Problem Set 3 - Experiments and Causality

Mike Gruzynski

```
# load packages
library(data.table)
library(foreign)
library(lmtest)
```

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

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

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

```
library(sandwich)
library(multiwayvcov)
```

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 paper on the effects of Facebook ads and download an anonymized version of the data for Facebook users only.

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

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.

```
# filter data to just study 1 information
d_study1 = d[d$studyno == 1,]
# fit a linear regression model for the effect of the ad (treatment or control) on candidate name recog
q1a_lm = lm(name_recall ~ as.factor(treat_ad), data = d_study1)
# use R function to compute 95% confidence interval for both y-intercept and dummy variable if subject
# control or treatment
confint(q1a_lm, level = 0.95)[2, ]
```

```
##       2.5 %      97.5 %
## -0.05101765  0.03142188
```

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

The clusters refer to individuals who share the same age (or ranges of age), gender and location. When taking clustering into account, the standard error will increase. This is because instead of the sample size equal to the total number of individuals, the sample size is the number of clusters. The clusters compresses the data with individuals with similar statistics into one cluster. Ex, if you had 5 samples (2 from Chicago and 3 from Seattle) and you were clustering samples on city, the cluster sample size will be 2 (1 cluster with Chicago data and 1 cluster with Seattle data) instead of a sample size of 5 (when clustering is not performed). The standard error calculation has sample size in the denominator, so when cluster sample size (2) is used vs unclustered sample size sample size (5), the standard error divided by a smaller value for the denominator will yield a larger overall value. The clustering mechanism reduces the effects of added data that dont contribute unique information.

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`.

```
q1c_lm_clustered = cluster.vcov(q1a_lm, ~ cluster)
q1c_lm_clustered.se <- sqrt(diag(q1c_lm_clustered))

data.frame("2.5 %" = c(q1a_lm$coefficients[2] - 1.96 * q1c_lm_clustered.se[2]),
           "97.5 %" = c(q1a_lm$coefficients[2] + 1.96 * q1c_lm_clustered.se[2]),
           check.names=FALSE)
```

```
##                              2.5 %     97.5 %
## as.factor(treat_ad)1 -0.05635499 0.03675922
```

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

```
# filter data to just study 2 information
d_study2 = d[d$studyno == 2,]
# fit a linear regression model for the effect of the ad (treatment or control) on candidate name recog
q1d_lm = lm(name_recall ~ as.factor(treat_ad), data = d_study2)
# use R function to compute 95% confidence interval for both y-intercept and dumy variable if subject w
# control or treatment
confint(q1d_lm, level = 0.95)[2, ]
```

```
##      2.5 %     97.5 %
## -0.0633702  0.0577635
```

```
q1d_lm_clustered = cluster.vcov(q1d_lm, ~ cluster)
q1d_lm_clustered.se <- sqrt(diag(q1d_lm_clustered))

data.frame("2.5 %" = c(q1d_lm$coefficients[2] - 1.96 * q1d_lm_clustered.se[2]),
           "97.5 %" = c(q1d_lm$coefficients[2] + 1.96 * q1d_lm_clustered.se[2]),
           check.names=FALSE)
```

```
##                              2.5 %    97.5 %
## as.factor(treat_ad)1 -0.0723899 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 omnibus regression. What is the treatment effect estimate and associated p-value?

```
q1d_lm = lm(name_recall ~ as.factor(treat_ad), data = d)
```

```
q1d_lm_clustered = cluster.vcov(q1d_lm, ~ cluster)
q1d_lm_clustered.se <- sqrt(diag(q1d_lm_clustered))

data.frame("2.5 %" = c(q1d_lm$coefficients[2] - 1.96 * q1d_lm_clustered.se[2]),
           "97.5 %" = c(q1d_lm$coefficients[2] + 1.96 * q1d_lm_clustered.se[2]),
           check.names=FALSE)
```

```
##                           2.5 %     97.5 %
## as.factor(treat_ad)1 -0.207465 -0.1026815
```

```
q1e_coef = q1d_lm$coefficients[2]
q1e_p_value = coef(summary(q1d_lm))[, "Pr(>|t|)"][2]

cat("The treatment effect of the pooled data is:", q1e_coef, "with an associative p-value of:", q1e_p_va
```

```
## The treatment effect of the pooled data is: -0.1550732 with an associative p-value of: 2.160639e-16
```

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?

