Problem Set 3

Pavan Kurapati

June 26, 2018

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

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

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

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

```
library(multiwayvcov)
```

0 Write Functions

You're going to be doing a few things a *number* of times – calculating robust standard errors, calculating clustered standard errors, and then calculating the confidence intervals that are built off these standard errors.

After you've worked through a few of these questions, I suspect you will see places to write a function that will do this work for you. Include those functions here, if you write them.

```
# Calculate the confidence interval of a given parameter (index) and return lower
# and upper bound
ci_calc <- function(model,level,index,v_cov){
  if(missing(v_cov)) {
    p<-data.frame(confint(model,level=level)[index,])
  } else {
    p<-data.frame(coefci(model,level=level,vcov=v_cov)[index,])
  }
  return(c(p[1,],p[2,]))
}
```

1 Replicate Results

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

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

a. Using regression without clustered standard errors (that is, ignoring the clustered assignment), compute a confidence interval for the effect of the ad on candidate name recognition in Study 1 only (the dependent variable is "name_recall").

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

```r
# Create subsets for study 1 and study 2
d_study1 <- subset(d,d$studyno==1)
d_study2 <- subset(d,d$studyno==2)
```

```r
# Linear regression model
model1a = lm(name_recall~treat_ad,data=d_study1)
# treatment coefficient
summary(model1a)$coefficients[2,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -0.009797887  0.021012191 -0.466295336  0.641078683
```

```r
se_1a <- summary(model1a)$coefficients[2,2]
```

```r
# Confidence Interval
ci_1a <- ci_calc(model1a,0.95,2)
ci_1a
```

```
## [1] -0.05101765  0.03142188
```

From the co-efficients above, it is evident that the probability of name_recall drops by 0.0097 with treatment. Confidence Interval of slope co-efficient is -0.0510177 to 0.0314219, to be interpreted as probability effect on name_recall with treatment.The result does not have statistical significance

For sanity check, let us use Bayes theorem to validate. This is an extra step. From Bayes theorem: $P(name\_recall|treat = 1) = \frac{P(treat=1|name\_recall)*P(name\_recall)}{P(treat=1)}$

$P(name\_recall|treat = 0) = \frac{P(treat=0|name\_recall)*P(name\_recall)}{P(treat=0)}$

Let us fill these values

```r
n1 = nrow(d_study1)
nr1 = nrow(d_study1[d_study1$name_recall==1,])
nr0 = nrow(d_study1[d_study1$name_recall==0,])
nt1 = nrow(d_study1[d_study1$treat_ad==1,])
nt0 = nrow(d_study1[d_study1$treat_ad==0,])

# P(T=1|NAME_RECALL=1)
p_t1_nr1 <- nrow(d_study1[(d_study1$treat_ad==1) & (d_study1$name_recall==1),])/nr1
# P(NAME_RECALL=1)
p_nr1 <- nr1/n1
#P(Treat=1)
p_t1 <- nt1/n1

p_nr1_t1 = p_t1_nr1*p_nr1/p_t1

# P(T=0|NAME_RECALL=1)
p_t0_nr1 <- nrow(d_study1[(d_study1$treat_ad==0) & (d_study1$name_recall==1),])/nr1

#P(Treat=0)
```

```
p_t0 <- nt0/n1

p_nr1_t0 = p_t0_nr1*p_nr1/p_t0

# Treatment effect
p_nr1_t1-p_nr1_t0
```

```
## [1] -0.009797887
```

This matches with the co-efficient we obtained above which validates the Linear regression model

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

Clusters are individuals with unique combination of age,gender and location. Number of clusters are generally less than the number of samples in full model and hence it increases the standard errors compared to full model.

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`.

```
model1a$cluster.vcov <- cluster.vcov(model1a, d_study1$cluster)
cf <- coeftest(model1a, model1a$cluster.vcov)
cf
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1824687  0.0184915  9.8677   <2e-16 ***
## treat_ad    -0.0097979  0.0237536 -0.4125   0.6801
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ci_1c <- ci_calc(model1a,0.95,2,model1a$cluster.vcov)
ci_1c
```

```
## [1] -0.05639555  0.03679977
```

There is a slight increase in standard error with clustering, which is 0.0237536. Standard error without clustering value in part 1a was 0.0210122. However, the coefficients did not change much. The CI has minor difference compared to 1a.

