Problem Set 3

Gaurav Khanna

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

```
## Warning: package 'data.table' was built under R version 3.4.4
```

```
library(foreign)
```

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

```
# Libraries for robust and clustered standard errors
library(lmtest)
```

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

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

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

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

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

```
library(sandwich)
```

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

```
library(multiwayvcov)
library(stargazer)
```

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

```
##
## 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
```

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 extend the OLS summary object
lr_extend <- function(lr1) {
  # OLS
  # Variance-Covariance matrix
  lr1$ols.vcov <- vcovHC(lr1, "const")
```

```
  # SE's
  lr1$ols.se   <- sqrt(diag(lr1$ols.vcov))
  # Calculating the confidence interval with the built in function
  lr1$ols.confint <- confint(lr1, level = 0.95)
  # Cluster Variance-Covariance matrix
  lr1$cluster.vcov <- cluster.vcov(lr1, ~ cluster)
  # coeftest(lr1, lr1$cluster.vcov)
  # Cluster standard errors
  lr1$cluster.se <- sqrt(diag(lr1$cluster.vcov))
  # Cluster confidence interval Beta +- 2 SE's
  lr1$cluster.confint <- c(lr1$coefficients['treat_ad'] - 2 * lr1$cluster.se['treat_ad'],
                           lr1$coefficients['treat_ad'] + 2 * lr1$cluster.se['treat_ad'])
  # Return the extended object
  return(lr1)
}
```

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.

```
d1 <- read.csv("./data/broockman_green_anon_pooled_fb_users_only.csv")
head(d1)
```

```
##   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
```

```
# Summary of the observations from the expriment
summary(d1)
```

```
##     studyno         treat_ad                              cluster
##  Min.   :1.000   Min.   :0.0000   Study 1, Cluster Number 799:  24
##  1st Qu.:1.000   1st Qu.:0.0000   Study 2, Cluster Number 333:  23
##  Median :1.000   Median :0.0000   Study 1, Cluster Number 781:  20
##  Mean   :1.496   Mean   :0.4213   Study 1, Cluster Number 800:  17
##  3rd Qu.:2.000   3rd Qu.:1.0000   Study 2, Cluster Number 425:  17
##  Max.   :2.000   Max.   :1.0000   Study 2, Cluster Number 501:  17
##                                   (Other)                    :2588
##   name_recall     positive_impression
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
```

```
##  Median :0.0000   Median :0.0000
##  Mean   :0.3887   Mean   :0.2603
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##  NA's   :5        NA's   :5
```

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.

```
# Separating out study 1
study1 <- d1[which(d1$studyno == 1),]
head(study1)
```

