Final Project

Yogesh Tahiliand and Krutika Nayak

4/21/2023

Goals

The goal of this project is to allow you to practice your analysis, theoretical, and computational skills on a new question. We will be using data from the Reproducibility Project: Psychology, which was an attempt to replicate experiments in Psychology. You can work in groups of up to three people for this project.

1. Read about the project at https://osf.io/447b3/ (https://osf.io/447b3/) (the first 28 pages).
2. Pick a study to examine: https://osf.io/ezcuj/wiki/Replicated%20Studies/ (https://osf.io/ezcuj/wiki/Replicated%20Studies/)

- You can consider examining this file: https://osf.io/fgjvw/ (https://osf.io/fgjvw/), which has all of the datasets listed out with the titles and the descriptors to help sort through a topic you might find interesting. Also, not all datasets are good datasets!

3. Read the paper for the replication, specifically focusing on the "target of the study" listed on the replication page. If you have trouble finding the article full-text, please contact me. Try using Google Scholar on the original citation listed in the wiki.
4. Go to the project page and download the data from the project. Make sure it's raw data that you can import into R.
5. Generate a hypothesis about the data using one of the analyses we've discussed (correlation, regression, t-tests, ANOVA).
6. Use the data screening we've discussed on the data you are using for the hypothesis (i.e., not the whole dataset).
7. Run the analysis that will answer your hypothesis test.
8. Fill in the following report based on your results.

Citation of the Study

https://bpb-us-w2.wpmucdn.com/u.osu.edu/dist/2/43662/files/2017/02/2008-Stinson-Logel-Zanna-Holmes-Cameron-Wood-Spencer-1rvryk7.pdf (https://bpb-us-w2.wpmucdn.com/u.osu.edu/dist/2/43662/files/2017/02/2008-Stinson-Logel-Zanna-Holmes-Cameron-Wood-Spencer-1rvryk7.pdf)

Summary of the Study

“The Cost of Lower Self-Esteem: Testing a Self- and Social-Bonds Model of Health” is a research paper published in the journal Personality and Social Psychology Review. The paper explores the relationship between self-esteem, social connections, and physical health outcomes. The authors propose a “Self- and Social-Bonds Model of Health” which suggests that both self-esteem and social connections play important roles in maintaining good health.

The paper reviews a variety of studies that have examined the link between self-esteem and physical health outcomes such as cardiovascular disease, chronic pain, and immune system functioning. The authors argue that low self-esteem can contribute to negative health behaviors such as smoking and poor diet, which in turn can lead to poor health outcomes.

The authors also discuss the importance of social connections for maintaining good health. They argue that social support can help buffer the negative effects of stress on health, and that social isolation can contribute to poor health outcomes.

Overall, the study highlights the importance of self-esteem and social connections in promoting good health outcomes. The findings suggest that interventions aimed at improving self-esteem and building social connections could have important health benefits.

Your Hypothesis

Given the available data, what is your hypothesis you would like to test on the data?

Null Hypothesis: H0: The number of classes missed is not dependent on the self-esteem, stress levels and external events that trigger stress. Alternative Hypothesis: H1: The number of classes are affected by the level of self-esteem, stress levels and external events that trigger stress.

The Data

Downloading all the required packages

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
library(Hmisc)
```

```
## Loading required package: lattice
```

```
## Loading required package: survival
```

```
## Loading required package: Formula
```

```
## Loading required package: ggplot2
```

```
##
## Attaching package: 'Hmisc'
```

```
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
```

```
## The following objects are masked from 'package:base':
##
##     format.pval, units
```

```
library(performance)
library(insight)
```

Import the data from the project here. You should reduce your data down to only the columns necessary for your analysis.

```
data <- read_excel("/Users/yogtah/Desktop/Day-1 CPT/Spring 2023/Assignments/Principle
s of Analytics I/Group Project/Project/FInalData.xlsx")

#Reldata here stands for Related Data for our analysis
reldata <- data[ , -c(1:8,11:17,19,50:67)]
summary(reldata)
```

```
##     Finished           PID                Gender      First Year Student
##  Min.   :0.0000   Min.   :1.075e+09   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.0000   1st Qu.:1.247e+09   1st Qu.:1.000   1st Qu.:1.000
##  Median :1.0000   Median :1.262e+09   Median :2.000   Median :1.000
##  Mean   :0.9539   Mean   :1.246e+09   Mean   :1.689   Mean   :1.146
##  3rd Qu.:1.0000   3rd Qu.:1.268e+09   3rd Qu.:2.000   3rd Qu.:1.000
##  Max.   :1.0000   Max.   :1.270e+09   Max.   :2.000   Max.   :2.000
##                                       NA's   :34      NA's   :34
##     Resident     Person of Worth Good Qualities  Failure feeling
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:6.000   1st Qu.:6.000   1st Qu.:2.000
##  Median :1.000   Median :8.000   Median :8.000   Median :3.000
##  Mean   :1.115   Mean   :7.234   Mean   :7.252   Mean   :3.972
##  3rd Qu.:1.000   3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:6.000
##  Max.   :2.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##  NA's   :167     NA's   :272     NA's   :272     NA's   :272
##  Getting things done   Not proud     Positive Attitude    Content
##  Min.   :1.000       Min.   :1.000   Min.   :1.000     Min.   :1.000
##  1st Qu.:6.000       1st Qu.:1.000   1st Qu.:5.000     1st Qu.:5.000
##  Median :7.000       Median :4.000   Median :7.000     Median :7.000
##  Mean   :7.041       Mean   :4.094   Mean   :6.677     Mean   :6.531
##  3rd Qu.:9.000       3rd Qu.:6.000   3rd Qu.:9.000     3rd Qu.:8.000
##  Max.   :9.000       Max.   :9.000   Max.   :9.000     Max.   :9.000
##  NA's   :272         NA's   :272     NA's   :272       NA's   :272
##   Self-respect    Feel useless   No good at all  Visit health services
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 1.000
##  1st Qu.:3.000   1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 1.000
##  Median :5.000   Median :5.000   Median :4.000   Median : 1.000
##  Mean   :5.189   Mean   :4.995   Mean   :4.497   Mean   : 2.798
##  3rd Qu.:8.000   3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.: 3.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :11.000
##  NA's   :272     NA's   :272     NA's   :272     NA's   :272
##  Classes missed   Number of friends in college
##  Min.   : 1.000   Min.   : 1.000
##  1st Qu.: 1.000   1st Qu.: 4.000
##  Median : 2.000   Median : 6.000
##  Mean   : 3.362   Mean   : 6.614
##  3rd Qu.: 5.000   3rd Qu.:10.000
##  Max.   :11.000   Max.   :11.000
```

