Problem Set 3

Atit Wongnophadol

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

# removes all objects from the current workspace (R memory)
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 to calculate confidence interval given dataset d

ci_cluster <- function(d){

  # fit a linear regression model assuming clustering
  mod <- lm.cluster(data=d, formula = recall ~ treat, cluster="cluster" )

  # confidence interval accounting for clustering
  sprintf("The 95 pct confidence interval accounting for clustering is [%.3f %.3f]", confint(mod)[2, 1]

}

ci <- function(tau, t_stat, se){
  upper <- tau + t_stat*se
  lower <- tau - t_stat*se
  sprintf("The 95 pct confidence interval is [%.3f %.3f]", lower, upper)
}
```

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.

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

# put the data into data.table
d <- data.table(d)

# inspect the column headers and first several rows
head(d)
```

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

```
# rename the columns
names(d) <- c('study','treat','cluster','recall','impressed')

# extract only observations from study 1
d1 <- d[study==1]

# number of observations in study 1
N1 <- d1[ , .N]

# number of clusters
k1 <- d1[ , length(unique(cluster))]

sprintf("In study 1, there are %.0f observations allocated across %.0f unique clusters", N1, k1)
```

```
## [1] "In study 1, there are 1364 observations allocated across 577 unique clusters"
```

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.

```
# fit a linear regression model of study 1
mod1a <- lm(recall ~ treat, data = d1)

# confidence interval in study 1
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod1a)[2, 1], confint(mod1a)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [-0.051 0.031]"
```

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

Answer: the clusters are unique combinations of age, gender, and location. Clustering can lead to a more homogenous pool of observations in each cluster, making comparisons among clusters more differentiated (heterogenous). The higher heterogeneity among clusters leads to a higher variation (thus standard errors) in an ATE estimate.

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`.

```
library(miceadds) # for lm.cluster
```

```
## Loading required package: mice
```

```
## Loading required package: lattice
```

```
##
## Attaching package: 'mice'
```

```
## The following objects are masked from 'package:base':
##
##     cbind, rbind
```

```
## * miceadds 3.0-16 (2018-12-11 10:39:14)
```

```
library(multiwayvcov) # for cluster.vcov

## Approach 1: make use of lm.cluster

# fit a linear regression model assuming clustering
mod1c <- lm.cluster( data=d1, formula = recall ~ treat, cluster="cluster" )

# confidence interval accounting for clustering
sprintf("The 95 pct confidence interval accounting for clustering is [%.3f %.3f]", confint(mod1c)[2, 1]
```

```
## [1] "The 95 pct confidence interval accounting for clustering is [-0.056 0.037]"
```

```
## Approach 2: make use of cluster.vcov

# cov matrix given clustering
cvcov1c <- cluster.vcov(mod1a, ~ cluster)

# standard error of coefficients
sec1c <- sqrt(diag(cvcov1c))

# standard error of the treatment
se1c <- sec1c[2]

# t-statistic of the confidence interval
t_stat <- qt(0.975, nrow(d1)-2)

# ETA
eta_1c <- coef(mod1a)[2]

# confidence interval = ETA +/ t_stat * se
upper_bound_1c <- eta_1c + t_stat * se1c
lower_bound_1c <- eta_1c - t_stat * se1c

# confidence interval accounting for clustering
sprintf("The 95 pct confidence interval accounting for clustering is [%.3f %.3f]", lower_bound_1c, uppe
```

```
## [1] "The 95 pct confidence interval accounting for clustering is [-0.056 0.037]"
```

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

```r
# extract only observations from study 2
d2 <- d[study==2]

# number of observations in study 1
N2 <- d2[ , .N]

# number of clusters
k2 <- d2[ , length(unique(cluster))]

sprintf("In study 2, there are %.0f observations allocated across %.0f unique clusters", N2, k2)
```

```
## [1] "In study 2, there are 1342 observations allocated across 448 unique clusters"
```

```r
# confidence interval of study 2
ci_cluster(d2)
```

```
## [1] "The 95 pct confidence interval accounting for clustering is [-0.072 0.067]"
```

