Name: Rawad Bader, Alex Pomraning

Instructor: Frank McGrade

Course: STAT 423

Date: 12/16/2021

R Project: Phase III

Summary

The purpose of this research was to see the time breakdown among different majors from the Stat 423 2021 class, how students from this class spend their time, and if pet ownership had an impact on that. First, we wanted to ascertain how students within this class spent their time. So, we asked if the average amount of time spent on schoolwork between early risers and non-early risers was different. Then we asked if the average time spent on social media was significantly higher for Computer Science majors vs other majors. To answer these two questions, we performed a two-sided, and one-sided t-test respectively. After analyzing both questions we concluded that, in both cases, we failed to reject the null hypothesis, because the p-value of both of the t-tests respectively (0.6981, 0.1433) turned out to be greater than the significant level ($\alpha = 0.05$). We know that both cases did not work due to the small sample and failing some of the assumptions such as normal data set in both the data sets.

We also wanted to know whether their pet ownership impacted circadian rhythm or food preferences. So, we asked if there was an association between being a pet owner and being an early riser as well as if there is an association between a person's preference for sushi and pet ownership. To test these questions, we performed a fisher's test and a chi-squared test respectively. After analyzing both questions we established that we failed to

reject the null hypothesis for both the t-test and the Fisher's test due to the p-values (p = 1, p t= 0.6382) being greater than the significant level ($\alpha = 0.05$). Therefore, we came to the conclusion that both tests failed due to the small sample size and other issues such as failed assumptions.

Lastly, to determine the work time break down across majors we asked if there was a significant difference in time spent on schoolwork between the three blocks: "Biology and Neuroscience", "Applied Mathematics, Engineering, and Computer Science", and "Botany and Environmental Science". To answer this question, we ran an ANOVA test where we failed to reject the null hypothesis. This led us to conclude that there is no significant difference in the time spent on schoolwork between these groups.

After running all three groups' tests, we concluded that we failed to reject the null hypothesis due to the p-values being greater than the significance level ($\alpha = 0.05$). Additionally, we failed to reject our null hypothesis due to the inconsistencies of our datasets, the limited sample size, the lack of random sampling, and having multiple non-normal distributions. Thus, there is a probability we have made a type II error and there is not enough evidence to accept the alternative hypothesis.

Methods

The data for all analysis was taken during a digital class survey that used convenience sampling. This survey had 11 questions, 26 observations and 11 variables as follows: sushi preference, pet ownership, self-identification as an early riser, age, chosen major, height, most recent hug length, total schoolwork hours, total social media hours, total current credit hours, and individual or group exam preference. Data were sampled from students across all campuses that were part of STAT 423 during 2021 with Frank Mcgrade. Data was manipulated by the instructor with exclusions of data that did not fit the assignment or that did not fit the observational cell. In such cases, points were replaced with an appropriate selection from an array of possibilities presented by students during the survey in cases where more than one possibility was presented. This was to maintain one observation per cell for reproducibility and to ease use. Two questions were selected by two individuals for each of the first two groups, and then a third group was composed of a single question. For this study, we were interested in the variable's social media hours, sushi preference, pet ownership, major, and whether subjects consider themselves early risers.

Questions Group One:

Group one had two questions. The first question asks if the average amount of time spent on schoolwork between early risers and non-early risers is different. Time spent was placed into a block with respective early or non-early risers. A two-sample t-test was used due to the small sample size, and data being continuous. To test the validity of our assumptions a QQ-plot (Fig. 2 & 4), and histogram (Fig.1 & 3) were used on each block to check for normality of the data before those variances were calculated for each block then divided and compared as a ratio to check if they were equal; additionally, we ran the f-test to check if there were significant differences between variances. Independency, random selection, and continuity assumptions were checked through observation of the data.

Question Two asked if the average time spent on social media was significantly higher for Computer Science and other majors. Data was placed into a new data frame with a column containing both the CS majors and other majors, and we checked for normality by visualizing the data with a Histogram (Fig. 5), box plot (Fig. 6), and QQ plot (Fig. 7). A one-tail Two-Sample T-test was performed on both groups to compare the two means. This method was selected because our two groups (computer science majors and other majors) indicated a parametric test, therefore this test was appropriate in this scenario. The validity of the four assumptions for the t-test was assessed. The two samples' independence was determined through observation. We summarized both mean and standard deviation for the groups, then the graphed QQ-plot (Fig. 8 & 10), and Histogram (Fig. 9 & 11) to check normality along groups. We determined that the CS major's data set was not normally distributed while the other majors had a normal distribution. Finally, two vectors were created to check our assumption regarding equal variances. Our approach was to check if there was any randomization done after data collection through observation of the sampling method.

Questions Group Two:

Question two asked if there was an association between being a pet owner and being an early riser. Data were arranged into a contingency table for pet owners and early risers. We visualized the percentage difference between both groups before we ran Fisher's test for the analysis (Fig. 12). This was chosen over a standard chi-squared test due to the higher power with smaller samples and cell sizes and its ability to work with discrete data. The assumption was that the rows were fixed, the sample was randomly selected, data were mutually exclusive (i.e. observations did not fall into more than one cell), and independence was all checked through observation of the data.

In group two's second question we used the same analysis as question two from group one. The question this time was if there is an association between a person's preference for sushi and pet ownership. The data was analyzed by creating two bar graphs to check the percentage of people who like sushi or pets (Fig. 13 & 14). Then observation of the data determined if there were two categorical variables present to implicate the use of a Chi-Square test of independence. The chi-squared test is used to determine whether there is a significant association between two categorical variables. The assumption that variables are categorical, independent, and that the contingency table data points are mutually exclusive with total cell counts that are 5 or greater in at least 80% of cells was checked through observation.

Questions Group Three:

In group three we are interested in seeing if there is a difference in time spent on schoolwork between the three groups of majors: “Biology and Neuroscience”, “Applied Mathematics, Engineering, and Computer Science”, and “Botany and Environmental Science”, and if there is, determine which grouping spends the most time on school work. We created a new data frame that includes a group column containing the three groups, then we summarized the mean and standard deviation for all groups. Additionally, we graphed the boxplot to check variability among the groups (Fig. 15). Then the assumption that we had random samples, and independent groups were determined through observation. Then a QQ-plot (Fig. 16) tested the assumption of normality before the assumption of equal variances was determined through a residual graph (Fig. 17) and Levene test. Finally, we ran an ANOVA test and visualized the data set in graphs.