```
##      studyno treat_ad                   cluster name_recall
## 1343       1        1 Study 1, Cluster Number 1           0
## 1344       1        1 Study 1, Cluster Number 1           0
## 1345       1        1 Study 1, Cluster Number 3           0
## 1346       1        1 Study 1, Cluster Number 4           0
## 1347       1        1 Study 1, Cluster Number 5           0
## 1348       1        1 Study 1, Cluster Number 9           0
##      positive_impression
## 1343                   0
## 1344                   0
## 1345                   0
## 1346                   0
## 1347                   0
## 1348                   0
```

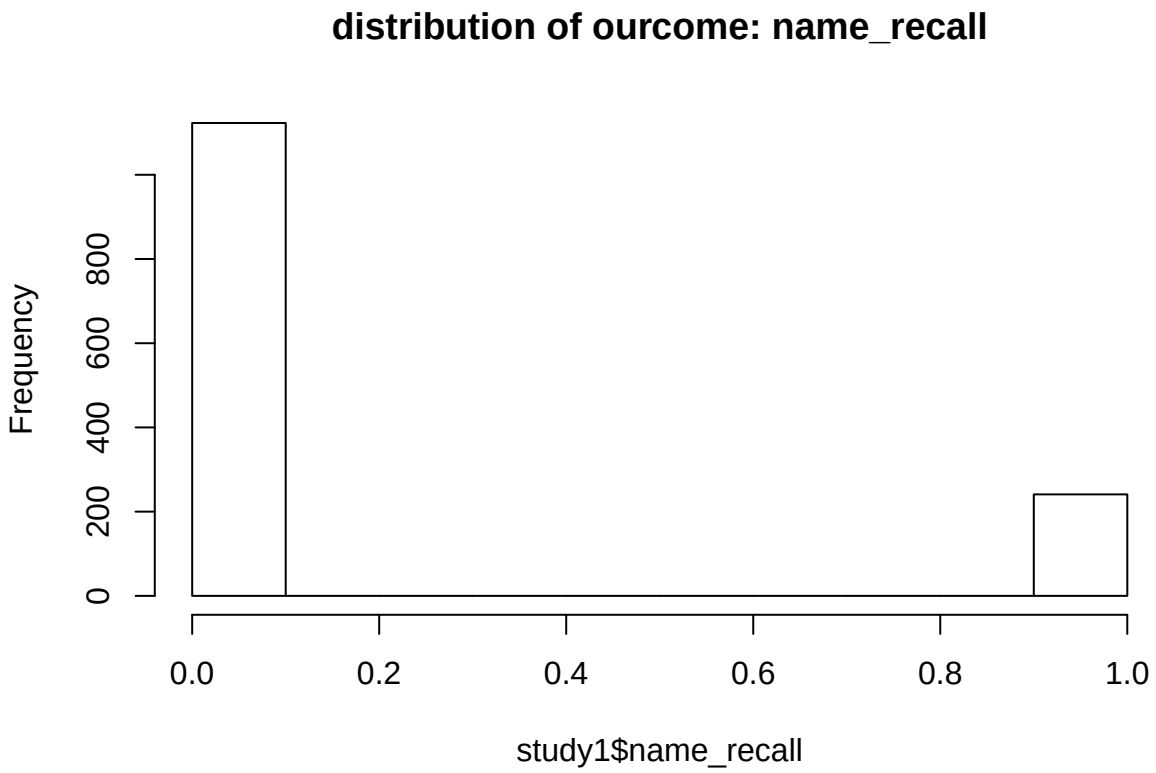
```
summary(study1)
```

```
##     studyno     treat_ad                               cluster
##  Min.   :1   Min.   :0.0000   Study 1, Cluster Number 799:  24
##  1st Qu.:1   1st Qu.:0.0000   Study 1, Cluster Number 781:  20
##  Median :1   Median :1.0000   Study 1, Cluster Number 800:  17
##  Mean   :1   Mean   :0.5902   Study 1, Cluster Number 743:  16
##  3rd Qu.:1   3rd Qu.:1.0000   Study 1, Cluster Number 801:  16
##  Max.   :1   Max.   :1.0000   Study 1, Cluster Number 779:  15
##                               (Other)                   :1256
##   name_recall     positive_impression
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1767   Mean   :0.1305
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

Note: We see 1276 observations in study 1

```r
# Regress name_recall on treat_ad

# Checking the outcome variable
hist(study1$name_recall, main = "distribution of ourcome: name_recall")
```



Note: Study 1 has a high propotion with name_recall = 0 Continuing with the linear regression

```r
# Linear regression
lr1 <- lm(name_recall ~ treat_ad, data=study1)
summary(lr1)
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad, data = study1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.1825 -0.1825 -0.1727 -0.1727  0.8273
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.182469   0.016142  11.304   <2e-16 ***
## treat_ad    -0.009798   0.021012  -0.466    0.641
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3817 on 1362 degrees of freedom
## Multiple R-squared:  0.0001596,  Adjusted R-squared:  -0.0005745
## F-statistic: 0.2174 on 1 and 1362 DF,  p-value: 0.6411
```

```
# Confidence interval
# Using the function defined above
lr1 <- lr_extend(lr1)
print('OLS Confidence Interval')
```

```
## [1] "OLS Confidence Interval"
```

```
lr1$ols.confint
```

```
##                    2.5 %     97.5 %
## (Intercept)  0.15080247 0.21413492
## treat_ad    -0.05101765 0.03142188
```

```
# lr1$ols.confint <- confint(lr1, level = 0.95)
# lr1$ols.confint
```

Answer: The confidence interval is: treat_ad -0.05101765 0.03142188

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

Answer: Clusters in the study are composed of individuals with the same age, gender and location The members of above clusters have little variance in study relevant attributes (age, gender, location). When assigned as a grouop, they suppress the variance in Y leading ot a smaller estimate for errors. When we take clustering into account, we correct for this, leading to an inflation in the SE's.