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

```
model1d = lm(name_recall~treat_ad,data=d_study2)
summary(model1d)$coefficients[2,]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## -0.002803349  0.030874006 -0.090799637  0.927665420
```

```
se_1d <- summary(model1d)$coefficients[2,2]
ci_1d <- ci_calc(model1d,0.95,2)
ci_1d
```

```
## [1] -0.0633702  0.0577635
```

```r
model1d$cluster.vcov <- cluster.vcov(model1d, d_study2$cluster)
cf_1d <- coeftest(model1d, vcov=model1d$cluster.vcov)
cf_1d
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.6057884  0.0181889  33.305   <2e-16 ***
## treat_ad    -0.0028033  0.0355033  -0.079   0.9371
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Confidence Interval
ci_1d <- ci_calc(model1d,0.95,2,model1d$cluster.vcov)
ci_1d
```

```
## [1] -0.07245176  0.06684507
```

Study2 also has higher standard error with clustering(0.0355033) compared to full model (0.030874), and hence the confidence interval has changed between cluster and no cluster

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
model1e = lm(name_recall~treat_ad,data=d)
model1e$cluster.vcov <- cluster.vcov(model1e, d$cluster)
cf_1e <- coeftest(model1e, vcov=model1e$cluster.vcov)
cf_1e
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  0.454196   0.018576 24.4504 < 2.2e-16 ***
## treat_ad    -0.155073   0.026730 -5.8014 7.344e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
ci_1e <- ci_calc(model1e,0.95,2,model1e$cluster.vcov)
ci_1e
```

```
## [1] -0.2074875 -0.1026589
```

The treatment effect estimate is -0.1550732 with standard error 0.0267305. The associated p-value is 7.3439543302794e-09

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

```r
d$study <- ifelse(d$studyno ==2, 1, 0)
model1f = lm(name_recall~treat_ad+study,data=d)
model1f$cluster.vcov <- cluster.vcov(model1f, d$cluster)
cf_1f <- coeftest(model1f, vcov=model1f$cluster.vcov)
cf_1f
```

```
##
## t test of coefficients:
```

```
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1806848  0.0169702 10.6472   <2e-16 ***
## treat_ad    -0.0067752  0.0204154 -0.3319     0.74
## study        0.4260988  0.0206970 20.5875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ci_1f <- ci_calc(model1f,0.95,2,model1f$cluster.vcov)
ci_1f
```

