Problem Set #3 - DATASCI W241

Greg Ceccarelli

October 21, 2015

Problem 1.

```r
library(dplyr)
library(stargazer)
library(memisc)

setwd("/Users/ceccarelli/MIDS/DATASCI_W241/Assignments/Problem Set 3/")
#clear variables
rm( list = ls() )
#read in tabular data
d <- read.csv("broockman_green_anon_pooled_fb_users_only.csv", sep=",", header = TRUE)

##Begin Problem 1 A##
#create new data frame only containing Study 1 data
d.s1 <- filter(d, grepl("Study 1",cluster)) %>%
  na.omit()

#simple plot
plot(x = d.s1$treat_ad, y=d.s1$name_recall)

#regress treatment against name recall
mod1_summary <- summary(lm(name_recall ~ treat_ad,d.s1))

#review line slope
abline(lm(name_recall ~ treat_ad,d))
```
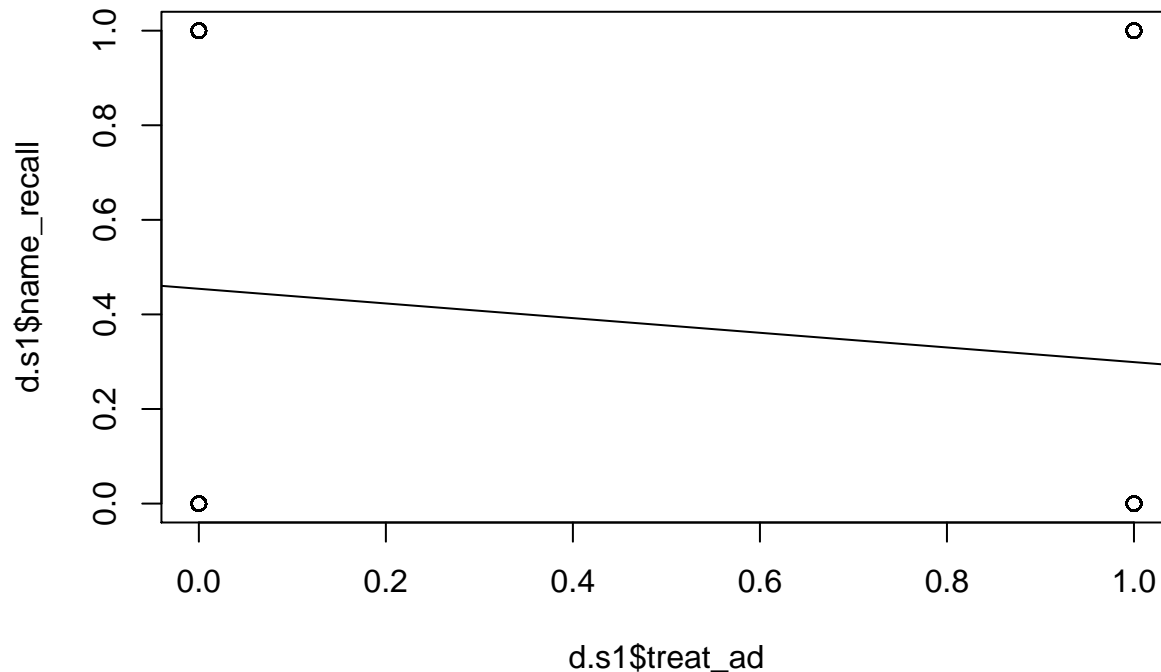
```r
#pull out data for confidence interval creation

ate <- mod1_summary$coefficients[2,1]
se <- mod1_summary$coefficients[2,2]

##End Problem 1 a.
ci <- c(ate-1.96 *se, ate+1.96*se)

##Begin Problem 1 c.

#initialize clustered standard errors function
cl <- function(fm, cluster){
  require(sandwich, quietly = TRUE)
    require(lmtest, quietly = TRUE)
    M <- length(unique(cluster))
    N <- length(cluster)
    K <- fm$rank
    dfc <- (M/(M-1))*((N-1)/(N-K))
    uj <- apply(estfun(fm),2, function(x) tapply(x, cluster, sum));
    vcovCL <- dfc*sandwich(fm, meat=crossprod(uj)/N)
    coeftest(fm, vcovCL)
}

#make sure to set cluster variable as a factor
d.s1$cluster <- factor(as.character(d.s1$cluster))

#run standard errors
c.se <- cl(lm(name_recall ~ treat_ad,d.s1), d.s1$cluster)

c.se.ate <- c.se[2,1]
c.se.se <- c.se[2,2]
```

```r
##End Problem 1 c.
c.se.ci <- c(c.se.ate -1.96 *c.se.se, c.se.ate+1.96*c.se.se)

#stargazer(cl(lm(name_recall ~ treat_ad,d.s1), d.s1$cluster), lm(name_recall ~ treat_ad,d.s1), type = "

#create new data frame only containing Study 2 data and removing NAs
d.s2 <- filter(d, grepl("Study 2",cluster)) %>%
  na.omit()

d.s2$cluster <- factor(as.character(d.s2$cluster))


#run standard errors
c2.se <- cl(lm(name_recall ~ treat_ad, d.s2), d.s2$cluster)

c2.se.ate <- c2.se[2,1]
c2.se.se <- c2.se[2,2]


##End Problem 1 d.
c2.se.ci <- c(c2.se.ate -1.96 *c2.se.se, c2.se.ate+1.96*c2.se.se)

##Begin Problem 1 e.

#create new data frame only containing Study 2 data and removing NAs
d.all <- d %>%
  na.omit()

d.all <- mutate(d.all, cluster =factor(as.character(d.all$cluster)))

#run standard errors
c3.se <- cl(lm(name_recall ~ treat_ad, d.all), d.all$cluster)

##End Problem 1 e.
c3.se.ate <- c3.se[2,1]
c3.se.pvalue <- c3.se[2,4]


##Begin Problem 1 f.

d.all <- mutate(d.all, dummy= factor(as.character(ifelse(studyno==1,0,ifelse(studyno==2,1,FALSE)))))
#filter(d.all, dummy ==5)

#d.all <- mutate(d.all, cluster =factor(as.character(d.all$cluster)))

c4.se <- cl(lm(name_recall ~ treat_ad+dummy, d.all), d.all$cluster)

c4.se
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.1806848  0.0169702 10.6472   <2e-16 ***
```