```
##  NA's   :272      NA's   :272
##  Number of friends outside college Amount of Stress with family
##  Min.   : 1.000                    Min.   :1.000
##  1st Qu.: 5.000                    1st Qu.:2.000
##  Median : 8.000                    Median :3.000
##  Mean   : 7.503                    Mean   :3.247
##  3rd Qu.:11.000                    3rd Qu.:4.000
##  Max.   :11.000                    Max.   :5.000
##  NA's   :272                       NA's   :272
##  Amount of Stress with close friend Amount of Stress with classmate
##  Min.   :1.000                       Min.   :1.0
##  1st Qu.:1.000                       1st Qu.:1.0
##  Median :2.500                       Median :2.0
##  Mean   :2.589                       Mean   :2.4
##  3rd Qu.:4.000                       3rd Qu.:3.0
##  Max.   :5.000                       Max.   :5.0
##  NA's   :272                         NA's   :272
##  Amount of Stress with romantic partner Amount of Stress with potential partner
##  Min.   :1.000                          Min.   :1.000
##  1st Qu.:1.000                          1st Qu.:1.000
##  Median :3.000                          Median :2.000
##  Mean   :2.661                          Mean   :2.495
##  3rd Qu.:4.000                          3rd Qu.:4.000
##  Max.   :5.000                          Max.   :5.000
##  NA's   :272                            NA's   :272
##   Job workload   Class Workload  Deal with stress with family
##  Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:3.000   1st Qu.:3.000
##  Median :3.000   Median :4.000   Median :4.000
##  Mean   :2.778   Mean   :3.466   Mean   :4.258
##  3rd Qu.:4.000   3rd Qu.:4.000   3rd Qu.:6.000
##  Max.   :5.000   Max.   :5.000   Max.   :6.000
##  NA's   :272     NA's   :272     NA's   :272
##  Deal with stress with close friend Deal with stress with classmate
##  Min.   :1.000                      Min.   :1.000
##  1st Qu.:4.000                      1st Qu.:4.000
##  Median :5.000                      Median :5.000
##  Mean   :4.642                      Mean   :4.645
##  3rd Qu.:6.000                      3rd Qu.:6.000
##  Max.   :6.000                      Max.   :6.000
##  NA's   :272                        NA's   :272
##  Deal with stress with romantic partner Deal with stress with potential partner
##  Min.   :1.000                          Min.   :1.000
##  1st Qu.:3.000                          1st Qu.:3.000
##  Median :5.000                          Median :5.000
##  Mean   :4.414                          Mean   :4.441
##  3rd Qu.:6.000                          3rd Qu.:6.000
```

```
##  Max.   :6.000                          Max.   :6.000
##  NA's   :272                            NA's   :272
##  Deal with job workload Deal with class Workload
##  Min.   :1.000          Min.   :1.000
##  1st Qu.:3.000          1st Qu.:3.000
##  Median :5.000          Median :4.000
##  Mean   :4.402          Mean   :4.208
##  3rd Qu.:6.000          3rd Qu.:5.000
##  Max.   :6.000          Max.   :6.000
##  NA's   :272            NA's   :272
```

```
##Filtering the data that we need for our regression analysis
reldata1 <- subset(reldata, reldata$Finished==1 & reldata$`First Year Student`==1 & r
eldata$Resident==1)
summary(reldata1)
```

```
##     Finished      PID                 Gender      First Year Student    Resident
##  Min.   :1   Min.   :1.079e+09   Min.   :1.000   Min.   :1          Min.   :1
##  1st Qu.:1   1st Qu.:1.247e+09   1st Qu.:1.000   1st Qu.:1          1st Qu.:1
##  Median :1   Median :1.263e+09   Median :2.000   Median :1          Median :1
##  Mean   :1   Mean   :1.247e+09   Mean   :1.696   Mean   :1          Mean   :1
##  3rd Qu.:1   3rd Qu.:1.268e+09   3rd Qu.:2.000   3rd Qu.:1          3rd Qu.:1
##  Max.   :1   Max.   :1.270e+09   Max.   :2.000   Max.   :1          Max.   :1
##  Person of Worth Good Qualities  Failure feeling Getting things done
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   :1.000
##  1st Qu.:6.000   1st Qu.:6.000   1st Qu.:2.000   1st Qu.:6.000
##  Median :8.000   Median :8.000   Median :3.000   Median :7.000
##  Mean   :7.259   Mean   :7.254   Mean   :3.958   Mean   :7.037
##  3rd Qu.:9.000   3rd Qu.:9.000   3rd Qu.:6.000   3rd Qu.:9.000
##  Max.   :9.000   Max.   :9.000   Max.   :9.000   Max.   :9.000
##    Not proud     Positive Attitude    Content       Self-respect
##  Min.   :1.000   Min.   :1.000     Min.   :1.000   Min.   :1.000
##  1st Qu.:1.000   1st Qu.:5.000     1st Qu.:5.000   1st Qu.:3.000
##  Median :4.000   Median :7.000     Median :7.000   Median :5.000
##  Mean   :4.071   Mean   :6.698     Mean   :6.547   Mean   :5.188
##  3rd Qu.:6.000   3rd Qu.:9.000     3rd Qu.:8.000   3rd Qu.:8.000
##  Max.   :9.000   Max.   :9.000     Max.   :9.000   Max.   :9.000
##   Feel useless   No good at all  Visit health services Classes missed
##  Min.   :1.000   Min.   :1.000   Min.   : 1.000        Min.   : 1.00
##  1st Qu.:2.000   1st Qu.:2.000   1st Qu.: 1.000        1st Qu.: 1.00
##  Median :5.000   Median :4.000   Median : 1.000        Median : 2.00
##  Mean   :4.963   Mean   :4.461   Mean   : 2.757        Mean   : 3.32
##  3rd Qu.:7.000   3rd Qu.:7.000   3rd Qu.: 3.000        3rd Qu.: 4.00
##  Max.   :9.000   Max.   :9.000   Max.   :11.000        Max.   :11.00
##  Number of friends in college Number of friends outside college
##  Min.   : 1.000               Min.   : 1.000
```