```
## [1] -0.0468067  0.0332562
```

The treatment effect estimate is -0.0067752 with standard error 0.0204154. The associated p-value is 0.7400138

g. Why did the results from parts (e) and (f) differ? Which result is biased, and why? (Hint: see pages 75-76 of Gerber and Green, with more detailed discussion optionally available on pages 116-121.)

Results in (1e) is biased because it did not take the effect of study group in consideration. Study group1 and group 2 are two blocks in the example. Group 2 and Group 1 has different treatment assignment probabilities within the groups. Group 2 has higher SE and lower ATE compared to Group 1. Hence, when considered as a pool (1e), the results will be biased.(1f) accounts for this by including study group as a variable, there by absorbing the residual errors.

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."

Facebook chose to run ads specifically in most populated counties in Florida, Dade and Broward, which have a combined population of 4.2 million. So the large percentage difference can be attributed to a huge population difference in areas where the advertisement was played compared to where it wasn't played.

- "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."

Facebook happened to be a great tool to identify specific personal characteristics of the individuals. In the campaign, they identified people by their work, their political orientation, likes/dislikes etc and targeted personal advertisements to those categories. The agency picked the most effective messages for each demographic group so that the marketing budget is efficiently used. I believe these specific tools helped achieve the increase in favorable votes compared to the experiment conducted by Broockman and Green which was more generic in nature.

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.

ATE = 0.187 kgs per week if recycling bin is provided

95%CI = 0.12428 kgs to 0.24972 kgs

The equation can be written as Y=0.187B-0.024S+0.105C+0.281Bl. Here 0.281 is the baseline (intercept), B is indicator variable for Bin, S is for SMS, C is for having cellphone and Bl is for Baseline for first 2 weeks. ATE(Bin) = 0.187kgs

Standard error for Any Bin = 0.032. 95% CI = 0.187 +- 0.032*1.96 = 0.12428 to 0.24972

b. In Column 3 of Table 4A, what is the estimated ATE of sending a text message reminder on the average weight of recyclables turned in per household per week? Provide a 95% confidence interval.

ATE = -0.024 Kgs per week if text message reminder is sent

95%CI is -0.1 to 0.0524 kgs (-0.024 +- 0.039*1.96)

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

Percentage of visits turned in a bag, Avg no. of bins turned in per week, Avg weight of recyclables turned in per week, Avg market value of recyclables given per week

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

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.

0.562 kgs

Since Baseline is taken as part of regression, its coefficient is 0.281. So the difference is 0.281x2 = 0.562kgs

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.

The baseline value has a positive correlation to the outcome "percentage of visits turned in bag". In other words, if people were turning in recycalbles more often, they would likely do it same or more after treatment. However, the baseline value does not appear to have any correlation to the treatment (bin/sms). So, omitting the baseline does not have any bias associated with it (as it is unrelated to the treatment). I would not expect a big change in ATE. However, because it has some correlation to outcome, omitting it will result in an increase in standard error of the regression coefficient.

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

I don't consider it as a bad control. For treatment effect of providing Bins, it is not a bad control although it has no significant impact. For SMS messages, people are not going to buy a cell phone just because SMS text messages about recycling is coming. In other words, the treatment is not impacting this covariate, so it is not a bad control. However, both are redundant because SMS messages are sent to only those that have cell phones.

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.

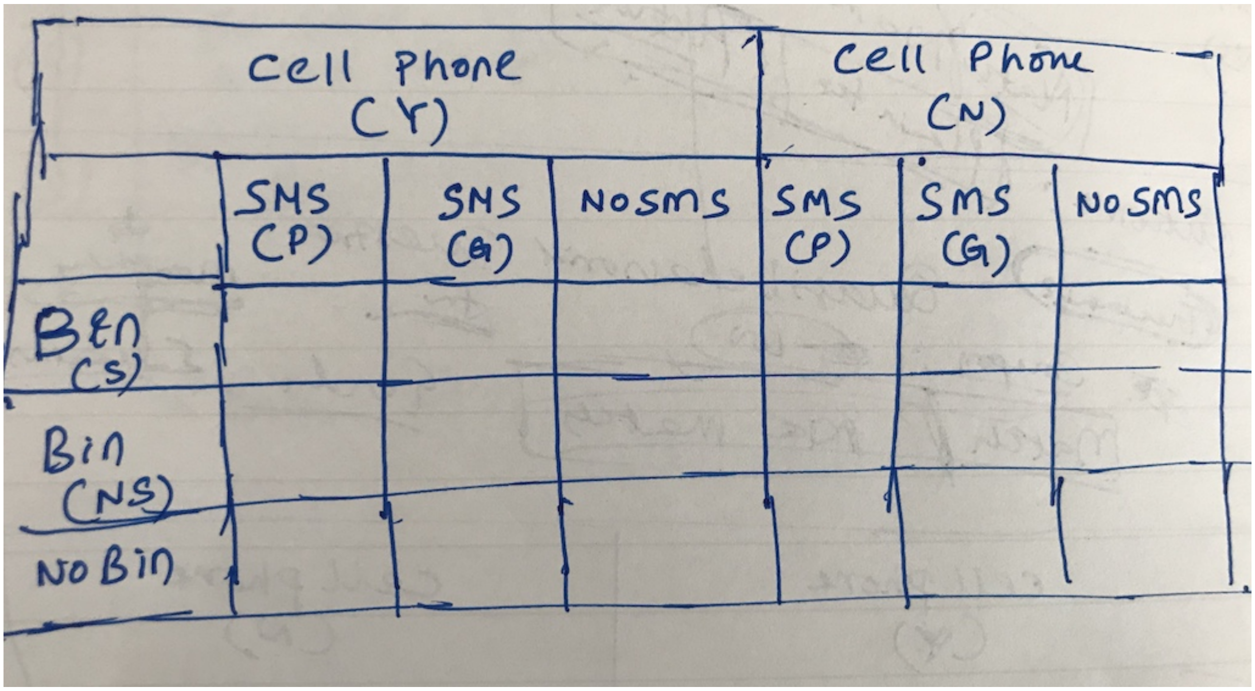
SMS message and cell phones are strongly correlated with each other. On the other hand, having cellphone does not seem to have direct correlation to the outcome. Hence, including the variable “has cell phone” results in multicollinearity, so it is better to exclude this variable. By removing the “has cell phone” variable, the coefficient would go up

3 Multifactor Experiments

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

a. What is the full experimental design for this experiment? Tell us the dimensions, such as 2x2x3. (Hint: the full results appear in Panel 4B.)

The experiment is a 3x3x2 design. We have three dimensions for Bin (No Bins, and Bins with and without stickers), three dimensions for SMS message (No SMS, SMS with General and Personal) and two dimensions for Cell Phones (With and without cell phones). The representation is as below:



	Cell Phone (Y)			Cell Phone (N)		
	SMS (P)	SMS (G)	No sms	SMS (P)	SMS (G)	No SMS
Bin (S)						
Bin (NS)						
No Bin						

b. In the results of Table 4B, describe the baseline category. That is, in English, how would you describe the attributes of the group of people for whom all dummy variables are equal to zero?

If all the dummy variables are equal to zero, the outcome will be the baseline value. The baseline value indicates the respective outcome variable in the two weeks prior to starting the treatment. The first two weeks were used to do the baseline measurement before the bins were distributed or SMS was sent.

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

For every new bin distributed without sticker, the “percentage of visits turned in bag” increases by 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?

Distributing recycling bins with sticker results in bigger treatment effect (5.5% increase) compared to the bins without stickers (3.5% increase). The magnitude of difference is about 2%

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

No, the difference is not statistically significant because the F-Test p-value is 0.31.

f. Notice that Table 4C is described as results from "fully saturated" models. What does this mean? Looking at the list of variables in the table, explain in what sense the model is "saturated."

The model is saturated because all the possible control variables and the interaction terms are included in the model.The fully saturated model results in perfect fit and is not generalized. The model can't absorb any more variations. There are certain variables that has no significance and are also redundant (Cell Phone for eg).

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
```

