Jaeson Valles

Introduction and Overview

For this report, I analyzed 2010-2019 school census data collected by Census at School. The data was collected throughout schools in California and included students from grades 4-12. This data can be used to see what factors can affect children's growth as well as performance. The data consists of 500 entries of a variety of information including hours slept on a school night and hours spent doing homework.

Research Questions

This report will examine the following three questions:

I. Do students consider internet access more important than reducing pollution?
II. Are students who receive more hours of sleep taller than those who receive less hours of sleep
III. Which activities have the most impact on hours spent doing homework

All hypothesis tests will be conducted at a 5% level of significance.

I. Importance of Reducing Pollution vs Importance of Internet Access

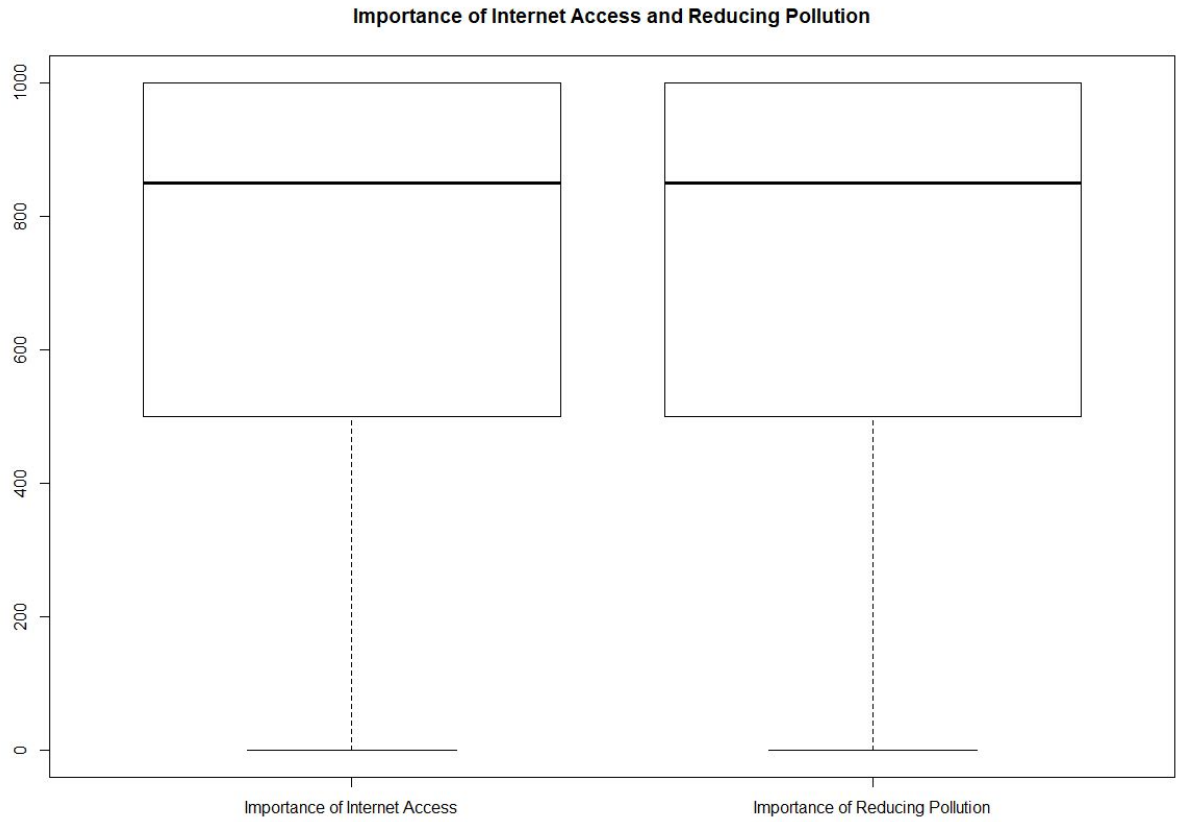
I tested if the means of the scores for the importance of reducing pollution and the importance of internet access were equal for all students in California using a t-test at a 5% level of significance. In order to conduct the t-test, I calculated the sample means for both scores.

$$H_0: \mu_1 = \mu_2$$
$$H_a: \mu_1 \neq \mu_2$$

The t-test indicated that there was sufficient evidence that the two means are equal. Therefore, students in California score reducing pollution as important as internet access. This conclusion is consistent with the confidence interval which indicates that at a 5% level of significance, there is no difference between the means.

