <- 0
combinedStudiesDummy_1f <- rbind(study1_1a,study2_1d)
lm_1f <- lm(name_recall ~ treat_ad + isStudy1, combinedStudiesDummy_1f)
lm <- lm_1f
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ cluster)))
for(ii in 1:length(lm$coefficients)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for (Intercept) is between 0.574 and 0.639"
## [1] "The 95% confidence interval (cluster-robust) for treat_ad is between -0.047 and 0.033"
## [1] "The 95% confidence interval (cluster-robust) for isStudy1 is between -0.467 and -0.386"
```

```
print(sprintf("The ATE of treatemnt x is %.3f",lm$coefficients[2]))
```

```
## [1] "The ATE of treatemnt x is -0.007"
```

```
print(sprintf("The p-value is %.3f",summary(lm)$coefficients[2,4]))
```

```
## [1] "The p-value is 0.709"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 5:

	Dependent variable:
	name_recall
treat_ad	−0.007
	(0.020)
isStudy1	−0.426***
	(0.021)
Constant	0.607***
	(0.017)
Observations	2,701
R^2	0.193
Adjusted R^2	0.193
Residual Std. Error	0.438 (df = 2698)
F Statistic	322.848*** (df = 2; 2698)
Note:	*p<0.1; **p<0.05; ***p<0.01

g. Why did the results from parts (e) and (f) differ? Which result is biased, and why? (Hint: see pages 75-76 of Gerber and Green, with more detailed discussion optionally available on pages 116-121.)

h. Skim this [Facebook case study](https://www.facebook.com/notes/us-politics-on-facebook/case-study-reaching-voters-with-facebook-ads-vote-no-on-8/10150257619200882) and consider two claims they make reprinted below. Why might their results differ from Broockman and Green's? Please be specific and provide examples.

+ "There was a 19 percent difference in the way people voted in areas where Facebook Ads ran versus areas where the ads did not run."

- The adds were targeted towards the two most populus counties in the state. This experiment fails to compare apples-to-apples becasuse in addition to the treatement, the voters in those two counties also had the impact of living in more populous counties than the rest of the state.

+ "In the areas where the ads ran, people with the most online ad exposure were 17 percent more likely to vote against the proposition than those with the least."

2 Peruvian Recycling

Look at [this article](https://drive.google.com/file/d/0BxwM1dZBYvxBVzQtQW9nbmd2NGM/view?usp=sharing) about encouraging recycling in Peru. The paper contains two experiments, a "participation study" and a "participation intensity study." In this problem, we will focus on the latter study, whose results are contained in Table 4 in this problem. You will need to read the relevant section of the paper (starting on page 20 of the manuscript) in order to understand the experimental design and variables. (*Note that "indicator variable" is a synonym for "dummy variable," in case you haven't seen this language before.*)

a. In Column 3 of Table 4A, what is the estimated ATE of providing a recycling bin on the average weight of recyclables turned in per household per week, during the six-week treatment period? Provide a 95

The ATE is 0.187kg more material is recycled by households recieving a recycling bin. With a standard error of 0.032kg, there is a 95% confidence interval around the ATE of 0.124kg to 0.250kg more recyled. This result is statistically significant.

b. In Column 3 of Table 4A, what is the estimated ATE of sending a text message reminder on the average weight of recyclables turned in per household per week? Provide a 95

The ATE is 0.024kg less material is recycled by households recieving a message. With a standard error of 0.039kg, there is a 95% confidence interval around the ATE of 0.100kg less to 0.052kg more recycled. This result is not statistically significant.

c. Which outcome measures in Table 4A show statistically significant effects (at the 5

The following outcome measures were effected by the recycling bin treatment at or above 95% confidence: Percentage of visits turned in bag, Avg. no. of bins turned in per week, Avg. weight (in kg) of recyclables turned in per week, Avg. market value of recyclables given per week. One other outcome variable, Avg. percentage of contamination per week, was not effected by this treatment to a statistically significant degree.

d. Which outcome measures in Table 4A show statistically significant effects (at the 5

None of the outcome measure were effected by the text message treatment to a statistically significant degree.

e. Suppose that, during the two weeks before treatment, household A turns in 2kg per week more recyclables than household B does, and suppose that both households are otherwise identical (including being in the same treatment group). From the model, how much more recycling do we predict household A to have than household B, per week, during the six weeks of treatment? Provide only a point estimate, as the confidence interval would be a bit complicated. This question is designed to test your understanding of slope coefficients in regression.

