GR5206 Midterm Exam

Quan and qy2205

10/19/2018

The STAT GR5206 Fall 2018 Midterm is open notes, open book(s), open computer and online resources are allowed. Students are **not** allowed to communicate with any other people regarding the exam. This includes emailing fellow students, using WeChat and other similar forms of communication. Before the exam, the students should **turn off** their cellphone and pass it to the left side of each row. At the same time, please **close** the mailbox and **log out** WeChat and all the other apps for messaging and chatting. If there is any suspicion of one or more students cheating, further investigation will take place. If students do not follow the guidelines, they will receive a zero on the exam and potentially face more severe consequences. The exam will be posted on Canvas at **2:50PM**. Students are required to submit both the .pdf and .Rmd files on Canvas (or .html if you must) by **4:30PM**. Late exams will not be accepted.

Part 1 (Google Play Store Apps Data - Split/Apply/Combine and R plot, 11 + 2 pts)

We work on the `apps` dataset which contains approximately $7,700$ Google Play Store apps. There are 13 features that describe a given app. They are:

- **App** – Application name.
- **Category** – Category the app belongs to.
- **Rating** – Overall user rating of the app (between 0 and 5).
- **Reviews** – Number of user reviews for the app.
- **Size** – Size of the app.
- **Installs** – Number of user downloads/installs for the app.
- **Type** – Paid of Free
- **Price** – Price of the app
- **Content Rating** Age group the app is targeted at
- **Genres** An app can belong to multiple genres (apart from its main category). For example, a musical family game will belong to Music, Game, Family genres.
- **Last Updated** Date when the app was last updated on Play Store
- **Current Ver** Current version of the app available on Play Store
- **Android Ver** Minimum required Android version

Read in the dataset using the following code:

```
apps<-read.csv("apps_v2.csv", header = T)
apps$Reviews<-as.numeric(apps$Reviews)
apps$Installs<-factor(apps$Installs, level= c("1+", "5+", "10+", "50+", "100+", "500+", "1,000+", "5,00
head(apps)
```

```
##   X       Category Rating Reviews Size    Installs Type Price
## 1 1 ART_AND_DESIGN    4.1     159 19.0     10,000+ Free     0
## 2 2 ART_AND_DESIGN    3.9     967 14.0    500,000+ Free     0
## 3 3 ART_AND_DESIGN    4.7   87510  8.7  5,000,000+ Free     0
```

```
## 4 4 ART_AND_DESIGN    4.5  215644 25.0 50,000,000+ Free     0
## 5 5 ART_AND_DESIGN    4.3     967  2.8    100,000+ Free     0
## 6 6 ART_AND_DESIGN    4.4     167  5.6     50,000+ Free     0
##   Content.Rating                     Genres     Last.Updated
## 1       Everyone                Art & Design  January 7, 2018
## 2       Everyone Art & Design;Pretend Play January 15, 2018
## 3       Everyone                Art & Design   August 1, 2018
## 4           Teen                Art & Design     June 8, 2018
## 5       Everyone   Art & Design;Creativity    June 20, 2018
## 6       Everyone                Art & Design   March 26, 2017
##          Current.Ver  Android.Ver
## 1              1.0.0 4.0.3 and up
## 2              2.0.0 4.0.3 and up
## 3              1.2.4 4.0.3 and up
## 4 Varies with device   4.2 and up
## 5                1.1   4.4 and up
## 6                1.0   2.3 and up
```

Problem 1.0

Check the dimension of `apps`, make sure that there are $7,726$ lines and 13 variables (features). [1 pt]

```
# code goes here
dim(apps)
```

```
## [1] 7726   13
```

