Problem Set #3

Experiment Design: Alex Smith

February 11, 2016

---

1 Replicate Results

Skim Broockman and Green's paper on the effects of Facebook ads and download an anonymized version of the data for Facebook users only.

```
# set the working directory, load the data and load the library datatable
library(data.table)
setwd("/Users/Alex/Documents/Berkeley/1601Spring/W241/PS3")
facebook_df <- read.csv("1BrookmanGreen.csv")
facebook_dt <- data.table(facebook_df)
head(facebook_dt)
```

```
##    studyno treat_ad                   cluster name_recall
## 1:       2        0 Study 2, Cluster Number 1           0
## 2:       2        0 Study 2, Cluster Number 2           1
## 3:       2        0 Study 2, Cluster Number 3           0
## 4:       2        0 Study 2, Cluster Number 4           1
## 5:       2        1 Study 2, Cluster Number 7           1
## 6:       2        1 Study 2, Cluster Number 7           0
##    positive_impression
## 1:                   0
## 2:                   0
## 3:                   0
## 4:                   0
## 5:                   1
## 6:                   0
```

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 variable is "name_recall").

- **Note**: Ignore the blocking the article mentions throughout this problem.
- **Note**: You will estimate something different than is reported in the study.

```
# subset the data into just study 1
fb_1 <- facebook_dt[studyno == 1]

# take just the columns of interest
fb_1 <- fb_1[,c("treat_ad","cluster","name_recall"), with=FALSE]

# check for any NAs in this dataset
fb_1$NAs <- apply(fb_1, 1, function(x) sum(is.na(x)) )
```

```r
# remove all the NAs from our limited dataset
fb_1 <- fb_1[NAs == 0]

# convert the cluster data into factors
fb_1$cluster <- as.factor(fb_1$cluster)

# create a linear model to describe the effect of an ad on a candidate's
# name recognition
linear_fb_1 <- lm(name_recall ~ treat_ad, data = fb_1)

# pull the 95% confidence interval
fb_1_bottom <- confint(linear_fb_1, level = 0.95)[2]
fb_1_top <- confint(linear_fb_1, level = 0.95)[4]
```

The 95% confidence interval for the effect of the ad on candidate name recognition is from **-0.0510177** to **0.0314219**.

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

The clusters in Broockman and Green's study are different demographic groups that Facebook allows for targeting. For example, a cluster of "24 year old males in San Francisco" might be assigned to treatment and a cluster of "24 year old males in Palo Alto" might be assigned to control. Due to the privacy limits of Facebook ad targeting, the researchers could only target broad demographic clusters and not individuals for random assignment.

Taking clustering into account will increase the standard errors because then we must consider the between cluster variance rather than the between individual variance. This between cluster variance is likely to be large because the clusters are by design, internally, demographically similar.

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.

```r
# load the libraries required by the cl function
library("sandwich")
library("lmtest")
```

```
## Loading required package: zoo

##
## Attaching package: 'zoo'

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

```r
cl <- function(fm, cluster){
  ## This function takes a fit model `fm` and a cluster `df$variable`
  ## and returns the cluster-correct standard errors.
  ##
```

```r
  ## This is really little more than an application of the sandwich
  ## estimator inside each of the clusters, but it isn't alwasy intuitive
  ## what is happening.
  ##
  ## Adapted from Mahmood Arai & Drew Dimmery
  ##
  ## Note: - This WON'T work with missing data; different vector lenghts
  ##       - I'd strongly recommend that you read your data the first time
  ##         without converting it to a factor.
  ##       - Instead, convert it to a factor after you have read-in the
  ##         data.

  require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
    coeftest(fm, vcovCL)
}

## Usage:
# my.lm <- lm(y~x, data)  # you're accustomed to doing this
# cl(my.lm, data$cluster) # cluster is a factor variable in the
                          # data frame that identifies the clusters

# re-factor the data to ensure that we don't have extraneous clusters
fb_1$cluster <- factor(fb_1$cluster)

# code to check that all clusters have been properly removed
# commented out here, but useful reference and check
# sum(is.na(tapply(fb_1$treat_ad, fb_1$cluster, length)))