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?

```r
# number of observations in both studies
N_all <- d[ , .N]

# number of clusters in both studies
k_all <- d[ , length(unique(cluster))]

sprintf("In both studies, there are %.0f observations allocated across %.0f unique clusters", N_all, k_a
```

```
## [1] "In both studies, there are 2706 observations allocated across 1025 unique clusters"
```

```r
# calculate 95% confidence interval given both studies combined
ci_cluster(d)
```

```
## [1] "The 95 pct confidence interval accounting for clustering is [-0.207 -0.103]"
```

```r
## extract treatment effect and associated p-value

# fit the model
mod1e <- lm.cluster(data=d, formula = recall ~ treat, cluster="cluster" )

# treatment effect for both studies
eta_1e <- coef(mod1e)[2]

# p-value
p_1e <- summary(mod1e)[8]
```

```
## R^2= 0.02469
##
##               Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)  0.4541960 0.01857624 24.450377 4.986122e-132
## treat       -0.1550732 0.02673048 -5.801363  6.577799e-09
```

```r
sprintf("In both studies, the treatment effect is %.3f and the p-value is %.3f", eta_1e, p_1e)
```

```
## [1] "In both studies, the treatment effect is -0.155 and the p-value is 0.000"
```

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?

```
## extract treatment effect and associated p-value

# replicate the dataset and create a dummy variable
dat <- d
dat[, dummy := as.numeric(NA)]
dat[study == 2, dummy := 1]
dat[study == 1, dummy := 0]

# fit the model
mod1f <- lm.cluster(data=dat, formula = recall ~ treat + dummy, cluster="cluster" )

# treatment effect
eta_1f <- coef(mod1f)[2]

# p-value
p_1f <- summary(mod1f)[11]
```

```
## R^2= 0.19311
##
##                 Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  0.180684806 0.01697018 10.6471921 1.797039e-26
## treat       -0.006775249 0.02041542 -0.3318692 7.399881e-01
## dummy        0.426098820 0.02069695 20.5875147 3.551241e-94
```

```
sprintf("When applying a dummy variable, the treatment effect is %.3f and the p-value is %.3f", eta_1f,
```

```
## [1] "When applying a dummy variable, the treatment effect is -0.007 and the p-value is 0.740"
```

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: Model (e) compares the treatment and control groups without controlloing for block, whereas Model (f) controls for block (i.e., via the dummy variable). The reulst from Model (e) is biased because it is ignoring the fact the potential outcomes in each block can be very different even if they may not vary much within. As shown in the summary table below, we should observe almost zero effect of the treatment in each block (i.e., study), the result that is observed in Model (f).

In addition, were the weight between control and treatment is equal for each block, the bias would not be a concern. However, given the different weight observed in each block (i.e., study), the treatment effect without controlling for block is doubtful. In this example, the bias is towards a negative treatment effect given a relatively higher weight in the control group (with its potential outcome that is greater than that of the treatment).

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:data.table':
##
##     between, first, last
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
d %>%
  group_by(study, treat) %>%
  summarise(mean = mean(recall, na.rm = T), count = n())
```

```
## # A tibble: 4 x 4
## # Groups:   study [?]
##   study treat  mean count
##   <int> <int> <dbl> <int>
## 1     1     0 0.182   559
## 2     1     1 0.173   805
## 3     2     0 0.606  1007
## 4     2     1 0.603   335
```

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: the target treatment in this study was not at all a randomization as it intentionally applied the treatment to the two most populous counties. It is doubtful to claim that the ad had an impact on voting for two reasons. One, the outcome from this study might be due to a chance. If it had applied the treatment in other counties, the results might have shown no impact from the treatment. Two, targeting the two most populous counties clearly would result in some bias in measurement because the characteristics of the population in a populous area would be different from those in a less populous area, and therefore any treatment being implied would have resulted in potential outcomes that are biased.

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.

```
# identify treatment effect, t_stat, and se
ate_bin <- 0.187
t_stat <- qt(0.975, 1781-2)
se_bin <- 0.032
```

```
# confidence interval
ci(ate_bin, t_stat, se_bin)
```

```
## [1] "The 95 pct confidence interval is [0.124 0.250]"
```

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.

```
# identify treatment effect, t_stat, and se
ate_sms <- -0.024
t_stat <- qt(0.975, 1781-2)
se_sms <- 0.039

# confidence interval
ci(ate_sms, t_stat, se_sms)
```

```
## [1] "The 95 pct confidence interval is [-0.100 0.052]"
```

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

Answer: (i) Percentage of visits turned in bag (4.5% increase),

(ii) Avg. no. of bins turned in per week (11.5 # of bins increase),

(iii) Avg. weight (in kg) of recyclables turned in per week (18.7 kg increase), and

(iv) Avg. market value of recyclables given per week (10.8 soles increase)

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

Answer: None.

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: 0.281x2 = 0.562.