Prblem 1.1

In order to get an overview of the dataset, we want to check some summary statistics of each variable, and this can be done by calling the `R` function `summary()`. Compute the summary statistics of all 13 variables and display the results in a **list**. To receive full credit, you must use a vectorized function from the `apply` family or `plyr` family. [2 pts]

```
# code goes here
# data.frame(unclass(summary(apps)), check.names = FALSE, stringsAsFactors = FALSE)
lapply(apps, summary)
```

```
## $X
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##       1    2691    5432    5436    8146   10841
##
## $Category
##      ART_AND_DESIGN   AUTO_AND_VEHICLES              BEAUTY
##                  59                  63                  37
## BOOKS_AND_REFERENCE            BUSINESS              COMICS
##                 144                 246                  48
##       COMMUNICATION              DATING           EDUCATION
##                 211                 173                 110
##       ENTERTAINMENT              EVENTS              FAMILY
##                  90                  38                1617
##             FINANCE      FOOD_AND_DRINK                GAME
##                 266                  84                 974
##  HEALTH_AND_FITNESS      HOUSE_AND_HOME  LIBRARIES_AND_DEMO
##                 223                  56                  62
##           LIFESTYLE MAPS_AND_NAVIGATION             MEDICAL
```

```
##                280                 95                324
##  NEWS_AND_MAGAZINES          PARENTING     PERSONALIZATION
##                169                 44                280
##         PHOTOGRAPHY        PRODUCTIVITY            SHOPPING
##                236                235                179
##              SOCIAL             SPORTS              TOOLS
##                177                246                633
##    TRAVEL_AND_LOCAL       VIDEO_PLAYERS             WEATHER
##                160                116                 51
##
## $Rating
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   4.300   4.174   4.500   5.000
##
## $Reviews
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##        1      107     2324   294777    38959 44893888
##
## $Size
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.008   5.300  14.000  22.959  33.000 100.000
##
## $Installs
##             1+             5+            10+            50+           100+
##              3              9             67             56            303
##           500+         1,000+         5,000+        10,000+        50,000+
##            197            690            420            969            436
##       100,000+       500,000+     1,000,000+     5,000,000+    10,000,000+
##           1037            490           1301            535            825
##    50,000,000+   100,000,000+   500,000,000+ 1,000,000,000+
##            147            201             30             10
##
## $Type
## Free Paid
## 7147  579
##
## $Price
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   0.000   0.000   1.128   0.000 400.000
##
## $Content.Rating
##     Everyone Everyone 10+   Mature 17+         Teen
##         6172          318          368          868
##
## $Genres
##                            Tools                    Entertainment
##                              633                              448
##                        Education                          Medical
##                              417                              324
##                           Action                  Personalization
##                              322                              280
##                        Lifestyle                          Finance
##                              279                              266
##                           Sports                         Business
```

```
##                             260                             246
##                     Photography                    Productivity
##                             236                             235
##                  Health & Fitness                   Communication
##                             223                             211
##                          Arcade                      Simulation
##                             186                             182
##                        Shopping                          Social
##                             179                             177
##                          Dating                  News & Magazines
##                             173                             169
##                          Casual                    Travel & Local
##                             160                             159
##                 Books & Reference           Video Players & Editors
##                             144                             115
##                          Puzzle                    Role Playing
##                             108                             103
##                        Strategy                  Maps & Navigation
##                              96                              95
##                     Food & Drink                         Racing
##                              84                              83
##                       Adventure                  Auto & Vehicles
##                              68                              63
##                  Libraries & Demo                    House & Home
##                              62                              56
##                     Art & Design                         Weather
##                              53                              51
##                          Comics              Education;Education
##                              47                              41
##                            Card                           Board
##                              39                              38
##                          Events                          Beauty
##                              38                              37
##                       Parenting             Educational;Education
##                              35                              33
##                          Casino                     Educational
##                              32                              31
##                Casual;Pretend Play                          Trivia
##                              30                              27
##                            Word             Education;Pretend Play
##                              24                              22
##           Educational;Pretend Play                Puzzle;Brain Games
##                              18                              18
##          Action;Action & Adventure         Entertainment;Music & Video
##                              16                              16
##          Casual;Action & Adventure                           Music
##                              15                              15
##                  Board;Brain Games           Racing;Action & Adventure
##                              14                              14
##       Adventure;Action & Adventure           Arcade;Action & Adventure
##                              13                              13
##                 Casual;Brain Games      Simulation;Action & Adventure
##                              13                              11
##                  Casual;Creativity           Art & Design;Creativity
```

```
