# 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
```

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

```
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
```

```
# For the scatterplot matrix
library(car)
```

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

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

## 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
```

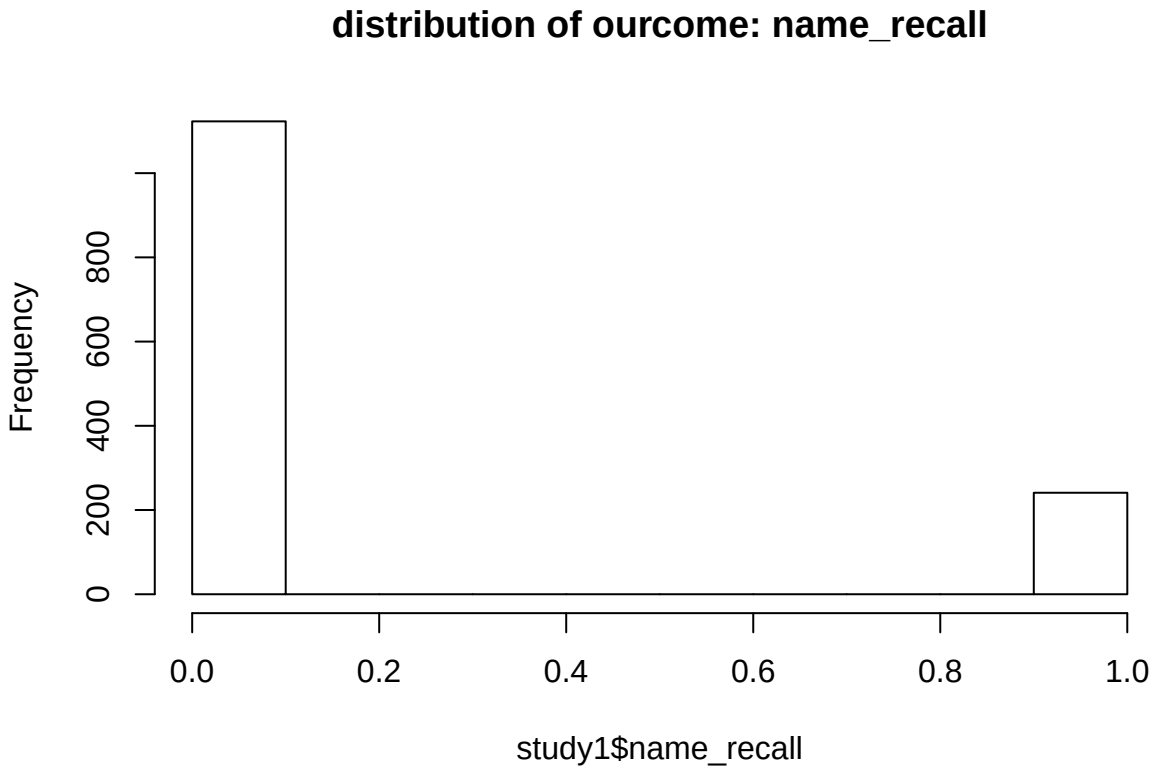
```
summary(study1)
```

```
##     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 1256 observations in study 1. Looking at the distribution of name_recall below*

```
# 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*

```
# 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 similar individuals (same age, gender and location) The members of above clusters have very little variance among themselves (age, gender, location) but there's significant variance between the clusters. When the subjects are assigned as clusters, it suppresses the variance in the outcome(y) leading to a smaller estimate for errors. When we take clustering into account, we correct for this, leading to bigger 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 [-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.**

```r
# 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
```

```r
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
```

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

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

```
lr2$cluster.se
```

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

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

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

```
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?**

```
# 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
```

```
# 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"
```

```
# t test of the coefficients
coeftest(lr3, lr3$cluster.vcov)
```

```
##
## 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
```

**Answer: The cluster confidence interval from study 1 and 2 combined is [-0.2085342 -0.1016123] The treatment effect is -0.1550732 The p value is close to 0 (7.344e-09 with robust SE's and 2e-16 with OLS SE's)**

**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"
```

```
# t test of the coefficients
coeftest(lr4, lr4$cluster.vcov)
```

```
##
## 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
## study10_study21  0.4260988  0.0206970 20.5875   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

**Answer: The cluster confidence interval from all studies combined is [-0.04760609 0.03405559] The treatment effect is -0.006775249 The p value is .74 with robust SE's and .709 with OLS SE's**

**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.)**

*CI of treatment effect from study 1 [-0.05101765 0.03142188] CI of treatment effect from study 2 [-0.07381003 0.06820333]*

**Answer: The confidence intervals, of the treatment effect, differ among the studies. The CI for study 2 is almost double the width of study 1. This indicates a meaningful difference in either the composition/baseline or context/environment/treatment of the subjects in each of the studies, which introduces a bias when we combine the studies. (e) is biased as it combines the studies.(f) captures the variability among the studies in the coefficient of the dummy variable. This creates a more accurate estimate of the treatment impact**

**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. In the FB case study, 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) were included as control while determining the effects of the treatment (ADS). If not included, the ATE claimed would be biased and will include noise due to these independent variables**
2. In places where the ads ran, exposure was determined by certain characteristics (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.**

```
# Coefficient/ATE of providing a recycling bin on the average weight of recyclables
lr_2a.coeff = .187
lr_2a.coeff
```

```
## [1] 0.187
```

```
se_2a = .032
se_2a
```

```
## [1] 0.032
```

```
# Confidence interval
c(lr_2a.coeff - 2 * se_2a, lr_2a.coeff + 2* se_2a)
```

```
## [1] 0.123 0.251
```

**Answer: The ATE is .187 The 95% confidence interval is [.123, .251]**

**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.**

```
# Coefficient/ATE
lr_2b.coeff = -.024
lr_2b.coeff
```

```
## [1] -0.024
```

```
se_2b = .039
se_2b
```

```
## [1] 0.039
```

```
# Confidence interval is +1 2 SE's
c(lr_2b.coeff - 2 * se_2b, lr_2b.coeff + 2* se_2b)
```

```
## [1] -0.102  0.054
```

**Answer: The ATE is -.024 The 95% confidence interval is [-.102, .054]**

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

**Answer: The following outcome measures show statististically significant effects:** 1. Percentage of visits turned in bag 2. Avg. no. of bins turned in per week 3. Avg. weight(in kg) turned in per week 4. Avg. market value per week

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

**Answer: No statistically significant effects (at 5% level)**

**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.**

```
# The difference in this outcome comes from the baseline weight difference
lr_2d.coeff = .281
lr_2d.baseline = 2
lr_2d.diff = lr_2d.coeff * lr_2d.baseline
lr_2d.diff
```