Results

RStudio was used to perform all analyses for the questions presented in groups one, two, and three.

Group One

Question One:

Question one performed a t-test to answer the question of if the average amount of time spent on schoolwork between early and non-early risers is different. The assumption of random sampling was determined through observation of the sampling method which was determined to be convenience sampling, a non-probability method. This assumption was unmet for all applicable tests we were expected to use. Thus, the analysis should be approached with skepticism. The assumption of independence was met as only one person answered per sample. Then the assumption that the variable is continuous was met as it was determined to be time spent which is a continuous variable. The t-test is a parametric test requiring normality to be assessed for early and non-early risers via a QQ-plot (Fig. 2 & 4) and Histogram (Fig. 1 & 3). We found that the data for early rises were not normally distributed while non-early risers were normally distributed. Thus, we concluded our data failed the normality assumption. Dividing the ratio of variances gave a value of 1.197098 (expected value roughly 1) and thus the assumption that the variances are equal was met. The t-test failed to reject the null hypothesis (p-value = 0.6981, a = 0.05, t = -0.3926, df = 24, CI = (-21.48218, 14.61552), m = (29.66667, 33.10000) that there was no difference in the time spent on schoolwork between early and non-early risers. Due to the failed assumptions, there is a high probability that a type II error has occurred, and there is not enough evidence to accept the alternative hypothesis.

Question Two:

For question two we asked if Computer Science majors have a higher average time spent on social media, than the other fields of study. The two groups independency was found to be true as they were tested for normality with QQ-plot (Fig. 8 & 10) and Histogram (Fig. 9 & 11) found that the first group (Computer Science majors) histogram showed a normal distributed data set but not the QQ plot and the second group (Other majors) histogram does not show any indication of a normal distributed data set. We also can clearly see that the data has a Right-skewed distribution. On the other hand, the QQ plot shows that the data set is normally distributed, therefore, we failed the normality assumption. The groups' variances' differences were verified by creating two vectors for both groups and testing their ratio which was equal to 0.984 indicating no significant difference between the two variances. In both groups, the samples were obtained not by random sampling because of our prior knowledge and the way how the data was collected. Finally, the T-test was used to check our hypothesis and discovered that the p-value of the test is ($p = 0.1574$), which is greater than the significance level ($\alpha = 0.05$), meaning that again we failed to reject the null hypothesis. Therefore, we came to the conclusion that the tests failed due to the small data sample and other factors such as failing the normal assumptions, and random sampling. Due to this, there is a high probability that a type II error has occurred, and there is not enough evidence to accept the alternative hypothesis.

Group Two

Question One:

Question one of group two performed a Fisher's test to answer the question of if there is an association between being a pet owner and being an early riser. The assumption of random sampling was determined through observation of the sampling method to be convenience sampling, a non-probability method. This assumption was unmet for all applicable tests we were expected to use. Thus, the test should be approached with skepticism. The assumption of independence was met as only one person answered per sample. The assumption of the rows being fixed was met because we had a 2x2 table. The final assumption for this type of Fisher's test is that the observations are mutually exclusive and were also met by observing the data and concluding that we cannot assign them to another cell. To run the Fisher's test, we created the contingency table from the data sets and visualized it with a bar stats graph (Fig. 12). The Fisher's test found that we failed to reject the null hypothesis (p-value = 1, CI = (0.1708393, 117.9170947), odds ratio = 2.087152) and that there is no significant association between being a pet owner and an early riser. Due to the failed assumptions, there is a high probability that a type II error has occurred, and there is not enough evidence to accept the alternative hypothesis.

Question Two:

For this question, we asked if there is a significant association between a person's preference for sushi and pet ownership. We started by graphing a separate bar graph (Fig. 13 & 14) for both groups to check how big the percentage difference is between them and found that there is little percent difference between the groups. The two groups were found to be independent through observation of the variable's categorical nature. Then through observation of the contingency table, we checked the assumption that the two groups are mutually exclusive and found that none of the individual cells belong to another individual's cell, in other words, they are mutually exclusive. It was determined through observation of the contingency table that the test did not pass the assumption that 80% of the cells have total counts of 5 or greater but no less than one. Finally, the Chi-squared test was conducted and we concluded that because the p-value (p = 0.6477) is greater than the significance level $(\alpha = 0.05)$, we failed to reject the null hypothesis and determined that the two variables are independent of each other. Due to the failed assumptions, there is a high probability that a type II error has occurred, and there is not enough evidence to accept the alternative hypothesis. An alternative test that might have found more conclusive results for this question is the fisher's test.

Group Three:

For this question we asked if there is a significant difference in time spent on schoolwork between the three groups of majors: "Biology and Neuroscience", "Applied Mathematics, Engineering, and Computer Science", and "Botany and Environmental Science", and determine which group spends the most time on schoolwork. We started by running the summary function on the groups to compare their mean, standard deviation, and then visualized the three groups' variability with a boxplot (Fig. 15) only to find there is no significant difference among groups. The groups were found to be independent when they were tested for normality with a QQ-plot (Fig. 16) which showed a normal distribution indicated by the close overlap of data points to a line of central tendency. Thus, we conclude that our data points came from a normal distribution. Then a graph (Fig. 17) of the residuals was used to observe if there was any variance. The graph showed that we have some variability in the third group but it is not significant which indicates homogeneity of variances across groups. We also checked the variance with the Levene test just to confirm that we have equal variances and found a p-value ($p = 0.1178327$, $\alpha = 0.05$), which is not significant. This means that there is no significant difference between variances across groups. Therefore, we can assume the homogeneity of variances in the different groups. Finally, an ANOVA test was used that found a p-value ($p = 0.506$, $\alpha = 0.05$), which is greater than the significance level, and thus we fail to reject the null hypothesis. Therefore, we conclude that there are no differences in the time the three groupings of majors work on schoolwork. Due to the failed assumptions, there is a high probability that a type II error has occurred, and there is not enough evidence to accept the alternative hypothesis.