```
0.281*2
```

```
## [1] 0.562
```

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: Removing the "baseline" variable from the regression shouldn't affect the estimated ATE much because the baseline supposedly is based on randomization in the experiment. However, this would lead to a higher standard error of the coefficient, as the standard error of the estimate increases.

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

Answer: "has cell phone" is not bad control as it is not affected by the treatment (i.e., providing a recycling bin). It is unlikely that providing a recycling bin would lead to someone either own a phone or not.

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: it probably doesn't change the coefficient on SMS message much because neither variable has strong relationship with the treatment (i.e., recycling).

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: 3 x 3 x 2 [(no bin, bin with sticker, bin without sticker) x (no SMS, personal SMS, generic SMS) x (no cell phone, has cell phone)].

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: Without the bin treatment, no cell phone and no SMS message, the baseline is the mean of the dependent variable of interest. For example, the percentage of visited turned in bag is 78% and the average weight of recyclables turned in per week is 0.76 kg.

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

Answer: With the bin without stciker, percentage of visits turned in bag increases 3.5%.

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: The bin with sticker has a stronger treatment effect with the difference of 2.0% percentage (i.e., 5.5% - 3.5%).

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

Answer: the difference is not statistically significant, as indicated in the F-test p-value of 0.31 that compares the difference between the two.

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: It means that all possible combinations of the variables of interest are included in the models. The list of variables show that all possible effects given these combinations are captured in the model.

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).

```
d4 <- read.dta("./data/karlan_data_subset_for_class.dta")
d4 <- data.table(d4)
head(d4)
```

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

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

## Exploratory data analysis

# summary of all variables
summary(d4)
```

```
##      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
##
```

```
nrow(d4)
```

```
## [1] 1785
```

```
# notice that there are 120 records with the street value of -999
d4[order(street), .(mean_treat = mean(avg_bins_treat), count = .N), by = .(street)][street < 0]
```

```
##    street mean_treat count
## 1:   -999  0.6704861   120
```

```
# It is unclear if these were keyed in by mistake. We may decide to exclude these from the analysis and
```

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.

```
# fit a model
mod4a <- lm(avg_bins_treat ~ bin, data = d4)

# confidence interval
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod4a)[2, 1], confint(mod4a)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [0.096 0.175]"
```

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.

```
# fit a model
mod4b <- lm(avg_bins_treat ~ bin + base_avg_bins_treat, data = d4)

# confidence interval
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod4b)[2, 1], confint(mod4b)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [0.092 0.157]"
```

Answer: The confidence interval shifted lower with slightly narrower band. This is probably because the treatment effect is explained more by the pre-treatment covariate. And because the pre-treatment covariate is prognostic of the outcome (as shown in its statistical significance when included in the regression), it makes the estimated outcome more precise, lowering the standard error of the treatment variable and hence a tighter confidence interval.

The regression table below compares the results between the two models.

```
library(stargazer)
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
## use stargazer to print formatted tables
stargazer(mod4a, mod4b, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          align = TRUE, no.space=TRUE)
```

```
##
## ================================================
##                          Dependent variable:
##                      ---------------------------
##                              avg_bins_treat
##                            (1)            (2)
## ------------------------------------------------
## bin                      0.135***      0.125***
##                          (0.020)        (0.017)
## base_avg_bins_treat                    0.393***
##                                         (0.013)
## Constant                 0.635***      0.350***
##                          (0.012)        (0.014)
## ------------------------------------------------
## Observations              1,785          1,785
## ================================================
## Note:                *p<0.1; **p<0.05; ***p<0.01
```

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.

```
# convert 'street' to factor
d4$street <- factor(d4$street)

# fit a model
mod4c <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street, data = d4)

# confidence interval
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod4c)[2, 1], confint(mod4c)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [0.080 0.147]"
```

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.