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`.

```
# Cluster standard error and  confidence interval
# Calculated in the function above
print('Cluster SE')
```

```
## [1] "Cluster SE"
```

```
lr1$cluster.se
```

```
## (Intercept)    treat_ad
##  0.01849151  0.02375363
```

```
print('Cluster Confidence interval')
```

```
## [1] "Cluster Confidence interval"
```

```
lr1$cluster.confint
```

```
##    treat_ad    treat_ad
## -0.05730514  0.03770936
```

Answer: The confidence interval based on Cluster standard errors is:

treat_ad treat_ad -0.05730514 0.03770936

```
# Visualizing the variance covariance matrix
stargazer(lr1, lr1,
          se = list(sqrt(diag(lr1$cluster.vcov))), header=F)
```

```
##
## \begin{table}[!htbp] \centering
##   \caption{}
##   \label{}
## \begin{tabular}{@{\extracolsep{5pt}}lcc}
## \\[-1.8ex]\hline
## \hline \\[-1.8ex]
##  & \multicolumn{2}{c}{\textit{Dependent variable:}} \\
## \cline{2-3}
## \\[-1.8ex] & \multicolumn{2}{c}{name\_recall} \\
## \\[-1.8ex] & (1) & (2)\\
## \hline \\[-1.8ex]
##  treat\_ad & $-$0.010 & $-$0.010 \\
##   & (0.024) & (0.021) \\
##   & & \\
##  Constant & 0.182$^{***}$ & 0.182$^{***}$ \\
##   & (0.018) & (0.016) \\
##   & & \\
## \hline \\[-1.8ex]
## Observations & 1,364 & 1,364 \\
## R$^{2}$ & 0.0002 & 0.0002 \\
## Adjusted R$^{2}$ & $-$0.001 & $-$0.001 \\
## Residual Std. Error (df = 1362) & 0.382 & 0.382 \\
## F Statistic (df = 1; 1362) & 0.217 & 0.217 \\
## \hline
## \hline \\[-1.8ex]
## \textit{Note:}  & \multicolumn{2}{r}{$^{*}$p$<$0.1; $^{**}$p$<$0.05; $^{***}$p$<$0.01} \\
## \end{tabular}
## \end{table}
```

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

```
# Separating out study 2
study2 <- d1[which(d1$studyno == 2),]
head(study2)
```

```
##   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
```

```
summary(study2)
```

```
##     studyno     treat_ad                              cluster
##  Min.   :2   Min.   :0.0000   Study 2, Cluster Number 333:  23
##  1st Qu.:2   1st Qu.:0.0000   Study 2, Cluster Number 425:  17
##  Median :2   Median :0.0000   Study 2, Cluster Number 501:  17
##  Mean   :2   Mean   :0.2496   Study 2, Cluster Number 546:  17
##  3rd Qu.:2   3rd Qu.:0.0000   Study 2, Cluster Number 516:  15
##  Max.   :2   Max.   :1.0000   Study 2, Cluster Number 541:  15
##                               (Other)                    :1238
##   name_recall     positive_impression
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :1.0000   Median :0.0000
##  Mean   :0.6051   Mean   :0.3927
##  3rd Qu.:1.0000   3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000
##  NA's   :5        NA's   :5
```

```
# Linear regression
lr2 <- lm(name_recall ~ treat_ad, data=study2)
summary(lr2)
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad, data = study2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6058 -0.6058  0.3942  0.3942  0.3970
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.605788   0.015454  39.199   <2e-16 ***
## treat_ad    -0.002803   0.030874  -0.091    0.928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4892 on 1335 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  6.176e-06,  Adjusted R-squared:  -0.0007429
## F-statistic: 0.008245 on 1 and 1335 DF,  p-value: 0.9277
```

```
# Extending the summary object to find the cluster SE and confidence interval
lr2 <- lr_extend(lr2)
print('OLS Confidence Interval')
```

```
## [1] "OLS Confidence Interval"
```

```
lr2$ols.confint
```

```
##                 2.5 %    97.5 %
## (Intercept)  0.5754710 0.6361058
## treat_ad    -0.0633702 0.0577635
```

```r
# Cluster standard error and  confidence interval
# Calculated in the function above
print('Cluster SE')
```

```
## [1] "Cluster SE"
```

```r
lr2$cluster.se
```

```
## (Intercept)    treat_ad
##  0.01818893  0.03550334
```

```r
print('Cluster Confidence interval')
```

```
## [1] "Cluster Confidence interval"
```

```r
lr2$cluster.confint
```

```
##    treat_ad    treat_ad
## -0.07381003  0.06820333
```