##                                 7                                 6
##       Education;Action & Adventure           Educational;Brain Games
##                                 6                                 6
##          Entertainment;Brain Games              Education;Creativity
##                                 6                                 5
##             Educational;Creativity           Parenting;Music & Video
##                                 5                                 5
##          Puzzle;Action & Adventure    Role Playing;Action & Adventure
##                                 5                                 5
##          Role Playing;Pretend Play     Educational;Action & Adventure
##                                 5                                 4
##           Board;Action & Adventure                  Casual;Education
##                                 3                                 3
##            Education;Music & Video   Entertainment;Action & Adventure
##                                 3                                 3
##                Music;Music & Video               Parenting;Education
##                                 3                                 3
##               Simulation;Education           Simulation;Pretend Play
##                                 3                                 3
##                Adventure;Education         Art & Design;Pretend Play
##                                 2                                 2
##        Books & Reference;Education           Card;Action & Adventure
##                                 2                                 2
##               Casual;Music & Video          Entertainment;Creativity
##                                 2                                 2
##         Entertainment;Pretend Play                 Puzzle;Creativity
##                                 2                                 2
##          Sports;Action & Adventure        Strategy;Action & Adventure
##                                 2                                 2
## Video Players & Editors;Creativity              Adventure;Brain Games
##                                 2                                 1
##                Arcade;Pretend Play                Board;Pretend Play
##                                 1                                 1
##                   Card;Brain Games                 Comics;Creativity
##                                 1                                 1
##              Education;Brain Games                           (Other)
##                                 1                                13
##
## $Last.Updated
##    August 3, 2018     July 31, 2018    August 1, 2018    August 2, 2018
##               205               189               178               173
##     July 30, 2018     July 25, 2018     July 26, 2018     July 27, 2018
##               130               124               114               101
##     July 24, 2018     July 16, 2018     July 18, 2018     July 23, 2018
##                98                93                84                84
##    August 6, 2018     July 11, 2018     July 17, 2018     July 12, 2018
##                81                80                73                71
##      July 3, 2018     July 19, 2018      July 5, 2018    August 4, 2018
##                71                68                66                63
##     July 20, 2018      July 9, 2018      May 24, 2018      July 6, 2018
##                63                58                58                56
##     July 13, 2018     June 27, 2018     June 26, 2018      May 25, 2018
##                55                48                46                46
##     June 13, 2018      June 6, 2018     June 19, 2018      July 2, 2018
```

```
##               41                41                38                37
##     June 29, 2018     August 5, 2018      June 25, 2018      July 28, 2018
##               37                36                36                35
##     June 20, 2018       July 4, 2018      July 10, 2018      June 21, 2018
##               35                34                33                33
##      June 5, 2018      June 12, 2018     August 7, 2018      July 29, 2018
##               33                32                30                30
##      June 8, 2018       May 23, 2018      June 15, 2018       May 21, 2018
##               29                29                28                28
##      May 31, 2018      June 18, 2018      June 28, 2018       June 7, 2018
##               27                26                26                26
##      May 30, 2018       June 1, 2018       May 22, 2018       May 29, 2018
##               26                25                25                25
##     July 15, 2018       July 8, 2018      June 11, 2018       May 18, 2018
##               24                24                24                24
##    April 26, 2018   February 5, 2017      June 14, 2018       May 28, 2018
##               23                22                22                22
##      July 1, 2018      March 6, 2018      June 22, 2018        May 4, 2018
##               21                20                19                19
##      July 7, 2018       May 15, 2018     April 23, 2018      April 3, 2018
##               18                18                17                17
##      June 4, 2018     March 16, 2018     March 20, 2018      March 5, 2018
##               17                17                17                17
##      May 10, 2018       May 17, 2018     April 11, 2018      July 21, 2018
##               17                17                16                16
##     July 22, 2018      April 5, 2018      April 9, 2018     March 28, 2018
##               16                15                15                15
##    March 29, 2018        May 3, 2018        May 9, 2018     April 17, 2018
##               15                15                15                14
## February 15, 2018       June 9, 2018     March 13, 2018     March 27, 2018
##               14                14                14                14
##       May 2, 2018     April 13, 2018     April 18, 2018     April 20, 2018
##               14                13                13                13
##  February 7, 2018   January 19, 2018    January 2, 2018           (Other)
##               13                13                13              3533
##
## $Current.Ver
##                1.0                1.1                1.2
##                458                195                126
##                1.3                2.0              1.0.1
##                119                118                 80
##                1.4 Varies with device                1.5
##                 77                 73                 72
##              1.0.0                1.6                2.1
##                 67                 56                 52
##              1.0.2              1.0.4                1.7
##                 51                 47                 45
##              1.0.3              1.0.6              1.2.1
##                 44                 43                 43
##              2.0.0                3.0                1.8
##                 41                 41                 39
##              1.0.5              1.1.0              1.2.0
##                 38                 36                 36
##                4.0                1.9              1.0.9
```

```
##                34              33              32
##             2.3.2           1.1.1             2.4
##                32              31              31
##               2.2           1.4.0             3.1
##                28              27              27
##             2.0.1             2.5             5.0
##                26              26              26
##             1.0.7           1.0.8           1.1.3
##                25              25              24
##             1.1.2           1.3.0           3.0.0
##                23              22              22
##             1.2.2           1.2.3           2.1.1
##                21              21              21
##               2.3           3.1.0             3.3
##                20              20              20
##               5.1             4.1           1.1.4
##                20              19              18
##               2.6             8.2           1.1.6
##                18              18              17
##             1.5.0           2.0.5             6.0
##                17              17              17
##             1.3.1           2.1.0           2.1.2
##                16              16              16
##               2.7             2.9               1
##                16              16              15
##              1.01           1.5.1           1.6.1
##                15              15              15
##             2.4.0           2.5.1           3.0.1
##                15              15              15
##             2.0.7           2.4.1           3.1.4
##                14              14              14
##              1.03           1.1.5           1.2.7
##                13              13              13
##               2.8             3.2           1.2.6
##                13              13              12
##             1.5.2           1.6.2           3.0.5
##                12              12              12
##             3.6.1           3.8.0             5.2
##                12              12              12
##               0.1          1.0.11           1.1.7
##                11              11              11
##              1.10            1.11           1.3.3
##                11              11              11
##             2.0.3           2.0.4           2.1.3
##                11              11              11
##             2.3.1           4.1.0         5.9.1.0
##                11              11              11
##               7.0           1.2.5           1.2.9
##                11              10              10
##           (Other)
##              4412
##
## $Android.Ver
##         1.0 and up         1.5 and up         1.6 and up
```