EDA

```
# First check for variables bin, sms and their subsets

n <- nrow(d)
# Check if bin count equals to bin_s and bin_g combined
nrow(d[d$bin==1,])
```

```
## [1] 603
```

```
nrow(d[(d$bin_g==1) | (d$bin_s==1),])
```

```
## [1] 603
```

```r
# Check if sms count equals to sms_p and sms_g combined
nrow(d[d$sms==1,])
```

```
## [1] 551
```

```r
nrow(d[(d$sms_p==1) | (d$sms_g==1),])
```

```
## [1] 551
```

```r
# Check for Cell variable
summary(d$havecell)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##  0.0000  0.0000  1.0000  0.5908  1.0000  1.0000       1
```

We can see NA in havecell variable that needs to be cleaned up

```r
# Remove the na value from the data frame
d <- na.omit(d)
```

```r
# Street
unique(d$street)
```

```
##   [1]    7    6    8    5    9   10 -999   11   17    3   45   46   47   63
##  [15]   62   64   78   80   70   77   66   81   73   88   86   91   89  124
##  [29]  138  109  125  132  121  131  149  136  106  166  196  198  188  191
##  [43]  216  233  225  222  221  241  244  243  236    2   22   21   20   23
##  [57]   37   40   41   38   61   60   75   82   67   69   74   85   79   83
##  [71]   84   94   96   93  137  111  115  134  105  113  112  118  110  133
##  [85]  107  128  130  117  126  160  153  154  157  158  156  152  155  164
##  [99]  163  172  171  170  180  183  182  192  189  185  197  200  193  207
## [113]  203  206  208  213  209  202  230  232  223  240  242  253  254  263
## [127]  261  260  262    4   15   44   43   42   68   72   98  119  148  151
## [141]  147  120  122  175  187  186  190  229  228  217  235  238  255  250
## [155]  248  249  247  256  259  258  257   26   53   32   58   99  103  100
## [169]  102  101  127  129  168  165  179  215  210  220  227  246
```