```
##  1st Qu.: 4.000              1st Qu.: 5.000
##  Median : 6.000              Median : 8.000
##  Mean   : 6.579              Mean   : 7.489
##  3rd Qu.:10.000              3rd Qu.:11.000
##  Max.   :11.000              Max.   :11.000
##  Amount of Stress with family Amount of Stress with close friend
##  Min.   :1.000               Min.   :1.000
##  1st Qu.:2.000               1st Qu.:1.000
##  Median :3.000               Median :2.000
##  Mean   :3.236               Mean   :2.577
##  3rd Qu.:4.000               3rd Qu.:3.750
##  Max.   :5.000               Max.   :5.000
##  Amount of Stress with classmate Amount of Stress with romantic partner
##  Min.   :1.000                   Min.   :1.00
##  1st Qu.:1.000                   1st Qu.:1.00
##  Median :2.000                   Median :3.00
##  Mean   :2.395                   Mean   :2.65
##  3rd Qu.:3.000                   3rd Qu.:4.00
##  Max.   :5.000                   Max.   :5.00
##  Amount of Stress with potential partner  Job workload   Class Workload
##  Min.   :1.000                           Min.   :1.000   Min.   :1.00
##  1st Qu.:1.000                           1st Qu.:1.000   1st Qu.:3.00
##  Median :2.000                           Median :3.000   Median :4.00
##  Mean   :2.482                           Mean   :2.775   Mean   :3.45
##  3rd Qu.:4.000                           3rd Qu.:4.000   3rd Qu.:4.00
##  Max.   :5.000                           Max.   :5.000   Max.   :5.00
##  Deal with stress with family Deal with stress with close friend
##  Min.   :1.000                Min.   :1.000
##  1st Qu.:3.000                1st Qu.:4.000
##  Median :4.000                Median :5.000
##  Mean   :4.251                Mean   :4.645
##  3rd Qu.:6.000                3rd Qu.:6.000
##  Max.   :6.000                Max.   :6.000
##  Deal with stress with classmate Deal with stress with romantic partner
##  Min.   :1.00                    Min.   :1.000
##  1st Qu.:4.00                    1st Qu.:3.000
##  Median :5.00                    Median :5.000
##  Mean   :4.64                    Mean   :4.418
##  3rd Qu.:6.00                    3rd Qu.:6.000
##  Max.   :6.00                    Max.   :6.000
##  Deal with stress with potential partner Deal with job workload
##  Min.   :1.000                           Min.   :1.000
##  1st Qu.:3.000                           1st Qu.:3.000
##  Median :5.000                           Median :5.000
##  Mean   :4.442                           Mean   :4.404
##  3rd Qu.:6.000                           3rd Qu.:6.000
##  Max.   :6.000                           Max.   :6.000
```

```
##  Deal with class Workload
##  Min.   :1.000
##  1st Qu.:3.000
##  Median :4.000
##  Mean   :4.215
##  3rd Qu.:5.000
##  Max.   :6.000
```

```
##By filtering the data all the missing data is automatically excluded and now we do not have to check for any missing data

## Adding new variables using rowmeans function and as there are more than 60 variables we tried to combine a couple of variables together in a new variable and taking the average of the ratings that was assigned to each of these variables -- A copy of our analysis will be attached in the form of excel file for future reference:

##HSE stands for High Self-Esteem
reldata1$HSE <- rowMeans(reldata1[, c("Person of Worth","Good Qualities","Getting things done", "Positive Attitude", "Content")])

##LSE stands for Low Self-Esteem
reldata1$LSE <- rowMeans(reldata1[, c("Failure feeling", "Not proud", "Self-respect", "Feel useless", "No good at all")])

##CST stands for Controllable Stress Triggers
reldata1$CST <- rowMeans(reldata1[, c("Amount of Stress with family", "Amount of Stress with close friend", "Amount of Stress with classmate", "Amount of Stress with romantic partner", "Amount of Stress with potential partner", "Job workload", "Class Workload")])

## SS Stands for Support System
reldata1$SS <- rowSums(reldata1[, c("Number of friends in college", "Number of friends outside college")])

## SR syands for Stress Response
reldata1$SR <- rowMeans(reldata1[, c("Deal with stress with family", "Deal with stress with close friend", "Deal with stress with classmate", "Deal with stress with romantic partner", "Deal with stress with potential partner", "Deal with job workload", "Deal with class Workload")])

## This is our dependent variable
reldata1$Impact <- reldata1$`Classes missed`

##Excluding the variables that were used above to find a new variable which is a mean of all of them:
regdata <- subset(reldata1[, c(1:5,34:39)])
summary(regdata)
```

```
##     Finished      PID                Gender      First Year Student    Resident
##  Min.   :1   Min.   :1.079e+09   Min.   :1.000   Min.   :1          Min.   :1
##  1st Qu.:1   1st Qu.:1.247e+09   1st Qu.:1.000   1st Qu.:1          1st Qu.:1
##  Median :1   Median :1.263e+09   Median :2.000   Median :1          Median :1
##  Mean   :1   Mean   :1.247e+09   Mean   :1.696   Mean   :1          Mean   :1
##  3rd Qu.:1   3rd Qu.:1.268e+09   3rd Qu.:2.000   3rd Qu.:1          3rd Qu.:1
##  Max.   :1   Max.   :1.270e+09   Max.   :2.000   Max.   :1          Max.   :1
##       HSE             LSE             CST             SS
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 2.00
##  1st Qu.:6.000   1st Qu.:2.600   1st Qu.:2.143   1st Qu.: 9.00
##  Median :7.400   Median :4.400   Median :2.714   Median :14.00
##  Mean   :6.959   Mean   :4.528   Mean   :2.795   Mean   :14.07
##  3rd Qu.:8.200   3rd Qu.:6.400   3rd Qu.:3.429   3rd Qu.:19.00
##  Max.   :9.000   Max.   :9.000   Max.   :5.000   Max.   :22.00
##        SR            Impact
##  Min.   :1.000   Min.   : 1.00
##  1st Qu.:3.714   1st Qu.: 1.00
##  Median :4.429   Median : 2.00
##  Mean   :4.431   Mean   : 3.32
##  3rd Qu.:5.286   3rd Qu.: 4.00
##  Max.   :6.000   Max.   :11.00
```

Data Screening

Do a complete data screening on the data provided from the project. You can add R chunks here to help separate out the different sections. You should comment in each section if the assumption has been met, what you did to fix errors, etc.