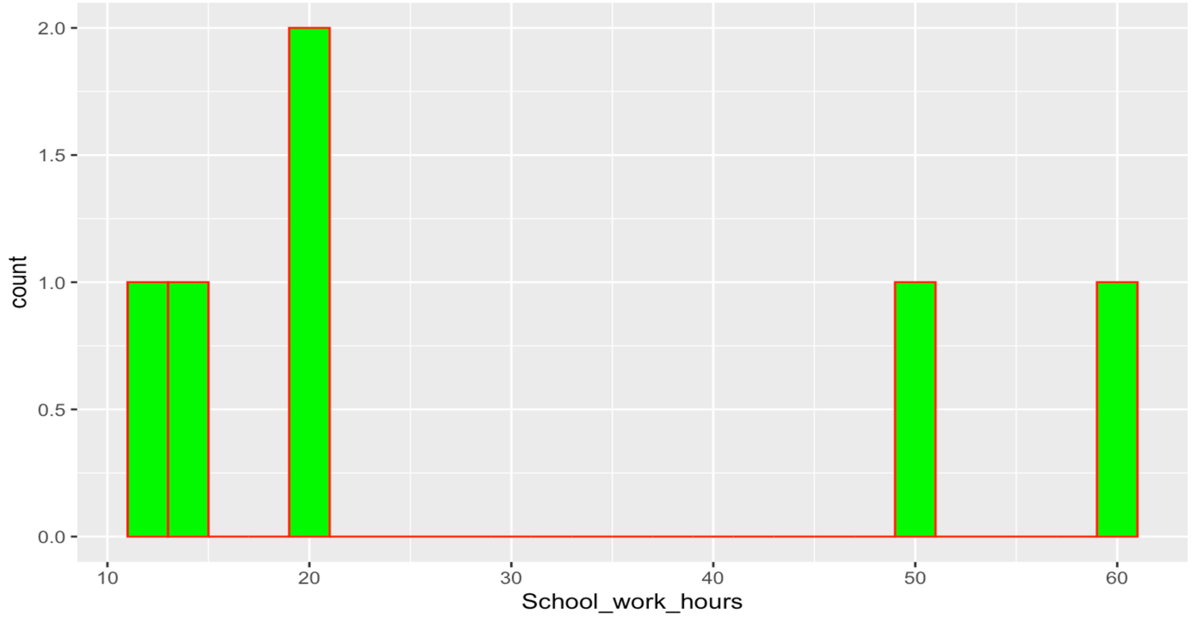

Figure 1

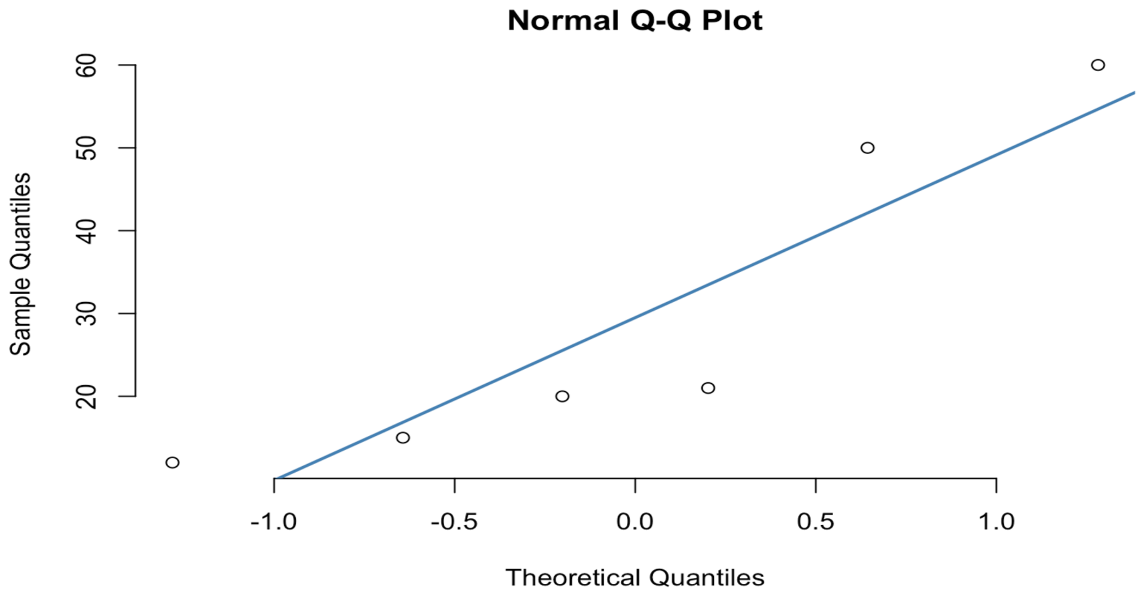

Figure 2

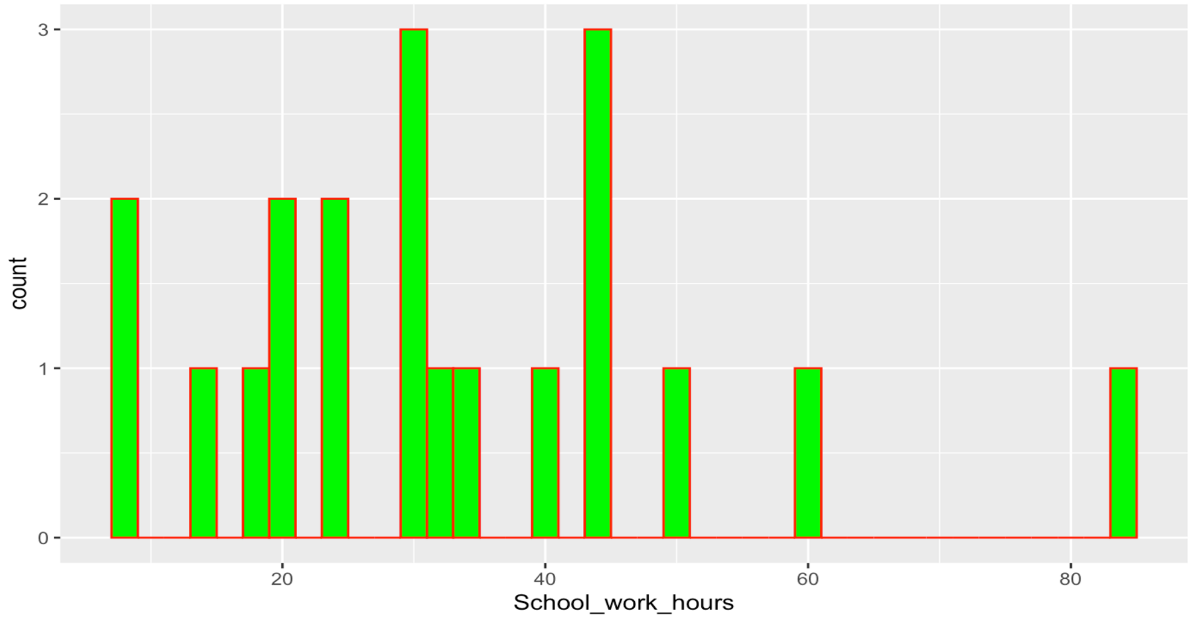

Figure 3