# compute the effect taking into account clustering
cluster_fb_1 <- cl(linear_fb_1, fb_1$cluster)

# store the treatment effect
fb_1_TE <- cluster_fb_1[2]

# store the standard error
fb_1_SE <- cluster_fb_1[4]

# set the constant multiplier for calculating the 95% confidence interval.
SE_MULTIPLIER <- 1.96

# store the 95% confidence interval bounds
fb_1_top_cluster <- fb_1_TE + SE_MULTIPLIER * fb_1_SE
fb_1_bottom_cluster <- fb_1_TE - SE_MULTIPLIER * fb_1_SE
```

For study 1, the 95% confidence interval *with* taking into account clustering is from **-0.056355** to **0.0367592**.

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

```r
# load the subsetting data for study 2 only
fb_2 <- facebook_dt[studyno == 2]

# take just the columns of interest
fb_2 <- fb_2[,c("treat_ad","cluster","name_recall"), with=FALSE]

# check for any NAs in this dataset
fb_2$NAs <- apply(fb_2, 1, function(x) sum(is.na(x)) )

# remove all the NAs so that we can compare apples to apples
# when we do cluster level analysis later
fb_2 <- fb_2[NAs == 0]

# convert the cluster data into factors
fb_2$cluster <- factor(fb_2$cluster)

# code to check that extraneous clusters have been properly removed
# sum(is.na(tapply(fb_2$treat_ad, fb_2$cluster, length)))

# create a linear model to describe the effect of an ad on a candidate's
# name recognition
linear_fb_2 <- lm(name_recall ~ treat_ad, data = fb_2)

# compute the effect taking into account clustering
cluster_fb_2 <- cl(linear_fb_2, fb_2$cluster)

# store the treatment effect
fb_2_TE <- cluster_fb_2[2]

# store the standard error
fb_2_SE <- cluster_fb_2[4]

# set the constant multiplier for calculating the 95% confidence interval.
SE_MULTIPLIER <- 1.96

# store the 95% confidence interval bounds
fb_2_top_cluster <- fb_2_TE + SE_MULTIPLIER * fb_2_SE
fb_2_bottom_cluster <- fb_2_TE - SE_MULTIPLIER * fb_2_SE
```

For study 2, the 95% confidence interval *with* taking into account clustering is from **-0.0723899** to **0.0667832**.

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 regression. What is the treatment effect estimate and associated p-value?

```r
# gather just the columns of interest
facebook_dt <- facebook_dt[,c("studyno","treat_ad","cluster","name_recall"),
                           with=FALSE]

# count the NAs by row
facebook_dt$NAs <- apply(facebook_dt, 1, function(x) sum(is.na(x)) )

# remove those rows with any NAs
```

```r
facebook_dt <- facebook_dt[NAs == 0]

# convert the clusters column into factors
facebook_dt$cluster <- factor(facebook_dt$cluster)

# run a linear regression ignoring the different between studies
linear_fb <- lm(name_recall ~ treat_ad, data = facebook_dt)

# take the effect of clustering into account
cluster_fb <- cl(linear_fb, facebook_dt$cluster)

# store the treatment effect
fb_TE <- cluster_fb[2]

# store the standard error
fb_SE <- cluster_fb[4]

# set the constant multiplier for calculating the 95% confidence interval.
SE_MULTIPLIER <- 1.96

# store the 95% confidence interval bounds
fb_top_cluster <- fb_TE + SE_MULTIPLIER * fb_SE
fb_bottom_cluster <- fb_TE - SE_MULTIPLIER * fb_SE

# store the p-value for the effect
fb_p_value <- cluster_fb[8]
```

Pooling the entire data and performing a regression *with* taking clustering into account (as in part c) results in an estimated treatment effect of **-0.1550732** (95% confidence interval from **-0.207465** to **-0.1026815**). The p-value is highly significant at near **0**, much lower than 0.05.

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?