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

**Answer: The difference in outcome among A and B (weight turned in) comes from the different baselines (2Kg). A would turn in 0.562kg more recyclables/week in the treatment period**

**f. Suppose that the variable "percentage of visits turned in bag, baseline" had been left out of the regression reported in Column 1. What would you expect to happen to the results on providing a recycling bin? Would you expect an increase or decrease in the estimated ATE? Would you expect an increase or decrease in the standard error? Explain your reasoning.**

**Answer: if the variable "percentage of visits turned in bag, baseline" was left out of the regression, we'll observe some if its effect to creep into the results of "providing a recycling bin". More precisely we may observe the following:** 1. An overestimate of the ATE of providing a recycling bin i.e. the estimated ATE of providing the bin would go up. "Visits in a bag" has a +ve impact on recycling and in the absense of the variable, some of that effect would show up in "providing the bin" treatment 2. We'll observe an increase in the Standard Error. Some of the variance related to the basline of bags turned in would show up in results of "providing a recyling bin"

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

**Answer: Per the expriment's design, "hascell" was used to identify the group that can be further subdivided for the SMS treatment. It was not meant to be a control or an independent variable of consequence. Since it was used for subdivision, additing it as control seems counter-intuitive and maybe bad control From regression, the effect of "hascell" looks inconsequential (coefficient is .022(.014)) also**

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

**Answer: The coefficient of "Any SMS message" would go up if we remove "has cell phone" form the regression. "hascell" was used to identify folks (with cellphone) that can be randomized for the 3 SMS treatments. Using it again as a control adds noise and probably takes some of the effect that'll otherwise show up for the SMS treatments**

---

# 3 Multifactor Experiments

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

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

**Answer: The experiment's design has the following factors:**

1. Bin (3 levels: with sticker, without stickers, no bin)
2. Cell phone (2 levels: has, has not)

- SMS (3 levels: No message, generic message, specific message)

**Answer: If we consider "has_cellphone" as a pre-treatment condition, the design is (3 * 3).** *We can consider "has_cellphone" as pre-treatment condition as it is used to identify the group that be further subdivided for the 3 SMS treatments* **Table 4A has more combinations than 3*3 though. It has (3*2*3 - 3) combinations. That is combinations of bin(3), SMS(3) and cellphone(2) - combinations of SMS with "no cell phone", which do not make sense anyway**

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

**Answer: "baseline" is a set of variables that capture the pre-experiment values (2 weeks before the expriment) of all the outcome variables. As the experimental treatments are not applied yet, all the dummy/indicator variables are 0 for this time period. The baseline variables are significant (theoratically and also via experiment's results) as can explain some of the effects seen in the outcome variables during the experiment**

- Below are the 5 Baseline (as independent variables) variables measured pre-treatment (2 weeks before experiment). These are included in the regression equation for individual outcomes (like baseline for "visits with bags" is included when measuring this outcome):
  - baseline for visits with bag
  - baseline of bins turned in
  - baseline of weight of recyclables
  - baseline for the market value of recyclables
  - baseline of % contamination per week

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

**Answer: The coefficient is .035(.015). The value is more than 2 * SE. Its marked as statistically significant at the 5% confidence level. The 95% Confidence interval based on these values would be [.005, .055] The interpretation would be that a bin (even a generic one without a sticker) creates a positive impact on recycling (education and lowering of the barrier to recycling). some of this impact shows up in visits that are not using the bin also. Its impact on recycling without the bin is not huge (The ATE is one SE bigger than the baseline), but is measurable (The 95% confidence interval is +ve [.005, .055])**

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

**Answer: The "recycling bin with sticker" has a stronger treatment effect. The magnitude of the difference is (.055 - .035 = .020) .020 Both the coefficients have the same SE, so that does not influence the difference in magnitudes calculation A thing to note is that recycling bin "with sticker"" is significant at a 1% confidence level. "without sticker" is significant at 5% confidence level**

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

**Answer: The difference above (.020) is not statistically significant. The SE (.015) for both measurement helps us with this conclusion If the estimates err in opposite directions (-.015 in one case and +.015 in the other) there could be cases where the difference in the outcomes (.020 for the point estimates) would reduce to 0 or change signs**

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

**Answer: The model is "saturated" as it includes all the independent variables and there interaction terms i.e. we measure the mean outcomes of all experimental conditions**

**The complete list of experimental conditions are:**

1. SMS message
2. No Bin
3. Bin without sticker
4. Bin with Sticker
5. Phone

*The equation for the model, gives us a coefficient for each of these conditions and there interactions (like "SMS message + bin" and "SMS message + No Bin")*

---

# 4 Now! Do it with data

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

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

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

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

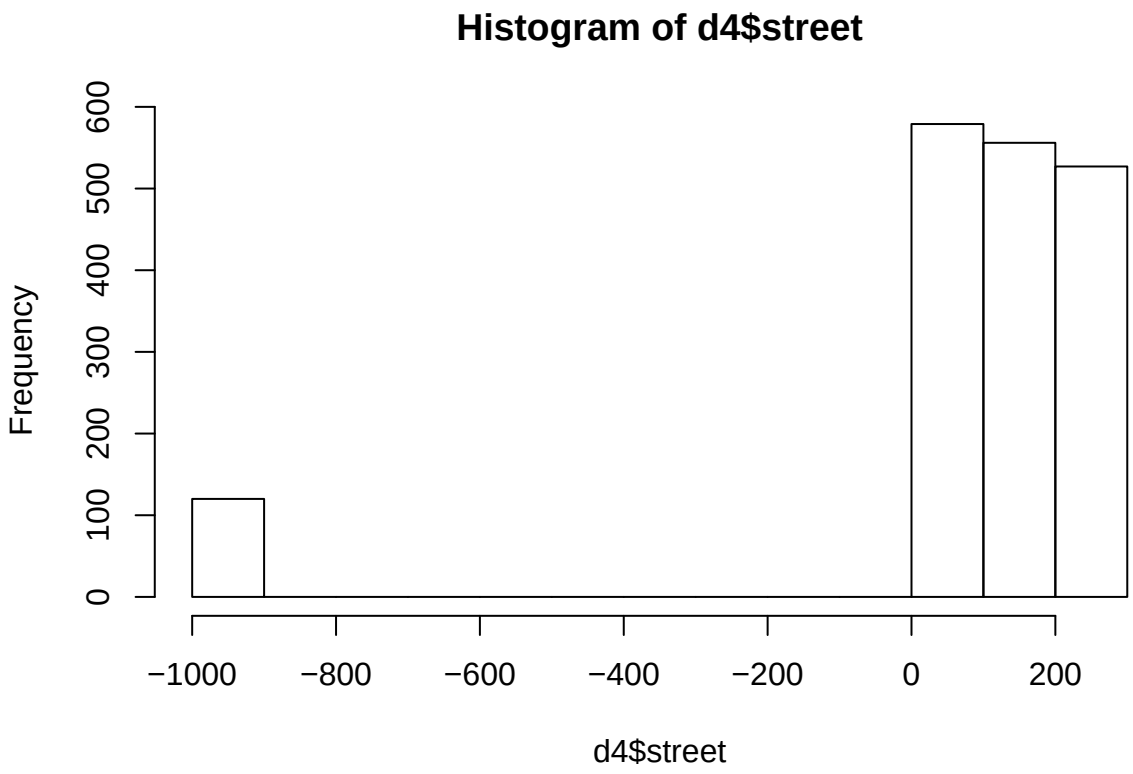
```
## Do some quick exploratory data analysis with this data. There are some values in this data that seem
```

```
# Summary
summary(d4)
```

```
##      street           havecell      avg_bins_treat   base_avg_bins_treat
##  Min.   :-999.00   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  69.00   1st Qu.:0.0000   1st Qu.:0.4167   1st Qu.:0.3750
##  Median : 131.50   Median :1.0000   Median :0.6250   Median :0.6250
##  Mean   :  68.81   Mean   :0.5908   Mean   :0.6811   Mean   :0.7363
##  3rd Qu.: 215.00   3rd Qu.:1.0000   3rd Qu.:0.8333   3rd Qu.:1.0000
##  Max.   : 263.00   Max.   :1.0000   Max.   :4.1667   Max.   :6.3750
##  NA's   :3         NA's   :1
##       bin              sms             bin_s            bin_g
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000   Median :0.0000   Median :0.0000
##  Mean   :0.3378   Mean   :0.3087   Mean   :0.1681   Mean   :0.1697
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##
##      sms_p            sms_g
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.0000   1st Qu.:0.0000
##  Median :0.0000   Median :0.0000
##  Mean   :0.1557   Mean   :0.1529
##  3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :1.0000   Max.   :1.0000
##
```

