# Problem Set 3

*Osmar Coronel*

*March 6, 2019*

```
# load packages
library(data.table)
library(foreign)
library(sandwich)
library(stargazer)
library(lmtest)
library(dplyr)
library(multiwayvcov)

options(digits=3)
rm(list = ls())
```

## 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.

```
# function that calculates confidence interval and p-value
# assumes a two sided-test
# it uses either robust standard error or
# cluster standard errors
# parameters are: regression model, degrees of freedom
# cluster name if clustered, alpha

CI_p_value <-function(m=NULL, n_observations=0,cluster=NULL, alpha=0.05){
  df <- n_observations - 1
  if (is.null(cluster)) {
    tc <- qt(1-alpha/2,df)
    CI <- coeftest(m, vcov=vcovHC(m))[2,1] + c(-1,1)*tc*coeftest(m, vcov = vcovHC(m))[2,2]
    p_value <- coeftest(m, vcov=vcovHC(m))[2,4]
    results <- list(CI=CI, p_value=p_value)
    return(results)
  }
  else if (!is.null(cluster)){
    tc <- qt(1-alpha/2,df)
    CI <- coeftest(m, vcov=cluster.vcov(m, ~ cluster))[2,1] + c(-1,1)*tc*coeftest(m, vcov = cluster.vco
    p_value <- coeftest(m, vcov=cluster.vcov(m, ~cluster))[2,4]
    results <- list(CI=CI, p_value=p_value)
    return(results)
  }
}
```

# 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")
d <- data.table(d)
head(d)
```

```
nrow(d)
```

```
## [1] 2706
```

```
str(d)
```

```
## Classes 'data.table' and 'data.frame':   2706 obs. of  5 variables:
##  $ studyno            : int  2 2 2 2 2 2 2 2 2 2 ...
##  $ treat_ad           : int  0 0 0 0 1 1 1 0 0 0 ...
##  $ cluster            : Factor w/ 1025 levels "Study 1, Cluster Number 1",..: 578 661 752 848 1001 1
##  $ name_recall        : int  0 1 0 1 1 0 1 1 0 1 ...
##  $ positive_impression: int  0 0 0 0 1 0 1 0 0 1 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(d)
```

```
##     studyno       treat_ad                              cluster
##  Min.   :1.0   Min.   :0.000   Study 1, Cluster Number 799:  24
##  1st Qu.:1.0   1st Qu.:0.000   Study 2, Cluster Number 333:  23
##  Median :1.0   Median :0.000   Study 1, Cluster Number 781:  20
##  Mean   :1.5   Mean   :0.421   Study 1, Cluster Number 800:  17
##  3rd Qu.:2.0   3rd Qu.:1.000   Study 2, Cluster Number 425:  17
##  Max.   :2.0   Max.   :1.000   Study 2, Cluster Number 501:  17
##                                (Other)                    :2588
##   name_recall   positive_impression
##  Min.   :0.00   Min.   :0.00
##  1st Qu.:0.00   1st Qu.:0.00
##  Median :0.00   Median :0.00
##  Mean   :0.39   Mean   :0.26
##  3rd Qu.:1.00   3rd Qu.:1.00
##  Max.   :1.00   Max.   :1.00
##  NA's   :5      NA's   :5
```

```
# Crates a cluster2 with the cluster names
# nlevels(d$cluster) 1025 clusters in total
d$cluster2 <- as.numeric(d$cluster)
# We omit the rows where we have omitted outcomes
# name_recall and positive_impression
d <- na.omit(d)
# summary(d)
# nrow(d) There are 2701 subjects
```

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.

```
title <- "Study 1 without Clustering"
m_1a <- lm(name_recall ~ treat_ad, data=d[studyno==1])
se.m_1a <- sqrt(diag(vcovHC(m_1a)))
```

```
stargazer( m_1a,
           title= title,
           type = "text",
           omit.stat="f",
           add.lines = list(c("Clustering Fixed Effects", "No")),
           se = list(se.m_1a),
           star.cutoffs =c(0.05, 0.01, 0.001),
           header = FALSE
           )
```

```
##
## Study 1 without Clustering
## ===================================================
##                               Dependent variable:
##                           -------------------------
##                                   name_recall
## ---------------------------------------------------
## treat_ad                            -0.010
##                                     (0.021)
##
## Constant                           0.182***
##                                     (0.016)
##
## ---------------------------------------------------
## Clustering Fixed Effects               No
## Observations                         1,364
## R2                                  0.0002
## Adjusted R2                         -0.001
## Residual Std. Error           0.382 (df = 1362)
## ===================================================
## Note:                   *p<0.05; **p<0.01; ***p<0.001
```

**Answer: 1a**
The 95% CI of the ATE= -0.010 +- 0.042

```
result <- CI_p_value(m_1a,1364, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.051" "0.032"
```

```
paste('p-value: ',round(result$p_value,3))
```

```
## [1] "p-value:  0.643"
```

b. *“What are the clusters in Broockman and Green’s study?* *Why might taking clustering into account increase the standard errors¿‘*

**Answer: 1b**
In Broockman and Green’s study the clusters are a grouping of people of the same sex, age, and county. The study mention 8 counties, 2 genders, 47 age values (each age 18-64), givin this 8x2x47

= 752 clusters. With the two studies combined we have a total of 1024 clusters after eliminating the cluster with missing outcome. Out of them, 577 clusters with 1364 participants in Study 1 and 447 clusters with 1337 participants in Study 2.
Taking clustering might increase the standard error because it fundamentaly reduces the total number of samples. In our case instead of having 2706 subjects we now only have 1024 groups, less than half the number of original subjects.