```r
# add a new column that binarizes the study number, where study 1 will be
# 0 and study 2 will be 1
facebook_dt$study[facebook_dt$studyno == 1] <- 0
facebook_dt$study[facebook_dt$studyno == 2] <- 1

# redo the linear regression but this time control for the study number
linear_fb_controlled <- lm(name_recall ~ treat_ad + study, data = facebook_dt)

# apply the clusters effect function
cluster_fb_controlled <- cl(linear_fb_controlled, facebook_dt$cluster)

# store the treatment effect
fb_TE_controlled <- cluster_fb_controlled[2]

# store the standard error
fb_SE_controlled <- cluster_fb_controlled[5]

# set the constant multiplier for calculating the 95% confidence interval.
SE_MULTIPLIER <- 1.96
```

```r
# store the 95% confidence interval bounds
fb_top_cluster_controlled <- fb_TE_controlled + SE_MULTIPLIER * fb_SE_controlled
fb_bottom_cluster_controlled <- fb_TE_controlled - SE_MULTIPLIER * fb_SE_controlled

# store the p-value for the effect
fb_p_value_controlled <- cluster_fb_controlled[11]
```

Pooling the entire data and performing a regression *with* taking clustering into account (as in part c) and controlling for the study with a dummy variable results in an estimated treatment effect of **-0.0067752** (95% confidence interval from **-0.0467895** to **0.033239**). The p-value is not significant at **0.7400138**.

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.)

The results from part (e) are biased because the the study is really from 2 different experiments that should not just be analyzed as one. In part (f) we control based on the study number. Controlling by the study number reduces the standard error and thereby improves the accuracy of our results because we compare each value to the mean of its group rather than to the mean of all the data.

h. Skim this Facebook case study 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."
- "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."

This case study likely had different results that Broockman and Green's because it was not a random experiment. They chose to target their ads to the two most populous counties in Florida. They compared the voting results of the people in these counties with the people in all other counties. There may be something very different about the people who live in the most populous counties compared to those who live in the rest of the state. These people may have voted against the proposition regardless of the Facebook ads. Because this case study failed to use random assignment, it cannot claim causality between the ads and the voting patterns.

2 Peruvian Recycling

Look at this article 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% confidence interval.

```r
# load the constants from table 4A for estimate ATE of providing a recycling bin
bin_ATE <- 0.187
bin_SE <- 0.032

# set the constant multiplier for calculating the 95% confidence interval.
SE_MULTIPLIER <- 1.96

# caculate the bottom and top of the confidence interval
bin_95_bottom <- bin_ATE - SE_MULTIPLIER * bin_SE
bin_95_top <- bin_ATE + SE_MULTIPLIER * bin_SE
```

The ATE of providing a recycling bin on the average weight of recyclable turned in per household per week is **0.187**. The 95% confidence interval stretches from **0.12428** to **0.24972**.

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% confidence interval.

```r
# load the constants from table 4A for estimating the ATE
sms_ATE <- -0.024
sms_SE <- 0.039

# caculate the bottom and top of the confidence interval
sms_95_bottom <- sms_ATE - SE_MULTIPLIER * sms_SE
sms_95_top <- sms_ATE + SE_MULTIPLIER * sms_SE
```

The ATE of providing a recycling bin on the average weight of recyclable turned in per household per week is **-0.024**. The 95% confidence interval stretches from **-0.10044** to **0.05244**.

c. Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of providing a recycling bin?

In table 4A, the outcome measures that show a statistically significant effect of providing a recycling bin are **percentage of visits turned in bag**, **average number of bins turned in per week**, **average weight (in kg) of recyclables turned in per week**, and **average market value of recyclables given per week**.

d. Which outcome measures in Table 4A show statistically significant effects (at the 5% level) of sending text messages?

None of the outcome measures in Table 4A show statistically significant effects (at the 5% level) of sending text messages.

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.