**Answer: "street" has some values that seem out of the normal range(-999). It may be that those that did not have any recordings for the street. Investigating further Summary shows that other variables have values within normal ranges for those variables**

```
# Looking closely at the street variable
hist(d4$street)
```



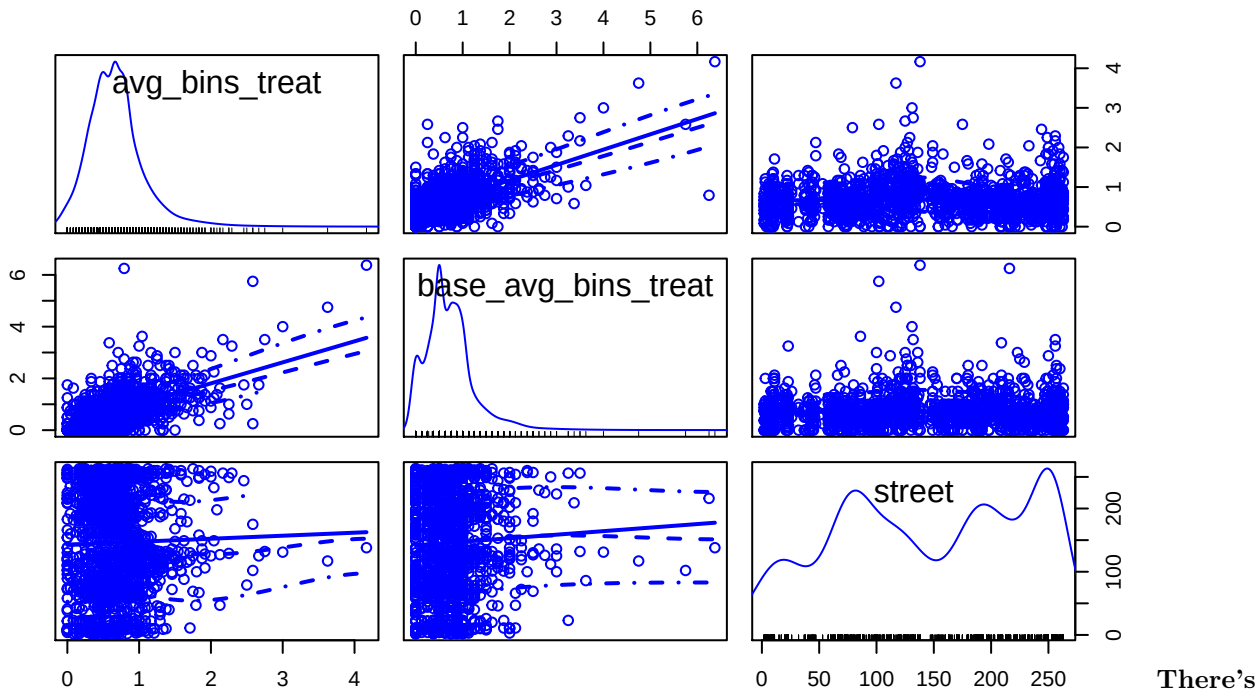

**So yes, just a couple of troubling values (-999) in "street". We can replace them with "NA"" to be consistent with rest of the document(there are other NA's for values that were not recorded)**

```
d4$street[d4$street == -999] <- NA
hist(d4$street, main = "Distribution of Street after replacing -999's with NA")
```


**The distribution of "Street" looks reasonable now The results of the regression do not dramatically change on fixing the input. We see a very slight lowering of SE in some of the regressions, especially the fully saturated model**

```
# Investigating the data some more
#scatterplotMatrix( ~ avg_bins_treat + base_avg_bins_treat + street + havecell + bin + sms + bin_s + bi
```