[-49.07,21.09]

This conclusion can also be seen in the boxplot in which both plots appear identical



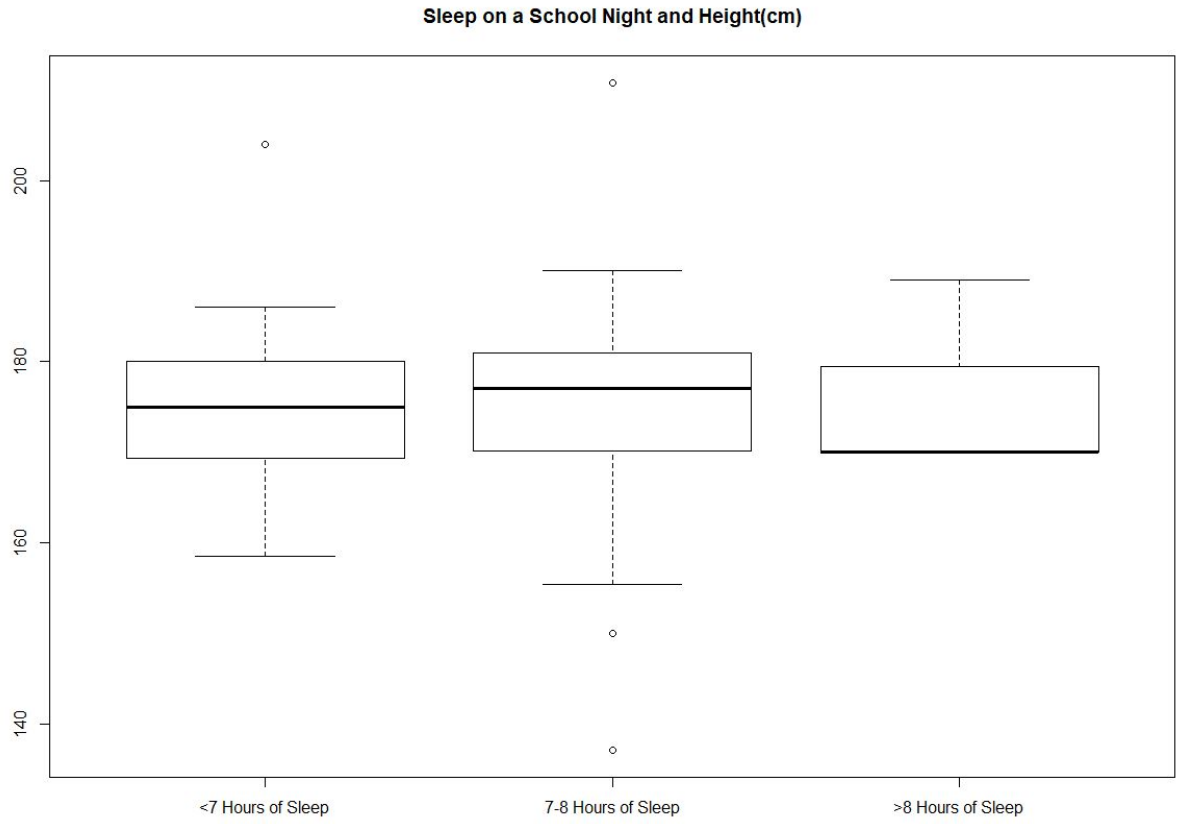
II. Effects of Hours Slept on a School Night and Height

I tested whether the means of three different groups of male students age 17 and up with different hours of sleep on a school night were all equivalent. The following 3 ranges of sleep were tested: Less than 7 hours of sleep, 7-8 hours of sleep, More than 8 hours of sleep.

$$H_0: \mu_1 = \mu_2 = \mu_3$$

$$H_a: \text{At least 1 mean is different}$$

The F-test indicated at a 5% level of significance that there was insufficient evidence to conclude that any of the groups were taller than the others on average. This is consistent with the boxplot which shoes the means and range do not differ by much.



Though the results conclude there is no significant difference between the group with less than 7 hours and 7-8 hours of sleep, the group with over 8 hours of sleep only had 3 data points so this test may not be enough to conclude that there was no difference in heights for that group.

III. Factors That Affect Hours Spent Doing Homework

I constructed a model for predicting how many hours were spent on doing homework based on the hours spent hanging out with friends, talking on the phone, doing things with family, doing outdoor activities, playing video games, browsing social media, text messaging, and using the computer. The model indicates that students spend at least 1 hours and 27 minutes on homework at a 0.1% level of significance assuming no time was spent doing any of the other activities used for the model. The factors that had an impact, as well as their level of significance, are doing things with family(1%), playing video games(5%), browsing social media(5%) and computer use(0.1%). The factors with positive impacts were doing things with family, browsing social media and computer use with computer use having the greatest impact. Playing video games had a negative impact on hours spent doing homework.