```
# printing the number of clusters and the number of subjects per study
paste("study 1, No Groups: ",length(unique(d[studyno==1]$cluster2)),
      "Participants: ",length(d[studyno==1]$cluster2))
```

```
## [1] "study 1, No Groups:  577 Participants:  1364"
```

```
paste("study 2, No Groups: ",length(unique(d[studyno==2]$cluster2)),
      "Participants: ",length(d[studyno==2]$cluster2))
```

```
## [1] "study 2, No Groups:  447 Participants:  1337"
```

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`.

```
m_1c <- lm(name_recall ~ treat_ad+cluster2, data=d[studyno==1])
cvcov1c <- cluster.vcov(m_1c, ~ cluster2)
sec1c <- sqrt(diag(cvcov1c))
title <- "Study 1 with Clustering"
stargazer( m_1c,
           title= title, omit = "cluster2",
           align=TRUE,
           type = "text",
           omit.stat="f",
           add.lines = list(c("Clustering Fixed Effects", "Yes")),
           se = list(sec1c),
           star.cutoffs =c(0.05, 0.01, 0.001)
           )
```

```
##
## Study 1 with Clustering
## ===================================================
##                               Dependent variable:
##                           -------------------------
##                                   name_recall
## ---------------------------------------------------
## treat_ad                            -0.008
##                                     (0.023)
##
## Constant                           0.113***
##                                     (0.026)
##
## ---------------------------------------------------
## Clustering Fixed Effects              Yes
## Observations                         1,364
## R2                                   0.009
## Adjusted R2                          0.007
## Residual Std. Error            0.380 (df = 1361)
```

```
## ================================================
## Note:                    *p<0.05; **p<0.01; ***p<0.001
```

**Answer: 1c**

```
result <- CI_p_value(m_1c,1364, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.049" "0.034"
```

```
paste('p-value: ',round(result$p_value,3))
```

```
## [1] "p-value:  0.72"
```

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

```
m_1d <- lm(positive_impression ~ treat_ad+cluster2, data=d[studyno==2])
cvcov1d <- cluster.vcov(m_1d, ~ cluster2)
sec1d <- sqrt(diag(cvcov1d))
title <- "Study 2 with Clustering"
stargazer( m_1d,
           title= title, omit = "cluster2",
           align=TRUE,
           type = "text",
           omit.stat="f",
           add.lines = list(c("Clustering Fixed Effects", "Yes")),
           se = list(sec1d),
           star.cutoffs =c(0.05, 0.01, 0.001)
           )
```

```
##
## Study 2 with Clustering
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                            positive_impression
## ------------------------------------------------
## treat_ad                          0.017
##                                  (0.031)
##
## Constant                          0.079
##                                  (0.095)
##
## ------------------------------------------------
## Clustering Fixed Effects            Yes
## Observations                       1,337
## R2                                 0.008
## Adjusted R2                        0.007
## Residual Std. Error          0.487 (df = 1334)
## ================================================
## Note:                    *p<0.05; **p<0.01; ***p<0.001
```

**Answer: 1d**

The 95% CI of Study 2 ATE= 0.017 +- 0.062

```
result <- CI_p_value(m_1d,n_observations = 1337,d$cluster2)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.045" "0.078"
```

```
paste('p-value: ',round(result$p_value,3))
```

```
## [1] "p-value:  0.592"
```

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?*

```
m_1e <- lm(positive_impression ~ treat_ad+cluster2, data=d)
cvcov1e <- cluster.vcov(m_1e, ~ cluster2)
sec1e <- sqrt(diag(cvcov1e))
title <- "Part (c) with the entire sample"
stargazer( m_1e,
           title= title, omit = "cluster2",
           align=TRUE,
           type = "text",
           omit.stat="f",
           add.lines = list(c("Clustering Fixed Effects", "Yes")),
           se = list(sec1e),
           star.cutoffs =c(0.05, 0.01, 0.001)
           )
```

```
##
## Part (c) with the entire sample
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                            positive_impression
## ------------------------------------------------
## treat_ad                        -0.006
##                                 (0.018)
##
## Constant                         0.008
##                                 (0.021)
##
## ------------------------------------------------
## Clustering Fixed Effects          Yes
## Observations                     2,701
## R2                               0.086
## Adjusted R2                      0.086
## Residual Std. Error        0.420 (df = 2698)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

**Answer: 1e**

```
result <- CI_p_value(m_1e,n_observations = 2701,d$cluster2)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.041" "0.03"
```

```
paste('p-value: ',round(result$p_value,3))
```

```
## [1] "p-value:  0.755"
```

(Here we are taking the entire sample, and we are applying clustering)
The p-level of 0.755 is larger than $\alpha = 0.05$
Hence, we fail to reject the hypothesis that the ATE is different from zero.