The recyclable weight coveriate has a slope coefficient of 0.281. So house A that turned in 2kg more recyclables during the baseline measurement would be expected to turn in $0.281 * 2kg = 0.562kg$ more recyclables per week than house B during the following 6 weeks of the experiment.

f. Suppose that the variable "percentage of visits turned in bag, baseline" had been left out of the regression reported in Column 1. What would you expect to happen to the results on providing a recycling bin? Would you expect an increase or decrease in the estimated ATE? Would you expect an increase or decrease in the standard error? Explain your reasoning.

In the case of the percentage of visits turned in bag covariate, the effect coefficient is nearly a full order of magnitude (10x) larger than the bin treatment variable. If the covariate was removed from the regression, then all of that variance moves back into the error term. This would increase the variance in the measurements of outcomes, which would likely reduce the size of the treatment effect and increase standard error.

g. In column 1 of Table 4A, would you say the variable "has cell phone" is a bad control? Explain your reasoning.

"has cell phone" is a good control. It would be a bad control if it was effected by the experiment in any way, but it appears to be measured before the experiment is conducted and there is no reason to believe that

giving a person a recycling bin or sending them a text would cause them to get a phone. On the other hand, not having a phone would effect whether or not the subject recieved a treatment of a text message. We are interested in knowing the effect of a text message conditional on whether or not the subject recieved the text, so "has cell phone" is a useful covariate in this regression.

h. If we were to remove the "has cell phone" variable from the regression, what would you expect to happen to the coefficient on "Any SMS message"? Would it go up or down? Explain your reasoning.

3 Multifactor Experiments

Staying with the same experiment, now lets think about multifactor experiments.

a. What is the full experimental design for this experiment? Tell us the dimensions, such as 2x2x3. (Hint: the full results appear in Panel 4B.)

This is a 3x3 experiment. Bin Treatment: + Subjects recieve a free bin with informational sticker + a free bin without a sticker + or no bin (control) SMS Treatment: + A personalized, weekly SMS message reminding the subject to recycle + a generic SMS reminder + or no SMS message (control)

b. In the results of Table 4B, describe the baseline category. That is, in English, how would you describe the attributes of the group of people for whom all dummy variables are equal to zero?

The group with all dummy variables false in table 4B are the control group subjects without cell phones. They did not recieve a recycling bin or a text message, but they were part of the study. For any of the outcome measures, the expected value for this group would be the same as the intercept (which isn't given in the table) plus the continuous covaraites below.

c. In column (1) of Table 4B, interpret the magnitude of the coefficient on "bin without sticker." What does it mean?

The coefficient for "bin without sticker" is 0.035. The positive sign indicates that the "bin without sticker" treatment produced an increase in the percentage of visits at which a subject turned in a bag of recyclables. The magnitude of 0.035 shows that the treatment group only turned in a bag per person at a rate of 0.035 greater than the control group. While small, this increase is enough to be statistically significant.

d. In column (1) of Table 4B, which seems to have a stronger treatment effect, the recycling bin with message sticker, or the recycling bin without sticker? How large is the magnitude of the estimated difference?

The recycling bin with a sticker has a larger treatment effect (0.055) than the bin without a sticker (0.035). The difference between these two treatment effects is 0.020.

e. Is this difference you just described statistically significant? Explain which piece of information in the table allows you to answer this question.