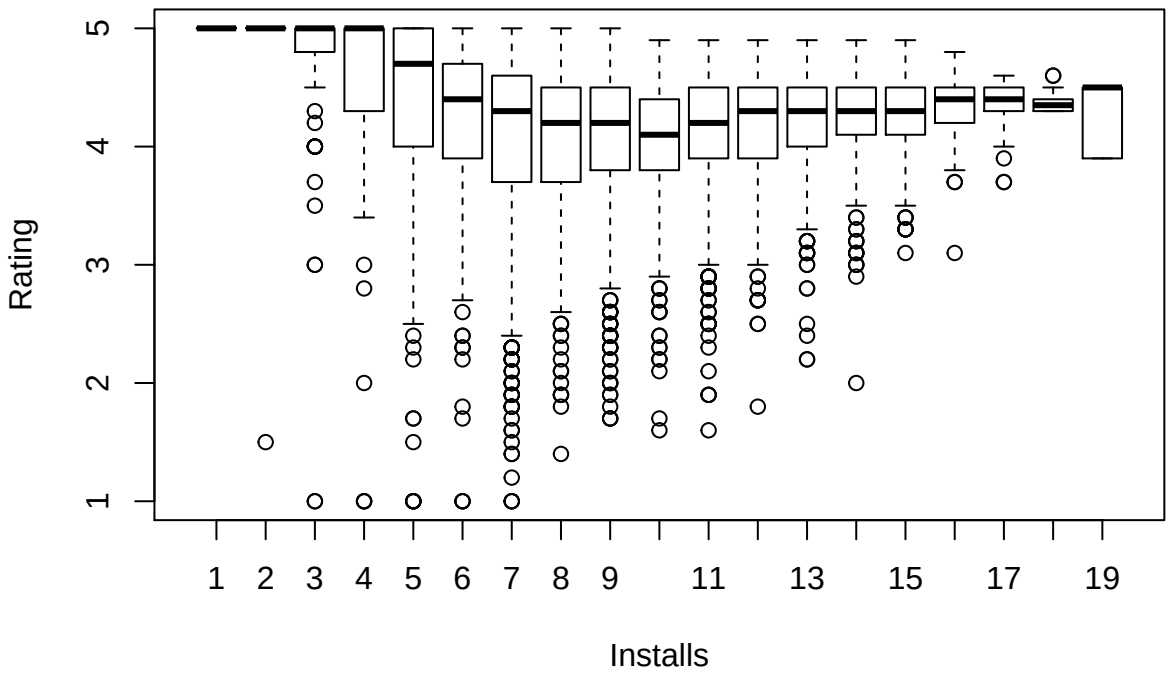
```
##                 2                15                87
##       2.0 and up      2.0.1 and up       2.1 and up
##                27                 7               113
##       2.2 and up       2.3 and up      2.3.3 and up
##               206               566               234
##       3.0 and up       3.1 and up       3.2 and up
##               211                 8                31
##       4.0 and up     4.0.3 - 7.1.1      4.0.3 and up
##              1109                 2              1194
##      4.1 - 7.1.1       4.1 and up       4.2 and up
##                1              1929               318
##       4.3 and up       4.4 and up      4.4W and up
##               195               805                 6
##        5.0 - 6.0        5.0 - 8.0       5.0 and up
##                1                 2               490
##       5.1 and up       6.0 and up      7.0 - 7.1.1
##               17                45                 1
##       7.0 and up       7.1 and up       8.0 and up
##               39                 2                 5
##              NaN Varies with device
##                2                56
```