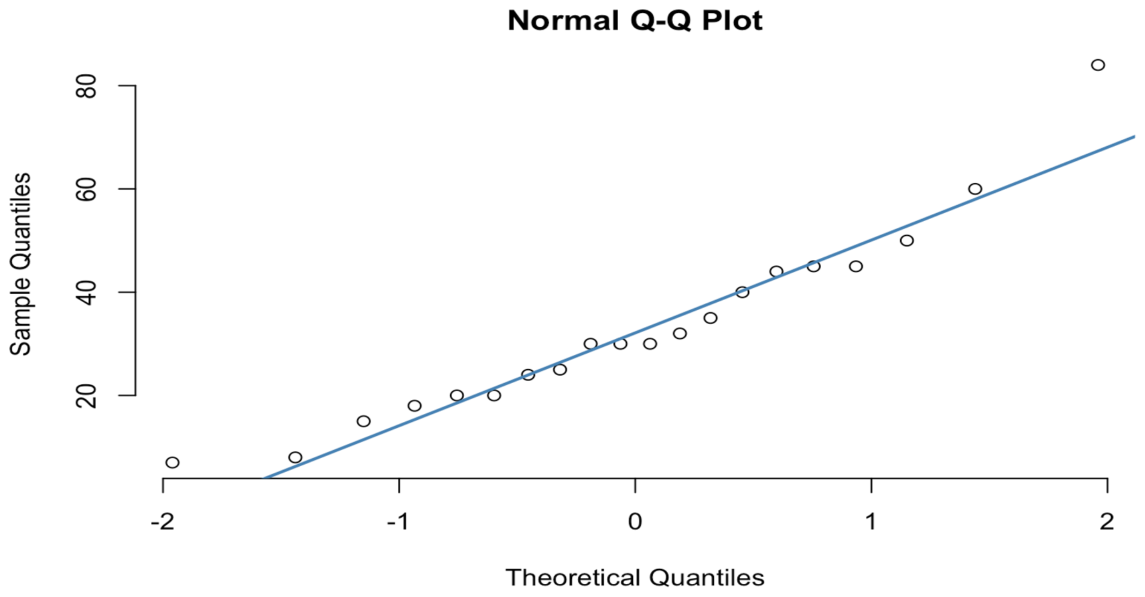

Figure 4

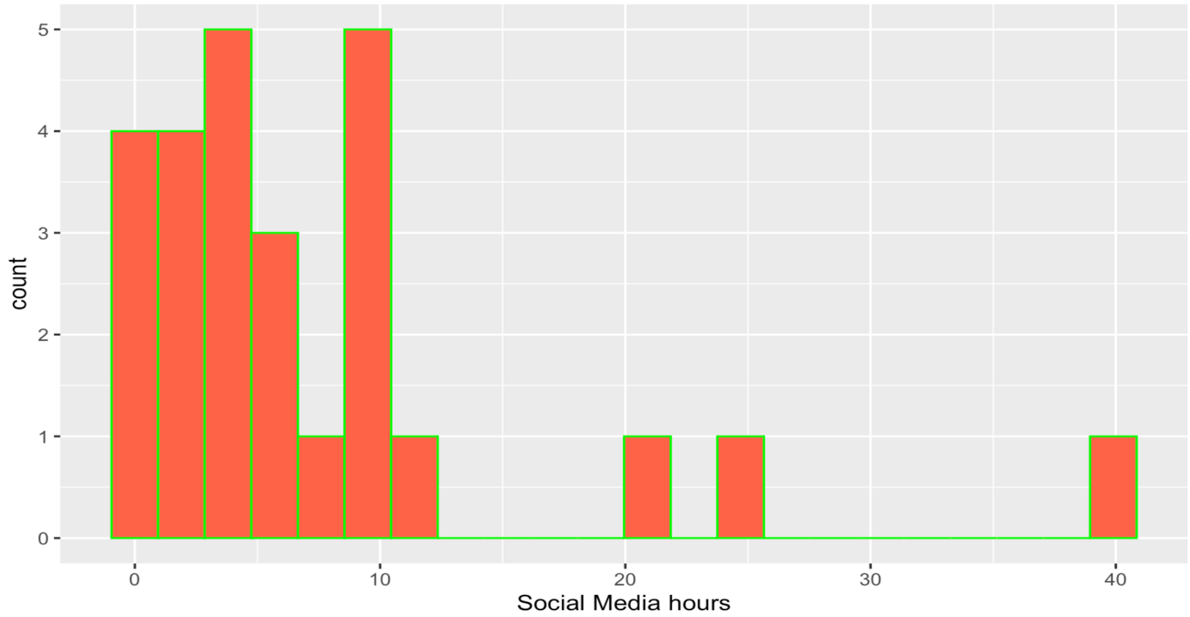

Figure 5

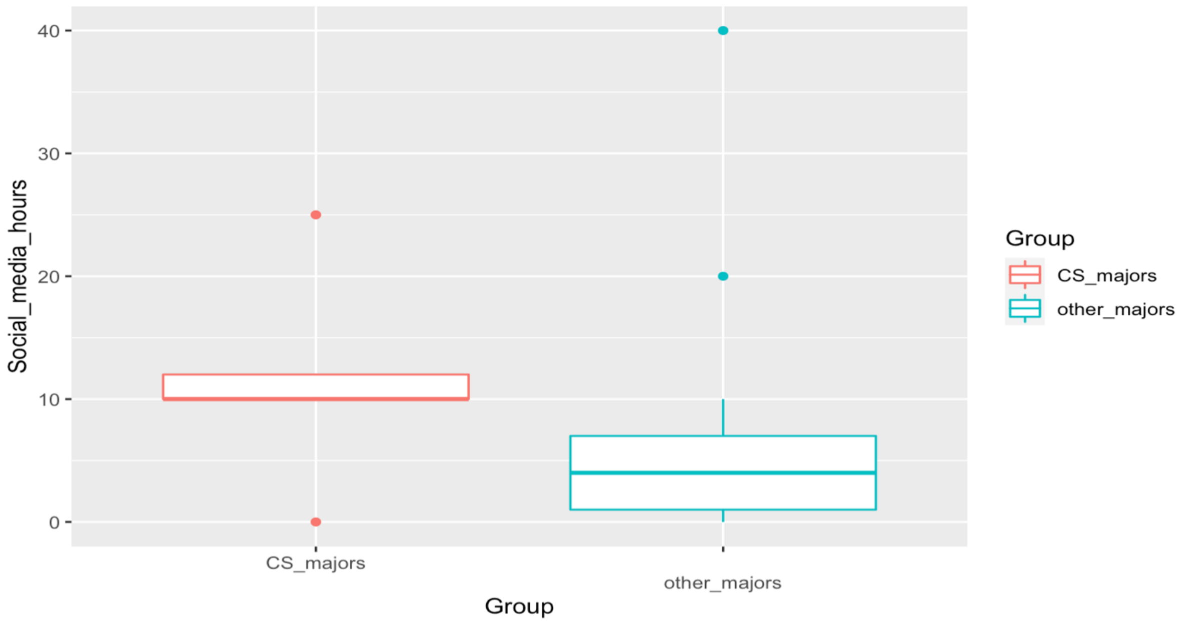

Figure 6

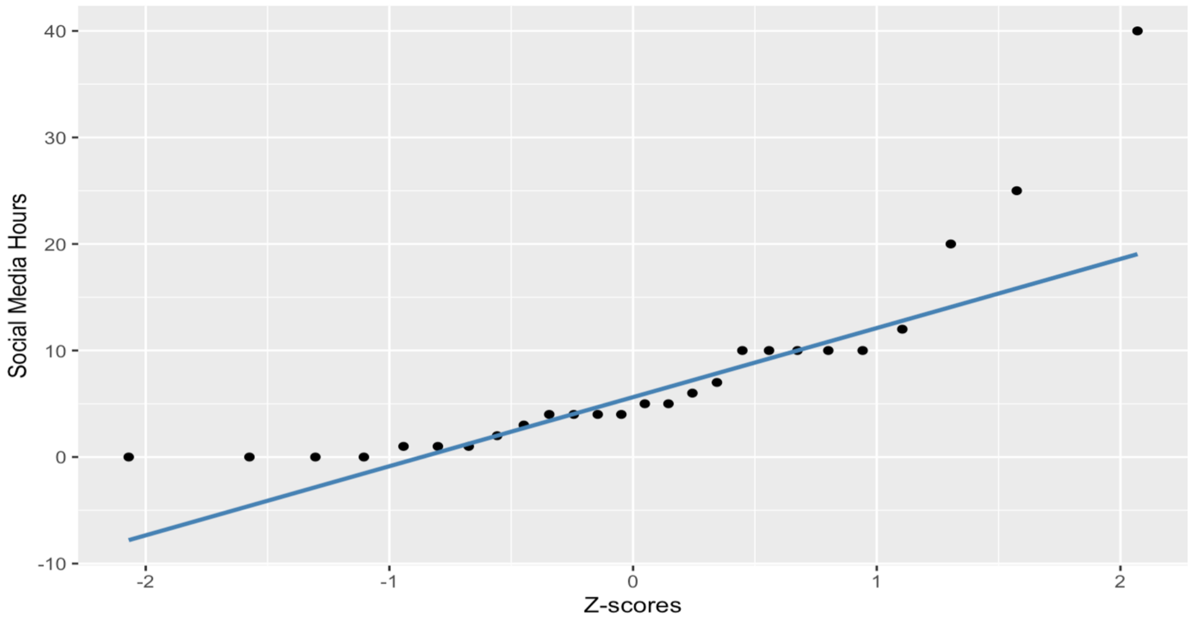