No it is not significant at the 95% level. 0.020 is less than 1.96*0.015, where 0.15 is the standard error of both original treatment effects.

f. Notice that Table 4C is described as results from "fully saturated" models. What does this mean? Looking at the list of variables in the table, explain in what sense the model is "saturated."

4 Now! Do it with data

```
d <- read.dta("./data/karlan_data_subset_for_class.dta")
head(d)
```

```
##   street havecell avg_bins_treat base_avg_bins_treat bin sms bin_s bin_g
## 1      7        1      1.0416666               0.750   1   1     1     0
## 2      7        1      0.0000000               0.000   0   1     0     0
## 3      7        1      0.7500000               0.500   0   0     0     0
## 4      7        1      0.5416667               0.500   0   0     0     0
## 5      6        1      0.9583333               0.375   1   0     0     1
## 6      8        0      0.2083333               0.000   1   0     0     1
##   sms_p sms_g
## 1     0     1
## 2     1     0
## 3     0     0
## 4     0     0
## 5     0     0
## 6     0     0
```
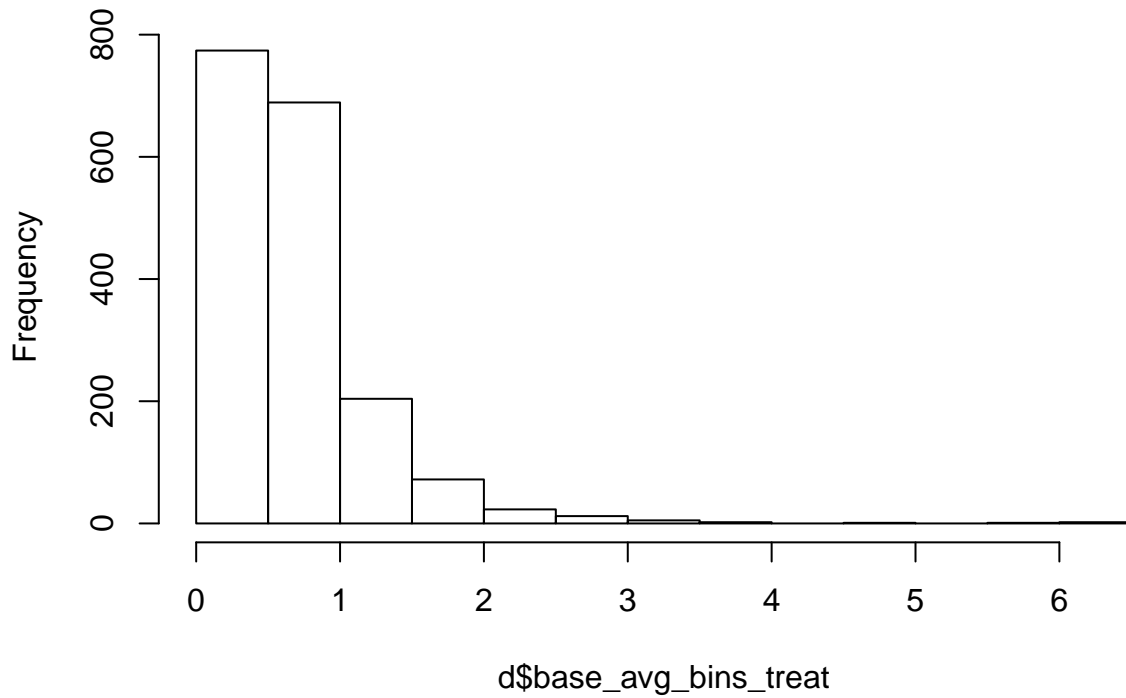
```
# distribution of continuous variables
hist(d$avg_bins_treat)
```
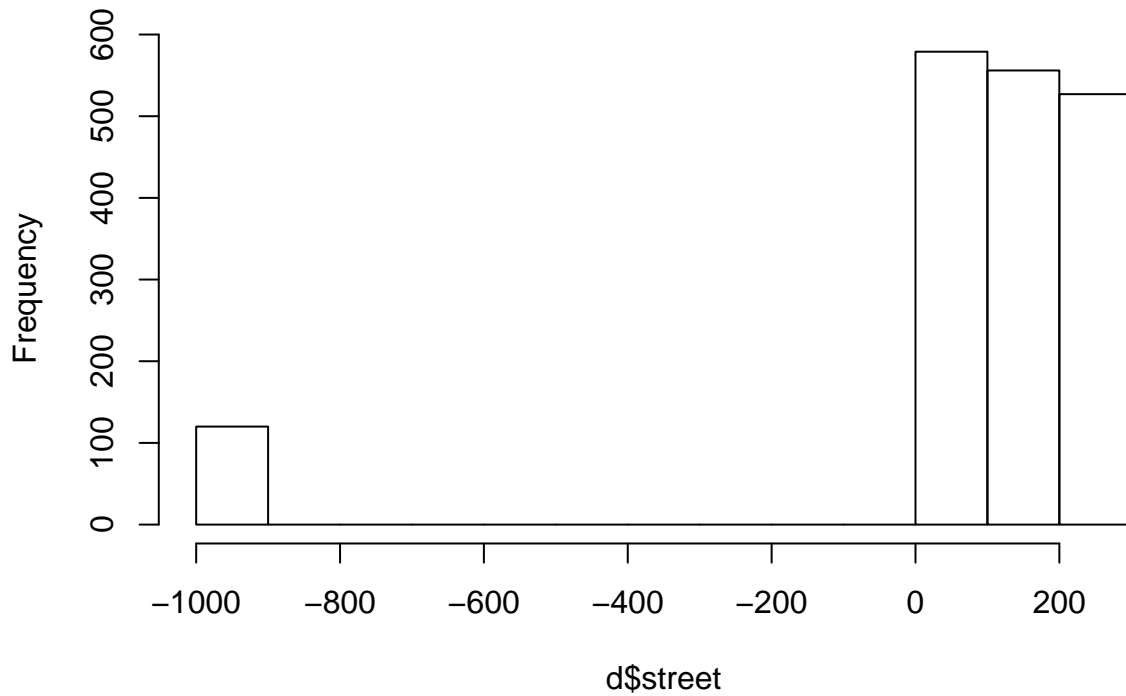
Histogram of d$avg_bins_treat
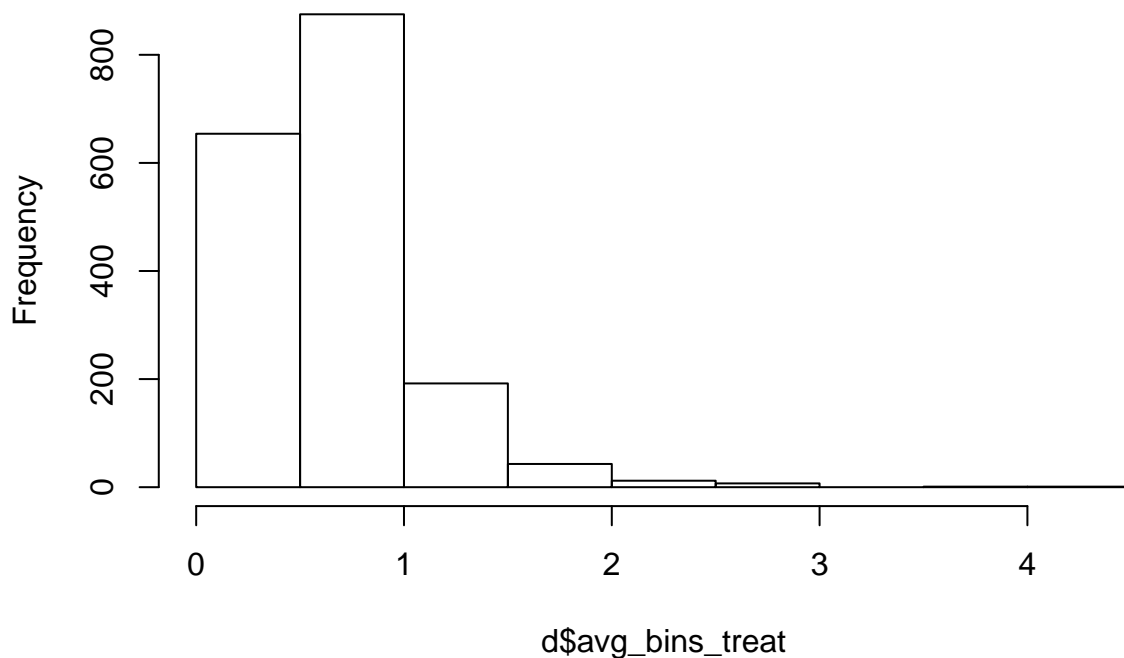
```
hist(d$base_avg_bins_treat)
```