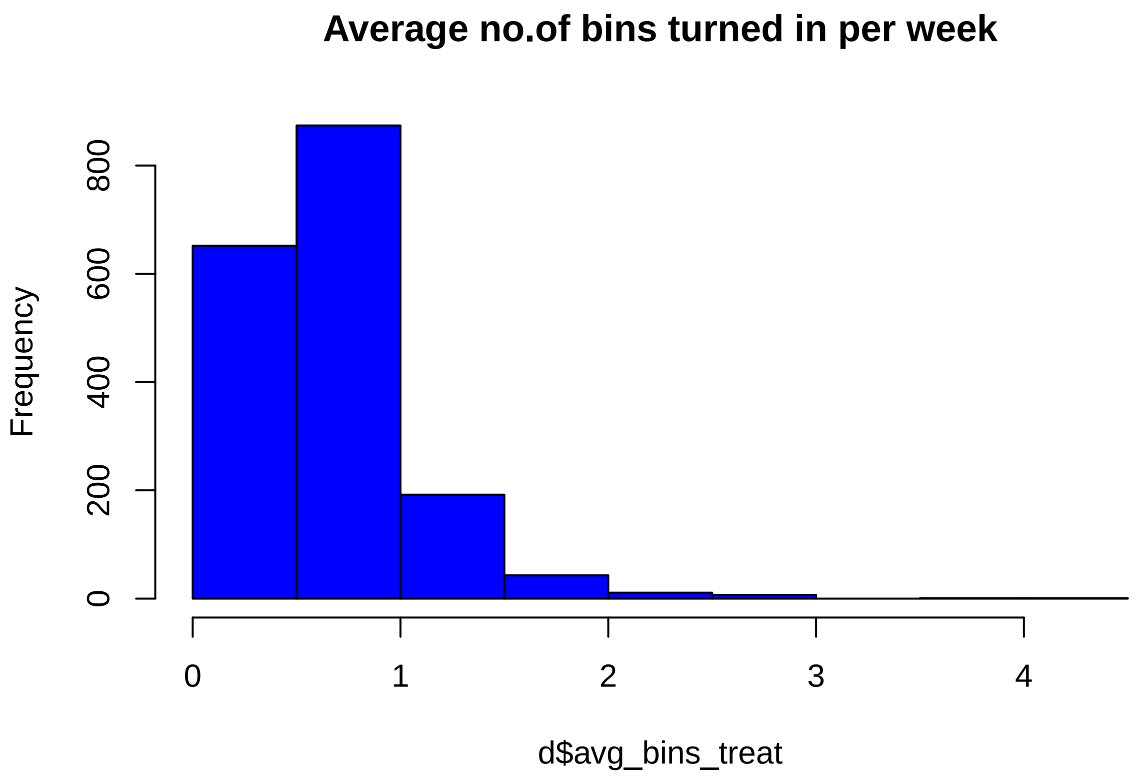
There is a negative street number that doesn't make sense.

```r
d_s <- subset(d,d$street==-999)
nrow(d_s)
```