Figure 7

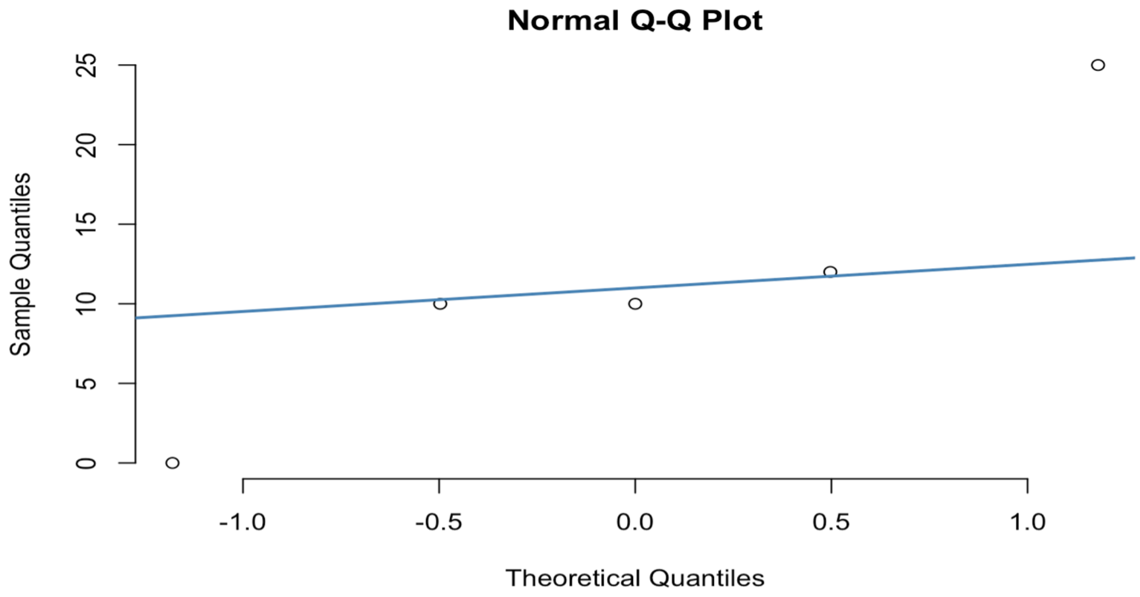

Figure 8

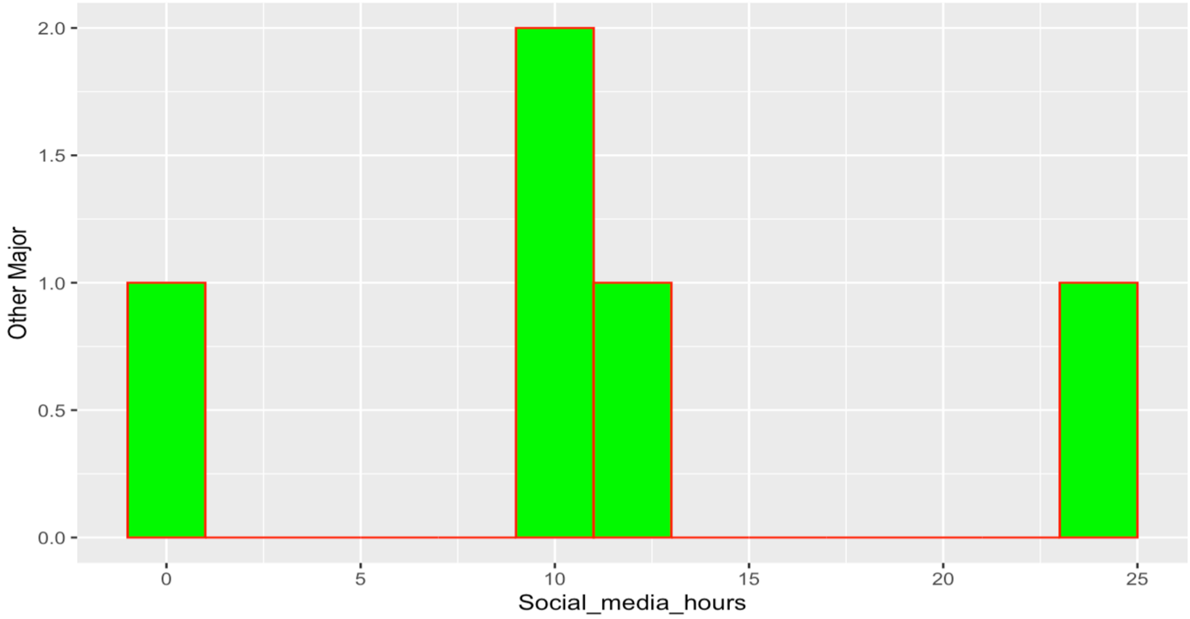

Figure 9

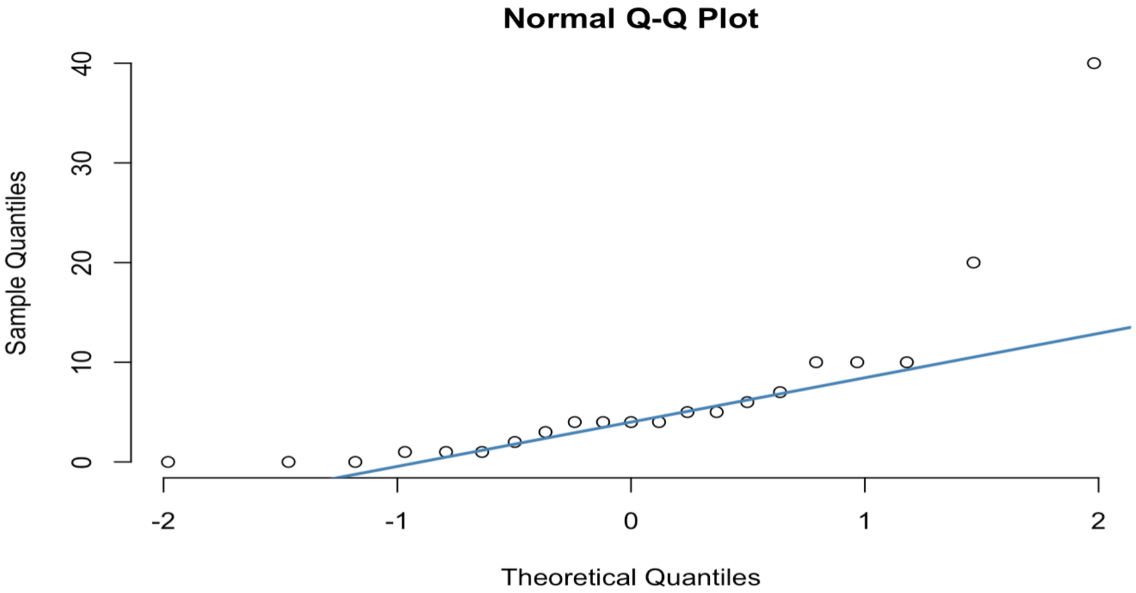