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 <- d[studyno==1,studyno2 := 0]
d <- d[studyno==2,studyno2 := 1]
```

```
m_1f <- lm(positive_impression ~ treat_ad+studyno2+cluster2, data=d)
cvcov1f <- cluster.vcov(m_1f, ~ cluster2)
sec1f <- sqrt(diag(cvcov1f))
title <- "Part (c) with the entire sample"
stargazer( m_1f,
           title= title, omit = "cluster2",
           align=TRUE,
           type = "text",
           omit.stat="f",
           add.lines = list(c("Clustering Fixed Effects", "Yes")),
           se = list(sec1f),
           star.cutoffs =c(0.05, 0.01, 0.001)
           )
```

```
##
## Part (c) with the entire sample
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                          positive_impression
## ------------------------------------------------
## treat_ad                       0.008
##                               (0.018)
##
## studyno2                     0.158***
##                               (0.032)
##
## Constant                      0.055*
##                               (0.022)
##
## ------------------------------------------------
## Clustering Fixed Effects        Yes
## Observations                   2,701
## R2                             0.094
## Adjusted R2                    0.093
## Residual Std. Error      0.418 (df = 2697)
## ================================================
```

```
## Note:                    *p<0.05; **p<0.01; ***p<0.001
```

**Answer: 1f**
The p-level is: 0.755 which is much larger than $\alpha = 0.05$ We fail to reject the hypothesis that the ATE is different from zero in this case.

```
result <- CI_p_value(m_1f,n_observations=2701,d$cluster2)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.027" "0.043"
```

```
paste('p-value: ',round(result$p_value,3))
```

```
## [1] "p-value:  0.655"
```

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.)

**Answer: 1g**
The results differ because Study 1 and Study 2 are two totally different experiments which we are mixing up in question (e) and question (f). In Study 2 the researchers collaborated with the most viable candidate (this was not the case in Study 1). In Study 2, adds were expanded with sponsored messages. In Study 1, the researchers randomized at the town level while in Study 2 they randomized at the county level. The towns in Study 1 were very disperse.

The answer in (e) is biased because because we are not controlling by the type of study like we are doing in (f). We can't assume that the variance is the same within each individual study cluster, then failing to appropriately account for the empirical variance migth lead us to biased results.

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."

**Answer: 1h**
I don't have the data from Facebook but very likely they did not correct for clustering and therefore their conclusions might be biased like the case of item (e) . Besides Facebook and Chong & Koster could be more than interested to prove that their "Vote No on 8" campaign was successful to showcase how successful their service is.

# 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.*

**Answer: 2a**
The estimated ATE with 95% Confidence Interval is: 0.187 +_ 0.064
It is significative because 0.187 much larger than the standard error

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.*

**Answer: 2b**
The ATE with 95% Confidence Interfal is: -0.024 +- 0.078
It is not very significative because because of the large standard error.

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

**Answer: 2c**
The treatment of providing the bin has a statistically significant effect because 0.187/0.032= 5.8 $> 2$.
The baseline of average weight of recyclables turned per week, had also a significant effect because 0.281/0.011=25.5 $> 2$.
Also, "Has cell phone" is statistically significative 0.105/0.038= 2.76 $> 2$

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

**Answer: 2d**
No outcome measures in Table 4A show any statistically significatn effect 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.

**Answer: 2e**
We could expect that household A will deliver $0.281 * 2Kgr = 0.56Kgr$ more recyclables than householdB per week during the 6 week period.

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.*

**Answer: 2f**
If the variable "percentage of visits turned in bag, baseline" had been left out, the estimated ATE of providing a recycling bin would be larger that the actual one in Column 1. Also, the standard error of "Any bin" would be larger than 0.012.

The reason of this is that the outcome is correlated with the variable "percentage of visits turned in bag,. . . ." so when we add this variable to the equation we reduce the value of the ATE to actual one.
Also, when the variable is added we decrease the standard error because the variable decreases the spurious errors due to this variable by absorving them when the variable is included.

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

**Answer: 2g**
Yes, "has a cell phone" is a bad control. You just don't want to drive inferences only about half the population. It is a bad control because only half of the people reported their cellphone number.

This creates two groups of subjects, the ones with cell phone who can have two treatments and the ones without a reported cellphone number that would have only one treatment. In fact both groups are two totally different groups.

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.*

**Answer: 2h**
If we remove the "Has a cell phone" covariate, the coefficient of "Any SMS message" *would increase* because "Has a cell phone" is a confound variable and the coefficient of "Has a cell phone" is positively correlated with the outcome and positively correlated with "Any SMS message".

# 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.)

**Answer: 3a**
I see the experiment with two matrices of 3x1. These are the features we are manipulating for all the subjects: Bin with sticker, Bin without sticker, No Bin, gives us a 3x1 matrix. At the same time, these are the features for people who reported their cellphone: Personal SMS message, Generic SMS Message, No SMS message. Hence, we also have a matrix 3x1.

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?

**Answer: 3b**
The baseline for Table 4B is the group of people that did not receive any bin and and also had not reported their cellphone.

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

**Answer: 3c**
The Bin without sticker coefficient of 0.035 which means that in average when the researches provided a bin without stickers their "percentage of visits turned in bag" increased in 0.035 in comparisson when they did not provide any 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?

**Answer: 3d**
The recycling bin with sticker seems to have a stronger treatment effect than the bin without sticker (0.055 vs 0.035 for bin with sticker vs bin without sticker respectively). See in the table "F-test p-value (1) = (2)"

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

**Anwser: 3e**
No. This difference is not statistically significant because the p-value of the null hypothesis Ho: (1)=(2) is 0.31 which is larger than our $\alpha = 0.05$.

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."

**Answer: 3f**
The model is saturated because in Table 4c, the author is presenting their result for all the possible combinations of treatment: Three different SMS treatments (Generic SMS, Personal SMS, No SMS) with the three bin treatments(Bin w/ sticker, Bin w/o sticker, and no bin). So far this provides (3x3= 9) options to report plus the treatments for the group with no cell phone. The treatments for the group with no cell phone are Bin w/ sticker, Bin w/o sticker (2x1). The default or baseline group is the group of subjects that received no bin and had no cell phone.

# 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_raw <- read.dta("./data/karlan_data_subset_for_class.dta")
d_raw <- data.table(d_raw)
d2 <- d_raw[ , .(ID                   = 1:nrow(d_raw),
                 Y                    = avg_bins_treat,
                 Any_bin              = bin,
                 Any_SMS              = sms,
                 Has_cell_phone       = havecell,
                 Avg_no_bins_turned_bl=base_avg_bins_treat,
                 Bin_with_sticker     = bin_s,
                 Bin_without_sticker  = bin_g,
                 Personal_SMS         = sms_p,
                 Generic_SMS          = sms_g,
                 Street = factor(street)
                 )
             ]
head(d2)
```

```
##    ID     Y Any_bin Any_SMS Has_cell_phone Avg_no_bins_turned_bl
## 1:  1 1.042       1       1              1                 0.750
## 2:  2 0.000       0       1              1                 0.000
## 3:  3 0.750       0       0              1                 0.500
## 4:  4 0.542       0       0              1                 0.500
## 5:  5 0.958       1       0              1                 0.375
## 6:  6 0.208       1       0              0                 0.000
##    Bin_with_sticker Bin_without_sticker Personal_SMS Generic_SMS Street
## 1:                1                   0            0           1      7
## 2:                0                   0            1           0      7
## 3:                0                   0            0           0      7
## 4:                0                   0            0           0      7
## 5:                0                   1            0           0      6
## 6:                0                   1            0           0      8
```

```
#ID                   = seq.int(nrow(d_raw)),

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