```
d_1f = d
d_1f$dummystudy = d_1f$studyno - 1

# fit a linear regression model for the effect of the ad (treatment or control) on candidate name recog
q1f_lm = lm(name_recall ~ as.factor(treat_ad) + as.factor(dummystudy), data = d_1f)
q1f_lm_clustered = cluster.vcov(q1f_lm, ~ cluster)
q1f_lm_clustered.se <- sqrt(diag(q1f_lm_clustered))

data.frame("2.5 %" = c(q1f_lm$coefficients[2] - 1.96 * q1f_lm_clustered.se[2]),
           "97.5 %" = c(q1f_lm$coefficients[2] + 1.96 * q1f_lm_clustered.se[2]),
           check.names=FALSE)
```

```
##                             2.5 %     97.5 %
## as.factor(treat_ad)1 -0.04678947 0.03323898
```

```
summary(q1f_lm)
```

```
##
## Call:
## lm(formula = name_recall ~ as.factor(treat_ad) + as.factor(dummystudy),
##     data = d_1f)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6068 -0.1807 -0.1739  0.3932  0.8261
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               0.180685   0.015994  11.297   <2e-16 ***
## as.factor(treat_ad)1     -0.006775   0.018177  -0.373    0.709
## as.factor(dummystudy)1    0.426099   0.017955  23.731   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4381 on 2698 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1931, Adjusted R-squared:  0.1925
```

```
## F-statistic: 322.8 on 2 and 2698 DF,  p-value: < 2.2e-16
```

The treatment effect with the added study1 and study2 dummy variable is -0.006775 with an associative p-value of 0.709, which is not statistically significant. The main effect of the data is whether the data is coming from study 1 or study 2 (with effect value of 0.426099 and p-value of <2e-16 - which is statistical significant).

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.)

```
N_control_study1 = nrow(d[d$studyno == 1 & d$treat_ad == 0,])
N_treatment_study1 = nrow(d[d$studyno == 1 & d$treat_ad == 1,])
N_control_study2 = nrow(d[d$studyno == 2 & d$treat_ad == 0,])
N_treatment_study2 = nrow(d[d$studyno == 2 & d$treat_ad == 1,])

percent_treatment_study1 = N_treatment_study1 / (N_treatment_study1 + N_control_study1)
percent_treatment_study2 = N_treatment_study2 / (N_treatment_study2 + N_control_study2)
percent_treatment_pooled = (N_treatment_study1 + N_treatment_study2) / (N_treatment_study1 + N_control_

cat(" Percent of data in treatment for study1 set: ", round(100 * percent_treatment_study1, 2), "\n",
    "Percent of data in treatment for study2 set: ", round(100 * percent_treatment_study2, 2), "\n",
    "Percent of data in treatment for pooled set: ", round(100 * percent_treatment_pooled, 2))
```

```
##  Percent of data in treatment for study1 set:  59.02
##  Percent of data in treatment for study2 set:  24.96
##  Percent of data in treatment for pooled set:  42.13
```

The above calculations show that the study1 percentage value for treatment assignment is different than the study2 assignment probability. When pooling data (in question 1e), all the control and treatment assignments get smeared together rather than blocking on study (like in question 1f). This is one of the reasons why the answers is different for question 1e vs 1f.

The pooled data will be biased because the probability of the random assignment is not the same for both studies, and the answer in question 1f has a factor to adjust for the differences in study assignment. Pooling observations across blocks that have variable probability of treatment assignment tends to yield higher potential outcomes for treatment than control group and in the end yields more biased estimates of coefficients for regression.

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."

There are key differences between the studies that cause differences in results between Broockman and Green study (BGS) and the facebook case study (FCS). The main difference is that the FCS used facebook to target users to show online advertisement in order to vote down a proposition that increases class size. The study states, ".. it targeted people who listed terms like"teacher," "pta" "math teacher" to reach educators. . . " and the article goes on to say ". . . "I love my son" and "I love my daughter" (and layered them with demographic targeting)." In this study they are clearly targeting population that is pro-student laws and grouped them by age groups and gender. The BGS on the other hand randomly assigned a cluster (based on age and gender and location) to an advertisement treatment or control group and blocked between two studies (one study was in a republican state legislative candidate in a non-battleground state and the other study was a democrative congessional candidate in a non-battleground state). The BGS studied the sample's candidate recognition and the candidates stance knowledge on the issue where as the FCS focused on targeting audience based on their profile characteristics and showing if ads helped or hurt the vote on reducing class size in florida.

The FCS is attacking samples bias towards children and school where as the BGS is focusing on candidate recognition and issue stance. The results will be different between the two studies (FCS and BGS) because of the experimental setup differences above.

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.

The estimated ATE of providing a recycling bin on the average weight of recyclables turned in per household per week is equal to : 0.187