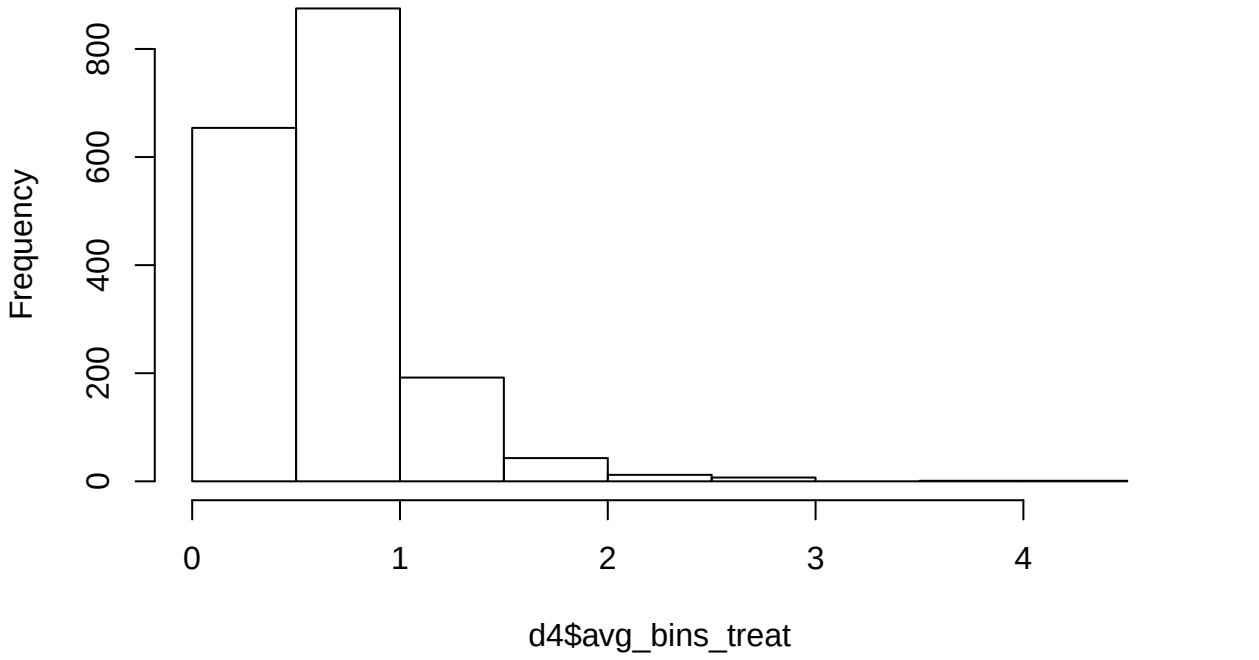
```
# Investigating with a  less crowded matrix
scatterplotMatrix( ~ avg_bins_treat + base_avg_bins_treat + street, data = d4, main = "Scatterplot Matri
```




There's correlation between baseline and the treatment effect. The "Street" has almost the same relationship with the baseline and treatment effect. This is intuitive

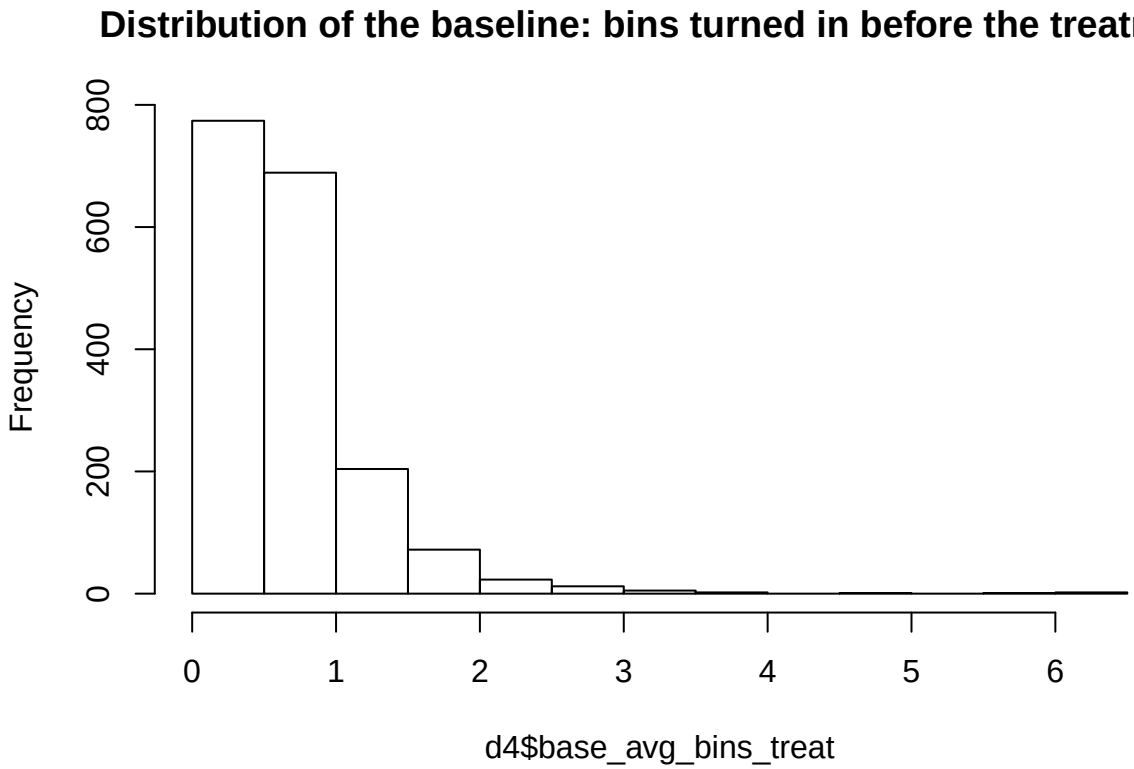
Next, investigating the distribution of the main outcome variable

```
hist(d4$avg_bins_treat, main = "Distribution of the outcome: bins turned in after the treatment")
```



**How was the baseline?**

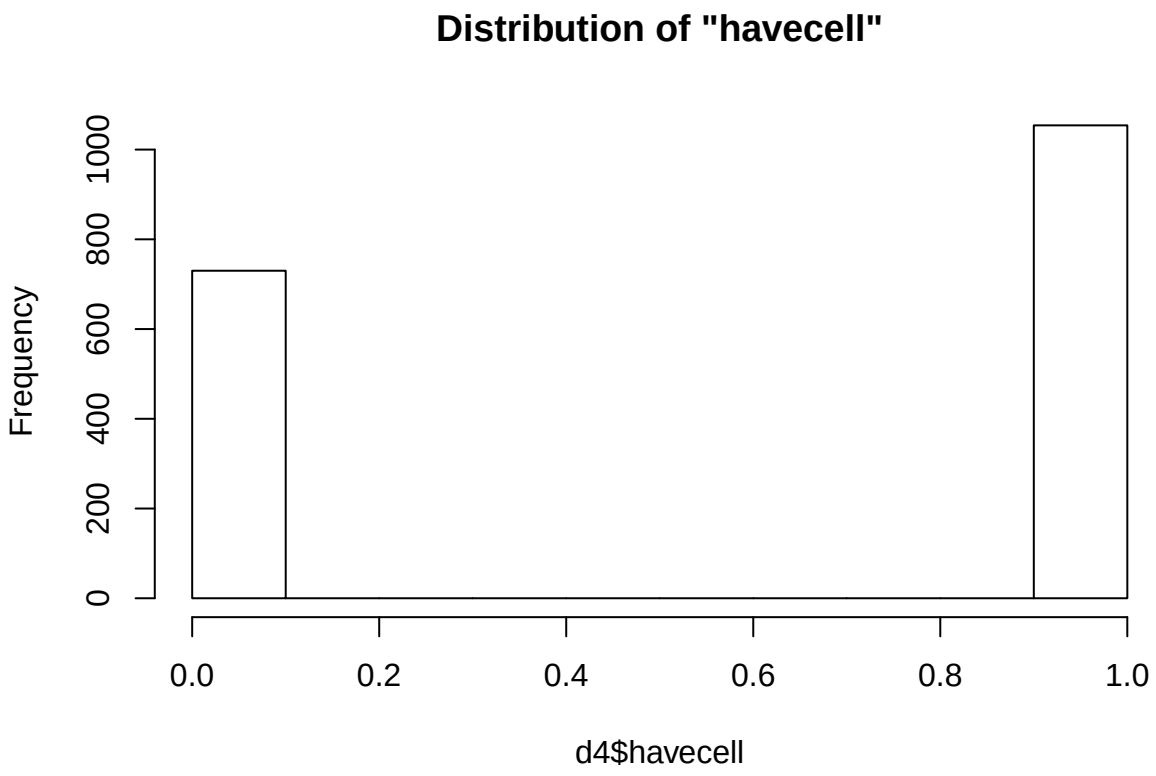
```
hist(d4$base_avg_bins_treat, main = "Distribution of the baseline: bins turned in before the treatment")
```