Figure 10

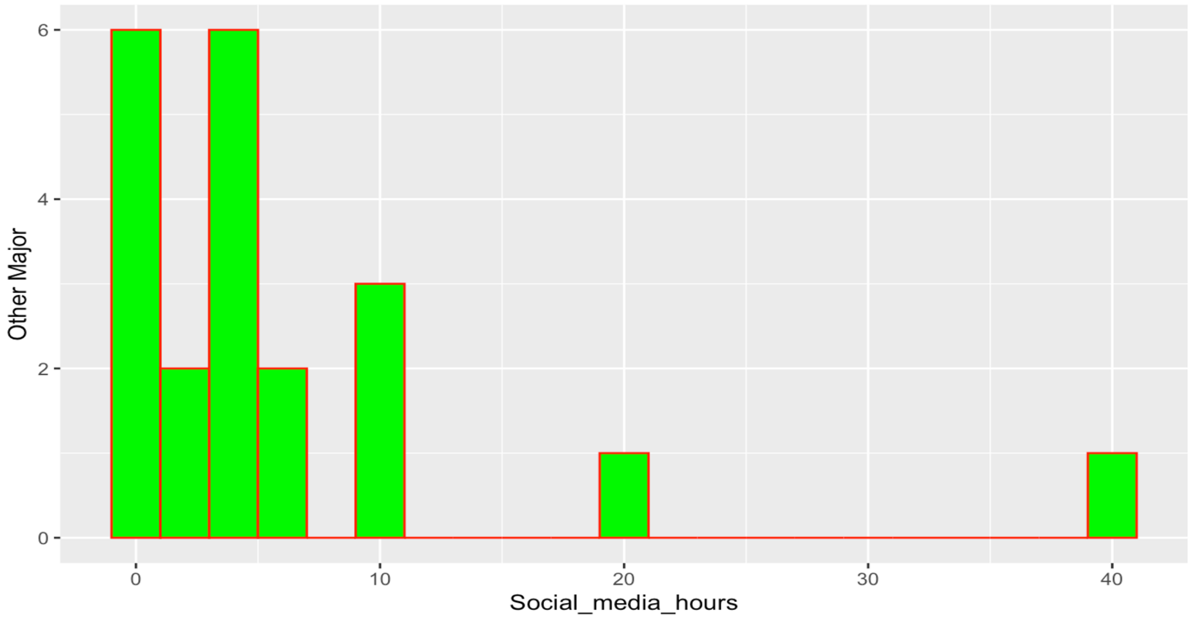

Figure 11

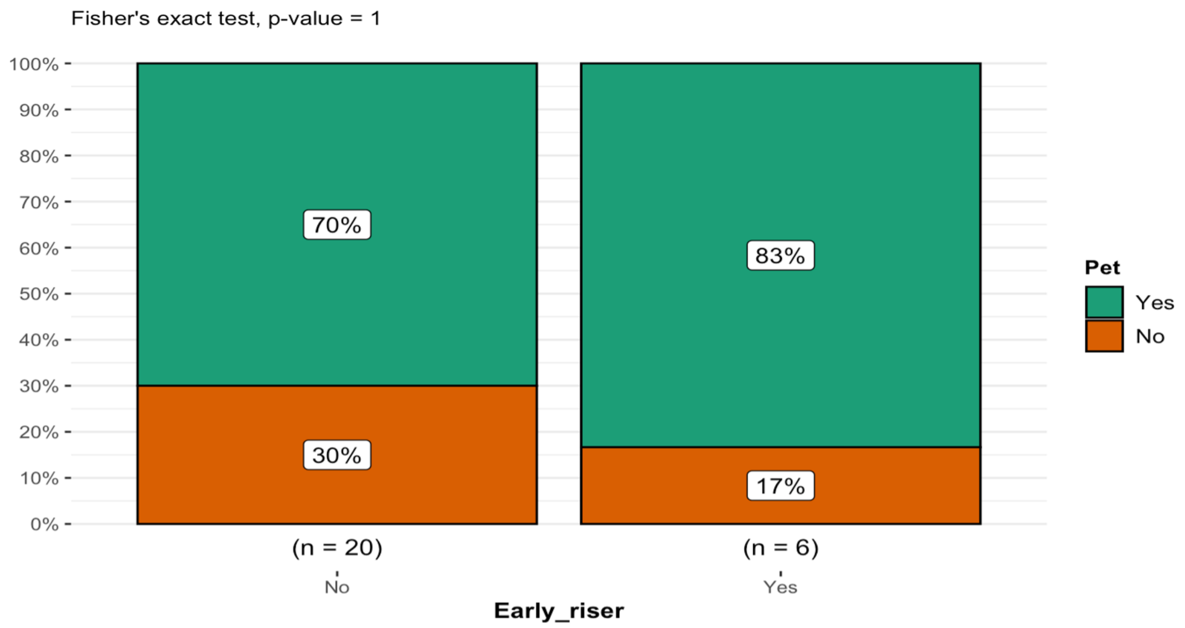

Figure 12

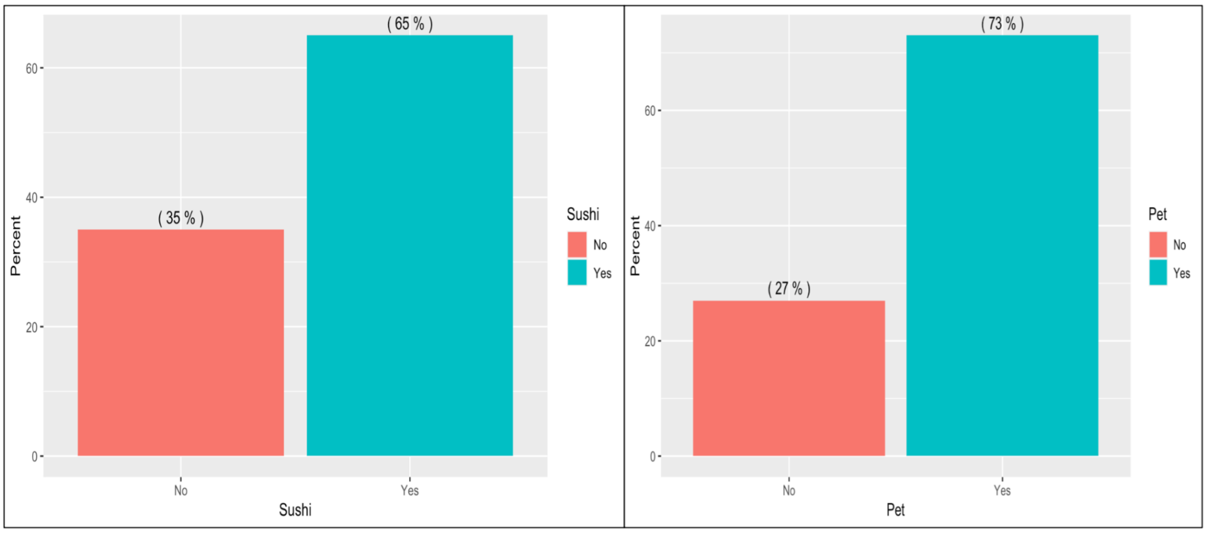