```
q2a_ate = 0.187
q2a_se = 0.032
q2a_lower = q2a_ate - 1.96 * q2a_se
q2a_upper = q2a_ate + 1.96 * q2a_se

cat("The 95% confidence interval for the estimate of ATE of providing a recycling bin on the average wei
```

```
## The 95% confidence interval for the estimate of ATE of providing a recycling bin on the average weigh
```

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.

The estimated ATE of sending a text message reminder on the average weight of recyclables turned in per household per week is equal to : -0.024

```
q2b_ate = -0.024
q2b_se = 0.039
q2b_lower = q2b_ate - 1.96 * q2b_se
q2b_upper = q2b_ate + 1.96 * q2b_se

cat("The 95% confidence interval for the estimate of ATE of sending a text message reminder on the avera
```

```
## The 95% confidence interval for the estimate of ATE of sending a text message reminder on the average
```

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

In table 4a, the columns that show significance level for any bin are:

column 1: Percentage of visits turned in bag column 2: Avg. no. of bins turned in per week column 3: Avg. weight (in kg) of recyclables turned in per week column 4: Avg. 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?

In table 4a, the columns that show significance level for any sending text messages are:

Not Available (no columns in this table show significance level of estimate for sending a text)

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 value for the slope coefficient for the baseline variable is equal to 0.281. The intepretation of this coefficient is for every 1 kg of recycling submitted during baseline measurement (per week) an increase of 0.281 kg will be added to the ATE during treatment weeks. Therefore, for the house that turns in 2kg more (household A), they will have an increase ATE of 0.562 kg over the value for household B.

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.

Taking out "percentage of visits turned in bag, baseline" from the regression" wont necessarily increase or decrease the ATE, but will be redistributed among the variables (some will go up and some will go down), the standard error will go up, because you are losing a very statistically significant variable and the variance captured from that term will be lost and bias the other terms.

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

Theoretically speaking, the control is not a post treatment variable, so it isnt a bad control. However, it is a poor control because the standard error is not small enough to make it statistically significant at a 0.05 alpha level. In addition, the precision of the estimate is not good and the 95% confidence interval crosses zero ([-0.06, 0.05]) and does not help support with confidence if the added indicator control variable helps or hurts the ATE.

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 you removed the "has cell phone" variable from column 1, I would expect that "Any SMS message" will go up in order to incorporate the information of "has cell phone". These two variables in theory contain positive correlation since if you have a cell phone brings along with it text message features and the capability of reading the messages.

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.)

The full experimental design was a 3x3 design. In the experimental design section explaining the experiment that produced table 4 data, the text stated:

Three equal-sized groups: a generic SMS message group, a personalized SMS message group, and a control group that received no text message. and three groups: to receive a plastic bin to store their recyclables, a plastic bin with a sticker, or no bin

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?

There are 5 baseline categories for table 4b:

1. Percentage of visits turned in bag, baseline

2. Avg. no. of bins turned in per week, baseline
3. Avg. weight (in kg) of recyclables turned in per week, baseline
4. Avg. market value of recyclables given per week, baseline
5. Avg. percentage of contamination per week, baseline

These are essentially a categorical based y-intercepts in the equation and in each column if all indicators are all 0 (control) then the regression will be predicted (results are recycling amount (kg) per week) by the baseline term only. The five different columns correspond to investigating the results of the experiment in terms of characteristics surrounding recycling.

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

The interpretation of this coefficient is: Holding everything else equal (ceteris paribus), when the subject recieves a bin without a sticker the percentage of visits turned in bag will go up by 3.5 % on average.

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?

I am defining stronger treatment effect as a coefficient that moves the baseline value the most, or in other words (the estimated value for the coefficient absolute value is largest).

Because of that definition, Bin with sticker is larger in absolute value than Bin without sticker. The size of the magnitude of the estimated difference is: 0.055 - 0.035 = 0.02 (or increase of 2% increase in visits with bag turned in.

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

The F-test null hypothesis is group means are equal.

From the table we see a F-test p-value (1) = (2) is equal to 0.31 (which is not statistically significant), which gives us statistical reason to fail to reject the null hypothesis. An F-test will tell you if a group of variables are jointly significant.

Another way to show if a value is significantly differnet from another is with a z test. THe below R code shows that the value of calculated z is below 1.96 (the threshold of being signigficant at a 5% level) - and the result is FALSE, Z_calculated is not greater than 1.96.