```r
# the baseline of recyclables turned in
past_interact <- 0.281

# household A's greater recycling
house_A <- 2

# more recycling prediction
more_house_A <- house_A * past_interact
```

Household A will produce **0.562** kg of recycling than household B based on the baseline coefficient of **0.281**.

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.

I would expect my estimated ATE to stay the same because the treatment group is randomly assigned. My pre-treatrement percent visits should have no correlation with the treatment and therefore no change on the estimed ATE. I would expect my standard error to increase as I control for less of the variability in my data. By adding the baseline visits variable, I'm able to control for those who come to program with high rates of recycling and those who come with low rates. Should I get rid of the baseline variable, my estimate would stay the same, but my percision would decrease.

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

No, I would not consider the variable "has cell phone" to be a bad control. A bad control is a post-treatment variable that is added to the regression and impacted by the predictor variables. However, in this study, the researchers calculated "has cell phone" before beginning the study so that they could randomly assign people to different text messaging groups (see Table 1, page 8).

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.

If I were to remove the "has cell phone" variable from the regression, I would expect the coefficient on "any SMS message" to rise because with the "has cell phone" variable, I only compare the effects of "any SMS message" among those individuals with cell phones. When I remove the "has cell phone" variable, I compare those who got the "any SMS message" blast to all study participants regardless of whether those participants had cell phones. It would be impossible for those without cell phones to show any effect because they would not have gotten the SMS blast.

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.)

I think it's most helpful to consider this design as two studies because individuals without a cell phone were not given the opportunity to enroll in the SMS blast treatment.

For individuals without cell phones, we have only 3 possible groups: "bins without sticker", "bins with sticker", and "no bin". This would produce a 3 x 1 table.

For individuals with cell phones, we have 3 possible possible groups for the bins: "bins without sticker", "bins with sticker", and "no bin" and 3 possible groups for the SMS blast: "no SMS message", "personal SMS message", "generic SMS message". This would produce a 3 x 3 table.

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 of people that make up the baseline category, those people for whom all dummy variables equal zero, are those people who did *not* get a bin (either with or without a sticker) and who did *not* receive a SMS message (either personal or generic).

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

The coefficient of 0.035 for "bin without sticker" means that those individuals who received a recycling bin without a sticker had a rate of turning in their recycling 3.5 percentage points greater than those individuals who did not receive a bin.

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 the message sticker has the stronger effect than the recycling bin without a sticker. The recycling bin with a sticker has a coefficient of 0.055, while the recycling bin without a sticker has a coefficient of 0.035. The recycling bin with a sticker also produced a more statistically significant effect.

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

```r
# calcualte the 95% confidence interval for both coefficients

# 95% confidence for the bin with sticker
bin_sticker_TE <- 0.055
bin_sticker_SE <- 0.015
bin_sticker_top <- bin_sticker_TE + SE_MULTIPLIER * bin_sticker_SE
bin_sticker_bottom <- bin_sticker_TE - SE_MULTIPLIER * bin_sticker_SE

# 95% confidence for the bin without sticker
bin_nosticker_TE <- 0.035
bin_nosticker_SE <- 0.015
bin_nosticker_top <- bin_nosticker_TE + SE_MULTIPLIER * bin_nosticker_SE
bin_nosticker_bottom <- bin_nosticker_TE - SE_MULTIPLIER * bin_nosticker_SE
```

The difference I've described is not statistically significant. I use the standard errors (provided in the table under the coefficients in parantheses) to calculate 95% confidence intervals. When I do so, I find that the 95% confidence intervals overlap. The 95% confidence interval for bin with sticker is from **0.0256** to **0.0844**. The 95% confidence interval for bin without sticker is from **0.0056** to **0.0644**.

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."

A fully saturated model accounts for all variables and variable combinations. This model is saturated because it includes every combination of treatment variables (bins and SMS messages). For example, it includes "generic SMS message + bin", "generic SMS message + bin with sticker", "generic SMS message + no bin", etc. It computes coefficients for every combination of SMS message and recycling bin.

4 Now! Do it with data

Download the data set for the recycling study in the previous problem, obtained from the authors. We'll be focusing on the outcome variable Y="number of bins turned in per week" (avg_bins_treat).

```
recycle <- read.csv("3PeruvianRecycling.csv")
recycle <- recycle[,-1]
head(recycle)
```

```
##   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
```

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% confidence interval for the treatment effect.