```
# nrow(d) There are 1785 rows
```

```
str(d2)
```

```
## Classes 'data.table' and 'data.frame':   1785 obs. of  11 variables:
##  $ ID                  : int  1 2 3 4 5 6 7 8 9 10 ...
```

```
##  $ Y                    : num  1.042 0 0.75 0.542 0.958 ...
##  $ Any_bin              : num  1 0 0 0 1 1 0 0 1 1 ...
##  $ Any_SMS              : num  1 1 0 0 0 0 0 1 1 0 ...
##  $ Has_cell_phone       : num  1 1 1 1 1 0 1 1 1 1 ...
##  $ Avg_no_bins_turned_bl: num  0.75 0 0.5 0.5 0.375 0 0.75 0.5 0.625 1 ...
##  $ Bin_with_sticker     : num  1 0 0 0 0 0 0 0 1 1 ...
##  $ Bin_without_sticker  : num  0 0 0 0 1 1 0 0 0 0 ...
##  $ Personal_SMS         : num  0 1 0 0 0 0 0 0 0 0 ...
##  $ Generic_SMS          : num  1 0 0 0 0 0 0 1 1 0 ...
##  $ Street               : Factor w/ 180 levels "-999","2","3",..: 7 7 7 7 6 8 8 8 5 9 ...
##  - attr(*, ".internal.selfref")=<externalptr>
```

```
summary(d2)
```

```
##        ID              Y           Any_bin         Any_SMS
##  Min.   :   1   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.: 447   1st Qu.:0.42   1st Qu.:0.000   1st Qu.:0.000
##  Median : 893   Median :0.62   Median :0.000   Median :0.000
##  Mean   : 893   Mean   :0.68   Mean   :0.338   Mean   :0.309
##  3rd Qu.:1339   3rd Qu.:0.83   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1785   Max.   :4.17   Max.   :1.000   Max.   :1.000
##
##  Has_cell_phone  Avg_no_bins_turned_bl Bin_with_sticker
##  Min.   :0.000   Min.   :0.00          Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.38          1st Qu.:0.000
##  Median :1.000   Median :0.62          Median :0.000
##  Mean   :0.591   Mean   :0.74          Mean   :0.168
##  3rd Qu.:1.000   3rd Qu.:1.00          3rd Qu.:0.000
##  Max.   :1.000   Max.   :6.38          Max.   :1.000
##  NA's   :1
##  Bin_without_sticker  Personal_SMS    Generic_SMS        Street
##  Min.   :0.00        Min.   :0.000   Min.   :0.000   -999   : 120
##  1st Qu.:0.00        1st Qu.:0.000   1st Qu.:0.000   250    :  35
##  Median :0.00        Median :0.000   Median :0.000   260    :  33
##  Mean   :0.17        Mean   :0.156   Mean   :0.153   256    :  32
##  3rd Qu.:0.00        3rd Qu.:0.000   3rd Qu.:0.000   215    :  31
##  Max.   :1.00        Max.   :1.000   Max.   :1.000   (Other):1531
##                                                      NA's   :   3
```

*Strange Values:* We can see that there are 120 streets associated with "-999". Very likely they did not have the name.
There are three missing values in the street name.
There is one missing value in "Has_cell_phone".

*Actions to be taken:*
Keep the "-999" for the unknown streets.
We will omit street and cell phone missing values.