Figure 13

Figure 14

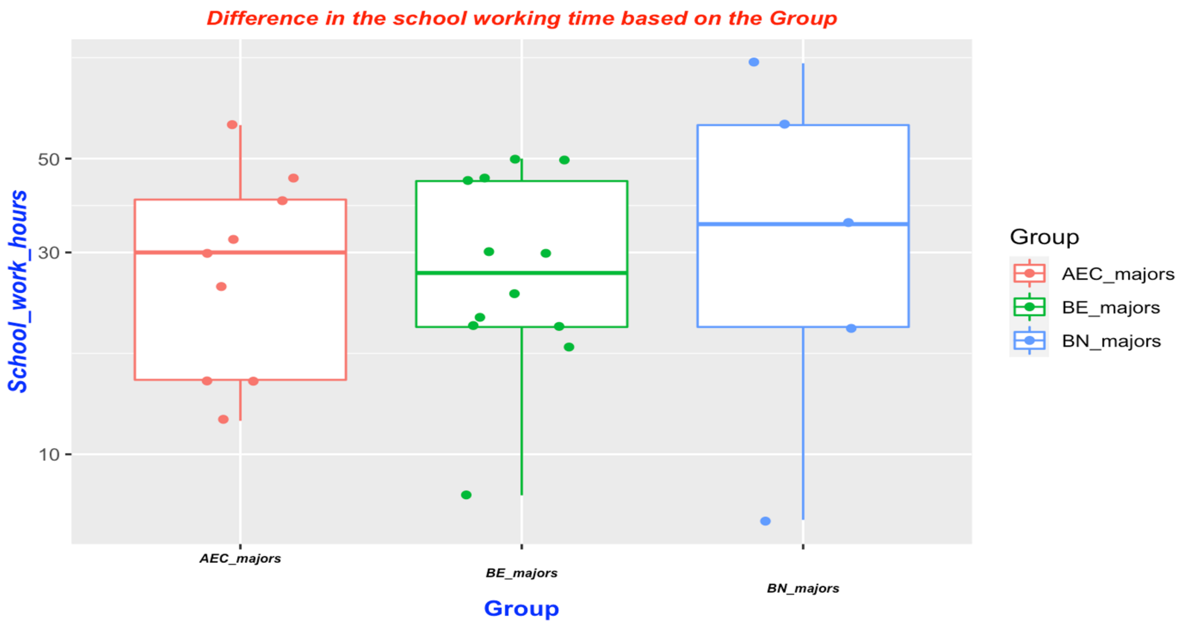
Difference in the school working time based on the Group
School_work_hours
50
30
10
AEC_majors
BE_majors