**Both distributions above are not normal. They are similar though and the sample size is big enough**

**Curious about the distribution of "havecell"**

```
hist(d4$havecell, main = "Distribution of \"havecell\"")
```




**This makes sense per the experiment's design. Folks that have a cell phone were further divided for the SMS treatment**

*Enough with the exploration*

---

**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.**

```
# Effect of providing a recycling bin
lr4a = lm(avg_bins_treat ~ bin, data = d4)
summary(lr4a)
```

```
##
## Call:
## lm(formula = avg_bins_treat ~ bin, data = d4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7707 -0.2603 -0.0520  0.1876  3.5313
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.63535    0.01179  53.874  < 2e-16 ***
## bin          0.13538    0.02029   6.672 3.36e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.4055 on 1783 degrees of freedom
## Multiple R-squared:  0.02436,    Adjusted R-squared:  0.02381
## F-statistic: 44.52 on 1 and 1783 DF,  p-value: 3.356e-11
```

```
lr4a.confint = confint(lr4a, level = .95)
print('95% confidence interval')
```

```
## [1] "95% confidence interval"
```

```
lr4a.confint
```

```
##                  2.5 %    97.5 %
## (Intercept) 0.61221964 0.6584797
## bin         0.09558421 0.1751758
```

**Answer: The ATE/Coefficient of bins is: 0.13538 (0.02029) 95% confidence interval is: [0.09558421 0.1751758]**

**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.**

```
# Adding pre-treatment value of Y as a covariate
lr4b = lm(avg_bins_treat ~ bin + base_avg_bins_treat, data = d4)
summary(lr4b)
```

```
##
## Call:
## lm(formula = avg_bins_treat ~ bin + base_avg_bins_treat, data = d4)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01396 -0.21275 -0.02647  0.16665  2.13549
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)          0.34960    0.01373  25.460  < 2e-16 ***
## bin                  0.12469    0.01667   7.481 1.15e-13 ***
## base_avg_bins_treat  0.39296    0.01339  29.356  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.333 on 1782 degrees of freedom
## Multiple R-squared:  0.3424, Adjusted R-squared:  0.3416
## F-statistic: 463.9 on 2 and 1782 DF,  p-value: < 2.2e-16
```

```
lr4b.confint = confint(lr4b, level = .95)
print('95% confidence interval')
```

```
## [1] "95% confidence interval"
```

```
lr4b.confint
```

```
##                         2.5 %    97.5 %
## (Intercept)         0.32267158 0.3765335
## bin                 0.09200378 0.1573822
```

```
## base_avg_bins_treat 0.36671039 0.4192189
```

**Answer: The ATE/Coefficient of bins is: 0.12469 (0.01667) 95% confidence interval is: [0.09200378 0.1573822] The confidence interval shrinks/becomes more precise when we add the baseline as a covariant. This is because some of the variance in the outcome is explained by the baseline (measurement prior to the experiment)**

**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.**

```
# adding the street fixed effects
# Converting the street to factor so the actual number does not count
d4$streetF = factor(d4$street)
head(d4$streetF)
```

```
## [1] 7 7 7 7 6 8
## 179 Levels: 2 3 4 5 6 7 8 9 10 11 15 17 20 21 22 23 26 32 37 38 40 ... 263
```

```
is.factor(d4$streetF)
```

```
## [1] TRUE
```

```
# Adding the street fixed effects into the regression
# Do we not need an interaction term? They may be already captured in the variable.
lr4c = lm(avg_bins_treat ~ bin + base_avg_bins_treat + streetF, data = d4)
lr4c.summary = summary(lr4c)
lr4c.summary$coefficients["bin",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.162529e-01 1.758668e-02 6.610282e+00 5.338231e-11
```

```
lr4c.confint = confint(lr4c, level = .95)
print('95% confidence interval')
```

```
## [1] "95% confidence interval"
```

```
lr4c.confint["bin",]
```

```
##      2.5 %     97.5 %
## 0.08175545 0.15075034
```

**Answer: The 95% confidence interval with the street effects is [0.08175545 0.15075034 ]**

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

**Answer: The 2 intervals differ very little (<.01) as the effect of the street is partially captured by the baseline. Its intuitive that certain neighborhoods may have higher inclination to recycle, which can be captured mostly in the preexperiment measurements (baseline)**

**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.**

**Answer: R implementation is below:**

```
# Add the column "no_cell" to the data
d4$no_cell = 1-d4$havecell
head(d4)
```

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

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

```
# Adding "no_cell to the regression"
# Without the interaction term first
lr4f = lm(avg_bins_treat ~ bin + base_avg_bins_treat + streetF + no_cell, data = d4)
lr4f.summary = summary(lr4f)
print('Coefficient for bin')
```

```
## [1] "Coefficient for bin"
```

```
lr4f.summary$coefficients["bin",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.171694e-01 1.758614e-02 6.662600e+00 3.784216e-11
```

```
print('Coefficient for no_cell')
```

```
## [1] "Coefficient for no_cell"
```

```
lr4f.summary$coefficients["no_cell",]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.04296723  0.01756186 -2.44662140  0.01453597
```

```
lr4f.confint = confint(lr4f, level = .95)
print('95% confidence interval')
```