```
##DEALING WITH THE OUTLIERS

mahal <- mahalanobis(regdata[ , 6:11],colMeans(regdata[ , 6:11], na.rm = T),cov(regda
ta[ , 6:11], use = "pairwise.complete.obs"))
cutoff <- qchisq(p = 1 - .001, df = ncol(regdata[ , 6:11]))
badmahal <- mahal > cutoff
table(badmahal)
```

```
## badmahal
## FALSE  TRUE
##   619     3
```

```
##Running an outlier's test using Leverage and Cooks Distance:

model <- lm(Impact ~ HSE+LSE+CST+SS+SR, data = regdata)
summary(model)
```

```
##
## Call:
## lm(formula = Impact ~ HSE + LSE + CST + SS + SR, data = regdata)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4368 -1.5520 -0.1671  1.2527  8.5634
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.33882    0.66117  -6.562 1.12e-10 ***
## HSE          0.07907    0.07374   1.072   0.2840
## LSE          0.44032    0.05441   8.093 3.12e-15 ***
## CST          1.26102    0.12414  10.158  < 2e-16 ***
## SS           0.03601    0.01692   2.128   0.0337 *
## SR           0.24452    0.10118   2.417   0.0160 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.342 on 616 degrees of freedom
## Multiple R-squared:  0.3637, Adjusted R-squared:  0.3585
## F-statistic: 70.42 on 5 and 616 DF,  p-value: < 2.2e-16
```

```
##Leverage Outliers Test
k <- length(coef(model))-1
leverage <- hatvalues(model)
cutleverage <- (2*k+2/nrow(regdata))
badleverage <- leverage > cutleverage
table(badleverage)
```

```
## badleverage
## FALSE
##   622
```

```
## Cooks Distance Outliers Test
cooks <- cooks.distance(model)
cutcooks <- 4/ (nrow(regdata)-k-1)-k-1
badcooks <- cooks > cutcooks
table(badcooks)
```

```
## badcooks
## TRUE
##  622
```

```
totalout <- badcooks + badleverage + badmahal
table(totalout)
```

```
## totalout
##   1   2
## 619   3
```

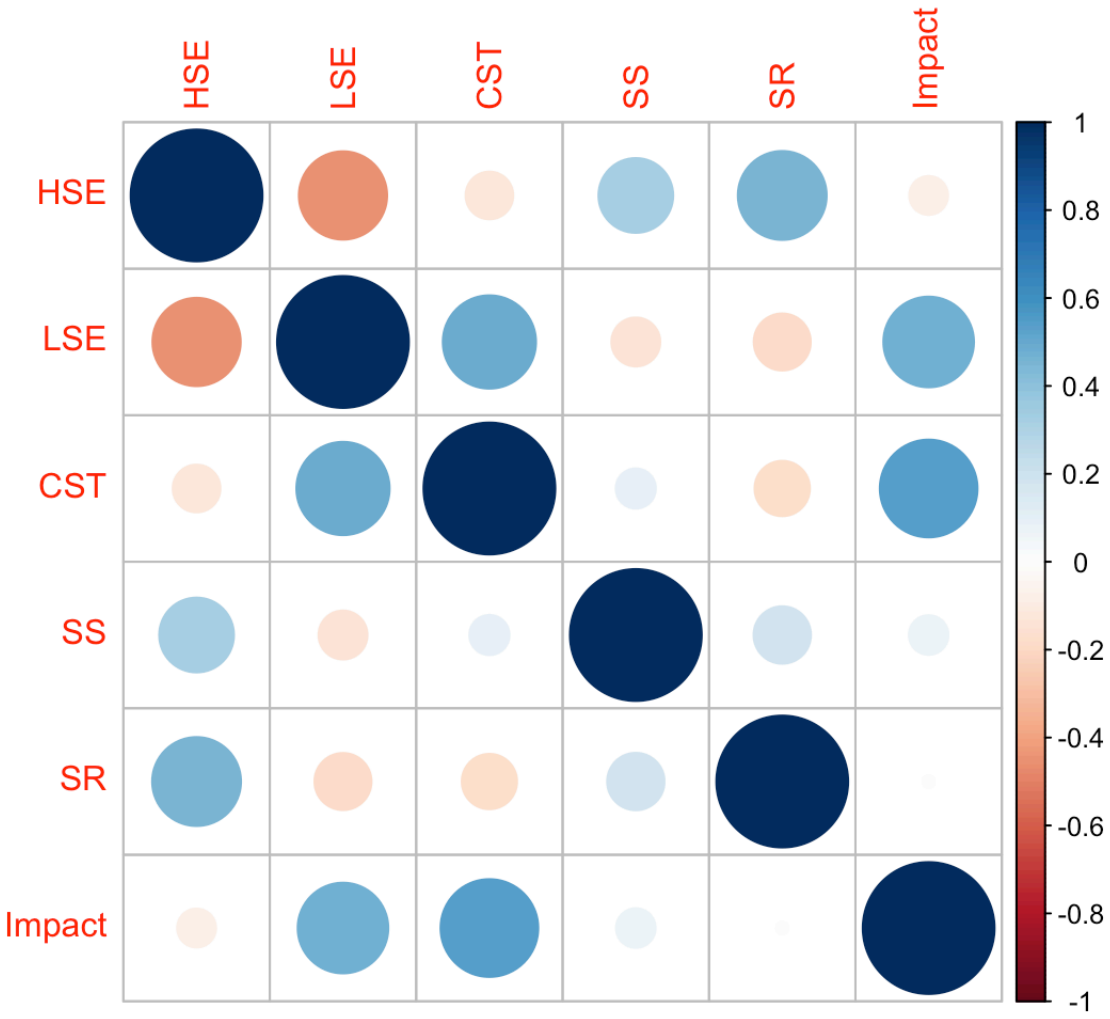
```
##Excluding the data points that have been identified as outliers by the above tests:
nooutliers <- subset(regdata, totalout < 2)
summary(nooutliers)
```

```
##     Finished      PID                Gender      First Year Student    Resident
##  Min.   :1   Min.   :1.079e+09   Min.   :1.000   Min.   :1          Min.   :1
##  1st Qu.:1   1st Qu.:1.248e+09   1st Qu.:1.000   1st Qu.:1          1st Qu.:1
##  Median :1   Median :1.263e+09   Median :2.000   Median :1          Median :1
##  Mean   :1   Mean   :1.247e+09   Mean   :1.698   Mean   :1          Mean   :1
##  3rd Qu.:1   3rd Qu.:1.268e+09   3rd Qu.:2.000   3rd Qu.:1          3rd Qu.:1
##  Max.   :1   Max.   :1.270e+09   Max.   :2.000   Max.   :1          Max.   :1
##       HSE             LSE             CST             SS
##  Min.   :1.000   Min.   :1.000   Min.   :1.000   Min.   : 2.00
##  1st Qu.:6.000   1st Qu.:2.600   1st Qu.:2.143   1st Qu.: 9.00
##  Median :7.400   Median :4.400   Median :2.714   Median :14.00
##  Mean   :6.966   Mean   :4.528   Mean   :2.796   Mean   :14.08
##  3rd Qu.:8.200   3rd Qu.:6.400   3rd Qu.:3.429   3rd Qu.:19.00
##  Max.   :9.000   Max.   :9.000   Max.   :5.000   Max.   :22.00
##        SR            Impact
##  Min.   :1.000   Min.   : 1.000
##  1st Qu.:3.714   1st Qu.: 1.000
##  Median :4.429   Median : 2.000
##  Mean   :4.427   Mean   : 3.309
##  3rd Qu.:5.286   3rd Qu.: 4.000
##  Max.   :6.000   Max.   :11.000
```