3

```
## treat_ad    -0.0067752  0.0204154 -0.3319     0.74
## dummy1       0.4260988  0.0206970 20.5875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

##End Problem 1 e.
c4.se.ate <- c4.se[2,1]
c4.se.pvalue <- c4.se[2,4]

##glimpse(d.all)
```

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").

```
ci
```

```
## [1] -0.05098178  0.03138601
```

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

In Study 1, 32,029 voters who were then assigned to 1,220 clusters across 18 age range, the 34 towns in the candidates' district, and 2 genders.

Clustering can increase standard errors when cluster-level noise is large or the number of clusters are small (and noise is large)

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.

```
c.se.ci
```

```
## [1] -0.05635499  0.03675922
```

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

```
c2.se.ci
```

```
## [1] -0.0723899  0.0667832
```

e. Repeat part (c), but using the entire sample from both studies. Do not take into account which study the data is from (more on this in a moment), but just pool the data and run one omnibus regression. What is the treatment effect estimate and associated p-value?

```
cat(" ATE = ", c3.se.ate, " pvalue = ", c3.se.pvalue)
```

```
##  ATE =  -0.1550732  pvalue =   7.343954e-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?

```
cat(" ATE = ", c4.se.ate, " pvalue = ", c4.se.pvalue)
```

```
##  ATE =  -0.006775249  pvalue =  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.)

They differ because we pooled the results of the study together. The pooled result (question e) is biased because the probability of being assigned to either study 1 or 2 is not exactly equal (i.e it's different).

   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.

In the facebook case study they use geographic location and user targeting. Additionally it is not clear that they actually employed any randomization. For example, they specifically state: "The agency relied on Facebook's Location 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." Accordingly, it's very likely their effect (19% difference in voting and 17% in no voting) claims are biased because its very likely that people in the ad-targeted counties were likely to vote a certain way regardless of the facebook ad saturation. This is very different than the Broockman and Green study which employed block and clustered randomization which allowed them to measure an unbiased treamtment effect.The facebook case study appears to be making a dubious causal claim based on their methodology or lackthereof.

Problem 2.

   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.

```
p2a.ate <- 0.187
p2a.se <- .032
```

```
cat(" ATE = ", p2a.ate, " conf interval = ", c(p2a.ate - 1.96 * p2a.se, p2a.ate + 1.96*p2a.se))
```

```
##  ATE =  0.187  conf interval =  0.12428 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.