Appendix

Code:

```
library(dplyr)
#reading file
df <- read.csv(file="project.csv")
#remove duplicate entries
df <- df %>% distinct()

pollution_internet <-
df[c("Importance_reducing_pollution","Importance_Internet_access
")]
pollution_internet <-
pollution_internet[!(pollution_internet$Importance_Internet_acce
ss %in% "" | pollution_internet$Importance_Internet_access %in%
NA | pollution_internet$Importance_reducing_pollution %in% "" |
pollution_internet$Importance_reducing_pollution %in% NA),]

#Research Question 1
#finding the difference between sample means
boxplot(pollution_internet$Importance_Internet_access,pollution_
internet$Importance_Internet_access,names=c("Importance of
Internet Access", "Importance of Reducing
Pollution"),main="Importance of Internet Access and Reducing
Pollution")
x_bar = mean(pollution_internet$Importance_reducing_pollution)
y_bar = mean(pollution_internet$Importance_Internet_access)
d = x_bar - y_bar
#Perform a t test to determine if the two population means are
equal
t.test(pollution_internet$Importance_reducing_pollution,pollutio
n_internet$Importance_Internet_access,conf=0.95)

#Research Question 2
#Create a subset of the data and remove outlier values that are
likely not due to sleep
#Restrict dataset to males ages 17 and up to restrict variance
due to puberty and sex
```

```
height_sleep <- df[!(df$Gender %in% "" | df$Gender %in% NA |
df$Gender %in% "Female" | df$Height_cm %in% "" | df$Height_cm
%in% NA | df$Height_cm < 137 | df$Sleep_Hours_Schoolnight %in%
"" | df$Sleep_Hours_Schoolnight %in% NA | df$Ageyears < 17),]
height_sleep <-
height_sleep[c("Height_cm","Sleep_Hours_Schoolnight")]
#Seperate into groups based on the recommended number of hours
of sleep
group_a <- height_sleep[!(height_sleep$Sleep_Hours_Schoolnight
>= 7),]
group_b <- height_sleep[!(height_sleep$Sleep_Hours_Schoolnight <
7 | height_sleep$Sleep_Hours_Schoolnight > 8),]
group_c <- height_sleep[!(height_sleep$Sleep_Hours_Schoolnight
<= 8),]
group_a$group="a"
group_b$group="b"
group_c$group="c"
height_group <- rbind(group_a,group_b,group_c)
height_group <- height_group[c("Height_cm", "group")]
#perform anova test and display summary
res.aov <- aov(Height_cm ~ group,data=height_group)
summary(res.aov)
boxplot(group_a$Height_cm,group_b$Height_cm,group_c$Height_cm,na
mes=c("<7 Hours of Sleep","7-8 Hours of Sleep",">8 Hours of
Sleep"),main="Sleep on a School Night and Height(cm)")

#Research Question 3
#Building a model for hours spent doing homework
hours <-
df[c("Hanging_Out_With_Friends_Hours","Talking_On_Phone_Hours","
Doing_Homework_Hours","Doing_Things_With_Family_Hours","Outdoor_
Activities_Hours","Video_Games_Hours","Social_Websites_Hours","T
exting_Messaging_Hours","Computer_Use_Hours")]
#Remove unusuable or erroneous entries
hours <- na.omit(hours)
hours$Row_Totals <- rowSums(hours)
hours <- hours[!(hours$Row_Totals > 168),]
#Build Linear Model
```

```
model <- lm(sqrt(Doing_Homework_Hours) ~
sqrt(Hanging_Out_With_Friends_Hours) +
sqrt(Talking_On_Phone_Hours) +
sqrt(Doing_Things_With_Family_Hours) +
sqrt(Outdoor_Activities_Hours) + sqrt(Video_Games_Hours) +
sqrt(Social_Websites_Hours) + sqrt(Texting_Messaging_Hours) +
sqrt(Computer_Use_Hours),data=hours)
```