```
c1 = 0.055
se1 = 0.015
c2 = 0.035
se2 = 0.015

z = c1 - c2 / sqrt(se1**2 + se2**2)
abs(z) > 1.96
```

```
## [1] FALSE
```

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."

The "fully saturated" model refers to a model that has indicator variables for every unique combination of the experiment design. Each indicator coefficient corresponds to each group in the experimental design in addition to indicator variables with no phone (which is an observation in th experiment not part of control or experiment).

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).

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

```
##      street           havecell      avg_bins_treat   base_avg_bins_treat
##  Min.   :-999.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  69.00   1st Qu.:0.0000   1st Qu.:0.4167   1st Qu.:0.3750
##  Median : 131.50   Median :1.0000   Median :0.6250   Median :0.6250
##  Mean   :  68.81   Mean   :0.5908   Mean   :0.6811   Mean   :0.7363
##  3rd Qu.: 215.00   3rd Qu.:1.0000   3rd Qu.:0.8333   3rd Qu.:1.0000
##  Max.   : 263.00   Max.   :1.0000   Max.   :4.1667   Max.   :6.3750
##  NA's   :3         NA's   :1
##       bin              sms             bin_s            bin_g
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3378   Mean   :0.3087   Mean   :0.1681   Mean   :0.1697
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##      sms_p            sms_g
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1557   Mean   :0.1529
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

From the summary table above, we see that the variables havecell, bin, sms, bin_s, bin_g, sms_p, sms_g are binary variables. The variables avg_bins_treat and base_avg_bins_treat appear to be continuous variables that represent ATE of treatment anf baseline values of of ATE. The street variables seems to be categorical variables, where each integer represents a different street.

Street variable has 3 NA values and havecell data has 1 NA value. In addition, there seems to be a code in value for street variable at -999. The below R code snippet shows that there is 123

```
nrow(d[d$street == -999, ])
```

```
## [1] 123
```

If we filter out all NA and -999 for all rows:

```
d4 = d[d$street != -999 & !is.na(d$street) & !is.na(d$havecell),]
nrow(d) - nrow(d4)
```

```
## [1] 124
```

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.

```
q4a_lm = lm(formula = avg_bins_treat ~ bin, data = d4)
confint(q4a_lm)[2:2,]
```

```
##      2.5 %     97.5 %
## 0.09377495 0.17561568
```

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.

```
q4b_lm = lm(formula = avg_bins_treat ~ bin + base_avg_bins_treat, data = d4)
confint(q4b_lm)[2:2,]
```

```
##      2.5 %     97.5 %
## 0.09333734 0.16087552
```

It appears the estimate for the indicator variable for bins has a slightly tighter confidence interval than the linear model created without the baseline covariate. The addition of the pre-treatment (baseline) variable allows us to explain more of the variation of the data. This additional capability of explaining the random variation leads us to a decreased standard error which will lead to a tighter confidence interval.