```
p2b.ate <- -0.024
p2b.se <- .039
```

```
cat(" ATE = ", p2b.ate, " conf interval = ", c(p2b.ate - 1.96 * p2b.se, p2b.ate + 1.96*p2b.se))
```

```
##  ATE =  -0.024  conf interval =  -0.10044 0.05244
```

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

   1. Percentage of visits turned in bag

2. Avg. no of bins turned in per week
3. Avg. weight of recyclables turned in per week
4. Avg. market value of recyclables given per week

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

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 a point estimate

```
cat("Over six weeks there would be", .187*2*6, "more KGs of recycling turned in")
```

```
## Over six weeks there would be 2.244 more KGs of recycling turned in
```

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.

Using a baseline allows us to compute a difference in difference. Without the baseline, I would expect the ate to decrease. We get additional power by using percentage of visits turn in bag, baseline as a covariate to predict the results on providing a recycling bin. We get additional power with the baseline so removing it would actually increase standard error.

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

No it's not. Because we're trying to measure recycling participation intensity, I don't think cell phone ownership is an outcome of recycling participation. It could have been argued in the past cell phone ownership (when it was a luxury) would be a good measure of socio economic status which thus creates a higher likelihood of being environmentally conscious but in today's day and age, cell phone usage is ubiquitous across all levels of SES.

Problem 3.

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

The full experimental design is: You receive a bin, a bin with a stick or no bin and Independent of bin randomization, If you have a phone, you receive a generic SMS, a personalized SMS or no SMS Thus it is a 3 x 3 design.

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?

The baseline measure provides a pretest measurement taken prior to bin and SMS distribution. Where dummy values are equal to zero indicates that the household did not receive the treatment.

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

Households that received a bin without a sticker were 3.5 percentage points more likely to turn in recyclables compared to those who did not receive a bin.

    d. In column (1) of Table 4B, which seems to have a stronger treatment effect, the recycling bin with message sticker, or the recycling bin without sticker? How large is the magnitude of the estimated difference?

Households receiving a bin with a sticker were 4.5 percentage points more likely to turn in recyclables compared to households receiving a bin without a sticker being 3.5 percetage points more likely. Thus those receiving a stickered bin seem to have a stronger treatment effect. The magintude of the difference is 1 % with results in a 1.2% over the sample mean of 78.

    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. The confidence intervals overlap and the standard errors for each coefficient allow you to answer this question.

    d. 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."

In brief, a fully saturated model is one where the right hand side of the equaiton is as flexible as possible. What this means is that there is a dummy variable set up for each condition with the intent to try and explain as much variance as possible.

Problem 4.

```
library(foreign)

d4 <- read.dta("karlan_data_subset_for_class.dta")

## define a function to compute 95% confidence intervals more easily (could also just grab the conf valu
conf.ints <- function(model,effect){
  mod <- summary(model)
  ate<-mod$coefficients[effect,1]
  se<-mod$coefficients[effect,2]
    return(c(ate-1.96 *se, ate+1.96*se))
}

##Begin Problem 4 A##


#regress bin against avg bins
p4a.mod <- summary(lm(avg_bins_treat ~ bin,d4))
```

```
#pull out data for confidence interval creation

p4a.ate <- p4a.mod$coefficients[2,1]
p4a.se <- p4a.mod$coefficients[2,2]

p4a.ci <- c(p4a.ate -1.96 * p4a.se, p4a.ate+1.96* p4a.se)

##Begin Problem 4 B##

#regress bin against avg bins using pre-treatment y as covariate
p4b.ci <- summary(lm(avg_bins_treat ~ bin + base_avg_bins_treat,d4))

## use function to compute conf
conf.ints(lm(avg_bins_treat ~ bin + base_avg_bins_treat,d4),2)
```

```
## [1] 0.09202538 0.15736058
```

```
##Problem C - H code all below
```

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.

```
conf.ints(lm(avg_bins_treat ~ bin,d4),2)
```

```
## [1] 0.09561049 0.17514950
```

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.

The confidence interval get's smaller because we're able to compute a difference in difference using this baseline value. It increases the power and thus decreases the standard error and the confidence intervals.

```
conf.ints(lm(avg_bins_treat ~ bin + base_avg_bins_treat,d4),2)
```

```
## [1] 0.09202538 0.15736058
```

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.

```
conf.ints(lm(avg_bins_treat ~ bin + base_avg_bins_treat + factor(as.character(street)),d4),2)
```

```
## [1] 0.08045342 0.14732014
```

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.

It doesn't vary much because since street was used in the randomization procedure, it's already inherent in the results and so adding fixed effects as a covariate doesn't increase power much.

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

```
d4 <- mutate(d4, nocell= factor(as.character(ifelse(havecell==1,0,ifelse(havecell==0,1,FALSE)))))
```

    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.

The reason adding no cell phone as a covariate doesn't cause the confidence interval to differ much is that it (having or not having a cell phone) is completely independent of the bin experiment.

```
conf.ints(lm(avg_bins_treat ~ bin + base_avg_bins_treat + factor(as.character(street)) + nocell,d4),2)
```

```
## [1] 0.08169263 0.14850886
```

    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.

It doesn't differ much because SMS messages ignores interaction effects (it's being treated independently).