```
dim(nooutliers)
```

```
## [1] 619  11
```

```
##Understanding the correlation between the variables on which we will be running reg
ression analysis:
corrplot(cor(nooutliers[, c(6:11)]))
```



##Hierarchical Regression:

```
m1 <- lm(nooutliers$Impact ~ nooutliers$HSE)
summary(m1)
```

```
##
## Call:
## lm(formula = nooutliers$Impact ~ nooutliers$HSE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.2375 -2.1476 -1.1164  0.9459  8.0081
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.39323    0.52164   8.422  2.6e-16 ***
## nooutliers$HSE -0.15571    0.07298  -2.134   0.0333 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.91 on 617 degrees of freedom
## Multiple R-squared:  0.007324,   Adjusted R-squared:  0.005715
## F-statistic: 4.552 on 1 and 617 DF,  p-value: 0.03327
```

```
## the p-value is not significant

m2 <- lm(Impact ~ LSE, data = nooutliers)
summary(m2)
```

```
##
## Call:
## lm(formula = Impact ~ LSE, data = nooutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1229 -1.7944 -0.3401  1.4047  8.6530
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.45901    0.23799   1.929   0.0542 .
## LSE          0.62932    0.04733  13.296   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.575 on 617 degrees of freedom
## Multiple R-squared:  0.2227, Adjusted R-squared:  0.2214
## F-statistic: 176.8 on 1 and 617 DF,  p-value: < 2.2e-16
```

```
## the p-value is significant and the Multiple R-squared:  0.2227

m3 <- lm(Impact ~ CST, data = nooutliers)
summary(m3)
```

```
##
## Call:
## lm(formula = Impact ~ CST, data = nooutliers)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.287 -1.645 -0.356  1.128  9.128
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.7384     0.3249  -5.351 1.23e-07 ***
## CST           1.8051     0.1108  16.297  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.442 on 617 degrees of freedom
## Multiple R-squared:  0.3009, Adjusted R-squared:  0.2998
## F-statistic: 265.6 on 1 and 617 DF,  p-value: < 2.2e-16
```

```
## the p-value is significant and the Multiple R-squared:  0.3009

m4 <- lm(nooutliers$Impact ~ nooutliers$SS)
summary(m4)
```

```
##
## Call:
## lm(formula = nooutliers$Impact ~ nooutliers$SS)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.653 -2.218 -1.131  1.130  8.130
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    2.69628    0.30090   8.961   <2e-16 ***
## nooutliers$SS  0.04349    0.01969   2.208   0.0276 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.909 on 617 degrees of freedom
## Multiple R-squared:  0.007842,   Adjusted R-squared:  0.006234
## F-statistic: 4.877 on 1 and 617 DF,  p-value: 0.02759
```

```
## the p-value is significant but the Multiple R-squared:  0.007842

m5 <- lm(nooutliers$Impact ~ nooutliers$SR)
summary(m5)
```

```
##
## Call:
## lm(formula = nooutliers$Impact ~ nooutliers$SR)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.3808 -2.2935 -1.2995  0.7141  7.7246
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.40189    0.50591   6.724 4.03e-11 ***
## nooutliers$SR -0.02108    0.11116  -0.190     0.85
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.92 on 617 degrees of freedom
## Multiple R-squared:  5.828e-05,  Adjusted R-squared:  -0.001562
## F-statistic: 0.03596 on 1 and 617 DF,  p-value: 0.8497
```

```
## the p-value is NOT significant but the Multiple R-squared:  5.828e-05

m6 <- lm(Impact ~ LSE + CST, data = nooutliers)
summary(m6)
```

```
##
## Call:
## lm(formula = Impact ~ LSE + CST, data = nooutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8496 -1.5609 -0.1467  1.2981  8.8283
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.11937    0.31738  -6.678 5.42e-11 ***
## LSE          0.35174    0.04988   7.052 4.74e-12 ***
## CST          1.37170    0.12307  11.146  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.351 on 616 degrees of freedom
## Multiple R-squared:  0.3532, Adjusted R-squared:  0.3511
## F-statistic: 168.2 on 2 and 616 DF,  p-value: < 2.2e-16
```

```
##stats
## Multiple R-squared:  0.3532, Adjusted R-squared:  0.3511; F-statistic: 168.2 on 2
and 616 DF, p-value: < 2.2e-16

m7 <- lm(Impact ~ LSE + CST + HSE, data = nooutliers)
summary(m7)
```

```
##
## Call:
## lm(formula = Impact ~ LSE + CST + HSE, data = nooutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.9067 -1.6179 -0.2107  1.2808  8.7551
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.72304    0.61637  -6.040 2.66e-09 ***
## LSE          0.42644    0.05535   7.704 5.29e-14 ***
## CST          1.32651    0.12317  10.770  < 2e-16 ***
## HSE          0.19980    0.06598   3.028  0.00257 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.335 on 615 degrees of freedom
## Multiple R-squared:  0.3627, Adjusted R-squared:  0.3595
## F-statistic: 116.6 on 3 and 615 DF,  p-value: < 2.2e-16
```

```
##Stats
## Multiple R-squared:  0.3627, Adjusted R-squared:  0.3595; F-statistic: 116.6 on 3 and 615 DF, p-value: < 2.2e-16

m8 <- lm(Impact ~ LSE + CST + HSE + SS, data = nooutliers)
summary(m8)
```

```
##
## Call:
## lm(formula = Impact ~ LSE + CST + HSE + SS, data = nooutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.0898 -1.5954 -0.1351  1.2035  8.7164
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.85911    0.61861  -6.238 8.24e-10 ***
## LSE          0.43455    0.05536   7.849 1.87e-14 ***
## CST          1.28683    0.12446  10.340  < 2e-16 ***
## HSE          0.16173    0.06851   2.360   0.0186 *
## SS           0.03377    0.01687   2.002   0.0457 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.33 on 614 degrees of freedom
## Multiple R-squared:  0.3668, Adjusted R-squared:  0.3627
## F-statistic: 88.92 on 4 and 614 DF,  p-value: < 2.2e-16
```

```
##Stats
##Multiple R-squared:  0.3668, Adjusted R-squared:  0.3627; F-statistic: 88.92 on 4 a
nd 614 DF, p-value: < 2.2e-16

m9 <- lm(Impact ~ LSE + CST + HSE + SS + SR, data = nooutliers)
summary(m9)
```

```
##
## Call:
## lm(formula = Impact ~ LSE + CST + HSE + SS + SR, data = nooutliers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5289 -1.5613 -0.1568  1.2375  8.6163
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.46489    0.66458  -6.718 4.20e-11 ***
## LSE          0.42017    0.05546   7.576 1.32e-13 ***
## CST          1.34023    0.12589  10.646  < 2e-16 ***
## HSE          0.08733    0.07478   1.168   0.2433
## SS           0.03054    0.01685   1.812   0.0705 .
## SR           0.24517    0.10080   2.432   0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.32 on 613 degrees of freedom
## Multiple R-squared:  0.3728, Adjusted R-squared:  0.3677
## F-statistic: 72.89 on 5 and 613 DF,  p-value: < 2.2e-16
```