Problem 1.2

We want to investigate the association between the user's overall rating (`Rating`) of an app and the total number of installs (`Installs`). Use function `plot()` to construct a multiple boxplot of the overall rating of an app split by number of installs (`Installs`). Can you see any relationship in the plot? [3 pts]

Adjust the labels of x-axis. Make sure that all levels of the variable `Installs` show in the plot. [2 extra pts]

```
# code goes here
boxplot(as.numeric(apps$Rating)~as.numeric(apps$Installs), xlab = 'Installs', ylab = 'Rating')
```




```
cat('Answer: based on the boxplot, we can see that when install between 2 and 12, the variance of Ratin
```

```
## Answer: based on the boxplot, we can see that when install between 2 and 12, the variance of Rating
```

Problem 1.3

We now investigate how the overall rating (`Rating`) is associated with the category of the app (`Category`) and its price (`Price`). Use **Split/Apply/Combine** strategy to split the dataset by `Category`. For each category, generate a plot of user's rating (`Rating`) against the app's price (`Price`). Display all plots in one figure. To receive full credit, you must use a vectorized function from the `plyr` family. Make sure your figure contains **33** subplots, with each of them corresponding to one category. [5 pts]

```
# code goes here
library(plyr)
# split
apps.split <- split(apps, apps$Category)
# apply
# ldply(apps.split, plot)
```

Part 2 (Basic Web Scraping, 15 + 2 pts)

In this part, we look at the voting record of the 2018 US Congress for roll call 274. The votes were compiled from http://clerk.house.gov. The raw data have been saved in the file `Roll_Call_274.xml`. We want to extract the voting results for 427 members of the House of Representatives. First, we read in data.

```
rollCall274<-readLines("Roll_Call_274.xml")
```

Problem 2.0

Check the number of lines contained in the `Roll_call_274.xml` file. There should be 485 lines. [1 pt]

```
# code goes here
length(rollCall274)
```

```
## [1] 485
```

Problem 2.1

Use the `grep()` function to find the lines in the file that correspond to the votes. Make sure `grep()` finds 427 lines. Hint: such a line starts with `<recorded-vote>`. [2 pts]

```
# code goes here
vote_pattern <- '<recorded-vote>'
votegrep <- grep(vote_pattern, rollCall274)
length(votegrep)
```

```
## [1] 427
```

Problem 2.2

Write a regular expression that will capture the ID of a member. Using it extract the ID of each member. Hint: you can use the fact that `name-id=` appears before the ID. The ID is inside a pair of quotes, and it consists of one capital latter and six digits. [2 pts]

```
# code goes here
reg_data <- function (pattern, data) {
  sgrep <- grep(pattern, data)
  matches <- gregexpr(pattern = pattern, text = data[sgrep])
  reg.data <- unlist(regmatches(data[sgrep], matches), use.names = FALSE)
  return(reg.data)
}
id_pattern <- 'name-id="[A-Z][0-9]{6}"'
id_data <- reg_data(id_pattern, rollCall274)
id_data <- substr(id_data, start = 10, stop = 16)
head(id_data, 5)
```

```
## [1] "A000374" "A000370" "A000055" "A000371" "A000372"
```

Problem 2.3

Using a regular expression extract the name of each member. Make sure that you can extract all names for 427 members. [2 extra pts]

```
# code goes here
name_pattern <- 'unaccented-name="[a-zA-Z(),.\' |-]+"'
name_data <- reg_data(name_pattern, rollCall274)
name_data <- sapply(strsplit(name_data, split = '"'), '[', 2)
length(name_data)
```

```
## [1] 427
```

```
head(name_data, 5)
```

```
## [1] "Abraham"  "Adams"    "Aderholt" "Aguilar"  "Allen"
```