```
conf.ints(lm(avg_bins_treat ~ bin + base_avg_bins_treat + factor(as.character(street)) + nocell + sms ,d
```

```
## [1] 0.0816336 0.1484737
```

    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.

It differs because instead of pooling the variables (bin and sms) we've broken out each into a subtreatment. We're able to observe the lesser effect the unaddorned bin has on behavior.

```
##reproducing column 2 in Table 4B
summary(lm(avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g + nocell + base_avg_bins_treat +factor(as.cha
```

```
##
## Call:
## lm(formula = avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g +
##     nocell + base_avg_bins_treat + factor(as.character(street)),
##     data = d4)
##
## Residuals:
##      Min      1Q   Median      3Q     Max
## -1.48037 -0.19632 -0.01729  0.15867  1.84297
```

9

```
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.849e-01  3.429e-02  11.225  < 2e-16 ***
## bin_s                            1.278e-01  2.224e-02   5.748 1.08e-08 ***
## bin_g                            1.032e-01  2.189e-02   4.714 2.64e-06 ***
## sms_p                           -8.041e-03  2.504e-02  -0.321 0.748122
## sms_g                            1.971e-02  2.520e-02   0.782 0.434271
## nocell1                         -4.638e-02  2.038e-02  -2.276 0.022972 *
## base_avg_bins_treat              3.739e-01  1.431e-02  26.119  < 2e-16 ***
## factor(as.character(street))10  -3.145e-02  7.981e-02  -0.394 0.693548
## factor(as.character(street))100  6.930e-02  1.065e-01   0.651 0.515230
## factor(as.character(street))101  2.090e-01  1.118e-01   1.869 0.061870 .
## factor(as.character(street))102 -5.290e-02  1.900e-01  -0.278 0.780742
## factor(as.character(street))103  8.721e-02  1.118e-01   0.780 0.435339
## factor(as.character(street))105  1.143e-01  1.475e-01   0.775 0.438630
## factor(as.character(street))106  3.892e-01  9.803e-02   3.970 7.50e-05 ***
## factor(as.character(street))107  4.557e-02  1.645e-01   0.277 0.781805
## factor(as.character(street))109 -2.444e-02  1.182e-01  -0.207 0.836159
## factor(as.character(street))11  -8.224e-02  9.805e-02  -0.839 0.401741
## factor(as.character(street))110  1.293e-01  1.479e-01   0.875 0.381913
## factor(as.character(street))111 -2.386e-02  1.478e-01  -0.161 0.871800
## factor(as.character(street))112  1.850e-01  1.183e-01   1.564 0.117915
## factor(as.character(street))113  5.636e-02  1.264e-01   0.446 0.655674
## factor(as.character(street))115  2.774e-01  9.800e-02   2.830 0.004706 **
## factor(as.character(street))117  5.661e-01  1.903e-01   2.975 0.002976 **
## factor(as.character(street))118  1.255e-02  1.891e-01   0.066 0.947100
## factor(as.character(street))119 -2.090e-02  1.183e-01  -0.177 0.859799
## factor(as.character(street))120 -8.037e-02  1.183e-01  -0.679 0.497005
## factor(as.character(street))121  1.699e-01  1.183e-01   1.436 0.151242
## factor(as.character(street))122 -1.687e-01  1.643e-01  -1.027 0.304680
## factor(as.character(street))124  2.700e-02  1.022e-01   0.264 0.791566
## factor(as.character(street))125  2.955e-01  8.648e-02   3.417 0.000649 ***
## factor(as.character(street))126 -1.183e-01  1.359e-01  -0.871 0.383949
## factor(as.character(street))127  6.774e-02  1.355e-01   0.500 0.617251
## factor(as.character(street))128  1.465e-01  9.799e-02   1.495 0.135152
## factor(as.character(street))129  3.768e-01  1.890e-01   1.993 0.046410 *
## factor(as.character(street))130  1.746e-01  1.181e-01   1.478 0.139531
## factor(as.character(street))131  2.700e-01  1.185e-01   2.278 0.022855 *
## factor(as.character(street))132  2.339e-01  8.880e-02   2.634 0.008520 **
## factor(as.character(street))133  3.148e-01  1.645e-01   1.913 0.055892 .
## factor(as.character(street))134 -8.563e-03  1.257e-01  -0.068 0.945691
## factor(as.character(street))136  1.601e-01  1.183e-01   1.354 0.176077
## factor(as.character(street))137  2.420e-01  1.890e-01   1.281 0.200462
## factor(as.character(street))138  1.092e-01  1.027e-01   1.064 0.287535
## factor(as.character(street))147  1.898e-01  1.891e-01   1.004 0.315694
## factor(as.character(street))148  1.030e-01  1.644e-01   0.626 0.531217
## factor(as.character(street))149  2.678e-01  1.125e-01   2.381 0.017404 *
## factor(as.character(street))15  -3.497e-02  1.646e-01  -0.212 0.831756
## factor(as.character(street))151 -2.765e-01  1.891e-01  -1.462 0.143888
## factor(as.character(street))152  1.418e-01  1.889e-01   0.751 0.453029
## factor(as.character(street))153 -1.611e-02  1.891e-01  -0.085 0.932110
## factor(as.character(street))154 -5.725e-02  1.064e-01  -0.538 0.590592
## factor(as.character(street))155 -2.708e-01  1.890e-01  -1.432 0.152272
```