Answer: The cluster confidence interval from study 2 is -0.07381003 0.06820333

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
# Linear regression on the entire sample
lr3 <- lm(name_recall ~ treat_ad, data=d1)
summary(lr3)
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4542 -0.4542 -0.2991  0.5458  0.7009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45420    0.01219  37.262   <2e-16 ***
## treat_ad    -0.15507    0.01876  -8.265   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4816 on 2699 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.02469,    Adjusted R-squared:  0.02432
## F-statistic: 68.31 on 1 and 2699 DF,  p-value: < 2.2e-16
```

```r
# Extending the summary object to find the cluster SE and confidence interval
lr3 <- lr_extend(lr3)
print('OLS Confidence Interval')
```

```
## [1] "OLS Confidence Interval"
```

```
lr3$ols.confint
```

```
##                  2.5 %     97.5 %
## (Intercept)  0.4302949  0.4780971
## treat_ad    -0.1918631 -0.1182834
```

```
# Cluster standard error and  confidence interval
# Calculated in the function above
print('Cluster SE')
```

```
## [1] "Cluster SE"
```

```
lr3$cluster.se
```

```
## (Intercept)    treat_ad
##  0.01857624  0.02673048
```

```
print('Cluster Confidence interval')
```

```
## [1] "Cluster Confidence interval"
```

```
lr3$cluster.confint
```

```
##   treat_ad   treat_ad
## -0.2085342 -0.1016123
```

```
# Treatment effect estimate
print('Treatment effect estimate')
```

```
## [1] "Treatment effect estimate"
```

```
lr3$coefficients['treat_ad']
```

```
##   treat_ad
## -0.1550732
```

```
print('p-value')
```

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

Answer: The cluster confidence interval from all studies combined is -0.2085342 -0.1016123 ** The treatment effect is -0.1550732 The p value of close to 0 (< 2.2e-16)**

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

```
# including a dummy variable for data in study 1 or 2
d1 <- within(d1, {study10_study21 = ifelse(studyno == 1, 0, 1)})
# Testing
head(d1)
```

```
##   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 study10_study21
## 1                   0               1
## 2                   0               1
## 3                   0               1
## 4                   0               1
## 5                   1               1
## 6                   0               1
```

```
summary(d1)
```

```
##     studyno         treat_ad                              cluster
##  Min.   :1.000   Min.   :0.0000   Study 1, Cluster Number 799:  24
##  1st Qu.:1.000   1st Qu.:0.0000   Study 2, Cluster Number 333:  23
##  Median :1.000   Median :0.0000   Study 1, Cluster Number 781:  20
##  Mean   :1.496   Mean   :0.4213   Study 1, Cluster Number 800:  17
##  3rd Qu.:2.000   3rd Qu.:1.0000   Study 2, Cluster Number 425:  17
##  Max.   :2.000   Max.   :1.0000   Study 2, Cluster Number 501:  17
##                                   (Other)                    :2588
##   name_recall     positive_impression study10_study21
##  Min.   :0.0000   Min.   :0.0000      Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000      1st Qu.:0.0000
##  Median :0.0000   Median :0.0000      Median :0.0000
##  Mean   :0.3887   Mean   :0.2603      Mean   :0.4959
##  3rd Qu.:1.0000   3rd Qu.:1.0000      3rd Qu.:1.0000
##  Max.   :1.0000   Max.   :1.0000      Max.   :1.0000
##  NA's   :5        NA's   :5
```