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.

```
q4c_lm = lm(formula = avg_bins_treat ~ bin + base_avg_bins_treat + as.factor(street), data = d4)
confint(q4c_lm)[2:2,]
```

```
##      2.5 %     97.5 %
## 0.08152104 0.15060719
```

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 blocking by street (or stratified at the street level) will theoretically make the ATE for each block group closer together. For instance not blocking will allow a house from higher social-economical background interact more with one from a lower social economical background. It would be like comparing poor quality apples to high quality apples instead of a apples to apples comparison. The experimental design creates a similar distribution of potential ATE. This is why we are seeing a similar confidence interval because we are adding a grouping mechanism to compare apples to apples and doesnt throw in much added variation to the model and therefore similar confidence intervals.

reduces the variation of the treatment effect if the variability between streets is larger than the variation within the street. This theoretically makes sense because houses on the same streets are more comparible (more similar social-economical characteristics) than comparing houses between streets. Recycling amount can be helped by the addition of street level indicator variables.

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.

```
d4$no_cell_phone = ifelse(d4$havecell==1,0,1)
```

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.

```
q4f_lm = lm(formula = avg_bins_treat ~ bin + base_avg_bins_treat + no_cell_phone + as.factor(street), da
confint(q4f_lm)[2:2,]
```

```
##     2.5 %    97.5 %
```

```
## 0.0826730 0.1516659
```

Having a cell phone or no cell phone does not help explain the additional variation in expected recycling (cell phones are not correlated with recycling). The other terms in the lm() equation explain the data well enough and the additional variable of "no cell phone" does not add much explaining power to the model.

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.

```
# went back to original data set without any filtered data to show same results as table 4a column 2 - 
d$no_cell_phone = ifelse(d$havecell==1,0,1)


q4g_lm = lm(formula = avg_bins_treat ~ bin + base_avg_bins_treat + no_cell_phone
            + sms + as.factor(street), data = d)

q4g_lm$coefficients[1:5]
```

```
##         (Intercept)                 bin base_avg_bins_treat
##         0.384642314         0.115053649         0.373482860
##       no_cell_phone                 sms
##        -0.046702054         0.005124375
```

```
confint(q4g_lm)[2:2,]
```

```
##      2.5 %     97.5 %
## 0.08160886 0.14849843
```

SMS did not seem to effect the overall amount of recycling (kg) turned in per week. The confidence interval has not changed much relative to the model with no SMS variable added to the model because the addition of the SMS does not add to explanation of variation in the model in addition you already have a cell phone related variable in the model (also the cell phone control is not a statistically strong variable), so the addition of a secondary one (which is correlated) does not add explainatory strength in to the new 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.

```
# went back to original data set without any filtered data to show same results as table 4b column 2 - 
q4f_lm = lm(formula = avg_bins_treat ~ bin_s + bin_g
            + sms_p + sms_g + no_cell_phone + base_avg_bins_treat + as.factor(street), data = d)

q4f_lm$coefficients[1:7]
```

```
##         (Intercept)               bin_s               bin_g
##         0.384943316         0.127812892         0.103190216
##               sms_p               sms_g       no_cell_phone
##        -0.008041152         0.019707117        -0.046383459
## base_avg_bins_treat
##         0.373852178
```

```
confint(q4f_lm)[3:3,]
```

```
##      2.5 %     97.5 %
## 0.06025627 0.14612416
```

We added more variation on the treatment effects (we added bin_s and bin_g instead of bin and sms_p and

sms_g instead of sms). This adds more confounding variables which will increase the standard error and decresases the precision on the estimate for the ATE on recycling.

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.

```
d5 <- read.csv("./data/ebola_rct2.csv")
head(d5)
```

```
##   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?

```
q5a_lm = lm(formula = vomiting_day14 ~ treat_zmapp, data = d5)

q5a_effect = q5a_lm$coefficients[2]
q5a_std_error = coef(summary(q5a_lm))[, "Std. Error"] [2]
q5a_p_value = coef(summary(q5a_lm))[, "Pr(>|t|)"][2]

cat("The estimated effect of ZMapp:", q5a_effect, ", with a standard error of (",  q5a_std_error, ") an
    that is equal to:", q5a_p_value, ". The p-value show statistical significance at a 0.05 alpha value
```

```
## The estimated effect of ZMapp: -0.2377015 , with a standard error of ( 0.08563161 ) and a p-value
##     that is equal to: 0.006595412 . The p-value show statistical significance at a 0.05 alpha value
```

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.