```
##Stats
##Multiple R-squared:  0.3728, Adjusted R-squared:  0.3677; F-statistic: 72.89 on 5 a
nd 613 DF,  p-value: < 2.2e-16

##Using Anova to compare the models whose p-value was significant
anova(m2, m3,m6,m7,m8,m9)
```

```
## Analysis of Variance Table
##
## Model 1: Impact ~ LSE
## Model 2: Impact ~ CST
## Model 3: Impact ~ LSE + CST
## Model 4: Impact ~ LSE + CST + HSE
## Model 5: Impact ~ LSE + CST + HSE + SS
## Model 6: Impact ~ LSE + CST + HSE + SS + SR
##   Res.Df    RSS Df Sum of Sq       F    Pr(>F)
## 1    617 4090.2
## 2    617 3678.6  0    411.61
## 3    616 3403.7  1    274.82 51.0477 2.568e-12 ***
## 4    615 3353.7  1     50.00  9.2872  0.002407 **
## 5    614 3332.0  1     21.76  4.0412  0.044839 *
## 6    613 3300.1  1     31.85  5.9156  0.015293 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Summary of the Anova Results:
## Model 6(m9): Impact ~ LSE + CST + HSE + SS + SR
## As p-value IS SIGNIFICANT, RSS(Residual sum of Squares) is the lowest which in ret
urn helps fit the model better.

##We additionally thought of using one more method in addition to Anova to compare th
e models

## Using package – performance to compare the models
result <- compare_performance(m2, m3, m6, m7, m8, m9)
print_md(result)
```

Comparison of Model Performance Indices

Name	Model	AIC (weights)	AICc (weights)	BIC (weights)	R2	R2 (adj.)	RMSE	Sigma
m2	lm	2931.5 (<.001)	2931.5 (<.001)	2944.7 (<.001)	0.22	0.22	2.57	2.57
m3	lm	2865.8 (<.001)	2865.8 (<.001)	2879.1 (<.001)	0.30	0.30	2.44	2.44
m6	lm	2819.7 (1.00e-03)	2819.8 (1.00e-03)	2837.5 (0.14)	0.35	0.35	2.34	2.35
m7	lm	2812.6 (0.04)	2812.7 (0.04)	2834.7 (0.56)	0.36	0.36	2.33	2.34
m8	lm	2810.6 (0.12)	2810.7 (0.12)	2837.1 (0.17)	0.37	0.36	2.32	2.33
m9	lm	2806.6 (0.84)	2806.8 (0.84)	2837.6 (0.13)	0.37	0.37	2.31	2.32

```
##m9 is MOST SIGNIFICANT with higher R-squared, low AIC and BIC values
```

Assumptions check for our regression model

```
standardized <- rstudent(m9)
fitvalues <- scale(m9$fitted.values)

## Additivity:
## The test is met as RS < .9

cor(nooutliers[ , 6:11])
```

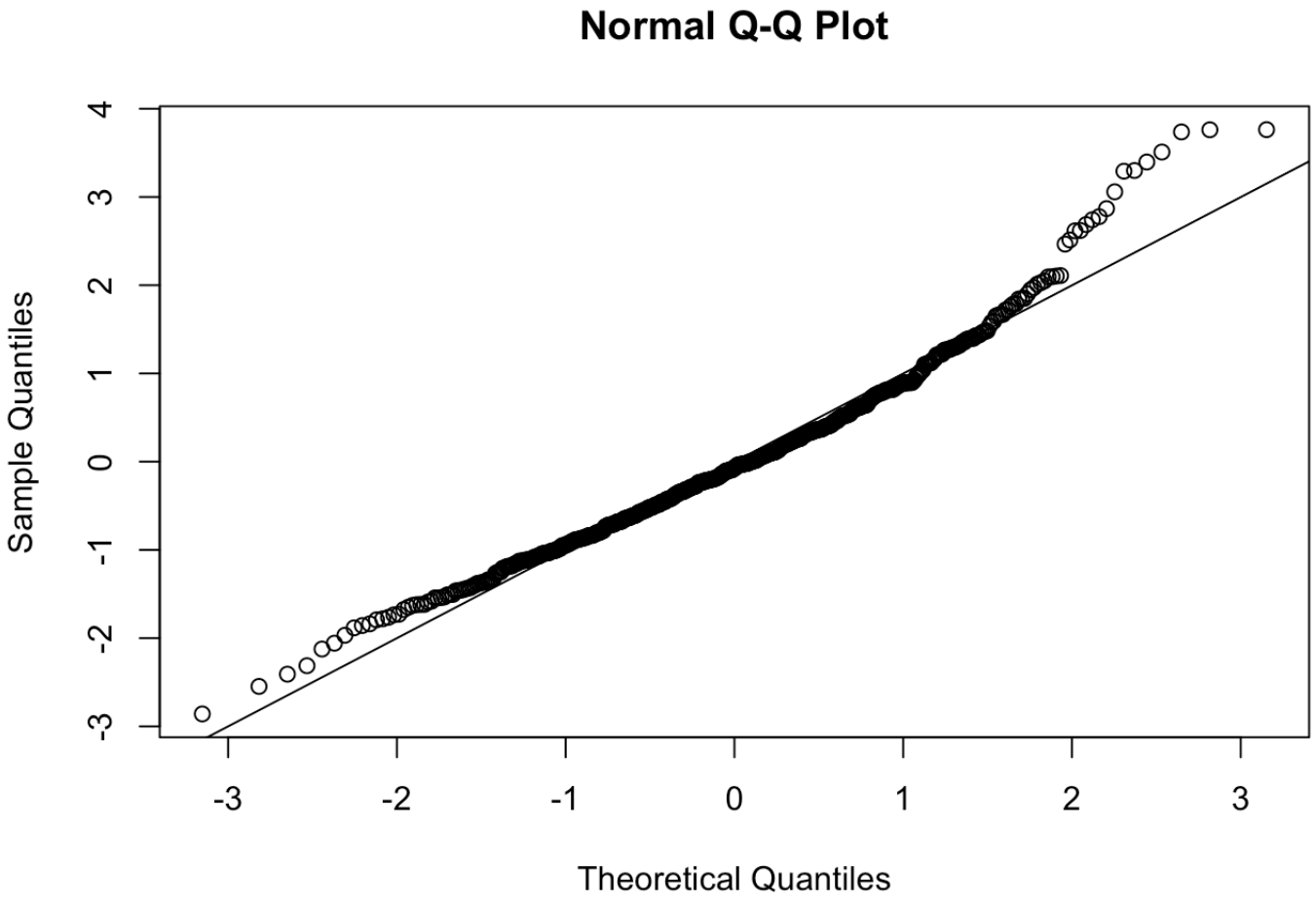
```
##                HSE        LSE        CST          SS           SR       Impact
## HSE     1.00000000 -0.4476151 -0.1296335  0.32109741  0.453682601 -0.085581939
## LSE    -0.44761513  1.0000000  0.4993287 -0.13703282 -0.186131059  0.471919421
## CST    -0.12963353  0.4993287  1.0000000  0.09177540 -0.175030917  0.548570680
## SS      0.32109741 -0.1370328  0.0917754  1.00000000  0.188692729  0.088555346
## SR      0.45368260 -0.1861311 -0.1750309  0.18869273  1.000000000 -0.007634376
## Impact -0.08558194  0.4719194  0.5485707  0.08855535 -0.007634376  1.000000000
```