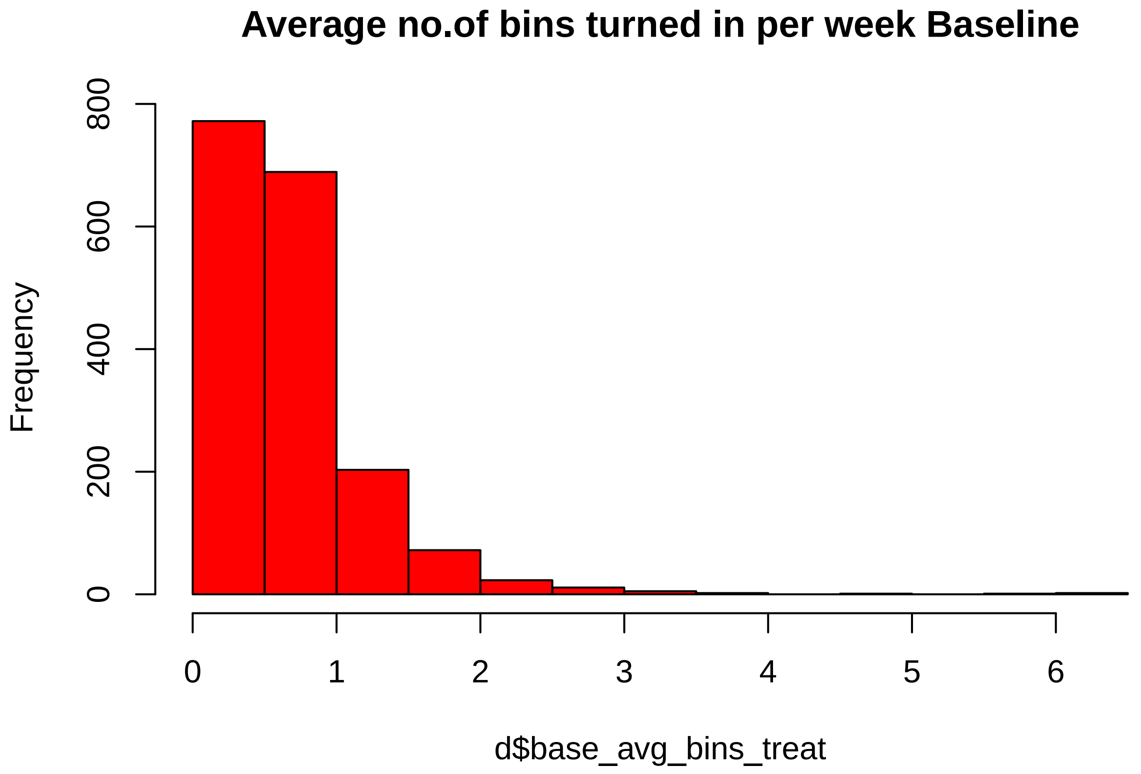
```
## [1] 120
```

There are 120 entries with this value, hence removing it is not a good idea. We will just keep it as-is

```r
hist(d$avg_bins_treat,col='blue',main="Average no.of bins turned in per week")
```



```r
hist(d$base_avg_bins_treat,col='red',main="Average no.of bins turned in per week Baseline")
```



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.

```
model4a <- lm(avg_bins_treat~bin,data=d)
summary(model4a)$coefficients
```

```
##              Estimate Std. Error  t value     Pr(>|t|)
## (Intercept) 0.6358877 0.01174877 54.12377 0.000000e+00
## bin         0.1330707 0.02024178  6.57406 6.415003e-11
```

```
ci_4a <- ci_calc(model4a,0.95,2)
ci_4a
```

```
## [1] 0.0933705 0.1727708
```

```
sum_4a <- summary(model4a)$coefficients[2,]
```

From the above model, adding a bin increases the average number of bins turned in per week by 0.1330707

95% Confidence interval of slope coefficient is 0.0933705 to 0.1727708

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.

```
model4b <- lm(avg_bins_treat~bin+base_avg_bins_treat,data=d)
summary(model4b)$coefficients
```

```
##                      Estimate Std. Error   t value      Pr(>|t|)
## (Intercept)         0.3520561 0.01376040 25.584727 3.445349e-123
## bin                 0.1242397 0.01668123  7.447874  1.471511e-13
## base_avg_bins_treat 0.3900001 0.01343737 29.023540 6.143352e-152
```

```
ci_4b <- ci_calc(model4b,0.95,2)
ci_4b
```

```
## [1] 0.09152279 0.15695654
```

```
sum_4b <- summary(model4b)$coefficients[2,]
```

The new 95%confidence interval is adjusted downwards from 0.0915228 to 0.1569565. When the baseline value was ommitted, the slope coefficient of the treatment was overstated. The reason is that there is a positive correlation between the baseline value and the outcome and there appears to be a positive correlation between baseline and treatment. People who were generally turning in bins would naturally do it same or more after treatment. This bias was reflected in 3a in the form of overstating the slope coefficient.

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.

```
model4c <- lm(avg_bins_treat~bin+base_avg_bins_treat+factor(street),data=d)
#coef(model_bin_pretreatment_s)
ci_4c <- ci_calc(model4c,0.95,2)
ci_4c
```

```
## [1] 0.08020966 0.14720847
```

```
#summary(model_bin_pretreatment_s)
```

The 95% confidence interval of treatment effect is now between 0.0802097 to 0.1472085

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.