```
## factor(as.character(street))156  2.156e-01  1.890e-01   1.141 0.254061
## factor(as.character(street))157  1.394e-01  1.650e-01   0.845 0.398293
## factor(as.character(street))158  3.892e-01  1.476e-01   2.636 0.008463 **
## factor(as.character(street))160 -1.186e-01  1.018e-01  -1.165 0.244259
## factor(as.character(street))163 -1.077e-01  1.263e-01  -0.853 0.393946
## factor(as.character(street))164 -9.032e-02  1.476e-01  -0.612 0.540549
## factor(as.character(street))165  1.326e-02  1.646e-01   0.081 0.935790
## factor(as.character(street))166 -1.371e-01  1.890e-01  -0.725 0.468334
## factor(as.character(street))168  1.671e-01  1.894e-01   0.882 0.377768
## factor(as.character(street))17   3.118e-02  1.353e-01   0.231 0.817727
## factor(as.character(street))170 -5.872e-03  8.852e-02  -0.066 0.947116
## factor(as.character(street))171 -3.679e-02  1.353e-01  -0.272 0.785815
## factor(as.character(street))172 -6.304e-03  1.182e-01  -0.053 0.957473
## factor(as.character(street))175  2.620e-01  9.446e-02   2.773 0.005617 **
## factor(as.character(street))179 -2.231e-02  1.645e-01  -0.136 0.892168
## factor(as.character(street))180 -9.261e-02  7.645e-02  -1.211 0.225970
## factor(as.character(street))182 -1.462e-01  1.352e-01  -1.081 0.279773
## factor(as.character(street))183 -7.243e-02  7.822e-02  -0.926 0.354594
## factor(as.character(street))185 -6.272e-02  1.353e-01  -0.464 0.642955
## factor(as.character(street))186 -4.780e-03  1.181e-01  -0.040 0.967715
## factor(as.character(street))187  6.120e-06  9.143e-02   0.000 0.999947
## factor(as.character(street))188 -3.000e-02  1.117e-01  -0.269 0.788224
## factor(as.character(street))189 -5.397e-02  7.652e-02  -0.705 0.480729
## factor(as.character(street))190  2.022e-02  1.183e-01   0.171 0.864342
## factor(as.character(street))191 -2.118e-02  1.117e-01  -0.190 0.849582
## factor(as.character(street))192 -1.017e-01  6.992e-02  -1.454 0.146123
## factor(as.character(street))193  7.828e-02  1.182e-01   0.663 0.507728
## factor(as.character(street))196 -2.395e-01  1.352e-01  -1.771 0.076736 .
## factor(as.character(street))197  4.531e-02  1.891e-01   0.240 0.810622
## factor(as.character(street))198  8.985e-02  8.851e-02   1.015 0.310163
## factor(as.character(street))2   -1.007e-01  1.018e-01  -0.989 0.322935
## factor(as.character(street))20  -2.310e-01  1.181e-01  -1.956 0.050613 .
## factor(as.character(street))200 -3.484e-02  1.258e-01  -0.277 0.781813
## factor(as.character(street))202  2.468e-02  1.477e-01   0.167 0.867348
## factor(as.character(street))203 -2.423e-01  1.258e-01  -1.926 0.054288 .
## factor(as.character(street))206 -2.395e-01  1.064e-01  -2.252 0.024479 *
## factor(as.character(street))207 -2.323e-01  1.643e-01  -1.414 0.157605
## factor(as.character(street))208 -1.061e-01  1.065e-01  -0.995 0.319673
## factor(as.character(street))209 -3.159e-01  9.436e-02  -3.347 0.000835 ***
## factor(as.character(street))21  -6.484e-02  1.357e-01  -0.478 0.632854
## factor(as.character(street))210 -2.059e-01  1.067e-01  -1.929 0.053850 .
## factor(as.character(street))213  5.142e-02  9.812e-02   0.524 0.600281
## factor(as.character(street))215 -5.381e-02  6.512e-02  -0.826 0.408791
## factor(as.character(street))216 -4.495e-01  1.488e-01  -3.020 0.002567 **
## factor(as.character(street))217 -1.898e-01  1.181e-01  -1.608 0.108126
## factor(as.character(street))22  -1.172e-01  6.887e-02  -1.702 0.088916 .
## factor(as.character(street))220 -1.982e-01  1.256e-01  -1.577 0.114913
## factor(as.character(street))221 -1.643e-01  1.118e-01  -1.470 0.141695
## factor(as.character(street))222 -1.524e-01  1.477e-01  -1.032 0.302381
## factor(as.character(street))223 -1.773e-01  1.180e-01  -1.502 0.133330
## factor(as.character(street))225 -9.061e-02  1.119e-01  -0.810 0.418247
## factor(as.character(street))227 -6.255e-02  1.477e-01  -0.424 0.671882
## factor(as.character(street))228  4.757e-02  1.890e-01   0.252 0.801306
## factor(as.character(street))229  5.036e-02  1.359e-01   0.371 0.710924
```