```
# Repeating the regression with the entire sample and a dummy variable for the studyno
lr4 <- lm(name_recall ~ treat_ad + study10_study21 , data=d1)
summary(lr4)
```

```
##
## Call:
## lm(formula = name_recall ~ treat_ad + study10_study21, data = d1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.6068 -0.1807 -0.1739  0.3932  0.8261
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.180685   0.015994  11.297   <2e-16 ***
## treat_ad        -0.006775   0.018177  -0.373    0.709
## study10_study21  0.426099   0.017955  23.731   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4381 on 2698 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.1931, Adjusted R-squared:  0.1925
## F-statistic: 322.8 on 2 and 2698 DF,  p-value: < 2.2e-16
```

```
# Extending the summary object to find the cluster SE and confidence interval
lr4 <- lr_extend(lr4)
print('OLS Confidence Interval')
```

```
## [1] "OLS Confidence Interval"
```

```
lr4$ols.confint
```

```
##                       2.5 %     97.5 %
## (Intercept)      0.14932337 0.21204624
## treat_ad        -0.04241695 0.02886645
## study10_study21  0.39089098 0.46130666
```

```
# Cluster standard error and  confidence interval
# Calculated in the function above
print('Cluster SE')
```

```
## [1] "Cluster SE"
```

```
lr4$cluster.se
```

```
##     (Intercept)        treat_ad study10_study21
##      0.01697018      0.02041542      0.02069695
```

```
print('Cluster Confidence interval')
```

```
## [1] "Cluster Confidence interval"
```

```
lr4$cluster.confint
```

```
##    treat_ad    treat_ad
## -0.04760609  0.03405559
```

```
# Treatment effect estimate
print('Treatment effect estimate')
```

```
## [1] "Treatment effect estimate"
```

```
lr4$coefficients['treat_ad']
```

```
##     treat_ad
## -0.006775249
```

```
print('p-value')
```

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

Answer: The cluster confidence interval from all studies combined is -0.04760609 0.03405559 ** The treatment effect is -0.006775249 The p value of close to 0 (< 2.2e-16)**

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: The number of observations in study 1 & 2 is about the same, so we can assume that the probability of assigment to each group is the same. The treatment effect (and the confidence intervals) are different among the 2 studies though, introducing a bias when we mix the observations from the studies (i.e. (e) is biased.(f) captures the impact of the studies in the coefficient of the dummy variable. This creates a better estimate of the treatment impact in (f)

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 claims would differ due to the following reasons: 1> The places where the ads ran were selected due to certain characteristics(Quote: "Targeting to reach people in two of the most populated counties in Florida, Dade and Broward, which have a combined population of 4.2 million"). We are not sure if these characteristics (like the population density or total population) we included as control while determining the effects of the treatment (ADS). If not included, the ATE claimed would be biased and included some of the noise due to these factors 2> In places where the ads ran, exposure was determined by selection characteristics by FB (Political interest, education, work, search phrases etc) (Quote: "Not only were our display ads based on the results of the Facebook research, but a lot of our ads ran to people who we originally aggregated on a remarketing list through the Facebook acquisition campaign."). Again, its not clear if the study controlled for, blocked or clustered around these distinguishing factors. Different treatment of these variables would lead to different estimates for ATE

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.

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.

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

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

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

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.

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

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

3 Multifactor Experiments

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

a. What is the full experimental design for this experiment? Tell us the dimensions, such as 2x2x3. (Hint: the full results appear in Panel 4B.)
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?
c. In column (1) of Table 4B, interpret the magnitude of the coefficient on "bin without sticker." What does it mean?
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?
e. Is this difference you just described statistically significant? Explain which piece of information in the table allows you to answer this question.
f. Notice that Table 4C is described as results from "fully saturated" models. What does this mean? Looking at the list of variables in the table, explain in what sense the model is "saturated."

4 Now! Do it with data

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

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

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

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

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.

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.

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.

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

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

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.

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.

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.

5 A Final Practice Problem

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

Here are the (fake) data.

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

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

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

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

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.

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

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

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

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?

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

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

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