```
# Eliminating missing values and keeping -999 as street name.
d2 <- na.omit(d2)
summary(d2)
```

```
##        ID              Y           Any_bin         Any_SMS
##  Min.   :   1   Min.   :0.00   Min.   :0.000   Min.   :0.000
##  1st Qu.: 446   1st Qu.:0.42   1st Qu.:0.000   1st Qu.:0.000
##  Median : 891   Median :0.62   Median :0.000   Median :0.000
```

```
##  Mean   : 891   Mean   :0.68   Mean   :0.337   Mean   :0.308
##  3rd Qu.:1337   3rd Qu.:0.83   3rd Qu.:1.000   3rd Qu.:1.000
##  Max.   :1782   Max.   :4.17   Max.   :1.000   Max.   :1.000
##
##  Has_cell_phone  Avg_no_bins_turned_bl Bin_with_sticker
##  Min.   :0.000   Min.   :0.00          Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.38          1st Qu.:0.000
##  Median :1.000   Median :0.62          Median :0.000
##  Mean   :0.591   Mean   :0.74          Mean   :0.167
##  3rd Qu.:1.000   3rd Qu.:1.00          3rd Qu.:0.000
##  Max.   :1.000   Max.   :6.38          Max.   :1.000
##
##  Bin_without_sticker  Personal_SMS    Generic_SMS        Street
##  Min.   :0.00        Min.   :0.000   Min.   :0.000   -999   : 120
##  1st Qu.:0.00        1st Qu.:0.000   1st Qu.:0.000   250    :  35
##  Median :0.00        Median :0.000   Median :0.000   260    :  33
##  Mean   :0.17        Mean   :0.156   Mean   :0.153   256    :  32
##  3rd Qu.:0.00        3rd Qu.:0.000   3rd Qu.:0.000   215    :  31
##  Max.   :1.00        Max.   :1.000   Max.   :1.000   22     :  27
##                                                      (Other):1503
```

```
# nrow(d2) The number of rows now is 1781
```

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.*

```
m4a <- lm(Y ~ Any_bin, data=d2 )
se.m4a <- sqrt(diag(vcovHC(m4a)))
title = "Model 4a, regression only with 'Any bin'"
stargazer(m4a,
          type = "text",
          title = title,
          omit.stat =  'f',
          add.lines = list(c("Street Fixed Effects", "No")),
          se = list(se.m4a),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## Model 4a, regression only with 'Any bin'
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  Y
## ------------------------------------------------
## Any_bin                        0.133***
##                                 (0.021)
##
## Constant                       0.636***
##                                 (0.011)
##
## ------------------------------------------------
## Street Fixed Effects              No
```

```
## Observations                     1,781
## R2                               0.024
## Adjusted R2                      0.023
## Residual Std. Error        0.404 (df = 1779)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(m4a,n_observations = 1781,cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "0.092" "0.174"
```

**Answer: 4a**
The ATE is 0.133 bins turned per week, with 95% CI 0.092, 0.174

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.*

```
m4b <- lm(Y ~ Any_bin+Avg_no_bins_turned_bl, data=d2 )
se.m4b <- sqrt(diag(vcovHC(m4b)))
# We are adding the previous model for comparisson purposes
title <- "Adding pre-treatment value"
stargazer(m4b,
          type = "text",
          title = title,
          omit.stat =  'f',
          se = list(se.m4b),
          add.lines = list(c("Street Fixed Effects", "No")),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## Adding pre-treatment value
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                    Y
## ------------------------------------------------
## Any_bin                          0.124***
##                                   (0.017)
##
## Avg_no_bins_turned_bl            0.390***
##                                   (0.031)
##
## Constant                         0.352***
##                                   (0.021)
##
## ------------------------------------------------
## Street Fixed Effects                No
## Observations                       1,781
## R2                                 0.338
## Adjusted R2                        0.337
## Residual Std. Error          0.333 (df = 1778)
```

```
## ================================================
## Note:                  *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(m4b,n_observations = 1785,cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "0.091" "0.158"
```

**Answer: 4b**
The ATE is 0.124 bins turned per week, with 95% CI 0.091, 0.158bins turned per week. This confidence in terval is smaller than the previous one by adding the pre-treatment value as a baseline.
This is because in the previous regression the "Avg_no_bins_turned_bl" was omitted and becomed a confound variable. By adding this variable, it is taking part of of the spurious variations that had absorved the treatment variable. Hence, reducing the 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.*

```
# We already used the factor() command at the beginning.
# We are adding previous models only for comparisson purposes
#cvcov4a <- cluster.vcov(m4a, ~ Street)
cvcov4b <- cluster.vcov(m4b, ~ Street)
#sec4a <- sqrt(diag(cvcov4a))
sec4b <- sqrt(diag(cvcov4b))
title <- "With Street effects"
stargazer(m4b,
          type = "text", omit = "Street",
          title = title,
          omit.stat = "f",
          add.lines = list(c("Street Fixed Effects","Yes")),
          se = list(sec4b),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## With Street effects
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                     Y
## ------------------------------------------------
## Any_bin                          0.124***
##                                   (0.018)
##
## Avg_no_bins_turned_bl            0.390***
##                                   (0.030)
##
## Constant                         0.352***
##                                   (0.021)
##
## ------------------------------------------------
## Street Fixed Effects                Yes
```

```
## Observations                      1,781
## R2                                0.338
## Adjusted R2                       0.337
## Residual Std. Error         0.333 (df = 1778)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(m4b,n_observations = 1785,cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "0.091" "0.158"
```

**Answer: 4c**
The new ATE 95% CI is 0.091, 0.158 bins turned per week, when clustering per street name.

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.

```
d2[,.(sd(Any_bin)), by=Street]
```

```
##      Street    V1
##   1:      7 0.447
##   2:      6 0.535
##   3:      8 0.426
##   4:      5 0.441
##   5:      9 0.452
##  ---
## 176:    215 0.445
## 177:    210 0.483
## 178:    220 0.488
## 179:    227 0.548
## 180:    246 0.516
```

**Answer: 4d**
Effectively the ATE 95% CI does not change much at all between (b) and (c), this is happening because as we can see above actually all clusters have approximately the same variation.

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.