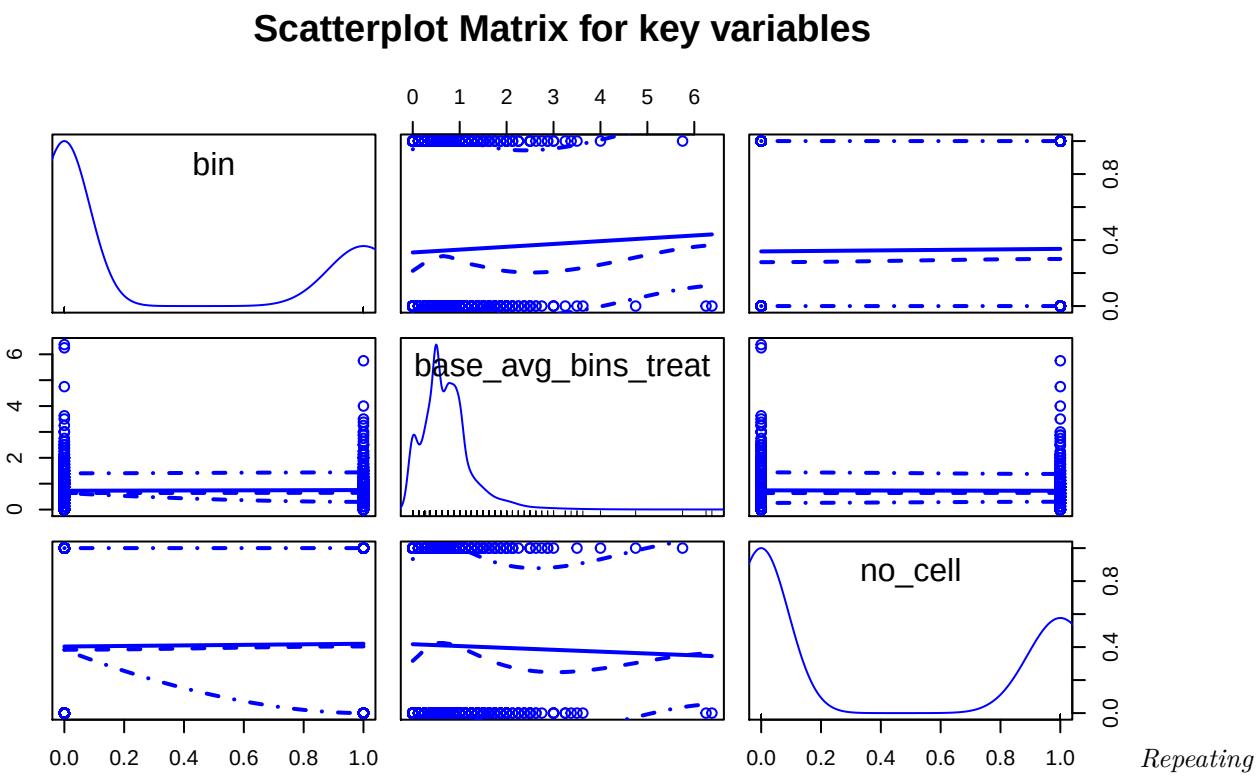
```
## [1] "95% confidence interval"
```

```
lr4f.confint["bin",]
```

```
##     2.5 %    97.5 %
## 0.0826730 0.1516659
```

*Correlation between bin, baseline, and cell phone would be good to know before answering this question*

```
# Correlation between  bin, baseline, and cell phone would be good to know before answering this questi
scatterplotMatrix( ~ bin + base_avg_bins_treat + no_cell, data = d4, main = "Scatterplot Matrix for key
```




*Repeating the regression above with the interaction term*

```
# And to make sure we account for the interaction, we can inlude the interaction term
lr4f = lm(avg_bins_treat ~ bin + base_avg_bins_treat + streetF + no_cell + no_cell * bin, data = d4)
lr4f.summary = summary(lr4f)
print('Coefficient for bin')
```

```
## [1] "Coefficient for bin"
```

```
lr4f.summary$coefficients["bin",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.254210e-01 2.291776e-02 5.472654e+00 5.200633e-08
```

```
print('Coefficient for no_cell')
```

```
## [1] "Coefficient for no_cell"
```

```
lr4f.summary$coefficients["no_cell",]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.03629090  0.02120942 -1.71107436  0.08727721
```

```
lr4f.confint = confint(lr4f, level = .95)
print('Coefficient for no_cell*bin')
```

```
## [1] "Coefficient for no_cell*bin"
```

```
lr4f.summary$coefficients["bin:no_cell",]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## -0.01996593  0.03554587 -0.56169471  0.57440920
```

```
print('95% confidence interval')
```

```
## [1] "95% confidence interval"
```

```
lr4f.confint["bin",]
```

```
##      2.5 %     97.5 %
## 0.08046617 0.17037576
```

**Answer:**

1. The 95% confidence interval of the treatment effect with "no cell" as the co-variate is [0.0826730 0.1516659] without the interaction term
2. The 95% confidence interval of the treatment effect with "no cell" as the co-variate is [0.08046617 0.17037576] with the interaction term

**The confidence interval almost matches the previous one as having a cell phone (or not) does not explain any of the variance in the primary treatment effect (bin). The variables are not correlated. The correlation charts and the coefficient for the interation term (-0.01996593 (0.03554587)) show the same**

**Answer: Per the experiment's design, "hascell" is a pre-treatment condition used to identify the group that was further divided for the SMS treatments. "hascell" was not considered as one of the independent variables having an effect on the outcome. With that being true, we were not measuring or expecting it to have an impact on the outcome**

```
# lr4f.summary$coefficients
```

**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.**

```
# Adding SMS to the regression
# Do we not need an interaction term? Yes, adding in the next code snippet
lr4g = lm(avg_bins_treat ~ bin + base_avg_bins_treat + streetF + no_cell + sms, data = d4)
lr4g.summary = summary(lr4g)
print('Coefficient for bin')
```

```
## [1] "Coefficient for bin"
```

```
lr4g.summary$coefficients["bin",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.169678e-01 1.759005e-02 6.649659e+00 4.122683e-11
```

```
print('Coefficient for sms')
```

```
## [1] "Coefficient for sms"
```

```
lr4g.summary$coefficients["sms",]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.01740018 0.02166596 0.80311146 0.42203945
```

```
lr4g.confint = confint(lr4g, level = .95)
print('95% confidence interval')
```