```
q5b_lm = lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0, data = d5)

q5b_effect = q5b_lm$coefficients[2]
q5b_std_error = coef(summary(q5b_lm))[, "Std. Error"] [2]
```

```
q5b_p_value = coef(summary(q5b_lm))[, "Pr(>|t|)"][2]

cat("The ATE of ZMapp:", q5b_effect, ", with a standard error of (",  q5b_std_error, ") and a p-value
    that is equal to:", q5b_p_value, ". The p-value show statistical significance at a 0.05 alpha value
```

```
## The ATE of ZMapp: -0.1655367 , with a standard error of ( 0.07567142 ) and a p-value
##     that is equal to: 0.03112852 . The p-value show statistical significance at a 0.05 alpha value
```

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

I would prefer to use the ATE reported in part b. Looking at the data between the model in part a and b, the standard error is lower in b, making for a more precise estimate for the ATE. The ATE in a went down by ~0.07 (which is a reduction by 30% which is significant). This points to the concern that model a shows some omitted variable bias (OVB) relative to model b (the estimate changed significantly with a reduction of the estimates standard error). In addition to the above reasons, below shows that model b adjusted R-squared and F-statistic has higher values which means variation is better explained in the model and more statistically significant (or higher liklihood to reject null hypothesis that the estimates are equal to zero).

```
model_a_b_df = data.frame(adj_r_squared = c(summary(q5a_lm)$adj.r.squared, summary(q5b_lm)$adj.r.squared
           f_statistic = c(summary(q5a_lm)$f[1], summary(q5b_lm)$f[1]))
rownames(model_a_b_df) = c("model_a", "model_b")
model_a_b_df
```

```
##         adj_r_squared f_statistic
## model_a    0.06343488     7.70541
## model_b    0.28951332    14.44704
```

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.

```
q5d_lm = lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 +
             temperature_day14, data = d5)

q5d_effect = q5d_lm$coefficients[2]
q5d_std_error = coef(summary(q5d_lm))[, "Std. Error"] [2]
q5d_p_value = coef(summary(q5d_lm))[, "Pr(>|t|)"][2]

cat("The ATE of ZMapp:", q5d_effect, ", with a standard error of (",  q5d_std_error, ") and a p-value
    that is equal to:", q5d_p_value, ". The p-value does not show statistical significance at a 0.05 al
```

```
## The ATE of ZMapp: -0.1201006 , with a standard error of ( 0.07767979 ) and a p-value
##     that is equal to: 0.1254056 . The p-value does not show statistical significance at a 0.05 alpha
```

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

```
model_b_d_df = data.frame(adj_r_squared = c(summary(q5b_lm)$adj.r.squared, summary(q5d_lm)$adj.r.squared
           f_statistic = c(summary(q5b_lm)$f[1], summary(q5d_lm)$f[1]))
rownames(model_b_d_df) = c("model_b", "model_d")
model_b_d_df
```

```
##         adj_r_squared f_statistic
## model_b     0.2895133    14.44704
## model_d     0.3123855    12.24400
```

```
confint(q5b_lm)[2,]
```

```
##       2.5 %      97.5 %
## -0.31574331 -0.01533017
```

```
confint(q5d_lm)[2,]
```

```
##       2.5 %      97.5 %
## -0.27431451  0.03411325
```

From the above data, we see that the F-statistic declines from ~14.4 to about ~12.2 (smaller F-statistic means estimate has more liklihood to fail to rehect null hypothesis of estimate being equal to zero) and the confidence interval of model d crosses, where as the model for question 5b does not. The adjusted R-squared will go up with the addition of more covatiates, however does not mean its a better explained model. Because of the above information, I prefer model b over model d for estimating this data.

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?