```
d2$No_cell_phone <- as.numeric(!d2$Has_cell_phone)
head(d2)
```

```
##    ID     Y Any_bin Any_SMS Has_cell_phone Avg_no_bins_turned_bl
## 1:  1 1.042       1       1              1                 0.750
## 2:  2 0.000       0       1              1                 0.000
## 3:  3 0.750       0       0              1                 0.500
## 4:  4 0.542       0       0              1                 0.500
## 5:  5 0.958       1       0              1                 0.375
## 6:  6 0.208       1       0              0                 0.000
##    Bin_with_sticker Bin_without_sticker Personal_SMS Generic_SMS Street
## 1:                1                   0            0           1      7
```

```
## 2:                0                   0           1           0      7
## 3:                0                   0           0           0      7
## 4:                0                   0           0           0      7
## 5:                0                   1           0           0      6
## 6:                0                   1           0           0      8
##    No_cell_phone
## 1:             0
## 2:             0
## 3:             0
## 4:             0
## 5:             0
## 6:             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.*

```
m4f <- lm(Y ~ Any_bin+No_cell_phone+Avg_no_bins_turned_bl, data=d2 )
cvcov4f <- cluster.vcov(m4f, ~ Street)
sec4f <-sqrt(diag(cvcov4f))
title <- "Adding 'No cell phone' "
stargazer(m4f,
          type = "text", omit = "Street",
          title = title,
          omit.stat = "f",
          add.lines = list(c("Street Fixed Effects","Yes")),
          se = list(sec4f),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## Adding 'No cell phone'
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                    Y
## ------------------------------------------------
## Any_bin                          0.125***
##                                   (0.018)
##
## No_cell_phone                    -0.049**
##                                   (0.016)
##
## Avg_no_bins_turned_bl            0.390***
##                                   (0.030)
##
## Constant                         0.372***
##                                   (0.023)
##
## ------------------------------------------------
## Street Fixed Effects                Yes
## Observations                       1,781
## R2                                 0.341
## Adjusted R2                        0.340
## Residual Std. Error          0.332 (df = 1777)
```

```
## ================================================
## Note:                 *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(m4f,n_observations = 1781,cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "0.092" "0.159"
```

**Answer: 4f**
The new ATE 95% CI is 0.092, 0.159 bins turned per week, now if (4f) is very similar to the previous one in (4c) because the variable "No cell phone" is totally independent of the treatment "Any_bin" so it cannot increase the standard error of Any_bin. Since No_cell_phone is negatively correlated with the outcome, so when it was not present the coefficient of Any_bin was slightly underestimated.

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.*

```
m4g <- lm(Y ~ Any_bin+Any_SMS+No_cell_phone +Avg_no_bins_turned_bl, data=d2 )
cvcov4g <- cluster.vcov(m4g, ~ Street)
sec4g <-sqrt(diag(cvcov4g))
title <- "Adding SMS treatment"
stargazer(m4g,
          type = "text", omit = "Street",
          title = title,
          omit.stat = "f",
          add.lines = list(c("Street Fixed Effects","Yes")),
          se = list(sec4g),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## Adding SMS treatment
## ================================================
##                            Dependent variable:
##                        ----------------------------
##                                    Y
## ------------------------------------------------
## Any_bin                          0.125***
##                                   (0.018)
##
## Any_SMS                           -0.031
##                                   (0.023)
##
## No_cell_phone                    -0.065**
##                                   (0.021)
##
## Avg_no_bins_turned_bl            0.388***
##                                   (0.030)
##
## Constant                         0.389***
##                                   (0.026)
```

```
##
## ------------------------------------------------
## Street Fixed Effects                Yes
## Observations                       1,781
## R2                                 0.342
## Adjusted R2                        0.340
## Residual Std. Error          0.332 (df = 1776)
## ================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(m4g,n_observations = 1781, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "0.092" "0.159"
```

**Answer: 4g**
The new ATE 95% CI is 0.092, 0.159 bins turned per week. The confidence interval does not change because the variable Any_SMS is independent of the treatment Any_bin and they don't have any interaction.

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.*

```
m4h <- lm(Y ~ Bin_with_sticker+Bin_without_sticker+Personal_SMS+Generic_SMS+Has_cell_phone, data=d2 )
cvcov4h <- cluster.vcov(m4h, ~ Street)
sec4h <-sqrt(diag(cvcov4h))
stargazer(m4h,
          type = "text", omit = "Street",
          title = "   ",
          omit.stat = "f",
          add.lines = list(c("Street Fixed Effects","Yes")),
          se = list(sec4h),
          star.cutoffs =c(0.05, 0.01, 0.001)
          )
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                                  Y
## ------------------------------------------------
## Bin_with_sticker              0.137***
##                                (0.026)
##
## Bin_without_sticker           0.131***
##                                (0.028)
##
## Personal_SMS                   -0.076*
##                                (0.032)
##
## Generic_SMS                    -0.060*
```

```
##                                 (0.030)
##
## Has_cell_phone                 0.090***
##                                 (0.026)
##
## Constant                       0.603***
##                                 (0.017)
##
## -------------------------------------------------
## Street Fixed Effects              Yes
## Observations                     1,781
## R2                               0.032
## Adjusted R2                      0.029
## Residual Std. Error        0.402 (df = 1775)
## =================================================
## Note:                *p<0.05; **p<0.01; ***p<0.001
```

```
coef <- coeftest(m4h,vcov=cluster.vcov(m4h, ~Street))
tc <- qt(1-0.05/2, 1781-1)
CI <- coef[3,1]+c(-1,1)*tc*coef[3,2]
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(CI,3))
```

```
## [1] "0.077" "0.185"
```

**Answer: 4g**
The 95% CI of the treatment with bin without sticker is 0.077, 0.185
There is a difference in CI between the results in (h) and (g) mainly because we are comparing CI of the treatment "Any bin" vs the treatment "Bin without sticker" two different treatments.