```
## factor(as.character(street))23    -1.590e-01  7.233e-02  -2.199 0.028039 *
## factor(as.character(street))230    1.202e-02  1.256e-01   0.096 0.923770
## factor(as.character(street))232    8.263e-02  1.645e-01   0.502 0.615502
## factor(as.character(street))233    4.881e-02  9.788e-02   0.499 0.618107
## factor(as.character(street))235   -1.269e-01  1.479e-01  -0.858 0.391080
## factor(as.character(street))236   -5.401e-04  8.181e-02  -0.007 0.994734
## factor(as.character(street))238   -9.894e-02  1.118e-01  -0.885 0.376352
## factor(as.character(street))240   -3.037e-01  1.476e-01  -2.058 0.039765 *
## factor(as.character(street))241   -1.192e-01  1.645e-01  -0.725 0.468702
## factor(as.character(street))242    5.022e-02  1.890e-01   0.266 0.790466
## factor(as.character(street))243    1.387e-01  1.648e-01   0.842 0.399988
## factor(as.character(street))244    1.088e-01  9.149e-02   1.190 0.234329
## factor(as.character(street))246   -1.321e-01  1.355e-01  -0.975 0.329921
## factor(as.character(street))247   -2.047e-03  9.434e-02  -0.022 0.982688
## factor(as.character(street))248    3.572e-04  7.105e-02   0.005 0.995990
## factor(as.character(street))249    9.026e-02  8.612e-02   1.048 0.294730
## factor(as.character(street))250    4.957e-02  6.222e-02   0.797 0.425744
## factor(as.character(street))253    1.218e-01  8.384e-02   1.453 0.146517
## factor(as.character(street))254    1.136e-01  1.260e-01   0.901 0.367473
## factor(as.character(street))255    3.703e-02  8.867e-02   0.418 0.676314
## factor(as.character(street))256    3.462e-02  6.447e-02   0.537 0.591381
## factor(as.character(street))257   -1.050e-01  8.174e-02  -1.284 0.199327
## factor(as.character(street))258    9.395e-02  7.661e-02   1.226 0.220228
## factor(as.character(street))259    7.295e-01  1.892e-01   3.856 0.000120 ***
## factor(as.character(street))26    -7.901e-03  1.890e-01  -0.042 0.966659
## factor(as.character(street))260   -3.442e-02  6.351e-02  -0.542 0.587859
## factor(as.character(street))261   -3.585e-02  1.182e-01  -0.303 0.761720
## factor(as.character(street))262    4.910e-02  1.118e-01   0.439 0.660660
## factor(as.character(street))263    2.767e-02  8.375e-02   0.330 0.741165
## factor(as.character(street))3     -5.491e-02  1.064e-01  -0.516 0.605954
## factor(as.character(street))32     1.703e-01  1.893e-01   0.900 0.368373
## factor(as.character(street))37     1.467e-01  1.481e-01   0.991 0.322072
## factor(as.character(street))38    -8.602e-02  1.476e-01  -0.583 0.560255
## factor(as.character(street))4     -1.304e-01  1.118e-01  -1.166 0.243790
## factor(as.character(street))40    -5.438e-02  1.892e-01  -0.287 0.773799
## factor(as.character(street))41     1.374e-02  1.477e-01   0.093 0.925891
## factor(as.character(street))42    -1.168e-01  1.894e-01  -0.617 0.537635
## factor(as.character(street))43     3.722e-02  1.477e-01   0.252 0.801032
## factor(as.character(street))44     1.116e-01  1.184e-01   0.943 0.345912
## factor(as.character(street))45    -4.479e-02  1.643e-01  -0.273 0.785163
## factor(as.character(street))46     1.775e-03  1.352e-01   0.013 0.989530
## factor(as.character(street))47     4.017e-01  1.647e-01   2.439 0.014832 *
## factor(as.character(street))5     -1.041e-01  1.117e-01  -0.932 0.351245
## factor(as.character(street))53     1.040e-02  1.890e-01   0.055 0.956147
## factor(as.character(street))58    -3.232e-02  9.439e-02  -0.342 0.732054
## factor(as.character(street))6     -5.712e-02  1.181e-01  -0.484 0.628696
## factor(as.character(street))60     2.622e-02  8.603e-02   0.305 0.760553
## factor(as.character(street))61    -5.212e-03  1.893e-01  -0.028 0.978042
## factor(as.character(street))62     4.030e-02  1.260e-01   0.320 0.749053
## factor(as.character(street))63    -3.165e-02  1.181e-01  -0.268 0.788679
## factor(as.character(street))64     3.891e-02  1.352e-01   0.288 0.773576
## factor(as.character(street))66    -5.115e-02  1.257e-01  -0.407 0.683999
## factor(as.character(street))67    -8.914e-03  1.117e-01  -0.080 0.936428
## factor(as.character(street))68     8.255e-02  1.180e-01   0.699 0.484389
```