Block randomization improves precision only when there is enough variation of outcome within each block. In this case, blocks are formed using a covariate (Street) that fails to predict experimental outcome. However, the estimates remain unbiased and offer similar results as complete randomization. In general, blocking in worst case is at least as good as complete randomization

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.

```
d$no_cell_phone <- 1-d$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.

```
model4f <- lm(avg_bins_treat~bin+base_avg_bins_treat+factor(street)+no_cell_phone,data=d)
ci_4f <- ci_calc(model4f,0.95,2)
ci_4f
```

```
## [1] 0.08166792 0.14853357
```

The 95% confidence interval of treatment effect is now between 0.0816679 to 0.1485336.This is because the effect of having a cell phone or not does not have a significant correlation to the final outcome. It does not have any correlation to the treatment either, hence the coefficient is not biased. It does not help in reducing standard errors either. Hence, the effect on Confidence Interval did not vary much

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.

```
model4g <- lm(avg_bins_treat~bin+sms+no_cell_phone+base_avg_bins_treat+factor(street),data=d)
coef(model4g)[2:5]
```

```
##                 bin                 sms       no_cell_phone
##         0.115053649         0.005124375        -0.046702054
## base_avg_bins_treat
##         0.373482860
```

```
ci_4g <- ci_calc(model4g,0.95,2)
ci_4g
```

```
## [1] 0.08160886 0.14849843
```

The 95% confidence interval of treatment effect is now between 0.0816089 to 0.1484984.SMS has very little/no correlation to potential outcome and had no significance in model 4A. Neither does it have any correlation to the treatment "bin". Hence, excluding SMS had no bias (in 4f) on the slope coefficient of bin, and it did not help explain the outcome any better. The standard error for treatment variable bin remains same irrespective of the presence of this covariate.

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.

```
model4h <- lm(avg_bins_treat~bin_s+bin_g+sms_p+sms_g+no_cell_phone+base_avg_bins_treat+factor(street),d
coef(model4h)[2:7]
```

```
##               bin_s               bin_g               sms_p
##         0.127812892         0.103190216        -0.008041152
##               sms_g       no_cell_phone base_avg_bins_treat
##         0.019707117        -0.046383459         0.373852178
```

```
ci_4h <- ci_calc(model4h,0.95,3)
ci_4h
```

```
## [1] 0.06025627 0.14612416
```

The 95% confidence interval of unadorned recycling bin coefficient is between 0.0602563 to 0.1461242. The coefficient is lower and the confidence interval width is higher. In part 4g the treatment variable was entire "bin" whereas in this question we separated it further to bin with sticker, and bin without sticker. In part 4g, variable bin absorbed the explanation for both the sub-variables in this model.

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?

```
model_5a <- lm(vomiting_day14~treat_zmapp,data=d)
summary(model_5a)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.84746 -0.03803  0.15254  0.21197  0.39024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.84746    0.05483  15.456   <2e-16 ***
## treat_zmapp -0.23770    0.08563  -2.776   0.0066 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4212 on 98 degrees of freedom
## Multiple R-squared:  0.0729, Adjusted R-squared:  0.06343
## F-statistic: 7.705 on 1 and 98 DF,  p-value: 0.006595
```

The probability of vomiting on day 14 reduces by 23% when ZMapp is administered. The standard error is 8.5%. The p-value for treatment is 0.0066 which is statistically significant

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.

```
model_5b <- lm(vomiting_day14~treat_zmapp+vomiting_day0+temperature_day0,data=d)
summary(model_5b)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,
##     data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.79643 -0.18106  0.04654  0.23122  0.68413
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -19.46966    7.44095  -2.617  0.01032 *
## treat_zmapp       -0.16554    0.07567  -2.188  0.03113 *
## vomiting_day0      0.06456    0.14635   0.441  0.66013
## temperature_day0   0.20555    0.07634   2.693  0.00837 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3668 on 96 degrees of freedom
## Multiple R-squared:  0.311,  Adjusted R-squared:  0.2895
## F-statistic: 14.45 on 3 and 96 DF,  p-value: 7.684e-08
```

In the new model, the probability of vomiting on day 14 reduces by ~16.5% when ZMapp is administered. The standard error is 7.5%. The p-value for treatment is 0.03 which is statistically significant

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

I prefer part(b) to part(a) as day 0 temperature seem to have correlation to the vomiting on

day14. Omitting the variable has resulted in overstating the treatment effect in part(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.