```
symnum(cor(nooutliers[ , 6:11]))
```

```
##        H L C SS SR I
## HSE    1
## LSE    . 1
## CST      . 1
## SS     .     1
## SR     .        1
## Impact   . .       1
## attr(,"legend")
## [1] 0 ' ' 0.3 '.' 0.6 ',' 0.8 '+' 0.9 '*' 0.95 'B' 1
```

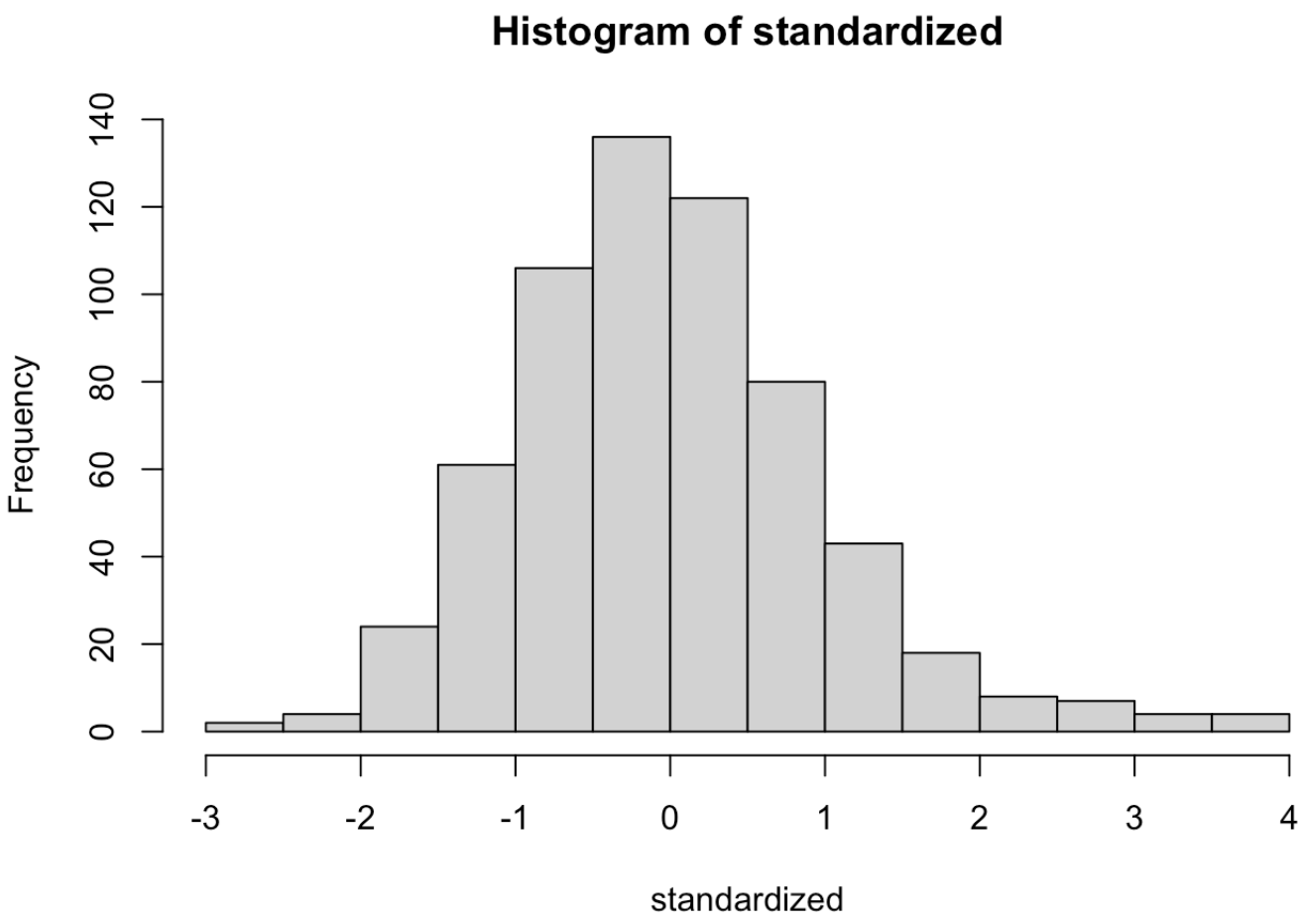
```
## Linearity:
## The test is met because most of the data points/dots are on the line.