```
# create a linear regression model where the outcome is avg_bins_treat and the
# predictor is whether the individual received a bin or not
bin_effect <- lm(avg_bins_treat ~ bin, data = recycle)

# store the standard error
bin_effect_SE <- coef(summary(bin_effect))[4]

# store the coefficient for bin effect
bin_effect_TE <- coef(summary(bin_effect))[2]

# store the p-value
bin_effect_p <- coef(summary(bin_effect))[8]

# store the top and bottom of the 95% confidence interval
```

```r
bin_effect_top <- bin_effect_TE + SE_MULTIPLIER * bin_effect_SE
bin_effect_bottom <- bin_effect_TE - SE_MULTIPLIER * bin_effect_SE
```

The estimated average treatment for getting a recycling bin is **0.13538** with a standard error of **0.0202906**. The 95% confidence interval for this treatment effect ranges from **0.0956105** to **0.1751495**.

b. Now add the pre-treatment value of Y as a covariate. Provide a 95% confidence interval for the treatment effect. Explain how and why this confidence interval differs from the previous one.

```r
# create a linear model that also controls for pre-treatment levels
bin_effect2 <- lm(avg_bins_treat ~ bin + base_avg_bins_treat, data = recycle)

# store the standard error
bin_effect_SE2 <- coef(summary(bin_effect2))[5]

# store the coefficient for bin effect
bin_effect_TE2 <- coef(summary(bin_effect2))[2]

# store the top and bottom of the 95% confidence interval
bin_effect_top2 <- bin_effect_TE2 + SE_MULTIPLIER * bin_effect_SE2
bin_effect_bottom2 <- bin_effect_TE2 - SE_MULTIPLIER * bin_effect_SE2
```

The estimated average treatment for getting a recycling bin is **0.124693** with a standard error of **0.0166671**. The 95% confidence interval for this treatment effect ranges from **0.0920254** to **0.1573606**. The confidence interval has shrunk because now I'm explaining more of the variation in the data. In the previous model, I was comparing those who recycle a lot anyway with those who do not recycle a lot. In this model, I'm controlling for indvidiuals' pre-treatment recycling.

c. Now add the street fixed effects. (You'll need to use the R command factor().) Provide a 95% confidence interval for the treatment effect.

```r
# store the street levels as a factor
recycle$street = factor(recycle$street)

# create a linear model that also controls for street effect
bin_effect3 <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street, data = recycle)

# store the standard error
bin_effect_SE3 <- coef(summary(bin_effect3))[184]

# store the coefficient for bin effect
bin_effect_TE3 <- coef(summary(bin_effect3))[2]

# store the top and bottom of the 95% confidence interval
bin_effect_top3 <- bin_effect_TE3 + SE_MULTIPLIER * bin_effect_SE3
bin_effect_bottom3 <- bin_effect_TE3 - SE_MULTIPLIER * bin_effect_SE3
```

The estimated average treatment for getting a recycling bin is **0.1138868** with a standard error of **0.0170578**. The 95% confidence interval for this treatment effect ranges from **0.0804534** to **0.1473201**.

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.

The confidence interval with fixed effects did not differ much because in this case, the fixed effects add no predictive power for the outcome variable. Adding them has little effect.

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.

```
# create a no cell phone vector and append to the data frame
recycle$nocell = 1 - recycle$havecell
```

f. Now add "no cell phone" as a covariate to the previous regression. Provide a 95% confidence interval for the treatment effect. Explain why this confidence interval does not differ much from the previous one.

```
# create a linear model that also controls for no cell phone
bin_effect4 <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell,
                  data = recycle)

# store the standard error
bin_effect_SE4 <- coef(summary(bin_effect4))[185]

# store the coefficient for bin effect
bin_effect_TE4 <- coef(summary(bin_effect4))[2]

# store the top and bottom of the 95% confidence interval
bin_effect_top4 <- bin_effect_TE4 + SE_MULTIPLIER * bin_effect_SE4
bin_effect_bottom4 <- bin_effect_TE4 - SE_MULTIPLIER * bin_effect_SE4
```

The estimated average treatment for getting a recycling bin is **0.1151007** with a standard error of **0.017045**. The 95% confidence interval for this treatment effect ranges from **0.0816926** to **0.1485089**. The variable no cell phone does not change the confidence intervals much because it hardly explains any more of the treatment effect. The variance in treatment effects can already largely be explained by the existing variables. Adding the no cell phone variable adds little value.

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% confidence interval for the treatment effect of the recycling bin. Explain why this confidence interval does not differ much from the previous one.