```
model_5d <- lm(vomiting_day14~treat_zmapp+vomiting_day0+temperature_day0+temperature_day14,data=d)
summary(model_5d)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 +
##     temperature_day14, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.87745 -0.27436  0.04701  0.24801  0.66445
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       -22.59159    7.47727  -3.021  0.00323 **
## treat_zmapp        -0.12010    0.07768  -1.546  0.12541
## vomiting_day0       0.04604    0.14426   0.319  0.75033
## temperature_day0    0.17664    0.07642   2.312  0.02296 *
## temperature_day14   0.06015    0.02937   2.048  0.04335 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3609 on 95 degrees of freedom
## Multiple R-squared:  0.3402, Adjusted R-squared:  0.3124
## F-statistic: 12.24 on 4 and 95 DF,  p-value: 4.545e-08
```

In the new model, the probability of vomiting on day 14 reduces by ~12% when ZMapp is administered. The standard error is 7.7%. The p-value for treatment is 0.12 which is not statistically significant

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

I prefer the estimate of the ATE in part (b) over part (d). temperature_day14 is a bad control because it is influenced by the ZMapp treatment itself.

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?

```
# Let us first use the model without any interaction term
model_5f_1 <- lm(temperature_day14~treat_zmapp+vomiting_day0+temperature_day0,data=d)
summary(model_5f_1)
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp + vomiting_day0 +
##     temperature_day0, data = d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7448 -0.9722 -0.3328  0.7384  2.6852
##
```

```
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        51.9039    25.4354   2.041  0.04403 *
## treat_zmapp        -0.7554     0.2587  -2.920  0.00436 **
## vomiting_day0       0.3079     0.5003   0.615  0.53973
## temperature_day0    0.4806     0.2610   1.842  0.06861 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.254 on 96 degrees of freedom
## Multiple R-squared:  0.2597, Adjusted R-squared:  0.2365
## F-statistic: 11.22 on 3 and 96 DF,  p-value: 2.227e-06
```

The treatment effect is a reduction of temperature by 0.75 degrees

Now let us use the interaction term to find if the effect is more on male or female

```
# Let us first use the model without any interaction term
model_5f_2 <- lm(temperature_day14~treat_zmapp+vomiting_day0+temperature_day0+treat_zmapp*male,data=d)
summary(model_5f_2)
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp + vomiting_day0 +
##     temperature_day0 + treat_zmapp * male, data = d)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70157 -0.37725 -0.02702  0.34687  0.73968
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       48.71269    9.26618   5.257 9.14e-07 ***
## treat_zmapp       -0.23087    0.11871  -1.945   0.0548 .
## vomiting_day0      0.04113    0.18208   0.226   0.8218
## temperature_day0   0.50480    0.09508   5.309 7.34e-07 ***
## male               3.08549    0.12644  24.403  < 2e-16 ***
## treat_zmapp:male  -2.07669    0.19164 -10.836  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4518 on 94 degrees of freedom
## Multiple R-squared:  0.9059, Adjusted R-squared:  0.9009
## F-statistic:   181 on 5 and 94 DF,  p-value: < 2.2e-16
```

For female, the ATE is a reduction of temperature by 0.23 degrees. For male, an additional reduction of 2.07 degrees happens. This implies that male patient ATE is a reduction of 2.3 degrees in temperature. Here, the interaction term has high statistical significance so we can safely say that males have higher effect on taking ZMapp than females

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.)

This is a fishing exercise and is a multiple-comparisons problem. Including more covariates and rerunning regression will eventually result in getting a highly significant result. Bonferroni correction would have identified the true statistical significance of this exercise. This result cannot be trusted otherwise.

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?

Yes, because we performed this test on a model that we designed and did not test it on all possible covariates. Since it was not repeated, this model can be trusted more than the previous model.

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.*)

An ideal experiment involving race is always FUQ'd (Fundamentally unidentified questions). It is impossible to conduct a field experiment that involves finding a causal effect of a race. We cannot manipulate the treatment in a controlled manner (Cannot control race). For example, to conduct this experiment, we would need to find two person groups who have lived in same place and same environment from the time of birth, except one having African descent and other group not. It is very difficult to conduct such a field experiment and that is why it is FUQ