# 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.

```
d3 <- read.csv("./data/ebola_rct2.csv")
head(d3)
```

```
##   temperature_day0 vomiting_day0 treat_zmapp temperature_day14
## 1             99.5             1           0               98.6
## 2             97.4             0           0               98.0
## 3             97.0             0           1               97.9
## 4             99.7             1           0               98.4
## 5             99.6             1           1               99.3
## 6             98.3             1           1               99.8
##   vomiting_day14 male
## 1              1    0
## 2              1    0
## 3              0    1
## 4              1    0
```

```
## 5              1    0
## 6              1    1
```

```
str(d3)
```

```
## 'data.frame':    100 obs. of  6 variables:
##  $ temperature_day0 : num  99.5 97.4 97 99.7 99.6 ...
##  $ vomiting_day0    : int  1 0 0 1 1 1 1 0 1 1 ...
##  $ treat_zmapp      : int  0 0 1 0 1 1 1 1 0 1 ...
##  $ temperature_day14: num  98.6 98 97.9 98.4 99.3 ...
##  $ vomiting_day14   : int  1 1 0 1 1 1 1 1 1 0 ...
##  $ male             : int  0 0 1 0 0 1 1 1 0 1 ...
```

```
summary(d3)
```

```
##  temperature_day0 vomiting_day0   treat_zmapp   temperature_day14
##  Min.   : 97.0    Min.   :0.00   Min.   :0.00   Min.   : 97.1
##  1st Qu.: 97.7    1st Qu.:0.00   1st Qu.:0.00   1st Qu.: 98.1
##  Median : 98.6    Median :1.00   Median :0.00   Median : 98.7
##  Mean   : 98.5    Mean   :0.66   Mean   :0.41   Mean   : 99.1
##  3rd Qu.: 99.2    3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.: 99.7
##  Max.   :100.0    Max.   :1.00   Max.   :1.00   Max.   :102.5
##  vomiting_day14      male
##  Min.   :0.00   Min.   :0.00
##  1st Qu.:0.75   1st Qu.:0.00
##  Median :1.00   Median :0.00
##  Mean   :0.75   Mean   :0.37
##  3rd Qu.:1.00   3rd Qu.:1.00
##  Max.   :1.00   Max.   :1.00
```

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?*

```
model5a <- lm(vomiting_day14 ~ treat_zmapp,data=d3)
se.model5a <- sqrt(diag(vcovHC(model5a)))
title <- "ZMapp vomiting vs treatment"
stargazer( model5a,
           title= title,
           align=TRUE,
           type = "text",
           omit.stat="f",
           se = list(se.model5a),
           star.cutoffs =c(0.05, 0.01, 0.001)
)
```

```
##
## ZMapp vomiting vs treatment
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                             vomiting_day14
## ------------------------------------------------
```

```
## treat_zmapp                    -0.238**
##                                 (0.091)
##
## Constant                       0.847***
##                                 (0.048)
##
## ------------------------------------------------
## Observations                      100
## R2                               0.073
## Adjusted R2                      0.063
## Residual Std. Error        0.421 (df = 98)
## ================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(model5a,n_observations = 100, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.419" "-0.056"
```

```
paste(round(result$p_value,3))
```

```
## [1] "0.011"
```

**Answer: 5a**
The effect of zmapp is to reduce vomiting in day 14 with and ATE= -0.238+-0.182. The p-value is 0.011 which is lower than $\alpha = 0.05$

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.

```
model5b <- lm(vomiting_day14 ~ treat_zmapp + temperature_day0 + vomiting_day0,data=d3)
se.model5b <- sqrt(diag(vcovHC(model5b)))
title <- "Model 5b with Model 5a for comparisson"
stargazer( model5a, model5b,
           title= title,
           type = "text",
           omit.stat="f",
           se = list(se.model5a, se.model5b),
           star.cutoffs =c(0.05, 0.01, 0.001)
)
```

```
##
## Model 5b with Model 5a for comparisson
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                            vomiting_day14
##                          (1)            (2)
## ------------------------------------------------
## treat_zmapp           -0.238**        -0.166*
##                        (0.091)        (0.082)
##
## temperature_day0                      0.206**
##                                       (0.078)
```

```
##
## vomiting_day0                              0.065
##                                           (0.178)
##
## Constant                0.847***         -19.500*
##                          (0.048)          (7.610)
##
## ------------------------------------------------
## Observations               100              100
## R2                        0.073            0.311
## Adjusted R2               0.063            0.290
## Residual Std. Error 0.421 (df = 98) 0.367 (df = 96)
## ================================================
## Note:                 *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(model5b,n_observations = 100, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.328" "-0.003"
```

```
paste('p_level: ',round(result$p_value,3))
```

```
## [1] "p_level:  0.046"
```

**Anwer: 5b**
zmapp will have a reduction of vomiting in day 14 with an ATE= -0.166 +-0.164.
We see a decreased "zmapp effect". The p-value is 0.046 which is just below $\alpha = 0.05$
We might say that this is statistically significative, hence we reject the hypothesis that ZMapp treatment has no effect.