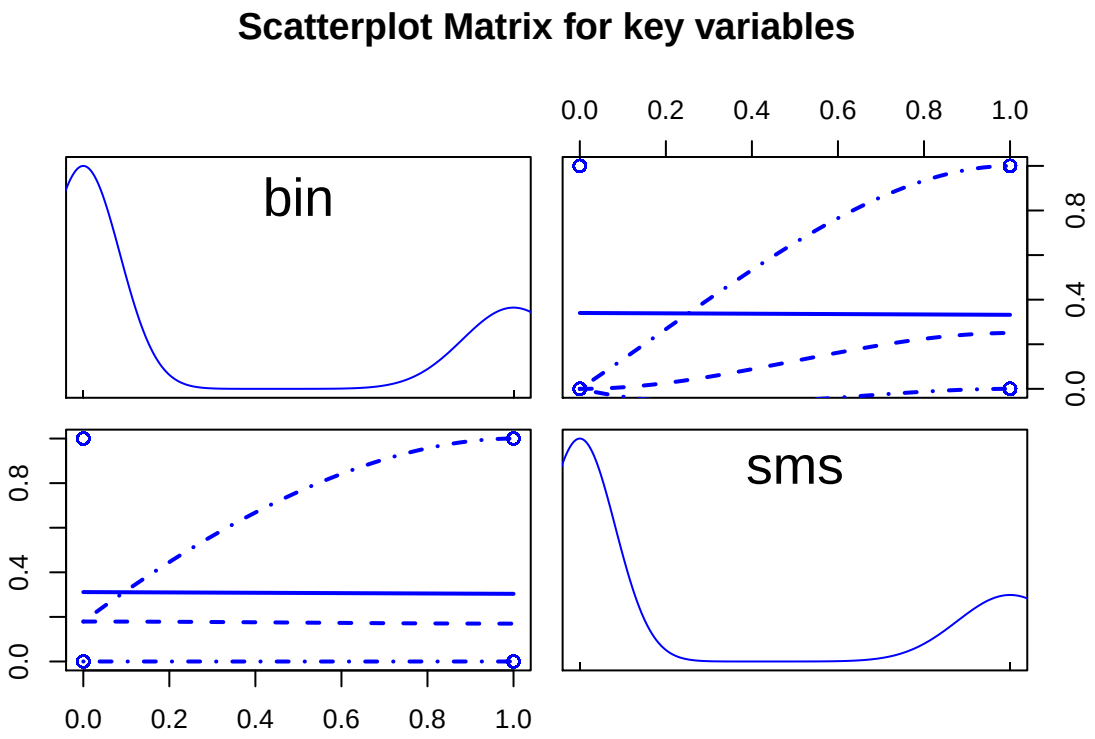
```
## [1] "95% confidence interval"
```

```
lr4g.confint["bin",]
```

```
##     2.5 %    97.5 %
## 0.0824637 0.1514719
```

*Checking for correlation among bin and sms. We're not expecting any*

```
# Correlation between  bin, baseline, and cell phone would be good to know before answering this questi
scatterplotMatrix( ~ bin + sms, data = d4, main = "Scatterplot Matrix for key variables")
```




*Repeating the regression with the interaction terms*

```
# Adding SMS to the regression
# Do we not need an interaction term?
lr4g = lm(avg_bins_treat ~ bin + base_avg_bins_treat + streetF + no_cell + sms + bin*sms, data = d4)
lr4g.summary = summary(lr4g)
print('Coefficient for bin')
```

```
## [1] "Coefficient for bin"
```

```
lr4g.summary$coefficients["bin",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.135595e-01 2.121844e-02 5.351924e+00 1.007610e-07
```

```
print('Coefficient for sms')
```

```
## [1] "Coefficient for sms"
```

```
lr4g.summary$coefficients["sms",]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.01372789 0.02515845 0.54565732 0.58538391
```

```
print('Coefficient for bin:sms')
```

```
## [1] "Coefficient for bin:sms"
```

```
lr4g.summary$coefficients["bin:sms",]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.01093590 0.03804816 0.28742262 0.77382906
```

```
lr4g.confint = confint(lr4g, level = .95)
print('95% confidence interval')
```

```
## [1] "95% confidence interval"
```

```
lr4g.confint["bin",]
```

```
##      2.5 %     97.5 %
## 0.07193798 0.15518094
```

**Answer: Yes, we get the same results as table 4A Answer: The 95% confidence interval with SMS included in the regression is [0.0824637 0.1514719 ] without the interaction term The 95% confidence interval with SMS included in the regression is [0.07193798 0.15518094 ] with the interaction term The confidence interval does not change again as "sms" does not explain much of the variance/noise in the treatment effect. SMS and bin do not look correlated.Another observation could be that sending an sms does not really help that much on top of providing a bin**

**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.**

```
# Adding 2 types of SMS treatment
# Do we not need an interaction term? Assuming not as we are trying to reproduce the results of table 4
lr4h = lm(avg_bins_treat ~ bin_g + bin_s + base_avg_bins_treat + streetF + no_cell + sms_p + sms_g, data
lr4h.summary = summary(lr4h)
print('Coefficient for bin_g')
```

```
## [1] "Coefficient for bin_g"
```

```
lr4h.summary$coefficients["bin_g",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.057386e-01 2.272342e-02 4.653288e+00 3.559514e-06
```

```
print('Coefficient for bin_s')
```

```
## [1] "Coefficient for bin_s"
```

```
lr4h.summary$coefficients["bin_s",]
```

```
##     Estimate   Std. Error      t value     Pr(>|t|)
## 1.288619e-01 2.272963e-02 5.669336e+00 1.721592e-08
```

```
print('Coefficient for sms_g')
```

```
## [1] "Coefficient for sms_g"
```

```
lr4h.summary$coefficients["sms_g",]
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
## 0.02950443 0.02616216 1.12775207 0.25960793
```

```
print('Coefficient for sms_p')
```

```
## [1] "Coefficient for sms_p"
```

```
lr4h.summary$coefficients["sms_p",]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## 0.006659619 0.026030012 0.255843882 0.798107038
```

```
lr4h.confint = confint(lr4h, level = .95)
print('95% confidence interval for the un-adorned bin')
```

```
## [1] "95% confidence interval for the un-adorned bin"
```

```
lr4h.confint["bin_g",]
```

```
##      2.5 %     97.5 %
## 0.06116499 0.15031225
```

**Answer: The results match column 2 of Table 4b**

1. The treatment effect of Bin without Sticker is: 1.057386e-01 (2.272342e-02)
2. The treatment effect of Bin with Sticker is is: 1.288619e-01 (2.272963e-02)
3. The treatment effect of generic SMS is: 0.02950443 (0.02616216)
4. The treatment effect of specific SMS is: 0.006659619 (0.026030012)

**Answer: The 95% confidence interval with generic recycling bin is [0.06116499 0.15031225 ] The confidence interval is a little tighter from that in part (g). The increased precision comes by breaking the "bins" variable into more precise "bin_g" and "bin_s". "bin_s" (bins with stickers) is able to explain some of the variance that was earlier in the confidence interval for generic "bin".**