```
# create a linear model that also controls for SMS treatment
bin_effect5 <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell + sms,
                  data = recycle)

# store the standard error
bin_effect_SE5 <- coef(summary(bin_effect5))[186]

# store the coefficient for bin effect
bin_effect_TE5 <- coef(summary(bin_effect5))[2]

# store the top and bottom of the 95% confidence interval
bin_effect_top5 <- bin_effect_TE5 + SE_MULTIPLIER * bin_effect_SE5
bin_effect_bottom5 <- bin_effect_TE5 - SE_MULTIPLIER * bin_effect_SE5
```

The estimated average treatment for getting a recycling bin is **0.1150536** with a standard error of **0.017051**. The 95% confidence interval for this treatment effect ranges from **0.0816336** to **0.1484737**. The variable SMS treatment does not change the confidence intervals much because, by now, we've already explained almost all of the variance in treatment effect. Most of the variance in the data can be explained by the already included variables. Each variable we add, explains less and adds less predictive power to the model.

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% confidence interval for the effect of the unadorned recycling bin. Explain how your answer differs from that in part (g), and explain why you think it differs.

```r
# load the library lmtest to use the coeftest function
library("lmtest")

# create a linear model modeled after column 2 in Talbe 4B
bin_effect6 <- lm(avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g +
                    nocell + base_avg_bins_treat + street,
                  data = recycle)

# print out the coefficients to show how closely we matched Table 2b
round(coeftest(bin_effect6)[-c(8:200),-c(3:5)],digits = 3)
```

```
##                     Estimate Std. Error
## (Intercept)            0.385      0.034
## bin_s                  0.128      0.022
## bin_g                  0.103      0.022
## sms_p                 -0.008      0.025
## sms_g                  0.020      0.025
## nocell                -0.046      0.020
## base_avg_bins_treat    0.374      0.014
```

```r
# store the coefficient for bin effect for the generic bin
bin_effect_TE6 <- coef(summary(bin_effect6))[3]

# store the standard error
bin_effect_SE6 <- coef(summary(bin_effect6))[189]

# store the top and bottom of the 95% confidence interval
bin_effect_top6 <- bin_effect_TE6 + SE_MULTIPLIER * bin_effect_SE6
bin_effect_bottom6 <- bin_effect_TE6 - SE_MULTIPLIER * bin_effect_SE6
```

The estimated average treatment for getting a generic recycling bin is **0.1031902** with a standard error of **0.0218889**. The 95% confidence interval for this treatment effect ranges from **0.0602881** to **0.1460924**.

My calculated treatment effect differs from part g because I am measuring a different variable. In part g, I am measuring the effect of giving an individual a bin (regardless of whether the bin has a sticker or not). In part h, my estimated treatment effect is lower than in part g, because in part h, I only measured the treatment effect of generic bins. In part g, the measured effect of bin treatment is skewed upwards by the effects of the bin with the sticker. I also have a wider confidence in part h than in part g because, in part h, I am estimated a more specific effect with a smaller sample size (only those with generic bins). In part g, because my sample size is larger (any bin) and my effect more general, I am able to produce a slightly more precise estimate.

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.

```
ebola <- read.csv("4Ebola.csv")
head(ebola)
```

```
##   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 associated with this estimate?

```
# create a linear model measures the effect of ZMapp on vomiting at day 14
zmap1 <- lm(vomiting_day14 ~ treat_zmapp, data = ebola)

# store the effect of zmap
zmap_TE1 <- coeftest(zmap1)[2,1]

# store the standard error of zmap
zmap_SE1 <- coeftest(zmap1)[2,2]

# store the associated p-value
zmap_p1 <- coeftest(zmap1)[2,4]
```

The estimated effect of ZMapp on whether someone is vomitting on day 14 is **-0.2377015 (0.0856316)**. Said in plainer English, those who took ZMapp were -23.770153% less likely to be vomiting at day 14. The p-value associated with this estimate is **0.0065954**.