Histogram of d$base_avg_bins_treat


```
hist(d$street)
```

Histogram of d$street

```
# which dummy variables have NAs
unique(d$havecell)
```

```
## [1]  1  0 NA
```

```
unique(d$bin)
```

```
## [1] 1 0
```

```
unique(d$sms)
```

```
## [1] 1 0
```

```
unique(d$bin_s)
```

```
## [1] 1 0
```

```
unique(d$bin_g)
```

```
## [1] 0 1
```

```
unique(d$sms_p)
```

```
## [1] 0 1
```

```
unique(d$sms_g)
```

```
## [1] 1 0
```

```
# remove na rows
d <- d[rowSums(!is.na(d)) > 0,]
```

```
## Do some quick exploratory data analysis with this data. There are some values in this data that seem
```

Everything seems resonable except for about 100 values in the street field which are NA. Those values were removed. There are also 3 NA values in "have cell" which were removed as well.

a. For simplicity, let's start by measuring the effect of providing a recycling bin, ignoring the SMS message treatment (and ignoring whether there was a sticker on the bin or not). Run a regression of Y on only the bin treatment dummy, so you estimate a simple difference in means. Provide a 95

```
lm_4a <- lm(avg_bins_treat ~ bin, d)
lm <- lm_4a
lm$robust_se <- sqrt(diag(vcovHC(lm)))
for(ii in 2:length(lm$coefficients)){
  print(sprintf("The 95%% confidence interval (robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (robust) for bin is between 0.095 and 0.176"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

b. Now add the pre-treatment value of Y as a covariate. Provide a 95

Table 6:

	Dependent variable:
	avg_bins_treat
bin	0.135***
	(0.021)
Constant	0.635***
	(0.011)
Observations	1,785
R^2	0.024
Adjusted R^2	0.024
Residual Std. Error	0.405 (df = 1783)
F Statistic	44.516*** (df = 1; 1783)

Note:	*p<0.1; **p<0.05; ***p<0.01

```r
lm_4b <- lm(avg_bins_treat ~ bin + base_avg_bins_treat, d)
lm <- lm_4b
lm$robust_se <- sqrt(diag(vcovHC(lm)))
for(ii in 2:length(lm$coefficients)){
  print(sprintf("The 95%% confidence interval (robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (robust) for bin is between 0.091 and 0.158"
## [1] "The 95% confidence interval (robust) for base_avg_bins_treat is between 0.334 and 0.452"
```

```r
stargazer(lm,se=list(lm$robust_se), header=F)
```

c. Now add the street fixed effects. (You'll need to use the R command factor().) Provide a 95

```r
d$street <- as.factor(d$street)
lm_4c <- lm(avg_bins_treat ~ bin + base_avg_bins_treat, d)
lm <- lm_4c
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ street)))
for(ii in list(2)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for bin is between 0.090 and 0.159"
```

```r
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 7:

	Dependent variable:
	avg_bins_treat
bin	0.125***
	(0.017)
base_avg_bins_treat	0.393***
	(0.030)
Constant	0.350***
	(0.021)
Observations	1,785
R^2	0.342
Adjusted R^2	0.342
Residual Std. Error	0.333 (df = 1782)
F Statistic	463.891*** (df = 2; 1782)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 8:

	Dependent variable:
	avg_bins_treat
bin	0.125***
	(0.018)
base_avg_bins_treat	0.393***
	(0.030)
Constant	0.350***
	(0.021)
Observations	1,785
R^2	0.342
Adjusted R^2	0.342
Residual Std. Error	0.333 (df = 1782)
F Statistic	463.891*** (df = 2; 1782)
Note:	*p<0.1; **p<0.05; ***p<0.01

d. Recall that the authors described their experiment as "stratified at the street level," which is a synonym for blocking by street. Explain why the confidence interval with fixed effects does not differ much from the previous one.

e. Perhaps having a cell phone helps explain the level of recycling behavior. Instead of "has cell phone," we find it easier to interpret the coefficient if we define the variable " no cell phone." Give the R command to define this new variable, which equals one minus the "has cell phone" variable in the authors' data set. Use "no cell phone" instead of "has cell phone" in subsequent regressions with this dataset.

```
d$no_cell_phone <- 1-d$havecell
d$havecell <- NULL
head(d)
```

street avg_bins_treat base_avg_bins_treat bin sms bin_s bin_g sms_p 1 7 1.0416666 0.750 1 1 1 0 0 2 7 0.0000000 0.000 0 1 0 0 1 3 7 0.7500000 0.500 0 0 0 0 0 4 7 0.5416667 0.500 0 0 0 0 0 5 6 0.9583333 0.375 1 0 0 1 0 6 8 0.2083333 0.000 1 0 0 1 0 sms_g no_cell_phone 1 1 0 2 0 0 3 0 0 4 0 0 5 0 0 6 0 1

f. Now add "no cell phone" as a covariate to the previous regression. Provide a 95

```
lm_4f <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + no_cell_phone, d)
lm <- lm_4f
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ street)))
for(ii in list(2)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for bin is between 0.091 and 0.160"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

g. Now let's add in the SMS treatment. Re-run the previous regression with "any SMS" included. You should get the same results as in Table 4A. Provide a 95

```
lm_4g <- lm(avg_bins_treat ~ bin + sms + base_avg_bins_treat + no_cell_phone, d)
lm <- lm_4g
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ street)))
for(ii in list(2,3)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for bin is between 0.091 and 0.160"
## [1] "The 95% confidence interval (cluster-robust) for sms is between -0.075 and 0.014"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 9:

	Dependent variable:
	avg_bins_treat
bin	0.126***
	(0.017)
base_avg_bins_treat	0.392***
	(0.030)
no_cell_phone	−0.050***
	(0.016)
Constant	0.370***
	(0.022)
Observations	1,784
R^2	0.346
Adjusted R^2	0.345
Residual Std. Error	0.332 (df = 1780)
F Statistic	313.598*** (df = 3; 1780)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 10:

	Dependent variable:
	avg_bins_treat
bin	0.126***
	(0.017)
sms	−0.030
	(0.023)
base_avg_bins_treat	0.391***
	(0.030)
no_cell_phone	−0.065***
	(0.021)
Constant	0.387***
	(0.026)
Observations	1,784
R^2	0.347
Adjusted R^2	0.345
Residual Std. Error	0.332 (df = 1779)
F Statistic	235.903*** (df = 4; 1779)
Note:	*p<0.1; **p<0.05; ***p<0.01