---

# 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.

```
d5 <- read.csv("./data/ebola_rct2.csv")
head(d5)
```

```
##   temperature_day0 vomiting_day0 treat_zmapp temperature_day14
## 1         99.53168             1           0          98.62634
## 2         97.37372             0           0          98.03251
## 3         97.00747             0           1          97.93340
## 4         99.74761             1           0          98.40457
```

```
## 5         99.57559             1           1          99.31678
## 6         98.28889             1           1          99.82623
##   vomiting_day14 male
## 1              1    0
## 2              1    0
## 3              0    1
## 4              1    0
## 5              1    0
## 6              1    1
```

```
# Getting more famiiliar with the data
summary(d5)
```

```
##  temperature_day0 vomiting_day0   treat_zmapp   temperature_day14
##  Min.   :97.01    Min.   :0.00   Min.   :0.00   Min.   : 97.09
##  1st Qu.:97.70    1st Qu.:0.00   1st Qu.:0.00   1st Qu.: 98.09
##  Median :98.57    Median :1.00   Median :0.00   Median : 98.74
##  Mean   :98.49    Mean   :0.66   Mean   :0.41   Mean   : 99.13
##  3rd Qu.:99.25    3rd Qu.:1.00   3rd Qu.:1.00   3rd Qu.: 99.67
##  Max.   :99.96    Max.   :1.00   Max.   :1.00   Max.   :102.53
##  vomiting_day14      male
##  Min.   :0.00   Min.   :0.00
##  1st Qu.:0.75   1st Qu.:0.00
##  Median :1.00   Median :0.00
##  Mean   :0.75   Mean   :0.37
##  3rd Qu.:1.00   3rd Qu.:1.00
##  Max.   :1.00   Max.   :1.00
```

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

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

```
# Regressing vomiting on day 14 on the treatment
lr_v1 = lm(vomiting_day14 ~ treat_zmapp, data = d5)
summary(lr_v1)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp, data = d5)
##
## 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
```

**Answer: The estimated effect is -0.23770(0.08563)** ** The p-value is 0.0066**

**b. Add covariates for vomiting on day 0 and patient temperature on day 0 to the regression from part (a) and report the ATE (with standard error). Also report the p-value.**

```
## Adding 2 covariates to the regression above
lr_v2 = lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0, data = d5)
summary(lr_v2)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0,
##     data = d5)
##
## 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
```

**Answer:The coefficient/ATE is -0.16554(0.07567) The p-value is .03113**

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

**Answer: I prefer the estimate in part (b). Its intuitive that patient's medical condition on day 14 would depend to some extent on there baseline condition on day 0. Leaving intuition aside, the analysis above shows that temperature on day 0 is statistically significant at a 5% confidence level. The coefficient of "temperature_day0" is .20 which is ~3 times the standard error of .07634. Also we get lower SE for the treatment effect estimate in the model with the baseline vomiting and temperature**

**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
## Adding temperature on day 14 a covariate to the regression above
lr_v3 = lm(vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + temperature_day14, data = d
summary(lr_v3)
```

```
##
## Call:
## lm(formula = vomiting_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 +
##     temperature_day14, data = d5)
##
## 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
```

**Answer: The ATE(SE) are -0.12010(0.07768). The p-value is .12541**

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

**Answer: I prefer the estimate in part(b). (d) adds temperature on day 14 as a covariate. This may be "bad control" as the "temperature on day 14" may be an effect of the treatment. The "temperature on day 14" could be impacted by both the treatment and the baseline variables (temperature and vomiting on day 0).**

**f. Now let's switch from the outcome of vomiting to the outcome of temperature, and use the same regression covariates as in part (b). Test the hypothesis that ZMapp is especially likely to reduce men's temperatures, as compared to women's, and describe how you did so. What do the results suggest?**

```r
# Regressing Day 14 temperature on the treatment and covariates

# First doing it for the entire sample. Sex is not a covariate
lr_v4 = lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0, data = d5)
summary(lr_v4)
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp + vomiting_day0 +
##     temperature_day0, data = d5)
##
## 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
```

*Note: We get an ATE estimate of -0.7554 (0.2587) with significance at the 5% confidence level. The p value is .00436*

```
# Trying the same regression with sex (male, Female) as a indicator variable Male = 1
# We also add the interaction term to measure the impact of "male" on treatment

lr_v5 = lm(temperature_day14 ~ treat_zmapp + vomiting_day0 + temperature_day0 + male + male*treat_zmapp
summary(lr_v5)
```

```
##
## Call:
## lm(formula = temperature_day14 ~ treat_zmapp + vomiting_day0 +
##     temperature_day0 + male + male * treat_zmapp, data = d5)
##
## 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
```

**Answer: One way to find out the relationship between sex and treatment would be to make "male" an indicator variable/covariate and add both the variable and the interaction term in the regression equation. The coefficient of the interaction term (male * treat_zmap) would tell us about the delta effect of being both "male" and in "treatment"**

**The coefficient of the interaction term is -2.07669(.1916) which is statistically significant at 1% confidence level**

**The result tells us that being "male" and in "treatment" is estimated to cause a -2.07 reduction (~ 10 times the standard error) in temperature from the case where this interaction is not present (i.e. women in treatment)**

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

**Answer: I would be sceptical of the efficacy of ZMapp (for curing Ebola in men) based on above paragraph alone. With 20M regressions (10K covariates, 2K health outcomes), it is likely that one finds an interaction with statistical significance by chance alone. From "Field Experiments": With 20 covariates, 1% confidence level, a statistically significant effect for a sub group can happen for 1 in 6 research studies Belief in such a claim would require a planned experiment where the treatment is varied along with the variable/sub-group under observation for the heterogenious effect**

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

**Answer: I would be more inclined to believe in the effect if it was tested by planning/design, as part of the experiment. The experiment in this case would have to vary the subgroup characteristic we are testing (sex in this case) along with the main treatment (Multi- factor experiment). If the experiment is not designed like above, the sub-group/co-variate analysis does not give us causal information. We know that treatment is more effective in groups of men, but we cannot establish 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.*)**

**Answer: This may be one of the FUQ's (defined in MHE chapter 1). It is impossible to come up with a practical experiment that can answer this question.**

**If we start with randomly chosen people and observe them for Ebola, we can create a record for who gets it. If the group can be divided into those of African descent and those not, we'll be able to observe if the there's more Ebola in one group over the other. Even if we find that more from Aftrican descent get Ebola in our study, we'll not be able to establish a causal relationship due to multiple reasons:**

1. There's no precise intervention which'll help us do an experiment to establish cause

2. There are too many confounding factors (Conditions at birth, Time spent in Africa, African parents but born outside. . . . . ) that are difficult to control for.

**The answer to this question does not really help also. One cannot change one's descent. The answer could help spark further investigation though i.e. there could be more precise questions on why a subgroup (of the population) is more prone to EBOLA**