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

**Answer: 5c**
I would prefer the ATE in part (b) because is closer to the real value because, in model (b) we are controlling for the vomiting and temperature in day 0. Thus, they are absorving some of the spurious noise affecting model (a)

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.

```
model5d <- lm(vomiting_day14 ~ treat_zmapp + temperature_day14 + temperature_day0 + vomiting_day0,data=d
se.model5d <- sqrt(diag(vcovHC(model5d)))
title <- "Model 5c, including previous models for comparisson"
stargazer( model5a, model5b, model5d,
           title= title,
           type = "text",
           omit.stat="f",
           se = list(se.model5a, se.model5b, se.model5d),
           star.cutoffs =c(0.05, 0.01, 0.001)
)
```

```
##
## Model 5c, including previous models for comparisson
## ============================================================================
```

```
##                                 Dependent variable:
##                     ---------------------------------------------
##                                       vomiting_day14
##                           (1)             (2)             (3)
## ------------------------------------------------------------------
## treat_zmapp            -0.238**        -0.166*         -0.120
##                         (0.091)         (0.082)         (0.086)
##
## temperature_day14                                      0.060*
##                                                        (0.026)
##
## temperature_day0                        0.206**         0.177*
##                                         (0.078)         (0.077)
##
## vomiting_day0                            0.065           0.046
##                                         (0.178)         (0.173)
##
## Constant               0.847***        -19.500*        -22.600**
##                         (0.048)         (7.610)         (7.750)
##
## ------------------------------------------------------------------
## Observations              100             100             100
## R2                       0.073           0.311           0.340
## Adjusted R2              0.063           0.290           0.312
## Residual Std. Error 0.421 (df = 98) 0.367 (df = 96) 0.361 (df = 95)
## ==================================================================
## Note:                                *p<0.05; **p<0.01; ***p<0.001
```

```
result <- CI_p_value(model5d,n_observations = 100, cluster=NULL)
paste('The 95% CI of the ATE:')
```

```
## [1] "The 95% CI of the ATE:"
```

```
paste(round(result$CI,3))
```

```
## [1] "-0.29" "0.05"
```

```
paste('p_level: ',round(result$p_value,3))
```

```
## [1] "p_level:  0.165"
```

**Anwer: 5d**
ZMapp will have a reduction of vomiting in day 14 with an ATE = -0.120 +-0.172. We can see that there is a big dispersion in the possible values of the ATE. The p-value is 0.165 which is larger than $\alpha > 0.05$
We fail to reject the Ho or that the ATE is different from zero.

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

**Answer: 5e**
I would prefer the estimate of model 2 part(b), because it has less variation and it is more precise and all the regressors are independent of the outcome. In part (d) we are introducing temperature in day 14 which is no longer independent of the outcome. Now the right side of the equation is a function of the outcome.

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?*

```r
modelt5f <- lm(temperature_day14 ~ treat_zmapp +  male*treat_zmapp + male+ temperature_day0+ vomiting_da
se.modelt5f <- sqrt(diag(vcovHC(modelt5f)))

title <- "Temperature Model"
stargazer( modelt5f,
           title= title,
           type = "text",
           omit.stat="f",
           se = list(se.modelt5f),
           star.cutoffs =c(0.05, 0.01, 0.001)
)
```

```
##
## Temperature Model
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                           temperature_day14
## ------------------------------------------------
## treat_zmapp                    -0.231
##                                (0.118)
##
## male                          3.080***
##                                (0.122)
##
## temperature_day0              0.505***
##                                (0.105)
##
## vomiting_day0                   0.041
##                                (0.195)
##
## treat_zmapp:male              -2.080***
##                                (0.198)
##
## Constant                      48.700***
##                                (10.200)
##
## ------------------------------------------------
## Observations                     100
## R2                              0.906
## Adjusted R2                     0.901
## Residual Std. Error        0.452 (df = 94)
## ================================================
## Note:               *p<0.05; **p<0.01; ***p<0.001
```

```r
m <- modelt5f
coeftest(m, vcov=vcovHC(m))
```

```
##
## t test of coefficients:
##
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       48.7127    10.1940    4.78  6.5e-06 ***
## treat_zmapp       -0.2309     0.1183   -1.95    0.054 .
## male               3.0855     0.1218   25.34  < 2e-16 ***
```

```
## temperature_day0   0.5048     0.1045    4.83  5.3e-06 ***
## vomiting_day0      0.0411     0.1945    0.21    0.833
## treat_zmapp:male  -2.0767     0.1984  -10.47  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Answer: 5f**
The treatment effect with ZMapp on men compared to women leaves men temperature 1.0.1 degrees higher than women, on day 14.
women ATE = -0.23 men ATE = -0.23+3.09-2.08 = -0.23 + 1.01 = 0.78 This results suggests that the drug ZMapp is especially likely 1.01 degree women temperate compared to 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.)

**Answer: 5g**
This way of analysis is WRONG. If the researcher runs 20,000,000 different regressions, he very likely went on a "fishing expedition". He should use as a decision criteria an $\alpha = 0.05/20,000,000$ which will be very challenging to beat.

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?

**Answer: 5h**
Yes, I will be inclined to think it might be true, because I did not go into a "fishing expedition". Hence, this result might not be a special case only found when going on a fishing expedition.

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.*)

**Answer: 5i**
I would tell him that we need to rethink and rephrase his idea.You cannot randomly assign someone to be african descent to implement the experiment. Putting some thought into the research question to be meaningful. What is the outcome? or What are the possible outcomes? Assuming a hypothetical experiment (and assuming there are no ethical limitations). Outcome: Getting sick with contact with the virus. Ho: If you are from America and you are exposed to the virus, you don't get the decease Ha: You get the dicesease