h. Now reproduce the results of column 2 in Table 4B, estimating separate treatment effects for the two types of SMS treatments and the two types of recycling-bin treatments. Provide a 95

```
lm_4g <- lm(avg_bins_treat ~ d$bin_g + d$bin_s + d$sms_g + d$sms_p + base_avg_bins_treat + no_cell_phon
lm <- lm_4g
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ street)))
for(ii in list(2,3,4,5)){
  print(sprintf("The 95%% confidence interval (cluster-robust) for %s is between %.3f and %.3f",
    names(lm$coefficients[ii]),
    lm$coefficients[ii]-lm$robust_se[ii]*1.96,
    lm$coefficients[ii]+lm$robust_se[ii]*1.96
    ))
}
```

```
## [1] "The 95% confidence interval (cluster-robust) for d$bin_g is between 0.065 and 0.156"
## [1] "The 95% confidence interval (cluster-robust) for d$bin_s is between 0.098 and 0.184"
## [1] "The 95% confidence interval (cluster-robust) for d$sms_g is between -0.066 and 0.037"
## [1] "The 95% confidence interval (cluster-robust) for d$sms_p is between -0.095 and 0.006"
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

5 A Final Practice Problem

Now for a fictional scenario. An emergency two-week randomized controlled trial of the experimental drug ZMapp is conducted to treat Ebola. (The control represents the usual standard of care for patients identified with Ebola, while the treatment is the usual standard of care plus the drug.)

Here are the (fake) data.

```
d <- read.csv("./data/ebola_rct2.csv")
head(d)
```

```
##   temperature_day0 vomiting_day0 treat_zmapp temperature_day14
## 1         99.53168             1           0          98.62634
## 2         97.37372             0           0          98.03251
## 3         97.00747             0           1          97.93340
## 4         99.74761             1           0          98.40457
## 5         99.57559             1           1          99.31678
## 6         98.28889             1           1          99.82623
##   vomiting_day14 male
## 1              1    0
## 2              1    0
## 3              0    1
## 4              1    0
## 5              1    0
## 6              1    1
```

You are asked to analyze it. Patients' temperature and whether they are vomiting is recorded on day 0 of the experiment, then ZMapp is administered to patients in the treatment group on day 1. Vomiting and temperature is again recorded on day 14.

a. Without using any covariates, answer this question with regression: What is the estimated effect of ZMapp (with standard error in parentheses) on whether someone was vomiting on day 14? What is the p-value

Table 11:

	Dependent variable:
	avg_bins_treat
bin_g	0.111***
	(0.023)
bin_s	0.141***
	(0.022)
sms_g	−0.014
	(0.026)
sms_p	−0.045*
	(0.026)
base_avg_bins_treat	0.392***
	(0.030)
no_cell_phone	−0.065***
	(0.021)
Constant	0.386***
	(0.026)
Observations	1,784
R^2	0.347
Adjusted R^2	0.345
Residual Std. Error	0.332 (df = 1777)
F Statistic	157.698*** (df = 6; 1777)
Note:	*p<0.1; **p<0.05; ***p<0.01

associated with this estimate?

```
lm_5a <- lm(vomiting_day14 ~ treat_zmapp, d)
lm <- lm_5a
lm$robust_se <- sqrt(diag(vcovHC(lm)))

stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 12:

	Dependent variable:
	vomiting__day14
treat_zmapp	−0.238***
	(0.091)
Constant	0.847***
	(0.048)
Observations	100
R^2	0.073
Adjusted R^2	0.063
Residual Std. Error	0.421 (df = 98)
F Statistic	7.705*** (df = 1; 98)
Note:	*p<0.1; **p<0.05; ***p<0.01

b. Add covariates for vomiting on day 0 and patient temperature on day 0 to the regression from part (a) and report the ATE (with standard error). Also report the p-value.

```
lm_5b <- lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0, d)
lm <- lm_5b
lm$robust_se <- sqrt(diag(vcovHC(lm)))

stargazer(lm,se=list(lm$robust_se), header=F)
```

c. Do you prefer the estimate of the ATE reported in part (a) or part (b)? Why?

d. The regression from part (b) suggests that temperature is highly predictive of vomiting. Also include temperature on day 14 as a covariate in the regression from part (b) and report the ATE, the standard error, and the p-value.

```
lm_5c <- lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + temperature_day14, d)
lm <- lm_5c
lm$robust_se <- sqrt(diag(vcovHC(lm)))

stargazer(lm,se=list(lm$robust_se), header=F)
```

e. Do you prefer the estimate of the ATE reported in part (b) or part (d)? Why?