```
## factor(as.character(street))69    6.694e-02  1.020e-01    0.656 0.511628
## factor(as.character(street))7    -7.541e-02  1.478e-01   -0.510 0.609882
## factor(as.character(street))70   -2.219e-02  9.141e-02   -0.243 0.808192
## factor(as.character(street))72   -4.595e-02  1.018e-01   -0.451 0.651851
## factor(as.character(street))73   -1.806e-01  1.181e-01   -1.530 0.126299
## factor(as.character(street))74   -7.358e-02  7.815e-02   -0.941 0.346604
## factor(as.character(street))75   -5.079e-02  9.804e-02   -0.518 0.604460
## factor(as.character(street))77   -3.874e-02  1.258e-01   -0.308 0.758101
## factor(as.character(street))78   -6.367e-02  8.616e-02   -0.739 0.460047
## factor(as.character(street))79    3.172e-01  1.646e-01    1.928 0.054050 .
## factor(as.character(street))8    -5.727e-02  9.127e-02   -0.627 0.530441
## factor(as.character(street))80    1.075e-02  6.886e-02    0.156 0.875937
## factor(as.character(street))81   -2.883e-02  1.644e-01   -0.175 0.860826
## factor(as.character(street))82    3.809e-02  1.893e-01    0.201 0.840566
## factor(as.character(street))83   -1.181e-01  1.262e-01   -0.936 0.349237
## factor(as.character(street))84   -1.779e-03  1.258e-01   -0.014 0.988720
## factor(as.character(street))85    6.364e-02  8.615e-02    0.739 0.460229
## factor(as.character(street))86   -7.689e-02  7.988e-02   -0.963 0.335879
## factor(as.character(street))88   -9.930e-02  1.064e-01   -0.933 0.350967
## factor(as.character(street))89   -9.898e-03  1.182e-01   -0.084 0.933271
## factor(as.character(street))9    -9.205e-02  9.785e-02   -0.941 0.346967
## factor(as.character(street))91   -2.393e-01  1.892e-01   -1.265 0.206188
## factor(as.character(street))93   -7.119e-02  1.257e-01   -0.566 0.571203
## factor(as.character(street))94   -5.101e-02  1.120e-01   -0.456 0.648754
## factor(as.character(street))96   -3.647e-02  1.257e-01   -0.290 0.771738
## factor(as.character(street))98    3.987e-02  9.827e-02    0.406 0.685016
## factor(as.character(street))99    2.324e-01  1.119e-01    2.077 0.037974 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.323 on 1595 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.4398, Adjusted R-squared:  0.3748
## F-statistic: 6.769 on 185 and 1595 DF,  p-value: < 2.2e-16
```

```
conf.ints(lm(avg_bins_treat ~ bin_s + bin_g + sms_p + sms_g + nocell + base_avg_bins_treat +factor(as.cl
```

```
## [1] 0.06028806 0.14609237
```

Problem 5.

```
##read in data
d5 <- read.csv("ebola_rct2.csv",header = T)
```

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?