Answer: this is because there are a large number of blocks, 180 in total. Further, blocking doesn't seem to have much impact on the regression results, as it doesn't change coefficients of the bin, baseline and the constant that much.

```
# number of street blocks
length(levels(d4$street))
```

```
## [1] 180
```

```
## use stargazer to print formatted tables
stargazer(mod4a, mod4b, mod4c, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          omit = "street",
          align = TRUE, no.space=TRUE)
```

```
##
## ================================================
##                          Dependent variable:
##                     ----------------------------
##                            avg_bins_treat
##                        (1)       (2)       (3)
## ------------------------------------------------
## bin                 0.135***  0.125***  0.114***
##                      (0.020)   (0.017)   (0.017)
## base_avg_bins_treat           0.393***  0.374***
##                                (0.013)   (0.014)
## Constant            0.635***  0.350***  0.368***
##                      (0.012)   (0.014)   (0.032)
## ------------------------------------------------
## Observations          1,785     1,785     1,782
## ================================================
## Note:                 *p<0.1; **p<0.05; ***p<0.01
```

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.

```
head(d4)
```

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

```
d4[ , nocell := 1-havecell]
```

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

```
# fit a model
mod4f <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell, data = d4)

# confidence interval
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod4f)[2, 1], confint(mod4f)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [0.082 0.149]"
```

Answer: Having a cell phone or not doesn't really have a relationship with whether the treatment (i.e., providing a bin) will be effective. Though cell phone is an indicator of wealth, education and social status which might in some way influence positive recycling behavior, and this relationship is reflected in the coefficient of the regression results.

```
## use stargazer to print formatted tables
stargazer(mod4a, mod4b, mod4c, mod4f, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          omit = "street",
          align = TRUE, no.space=TRUE)
```

```
##
## ===============================================================
##                               Dependent variable:
##                     ------------------------------------------
##                                   avg_bins_treat
##                        (1)       (2)       (3)       (4)
## ---------------------------------------------------------------
## bin                 0.135***  0.125***  0.114***  0.115***
##                     (0.020)   (0.017)   (0.017)   (0.017)
## base_avg_bins_treat           0.393***  0.374***  0.373***
##                               (0.013)   (0.014)   (0.014)
## nocell                                            -0.050***
##                                                    (0.017)
## Constant            0.635***  0.350***  0.368***  0.387***
##                     (0.012)   (0.014)   (0.032)   (0.032)
## ---------------------------------------------------------------
## Observations         1,785     1,785     1,782     1,781
## ===============================================================
## Note:                             *p<0.1; **p<0.05; ***p<0.01
```

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.

```
# fit a model
mod4g <- lm(avg_bins_treat ~ bin + base_avg_bins_treat + street + nocell + sms, data = d4)

# confidence interval
sprintf("The 95 pct confidence interval is [%.3f %.3f]", confint(mod4g)[2, 1], confint(mod4g)[2, 2])
```

```
## [1] "The 95 pct confidence interval is [0.082 0.148]"
```

Answer: SMS isn't a prognostic variable of the outcome of interest, so adding it to the regression doesn't do much in terms of explaining the outcome. If it does, it would have soaked up some effect of the treatment.

```
## use stargazer to print formatted tables
stargazer(mod4a, mod4b, mod4c, mod4f, mod4g, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
```

```
          omit = "street",
          align = TRUE, no.space=TRUE)
```

```
##
## ===============================================================
##                                 Dependent variable:
##                     ------------------------------------------
##                                     avg_bins_treat
##                       (1)      (2)      (3)       (4)      (5)
## ---------------------------------------------------------------
## bin                 0.135*** 0.125*** 0.114*** 0.115***  0.115***
##                     (0.020)  (0.017)  (0.017)   (0.017)  (0.017)
## base_avg_bins_treat          0.393*** 0.374*** 0.373***  0.373***
##                              (0.013)  (0.014)   (0.014)  (0.014)
## nocell                                         -0.050*** -0.047**
##                                                 (0.017)  (0.020)
## sms                                                       0.005
##                                                          (0.021)
## Constant            0.635*** 0.350*** 0.368*** 0.387***  0.385***
##                     (0.012)  (0.014)  (0.032)   (0.032)  (0.034)
## ---------------------------------------------------------------
## Observations         1,785    1,785    1,782     1,781    1,781
## ===============================================================
## Note:                                 *p<0.1; **p<0.05; ***p<0.01
```