BN_majors
Group
Group
AEC_majors
BE_majors
BN_majors

Figure 15

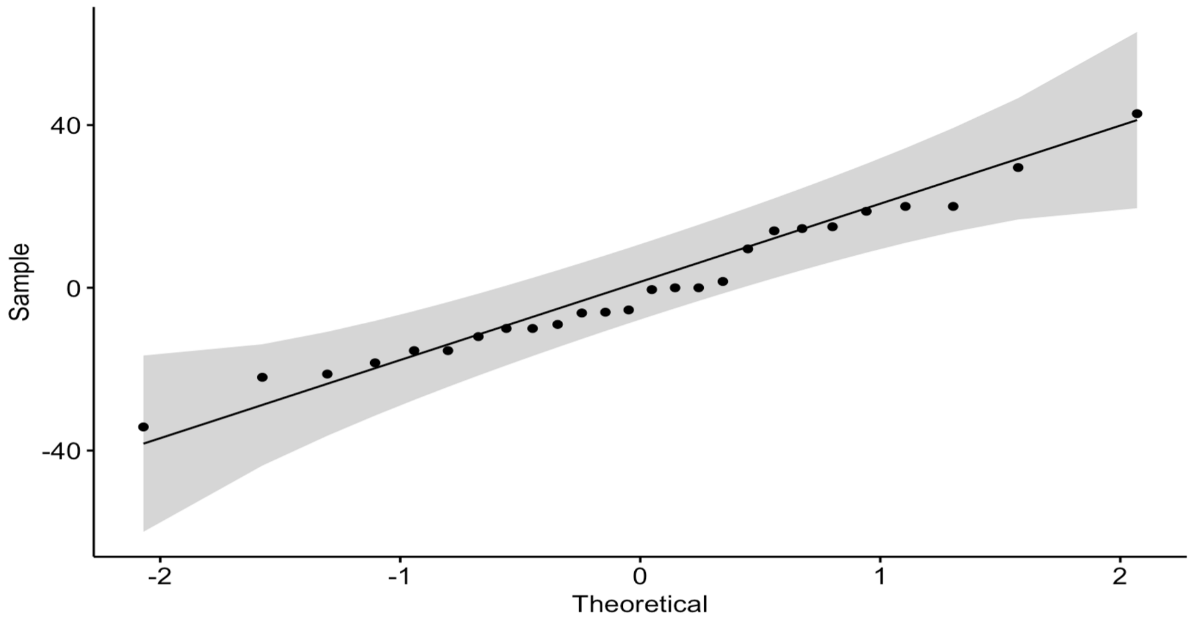

Figure 16

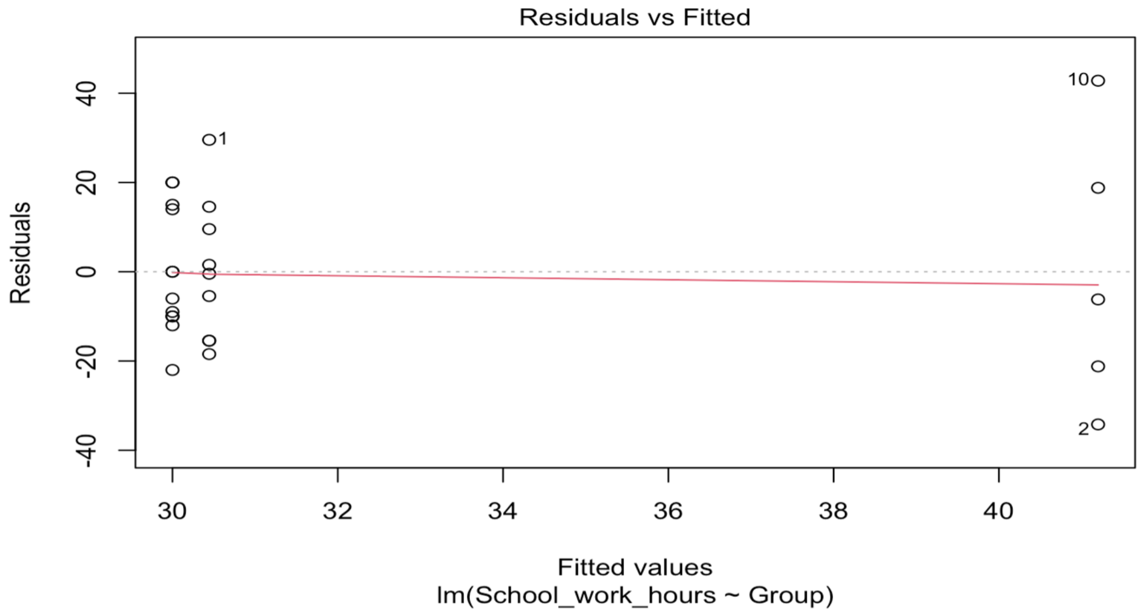

Figure 17