Problem 2.4

Extract the party of each member by using a regular expression. There should be 193 Democrats and 234 Republicans. [2 pts]

```
# code goes here
party_pattern <- 'party="[A-Z]"'
party_data <- reg_data(party_pattern, rollCall274)
party_data <- substr(party_data, start = 8, stop = 8)
head(party_data, 5)
```

```
## [1] "R" "D" "R" "D" "R"
```

```
table(party_data)
```

```
## party_data
##   D   R
## 193 234
```

Problem 2.5

Extract the state for each member by using a regular expression. [2 pts]

```
# code goes here
state_pattern <- 'state="[A-Z]{2}"'
state_data <- reg_data(state_pattern, rollCall274)
state_data <- substr(state_data, start = 8, stop = 9)
head(state_data, 5)
```

```
## [1] "LA" "NC" "AL" "CA" "GA"
```

Problem 2.6

Last, use a regular expression to extract the vote. [2 pts]

```
# code goes here
vote_pattern1 <- '<vote>[A-Za-z]{2,3}'
vote_data1 <- reg_data(vote_pattern1, rollCall274)
vote_data1 <- substr(vote_data1, start = 7, stop = 100)
head(vote_data1, 5)
```

```
## [1] "Aye" "No"  "Aye" "No"  "Aye"
```

Problem 2.7

Make the extracted vote as a factor, and check its levels. Make a new variable called `numeric.vote`, which takes value 1 if the member voted "Yes (Aye)", 0 if the vote is "No", and -1 for "Not Voting". [2 pts]

```
# code goes here
vote <- factor(vote_data1)
numeric.vote <- ifelse(vote == 'Aye', 1, ifelse(vote == 'No', 0, -1))
head(numeric.vote, 5)
```

```
## [1] 1 0 1 0 1
```

Problem 2.8

Create a dataframe `rollCall274`, which contains the following five variables: `name`, `state`, `party`, `vote`, `numeric.vote`. Use `id` to name the rows of this dataframe. [2 pts]

```r
# code goes here
rollCall274 <- data.frame(name_data, state_data, party_data, vote_data1, numeric.vote)
rownames(rollCall274) <- id_data
head(rollCall274, 5)
```

```
##         name_data state_data party_data vote_data1 numeric.vote
## A000374   Abraham         LA          R        Aye            1
## A000370     Adams         NC          D         No            0
## A000055  Aderholt         AL          R        Aye            1
## A000371   Aguilar         CA          D         No            0
## A000372     Allen         GA          R        Aye            1
```

Part 3 (Bootstrap, 4 + 2 pts)

We consider the `strikes` data which we used in Lecture 6. The data set is about strikes in 18 countries over 35 years (compiled by Bruce Western, in the Sociology Department at Harvard University). The measured variables are:

- **country, year** – country and year of data collection
- **strike.volume** – days on strike per 1000 workers
- **unemployment** – unemployment rate
- **inflation** – inflation rate
- **left.parliament** – leftwing share of the government
- **centralization** – centralization of unions
- **density** – density of unions

In this problem, we *only* look at the strikes in Italy. On this subset, we run a simple linear regression using `strike.volume` as the response (Y) and `left.parliament` as the predictor (X). Our model is

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

```r
strikes<-read.csv("strikes.csv", header = T)
italy.strikes<-strikes[strikes$country == "Italy", ]
dim(italy.strikes)
```

```
## [1] 35  8
```

```r
lm.fit<- lm(strike.volume ~ left.parliament, data = italy.strikes)
round(lm.fit$coefficients,3)
```

```
##     (Intercept) left.parliament
##        -738.745          40.291
```

Problem 3.1

Denote our estimate of β_1 as $\hat{\beta}_1$, which is 40.291 according to the analysis above. Use the Bootstrap method to estimate the variance of $\hat{\beta}_1$. Here, you may draw 100 Bootstrap samples. [4 pts]

```r
# B <- 100
# resampled_ests <- matrix(NA, nrow = B, ncol = 2)
# names(resampled_ests) <- c("Intercept_Est", "Slope_Est")
# for (b in 1:B) {
#   data <- strikes[resampled_values[b,],]
```

```
#   ests <- lm(data$strike.volume~data$left.parliament)$coefficients
#   resampled_ests[b, 1] <- ests[1]
#   resampled_ests[b, 2] <- ests[2]
# }
# head(resampled_ests)
# var(resampled_ests[,1])
# var(resampled_ests[,2])
```

Problem 3.2

Construct a 95% confidence interval of β_1 based on the result in part 3.1. [2 extra pts]

```
# code goes here
```