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

```
# fit a model
mod4h <- lm(avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g + nocell + base_avg_bins_treat + street, data

# confidence interval
sprintf("The 95 pct confidence interval of bin with sticker is [%.3f %.3f]", confint(mod4h)[2, 1], confi
```

```
## [1] "The 95 pct confidence interval of bin with sticker is [0.084 0.171]"
```

```
# confidence interval
sprintf("The 95 pct confidence interval of bin without sticker is [%.3f %.3f]", confint(mod4h)[3, 1], co
```

```
## [1] "The 95 pct confidence interval of bin without sticker is [0.060 0.146]"
```

```
## use stargazer to print formatted tables
stargazer(mod4a, mod4b, mod4c, mod4f, mod4g, mod4h, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          omit = "street",
          align = TRUE, no.space=TRUE)
```

```
##
## ========================================================================
##                                      Dependent variable:
##                     ---------------------------------------------------
##                                          avg_bins_treat
##                       (1)      (2)      (3)       (4)      (5)      (6)
## ------------------------------------------------------------------------
## bin                 0.135*** 0.125*** 0.114*** 0.115***  0.115***
```

```
##                       (0.020)   (0.017)   (0.017)    (0.017)   (0.017)
## bin_s                                                                    0.128***
##                                                                          (0.022)
## bin_g                                                                    0.103***
##                                                                          (0.022)
## sms_p                                                                     -0.008
##                                                                          (0.025)
## sms_g                                                                     0.020
##                                                                          (0.025)
## base_avg_bins_treat           0.393*** 0.374*** 0.373***  0.373*** 0.374***
##                               (0.013)  (0.014)   (0.014)  (0.014)  (0.014)
## nocell                                          -0.050*** -0.047** -0.046**
##                                                  (0.017)  (0.020)  (0.020)
## sms                                                        0.005
##                                                           (0.021)
## Constant              0.635*** 0.350*** 0.368*** 0.387***  0.385*** 0.385***
##                       (0.012)  (0.014)  (0.032)   (0.032)  (0.034)  (0.034)
## ------------------------------------------------------------------------------
## Observations           1,785    1,785    1,782     1,781    1,781    1,781
## ==============================================================================
## Note:                                          *p<0.1; **p<0.05; ***p<0.01
```

Answer: the results are close to the previous regression, except for that we are now breaking out the treatment effect of the bin into "with" and "without" a sticker. Where previously the effect could be thought of average between the two, and now we clearly see the effect of the two different treatments. The effect from bin with sticker is clearly higher than the effect from the bin without sticker, both are statistically significant, and are practically intuitive.

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.

```
# read the data
d5 <- read.csv("./data/ebola_rct2.csv")
d5 <- data.table(d5)

# see the header
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
```

```
# rename the columns
names(d5) <- c('temp_pre','vom_pre','treat','temp_post','vom_post','male')
```

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?

```
# fit a model
mod5a <- lm(vom_post ~ treat, data = d5)

# use stargazer to print formatted tables
stargazer(mod5a, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          align = TRUE, no.space=TRUE)
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                         vom_post
## ----------------------------------------
## treat                   -0.238***
##                          (0.086)
## Constant                0.847***
##                          (0.055)
## ----------------------------------------
## Observations              100
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

```
# p-value of the treatment
summary(mod5a)$coefficients[2,4]
```

```
## [1] 0.006595412
```

Answer: the estimated effect is -0.238, or a reductiong of 23.8 in percentage point, with the standard error of 0.086. The p-value is 0.0066.

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.

```
# fit a model
mod5b <- lm(vom_post ~ treat + vom_pre + temp_pre, data = d5)

# use stargazer to print formatted tables
stargazer(mod5a, mod5b, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          align = TRUE, no.space=TRUE)
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                         vom_post
##                     (1)            (2)
## ----------------------------------------
## treat           -0.238***      -0.166**
##                  (0.086)        (0.076)
## vom_pre                          0.065
##                                 (0.146)
## temp_pre                        0.206***
##                                 (0.076)
## Constant         0.847***      -19.470**
##                  (0.055)        (7.441)
## ----------------------------------------
## Observations       100            100
## ========================================
## Note:          *p<0.1; **p<0.05; ***p<0.01
```

```
# p-value of the treatment
summary(mod5b)$coefficients[2,4]
```

```
## [1] 0.03112852
```

Answer: the estimated effect is -0.166, or a reductiong of 16.6 in percentage point, with the standard error of 0.076. The p-value is 0.0311.

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

Answer: I prefer the estimate from part (b) because it has a better precision in the outcome due to a reduced variability in the outcome (i.e., vomit post treatment) and a lower standard error of the treatment coefficient. Further, accounting for the pre-treatment condition equalizes the comparison point between the control and treatment groups (i.e., you're really comparing the treatment effect on an apples-to-apples basis).

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.