```
d5a <- summary(lm(vomiting_day14 ~ treat_zmapp,d5))
cat(" ATE = ", d5a$coefficients[2,1], " se = ", d5a$coefficients[2,2], " p-value = ", d5a$coefficients[:
```

```
##  ATE =  -0.2377015  se =  0.08563161  p-value =  0.006595412
```

```r
conf.ints(lm(vomiting_day14 ~ treat_zmapp,d5),2)
```

```
## [1] -0.40553948 -0.06986358
```

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.

```r
d5b <- summary(lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,d5))
cat(" ATE = ", d5b$coefficients[2,1], " se = ", d5b$coefficients[2,2], " p-value = ", d5b$coefficients[
```

```
##  ATE =  -0.1655367  se =  0.07567142  p-value =  0.03112852
```

```r
conf.ints(lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,d5),2)
```

```
## [1] -0.31385272 -0.01722077
```

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

I like the estimate of ATE in part B better because we're able to add a baseline that reduces our standard error and improves our confidence interval

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.

```r
d5d <- summary(lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + temperature_day14,d5
cat(" ATE = ", d5d$coefficients[2,1], " se = ", d5d$coefficients[2,2], " p-value = ", d5b$coefficients[
```

```
##  ATE =  -0.1201006  se =  0.07767979  p-value =  0.03112852
```

```r
conf.ints(lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + temperature_day14, d5),2
```

```
## [1] -0.27235302  0.03215176
```

```r
stargazer(lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,d5), lm(vomiting_day14 ~ t
```

```
##
## ===================================================================
##                              Dependent variable:
##              ---------------------------------------------
##                             vomiting_day14
##                      (1)                      (2)
## -------------------------------------------------------------------
## treat_zmapp              -0.166**                 -0.120
##                      (0.076)                  (0.078)
##
```

14

```
## vomiting_day0                  0.065                   0.046
##                               (0.146)                 (0.144)
##
## temperature_day0              0.206***                0.177**
##                               (0.076)                 (0.076)
##
## temperature_day14                                     0.060**
##                                                       (0.029)
##
## Constant                     -19.470**               -22.592***
##                               (7.441)                 (7.477)
##
## ---------------------------------------------------------------
## Observations                    100                     100
## R2                             0.311                   0.340
## Adjusted R2                    0.290                   0.312
## Residual Std. Error    0.367 (df = 96)         0.361 (df = 95)
## F Statistic          14.447*** (df = 3; 96) 12.244*** (df = 4; 95)
## ===============================================================
## Note:                                   *p<0.1; **p<0.05; ***p<0.01
```

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

I prefer the estimate of ATE in part B rather than part D. Adding temperature_day14 is a bad control because it is actually an outcome of the treatment.

    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?

The results suggest that male's actually experience 2.17 degrees higher temperature than women when being treated with zmapp.

```r
summary(lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + male,d5))
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp + vomiting_day0 +
##     temperature_day0 + male, data = d5)
##
## Residuals:
##      Min      1Q   Median      3Q      Max
## -1.38138 -0.50233  0.08585  0.50550  1.26462
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       35.9694    13.7118   2.623   0.0101 *
## treat_zmapp       -1.0179     0.1401  -7.266 1.03e-10 ***
## vomiting_day0     -0.1393     0.2705  -0.515   0.6078
## temperature_day0   0.6383     0.1406   4.538 1.66e-05 ***
## male               2.1796     0.1415  15.402  < 2e-16 ***
## ---
```

15

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.674 on 95 degrees of freedom
## Multiple R-squared:  0.7883, Adjusted R-squared:  0.7794
## F-statistic: 88.44 on 4 and 95 DF,  p-value: < 2.2e-16
```

```
d5f <- summary(lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + male,d5))
cat(" ATE = ", d5f$coefficients[2,1], " se = ", d5f$coefficients[2,2], " p-value = ", d5f$coefficients[
```

```
##  ATE =  -1.017867  se =  0.1400866  p-value =  1.025292e-10
```

```
conf.ints(lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + male,d5),2)
```

```
## [1] -1.2924368 -0.7432973
```

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 ... etc

I think it is very unlikely based on the above. Specifically, the analysis of subgroups (men / women) conflates the effectiveness of the drug with the genders of people who receive them. A direct experimental test would be required rather than treatment by covariate interaction analysis to distinguish whether this is really the case.

f. 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?

If I had this theory and had conducted a direct experiment to test gender specifically (and not after the fact) then and only then could I be inclined to believe that the heterogenous treatment effect really existed (i.e. that men are more responsive to the drug than women, for example).

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.)

That's a fundamentally unanswerable question because you can't randomly manipulate people's races in isolation and you would need to do so in order to construct an examperiment to test this question.