{
     qqnorm(standardized)
     abline(0,1)
 }
```



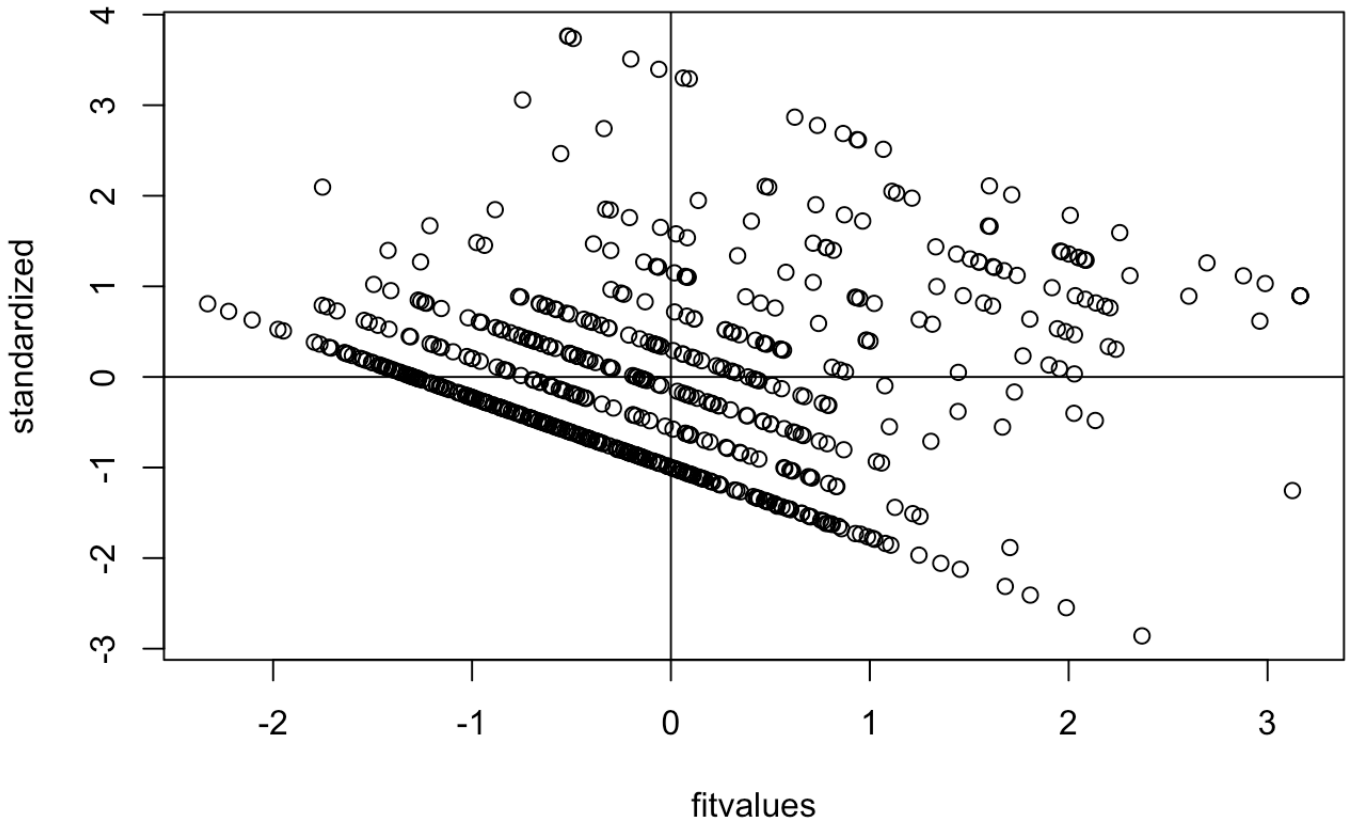

```
## Normality: The test is met as the residuals are centered around ZERO

hist(standardized)
```




```
##Homogeneity/Homoscedasticity:
##Yes, the test is met as the dots are evenly spread around the X-axis and Y-axis

{
  plot(fitvalues, standardized)
  abline(v=0)
  abline(h=0)
}
```




```
##Note:
## Though the graph looks weird and shows a pattern -  the homoscedasticity test is m
et as majority of the dots are bunching around the center and we do not see leaning t
hem either in the lower or higher residual bands and except some outliers most of the
dots are within the range 3 and -3.
```

The Analysis

Include your analysis for your hypothesis.

```
## Regression Analysis can be found above - wherein we ran Multiple Linear and Multiv
ariate Linear Regression Models (also known as Hierarchical Regression) and compared
the outcomes using Anova and Performance Package and picked up the model with the mos
t significant p-value and highest Adjusted R-squared
```

Summarize

In this section, you should write a summary of the results. Include your hypothesis, any issues you found in data screening, the type of test performed to answer your hypothesis, and the results of your test. Did it support your hypothesis?

Null Hypothesis: H0: The number of classes missed is not dependent on the self-esteem, stress levels and external events that trigger stress. Alternative Hypothesis: H1: The number of classes are affected by the level of self-esteem, stress levels and external events that trigger stress.

We used a regression analysis to understand the relationship between the variables and how these explanatory variables (HSE, LSE, CST, SS) impact our response variable (Number of Classes Missed). We started with understanding more about the questions(variables) that were asked to the population and how these questions were rated on different scales such as 1-9 or 1-5 and then understanding how the questions are impacting our response variable and then grouping them together so that we are not missing out any of the important variables that have a certain kind of relationship with our response variable be it +ve or -ve.

Results of our hypothesis: Our model incorporated how LSE + CST + HSE + SS + SR are influencing Impact and we achieved a significant results that can reject the NULL HYPOTHESIS: Multiple R-squared: 0.3728; Adjusted R-squared: 0.3677 –> 36.77% of the variance can be predicted by this model and this was the highest as compared to the other regression models that we ran; F-statistic: 72.89 on 5 and 613 DF; p-value: < 2.2e-16 –> which is significant.

When comparing all the models using Anova and Performance Package: Our model had the lowest Residuals for Sum of Squares(RSS) of 3300.1 with F-statistic of 5.9156 and p-value of 0.015293. Though these statistics were lower than the other models we tested - specifically Model 3: Impact ~ LSE + CST With RSS of 3403.7; F-statistic of 51.0477 and p-value of 0.000000000002568 *** –> we went ahead with our model as the RSS was the lowest of the all which means that the amount of variation in the dependent variable that was not explained by the regression model was the least in our model.

To ensure that we selected the right model, we ran a performance check as well using Performance package on all our models: And our model had the lowest AIC and BIC weights - 2806.6 (0.84) and 2837.6 (0.13) respectively and the R-squared be it Adjusted or Multiple was the highest across staying consistent at 37%. This furthermore strengthened our model selection.

To summarize our findings: As we were able to reject the null hypothesis that “classes missed is not affected by levels of low-esteem, social circle, and events that trigger stress,” it suggests that at least one of these factors is statistically significant in predicting the number of classes missed. In other words, there is evidence to suggest that low self-esteem, a small social circle, and stressful events could be related to higher levels of classes missed. We can also deduce that the hypothesis that these variables do not affect the number of classes missed is not supported by the data.

However, the regression analysis that we ran suggests that low self-esteem, limited social connections, and high levels of stress may have a significant impact on a student’s attendance in classes. This finding could have implications for educational interventions that aim to improve student attendance and academic performance by addressing these underlying factors.

Certain Shortcomings of the research: We did exclude couple of questions as highlighted in the excel file as these were just categorical variables answered in the format of Yes or No. It would have been great to account for the impact of these variables as well and it would have been better if the questions asked would have been framed in a different way like, In the past 2 weeks, if the candidate in the population used for the research paper broke up with the partner - how this impacted their mental well being on the scale of 1-9 and how this impacted their health and/or the academia in terms of classes missed, grade, etc. This attributes of the data would have aided to us to include all the questions/variables and analyse even further on how it is impacting our DV.