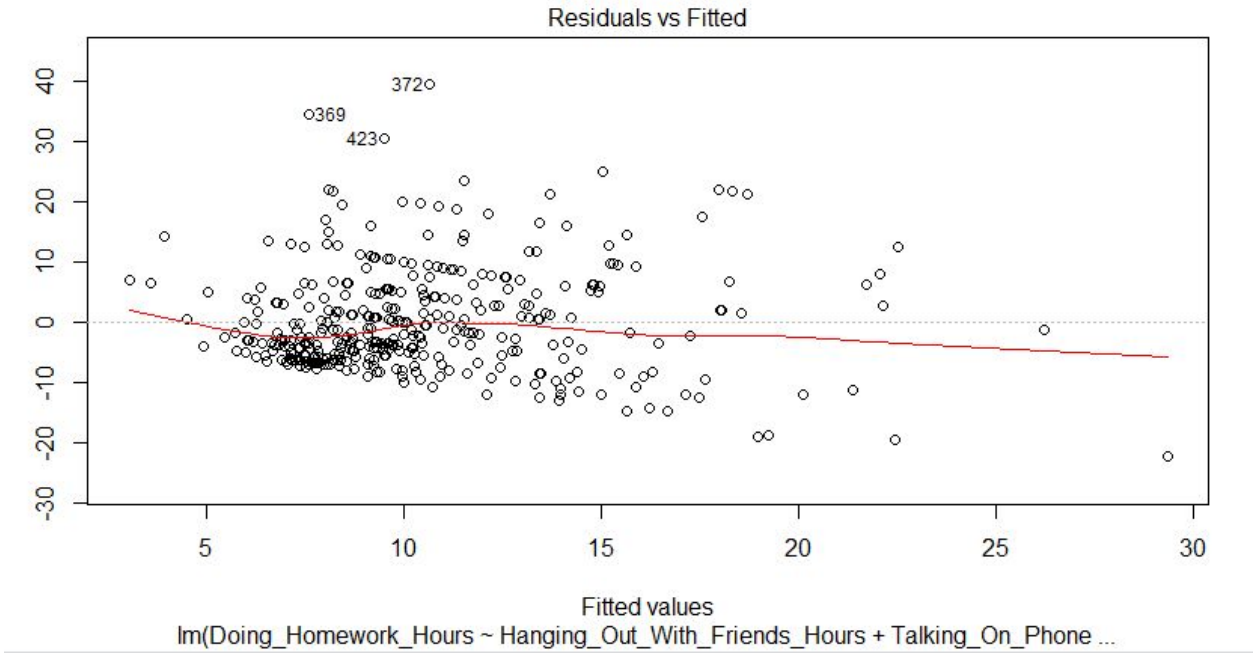
T-test Summary and Confidence Interval:

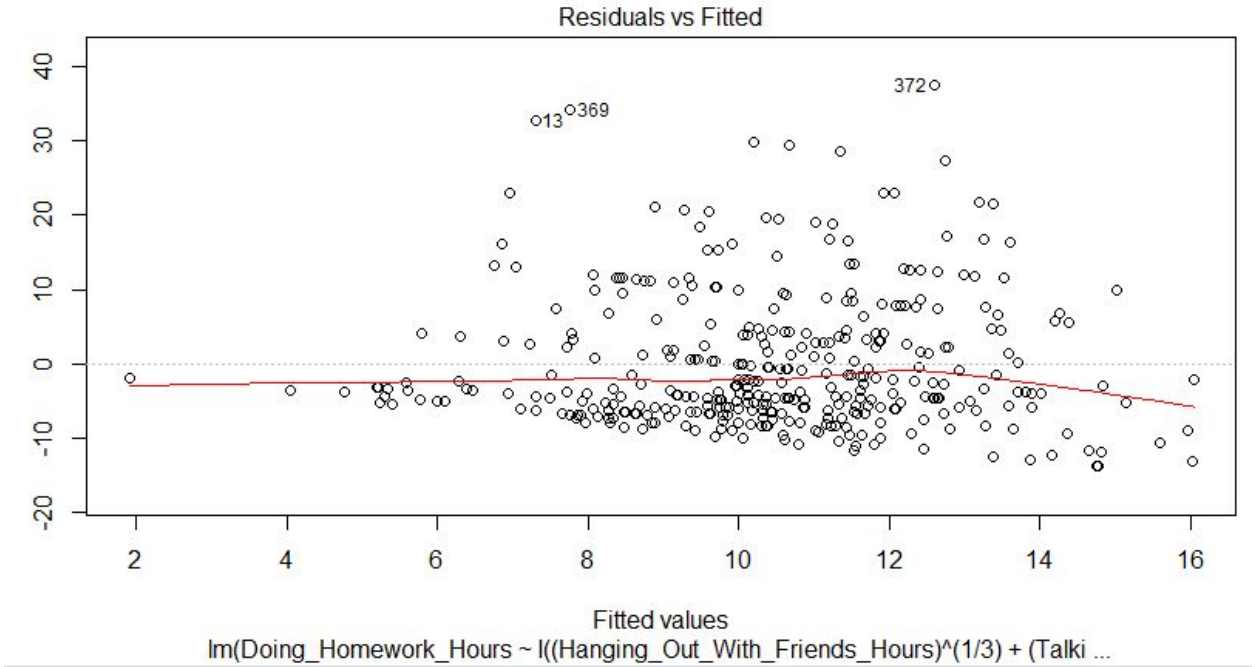
```
        Welch Two Sample t-test

data:  df$Importance_reducing_pollution and df$Importance_Internet_access
t = -0.58745, df = 908.32, p-value = 0.557
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -53.96645  29.10182
sample estimates:
mean of x mean of y
 695.0568  707.4891
```