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.

```r
# add covariates for vomiting at day 0 and patient temperature at day 0 to the model
zmap2 <- lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,
            data = ebola)

# store the effect of zmap
zmap_TE2 <- coeftest(zmap2)[2,1]

# store the standard error of zmap
zmap_SE2 <- coeftest(zmap2)[2,2]

# store the associated p-value
zmap_p2 <- coeftest(zmap2)[2,4]
```

The estimated effect of ZMapp on whether someone is vomitting on day 14, when controlling for vomiting and temperature at day 0, is **-0.1655367 (0.0756714)**. The p-value associated with this estimate is **0.0311285**.

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

I prefer the estimate in part b because it controls for the pre-treatment health of the individuals in the study. Even though we used random assignment, perhaps the sicker people ended up in the ZMapp treatment group. This would obscure the effect of the treatment (by making it appear smaller than it is in reality). By controlling for the temperature and vomiting at day 0, I can test more similar groups.

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.

```r
# add covariates for vomiting at day 0 and patient temperature at day 0 to the model
zmap3 <- lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 +
              temperature_day0 + temperature_day14,
            data = ebola)

# store the effect of zmap
zmap_TE3 <- coeftest(zmap3)[2,1]

# store the standard error of zmap
zmap_SE3 <- coeftest(zmap3)[2,2]

# store the associated p-value
zmap_p3 <- coeftest(zmap3)[2,4]
```

The estimated effect of ZMapp on whether someone is vomitting on day 14, when controlling for vomiting and temperature at day 0 and vomiting at day 14, is **-0.1201006 (0.0776798)**. The p-value associated with this estimate is **0.1254056**.

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

I prefer the estimate of the ATE in part b because in part d we add a bad control. Bad controls are variables that are impacted by the treatment. In this case, vomitting at day 14 is a bad control because it is likely to be impacted by treatment from ZMapp.

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?

```
# switch the model to consider the temperature at day 14 and not vomiting at day 14
zmap4 <- lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + male +
              male * treat_zmapp,
            data = ebola)

# store the interaction effect of ZMapp with gender
zmap_IE4 <- coeftest(zmap4)[6,1]

# store the standard error of zmap
zmap_SE4 <- coeftest(zmap4)[6,2]

# store the associated p-value
zmap_p4 <- coeftest(zmap4)[6,4]
```

To test the hypothesis that ZMapp is especially likely to reduce men's temperatures, I added an interaction variable to my model. I did this by modifying my linear model to not only control for the male variable but also the interaction of the male variable with the teatment variable (male x treatment). I found an interaction effect of **-2.0766863** with a standard error of **0.1916387**. In other words, I found that the treatment effect on temperature of ZMapp results in an extra **-2.0766863** drop in temperature for men.

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.)

My colleague has run into the classic *multiple comparisons problem* (page 300 of Field Experiments). When a researcher makes so many comparisons between subgroups, there's a pretty good chance that at least one will pop up as statistically significant. When running a single test of significance with a standard 0.05 p-value, there's a 5% chance that we'll declare a result statistically significant when it really isn't. In my colleague's case, the probability of finding at least one covariate that interacts significantly with the treatment is very nearly *100%*. I am actually surprised that he only found one statistically significant effect. My colleague went on a fishing expedition. In the future, he should adjust his p-value threshold for statistically significant with a Bonferroni correction.

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?

If I had only tested this single heterogreneous treatment effect, I would be much more inclined to believe that a heterogeneous treatment effect really does exist. In this case, I have made only a single comparison.

This single comparison has only a 5% chance of finding a statistically significant results where there really is not one. I only need to apply corrections (like the Bonferroni correction) when I'm making more than one comparison.

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.*)

My other colleague has posed a fundamentally unanswerable question. In order to truly test the causality of something, I need to be able to randomly assign individuals to treatment and control groups. To answer my colleague's question, I need to randomly assign individuals to be of African descent and to not be of African descent and then measure how likely they are to get Ebola. This is impossible because I cannot randomly assign someone's ancestry. People are born with their ancestries. I could attempt modified experiments (e.g. the effect of being born in Africa on the likelihood of getting Ebola); however, these experiments would not answer my colleague's question about the impact of African ancestry on the likelihood of getting Ebola.