```
# fit a model
mod5d <- lm(vom_post ~ treat + vom_pre + temp_pre + temp_post, data = d5)

# use stargazer to print formatted tables
stargazer(mod5a, mod5b, mod5d, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          align = TRUE, no.space=TRUE)
```

```
##
## ========================================
##                   Dependent variable:
##               ---------------------------
##                         vom_post
##                 (1)       (2)       (3)
```

```
## ------------------------------------------
## treat        -0.238*** -0.166**    -0.120
##               (0.086)   (0.076)   (0.078)
## vom_pre                  0.065     0.046
##                         (0.146)   (0.144)
## temp_pre               0.206***   0.177**
##                         (0.076)   (0.076)
## temp_post                         0.060**
##                                   (0.029)
## Constant     0.847***  -19.470** -22.592***
##               (0.055)   (7.441)   (7.477)
## ------------------------------------------
## Observations    100       100       100
## ==========================================
## Note:           *p<0.1; **p<0.05; ***p<0.01
```

```r
# p-value of the treatment
summary(mod5d)$coefficients[2,4]
```

```
## [1] 0.1254056
```

Answer: the estimated effect is -0.120, or a reductiong of 12.0 in percentage point, with the standard error of 0.078. The p-value is 0.125.

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

Answer: I still prefer the estimate from part (b) simply because the model in part (d) accounts for post-treatment covariate that can be influenced by the treatment. In other words, I believe that *temperature on day 14* is a bad control that shouldn't be included in the model as it would incorrectly cloud out the effect of the treatment on 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
# fit a model
mod5f <- lm(temp_post ~ treat + vom_pre + temp_pre + male + treat*male, data = d5)

# use stargazer to print formatted tables
stargazer(mod5b, mod5f, type = "text",
          omit.stat=c("LL","ser","f","adj.rsq","rsq"),
          object.names = TRUE,
          align = TRUE, no.space=TRUE)
```

```
##
## ========================================
##                  Dependent variable:
##              ---------------------------
##                  vom_post      temp_post
##                    (1)           (2)
##                   mod5b         mod5f
## ----------------------------------------
## treat           -0.166**       -0.231*
##                 (0.076)        (0.119)
```

```
## vom_pre          0.065          0.041
##                 (0.146)        (0.182)
## temp_pre        0.206***       0.505***
##                 (0.076)        (0.095)
## male                           3.085***
##                                (0.126)
## treat:male                     -2.077***
##                                (0.192)
## Constant       -19.470**       48.713***
##                 (7.441)        (9.266)
## ----------------------------------------
## Observations      100            100
## ========================================
## Note:         *p<0.1; **p<0.05; ***p<0.01
```

```
# p-value of the treatment
summary(mod5f)$coefficients[2,4]
```

```
## [1] 0.05478966
```

Answer: From the regression results above, which include the interaction term between male and treatment, we can see that being male who received the treatment results in a reduction in the termperature with statistical significance (i.e., treatment effect of -2.080 and standard error of 0.192). This tells us that the treatment on male is likely to reduce men's temperature as compared to women's, holding all else equal.

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 if ZMapp works especially well for curing Ebola in men, and why? (This question is conceptual can be answered without performing any computation.)

Answer: his claim would be doubtful because he went through a very comprehensive list of variables. Very likely, he would find a variable that is statistically significant to which he could claim that it is predictive of the outcome. This is technically a p-value hacking or 'fishing' technique that goes against a proper research methodology in which variables of interest are determined upfront. The statistical significance from the 'fishing' technique may be just by chance that doesn't mean much. Were the research to repeat the same experiment using the claimed variable, would he still get statistical significance most of the time?

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: I would be more inclined to believe the heterogeneous treatment effect given a proper research methodology. In this case, as I tested only this effect and nothing else, I would be more confident in the research results.

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: this is a FUQ (fundamentally unanswered question) that was discussed in async. It is very challenging (if not possible) to design an experiment that associates races with outcomes of a disease. There is no logical connection between the two. And thus hard to design a causal-effect kind of experimentation.