B is the preferable prediction because temperature_day14 is a bad control. It is measured after the treatment

Table 13:

	Dependent variable:
	vomiting__day14
treat_zmapp	-0.166^{**}
	(0.082)
vomiting__day0	0.065
	(0.178)
temperature_day0	0.206^{***}
	(0.078)
Constant	-19.470^{**}
	(7.608)
Observations	100
R^2	0.311
Adjusted R^2	0.290
Residual Std. Error	0.367 (df = 96)
F Statistic	14.447^{***} (df = 3; 96)
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 14:

	Dependent variable:
	vomiting__day14
treat_zmapp	-0.120
	(0.086)
vomiting__day0	0.046
	(0.173)
temperature_day0	0.177^{**}
	(0.077)
temperature_day14	0.060^{**}
	(0.026)
Constant	-22.592^{***}
	(7.746)
Observations	100
R^2	0.340
Adjusted R^2	0.312
Residual Std. Error	0.361 (df = 95)
F Statistic	12.244^{***} (df = 4; 95)
Note:	*p<0.1; **p<0.05; ***p<0.01

is given and it is extreamly likely that a persons temperature would be effected by medicine.

f. Now let's switch from the outcome of vomiting to the outcome of temperature, and use the same regression covariates as in part (b). Test the hypothesis that ZMapp is especially likely to reduce men's temperatures, as compared to women's, and describe how you did so. What do the results suggest?

```
lm_5f <- lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0, d)
lm <- lm_5f
lm$robust_se <- sqrt(diag(cluster.vcov(lm, ~ male)))
```

```
stargazer(lm,se=list(lm$robust_se), header=F)
```

Table 15:

	Dependent variable:
	vomiting_day14
treat_zmapp	−0.166***
	(0.046)
vomiting_day0	0.065
	(0.309)
temperature_day0	0.206*
	(0.112)
Constant	−19.470*
	(10.791)
Observations	100
R^2	0.311
Adjusted R^2	0.290
Residual Std. Error	0.367 (df = 96)
F Statistic	14.447*** (df = 3; 96)
Note:	*p<0.1; **p<0.05; ***p<0.01

g. Suppose that you had not run the regression in part (f). Instead, you speak with a colleague to learn about heterogenous treatment effects. This colleague has access to a non-anonymized version of the same dataset and reports that he had looked at heterogenous effects of the ZMapp treatment by each of 10,000 different covariates to examine whether each predicted the effectiveness of ZMapp on each of 2,000 different indicators of health, for 20,000,000 different regressions in total. Across these 20,000,000 regressions your colleague ran, the treatment's interaction with gender on the outcome of temperature is the only heterogenous treatment effect that he found to be statistically significant. He reasons that this shows the importance of gender for understanding the effectiveness of the drug, because nothing else seemed to indicate why it worked. Bolstering his confidence, after looking at the data, he also returned to his medical textbooks and built a theory about why ZMapp interacts with processes only present in men to cure. Another doctor, unfamiliar with the data, hears his theory and finds it plausible. How likely do you think it is ZMapp works especially well for curing Ebola in men, and why? (This question is conceptual can be answered without performing any computation.)

h. Now, imagine that what described in part (g) did not happen, but that you had tested this heterogeneous treatment effect, and only this heterogeneous treatment effect, of your own accord. Would you be more or less inclined to believe that the heterogeneous treatment effect really exists? Why?

i. Another colleague proposes that being of African descent causes one to be more likely to get Ebola. He asks you what ideal experiment would answer this question. What would you tell him? (*Hint: refer to Chapter 1 of Mostly Harmless Econometrics.*)