```
q5f_lm <- lm(temperature_day14 ~ treat_zmapp*male + vomiting_day0 + temperature_day0 + male, data = d5)

q5f_lm$vcovHC = vcovHC(q5f_lm)
q5f_lm_coef = coeftest(q5f_lm)

cat("Male Zmapp Information\n")
```

```
## Male Zmapp Information
```

```
q5f_lm_coef[2] + q5f_lm_coef[6]
```

```
## [1] -2.307552
```

```
cat("Female Zmapp Information\n")
```

```
## Female Zmapp Information
```

```
q5f_lm_coef[2]
```

```
## [1] -0.2308655
```

```
summary(q5f_lm)
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp * male + vomiting_day0 +
##     temperature_day0 + male, data = d5)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70157 -0.37725 -0.02702  0.34687  0.73968
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      48.71269    9.26618   5.257 9.14e-07 ***
## treat_zmapp      -0.23087    0.11871  -1.945   0.0548 .
## male              3.08549    0.12644  24.403  < 2e-16 ***
## vomiting_day0     0.04113    0.18208   0.226   0.8218
## temperature_day0  0.50480    0.09508   5.309 7.34e-07 ***
## treat_zmapp:male -2.07669    0.19164 -10.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 0.4518 on 94 degrees of freedom
## Multiple R-squared:  0.9059, Adjusted R-squared:  0.9009
## F-statistic:   181 on 5 and 94 DF,  p-value: < 2.2e-16
```

Looking at the above summary table, we can see that the treat_zmapp on average has a ~ -0.23 ATE in the model (not at a significant level of 0.05). Now when you look at the interaction of treat_zMapp and male, or in other words an indicator variable which is only turned on if the subject was both male and had treatment, we see a statistically significant interaction term of ~ -2.07. THis term shows that in this model being male and on treatment does statistically show the subjects having lower temperature. However, need to remeber that the study had unequal amount of male and female subjects at about a 1:2 ratio (shown below). In addition, do male and females actual have resting temperatures at the exact same degree level. A interesting investigation would be to get baseline temperatures of each subject and see the temperature effect deltas for treatment and control subjects.

```
cat("Male number of subjects:", nrow(d5[d5$male == 1,]), "\n")
```

```
## Male number of subjects: 37
```

```
cat("Female number of subjects:",nrow(d5[d5$male == 0,]))
```

```
## Female number of subjects: 63
```

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.)

using information from page 300 in DE

```
1 - 2000 * (1 - 0.001)**10000
```

```
## [1] 0.9096533
```

The probability of finding at least one covariate that significantly interacts with the treatment at a significant level of 0.001 (2000 different indicator variables) with 10,000 covariates is ~ 0.91. There are just too many covariates and explained potential variables investigated here and variable fishing is a relevant concern in this situation, so ZMapp may work better for men, however the statistics show that there is ahigh probability of it just happening to interact in males case better with the volume of independent and dependent variables. If the alpha value was at the standard 0.05 level the probability shoots to 1, meaning with these amount of covariates and independent outcome variables the probability dictates something will come back statistically significant.

*NOTE USED CRITICAL VALUE OF 0.001 TO SHOW EFFECT RATHER THN 0.05 WHICH ANSWER IS 1 (IN THE ABOVE EQUATION)

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 would be more inclined to believe the results of an experiment if we control for how many heterogeneous treatment effects and covariates to explain the models then statistically speaking, the liklihood of getting a statistically significant effect by chance goes down. For example if we run one model with 5 covariates the liklihood by chance of finding a statistically significant result (at 0.001 value) jumps down to ~ 0.005.

```
1 - (1 - 0.001)**5
```

```
## [1] 0.00499001
```

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.*)

I believe this is a FUQ'd question. There is no ideal experiment that randomly assigns the treatment of African descent to the subject (i have no idea how you can do that). Also, what would the control be (any other nation, what if African descent is entangled in the culture like Jamaica or Caribbean). This can be an observational study, but not an experiment becasue group assignments in experimental design are just not possible. You cant send people to go be birthed there because even if they were born in africa how long do they have to be there to be of African descent?...one...two generations more. It just is not possible. Also how does your colleague define African descent? This is a very vague question and can not be answered with an experiment.