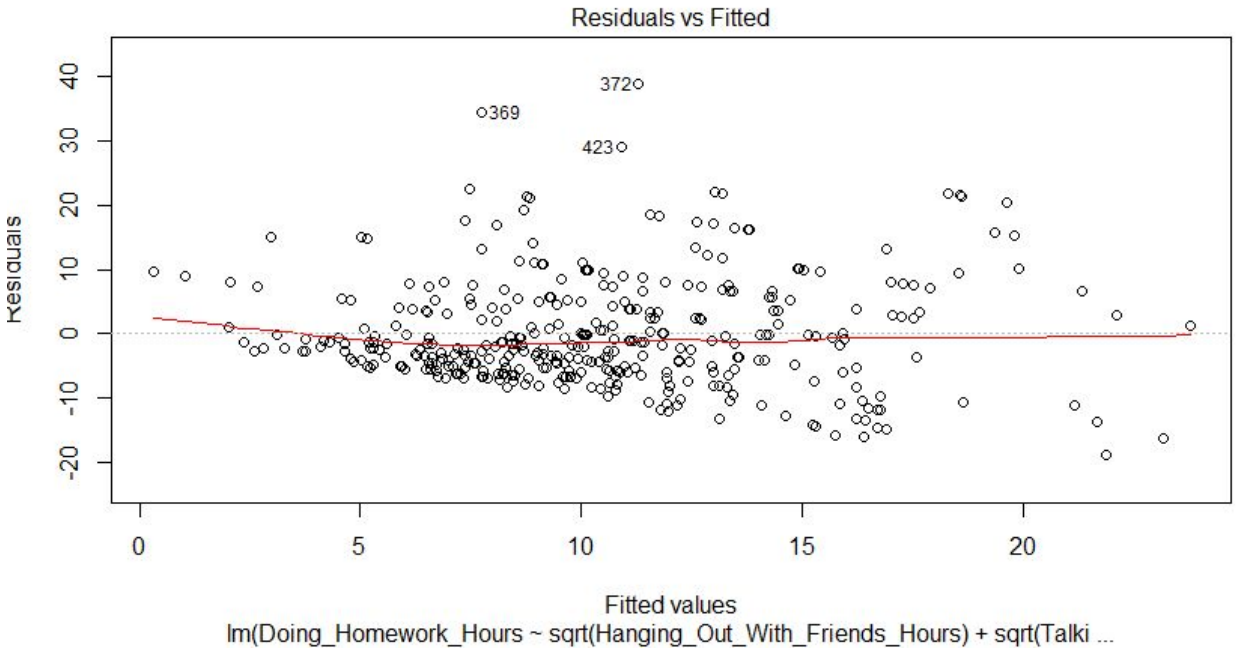
ANOVA results:

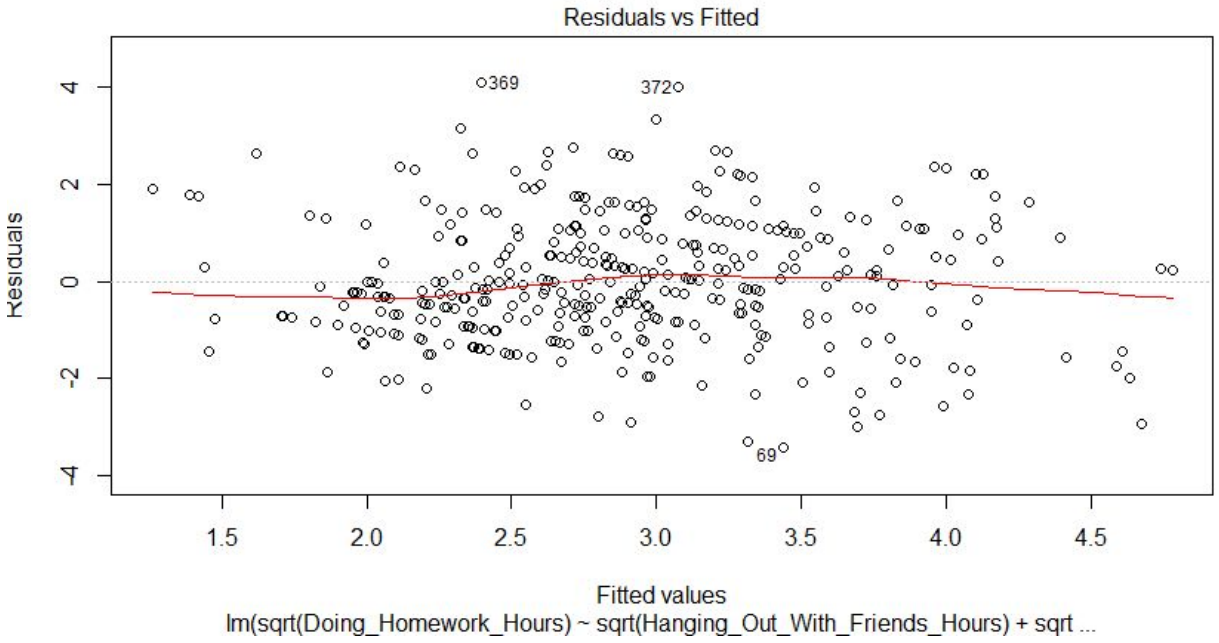
```
> summary(res.aov)
            Df Sum Sq Mean Sq F value Pr(>F)
group        2     37   18.26   0.188  0.829
Residuals   97   9413   97.04
```

Residual Plots for Models:

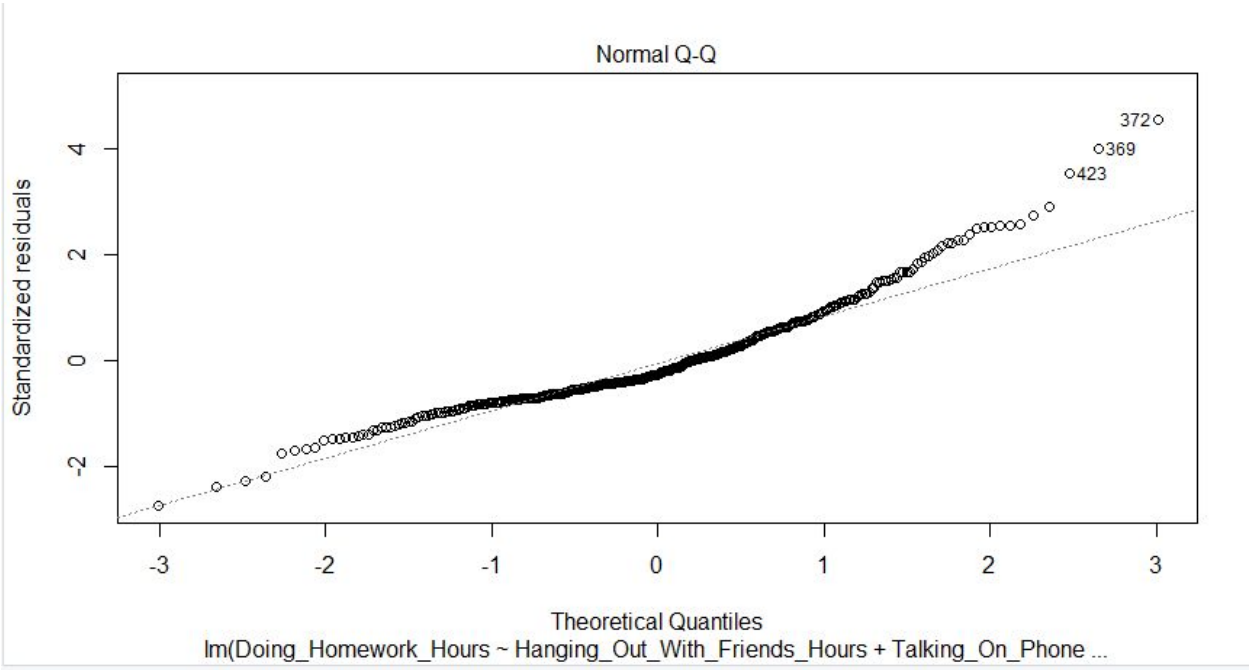


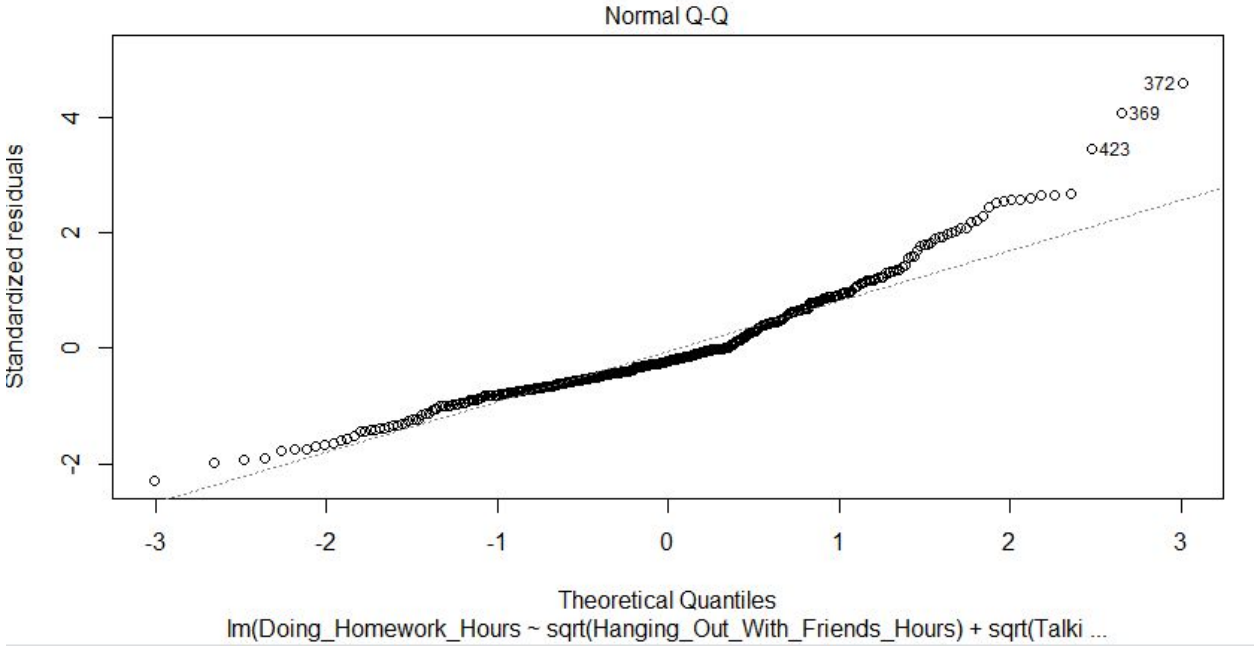


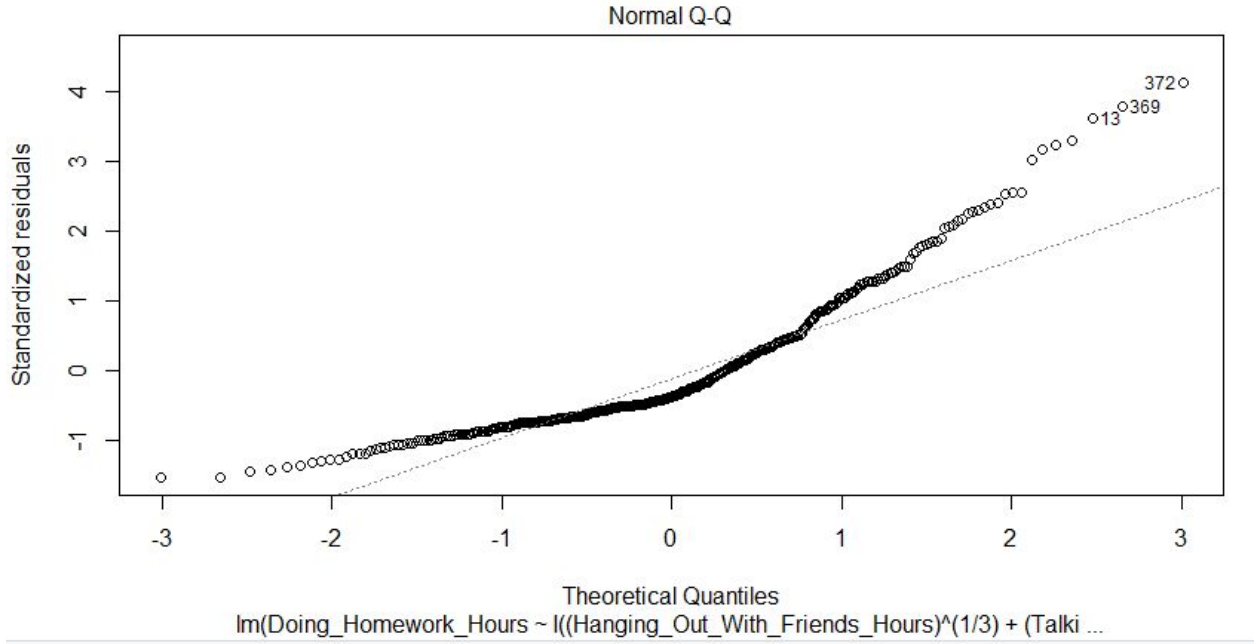

Residuals vs Fitted
Residuals
372
369
423
40
30
20
10
0
-10
-20
0
5
10
15
20
Fitted values
lm(Doing_Homework_Hours ~ sqrt(Hanging_Out_With_Friends_Hours) + sqrt(Talki ...

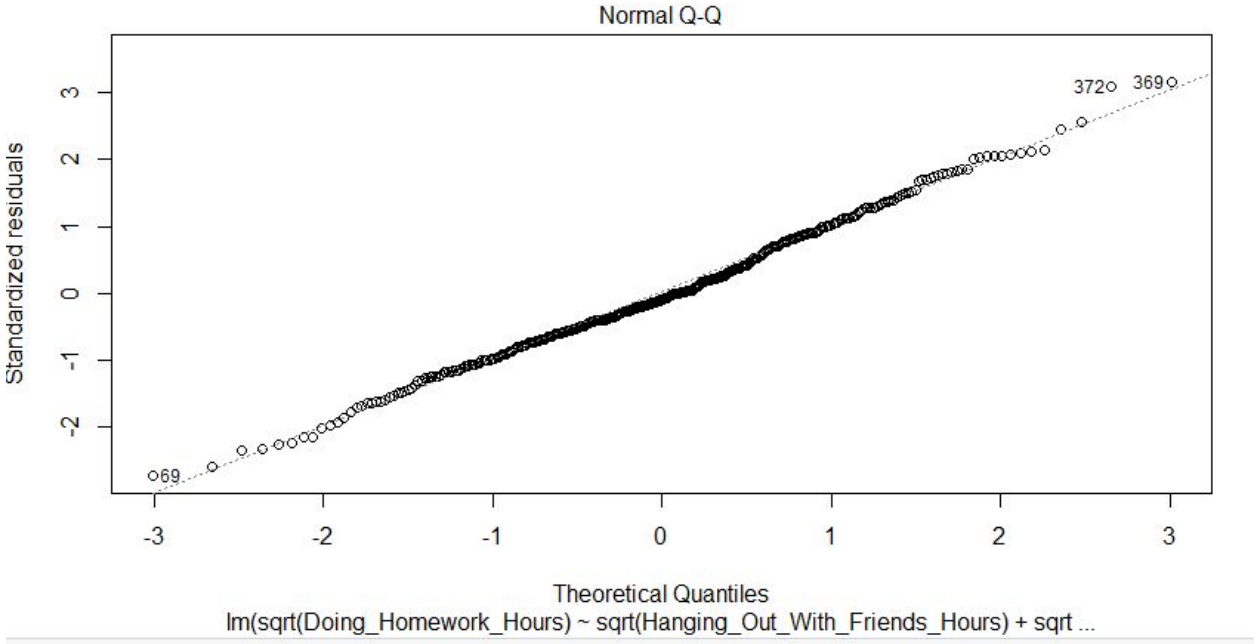

Residuals vs Fitted
Residuals
369
372
69
4
2
0
-2
-4
1.5
2.0
2.5
3.0
3.5
4.0
4.5
Fitted values
lm(sqrt(Doing_Homework_Hours) ~ sqrt(Hanging_Out_With_Friends_Hours) + sqrt ...

QQ Plots for Models:






Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(Doing_Homework_Hours ~ I((Hanging_Out_With_Friends_Hours)^(1/3) + (Talki ...
372
369
13


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(sqrt(Doing_Homework_Hours) ~ sqrt(Hanging_Out_With_Friends_Hours) + sqrt ...
372
369
69