STA302 Final Project Report

Yiqin Huang 1007162892, Yifan Cheng 1007226850, Ruiqi Kang 1006941017, Tony Chen 1006836480

2022/8/20

Introduction

Many university students put a lot of efforts in studying in aims to obtain higher grades for better opportunities in the job market and better chances in pursuing graduate education, but we still experienced the pandemic epidemic in 2022, the COVID situation has affected every student's academic life: more and more people will consider the harm caused by the epidemic, which can significantly affect everyone's learning time. It also comes with the delivery mode of course in online format where students might may not stick with the schedule closely and also find it hard to catch up with later. Here is a question: how factors may influence the Term Test score for a student? Our concerned factors include efforts of studying and time on other things (exclude studying).

In this report, our purpose is to build a good-fit linear model to show the relationship between Term Test Score and some selected preditor variables, thus we can use our model to estimate the Term Test score of a student in STA302 based on the available information.

Experiment Information

There were 4 surveys distributed weekly towards the students enrolled in the Summer 2022 (July - August) STA302 course. As of August 21, there are 154 students in this course, with the course limit being 200. In this study, questionnaires were filled out on Quercus where all the study related material can be found. Each survey asked the individual students about their studying time measured in hours where the expected minimum is 6, due to 3 hours per lecture, 2 lectures per week; their time spent thinking about COVID, measured in hours; their time spent in miscellaneous activities excluding sleeping, eating, etc.; their attendance in office hours; their familiarity about the material before the term test. The last two variables were accessed qualitatively. The experiment was conducted in aims to developing applausable for fitting and potentially predicting. And as of the data set, due to missing surveys and various reasons, there were only 110 valid observations.

Purpose of Developing Model

We aim to develop a model in order to see what kind of relation between the term test grade and the predictor variables mentioned previously. Such desired relation may not exist if we limit our scope of model selection with linear models. We restrict ourselves with this because we have mainly learned about linear models in this course, and verification and interpretation of the final selected model are also two main components in this course. In order to have a model that actually can both make sense and be understood, interpreted, evaluated in terms of statistical significance, linear models should be used.

The development of such model would be substantially beneficial to educators and students, in aspects like professors and coordinate and adjust the design of the course properly and make available resources like office

hours and maybe tutorials used to their fullest; on the other hand, students can learn from the model in order to maximize their use of time in pursuit of a higher mark in the course. We can also learn from how various factors affect the learning process of this course.

Explantory Data Analysis

1) Creating New Explanatory Data

Since the term test is accumlative, we find it more appropriate to ignore the time factor to encounter different study strategies. Mean_studying is the mean studying time of a student in one week.
Mean_covid is the mean time a student thinking about COVID in one week.
Mean_miscellaneous is the mean time a student spending on sport and other fun stuff in one week.
Mean_nonstudying is the mean non-studying time of a student in one week.
OH is a categorical variable, indicates whether a student had attended the office hour in the four weeks on average.

2) Checking Multicollinearity between Quantitive Variables (Studying Time and Non-Studying Time)

Based on the experience from personal lives, the mean studying time and mean non-studying time should be correlated.
We will check whether there exists multicollinearity.

```
VIF_mean
```

```
## [1] 1.086629
```

Because the mean VIF value is greater than 1, we have discovered the existence of serious multicollinearity. Thus, the estimated coefficients have a large change when a predictor variable is added or deleted, and the t-tests for regression coefficients would likely have non-significant results. There is also a high correlation between the explanatory variables, and therefore difficult to interpret them individually. Besides, it results in high variance and standard error of the coefficients.

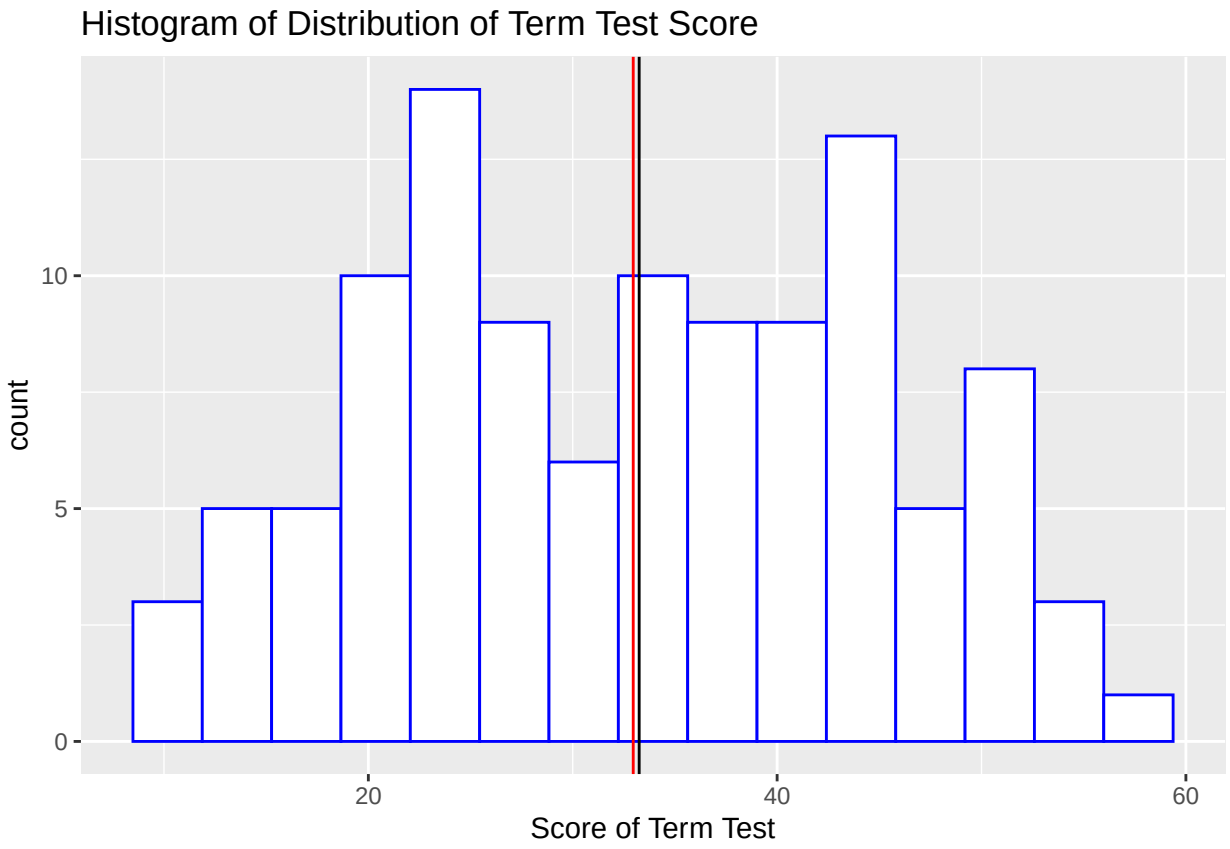
3) Description of Each Variables

The variables we will select are mean_studying, mean_nonstudying, and OH_attendance.
For the first variable, mean_studying, it stands for the average studying time of a student from week one to week four.
And mean_studying is a numerical variable.
For the second variable, mean_nonstudying, it stands for the average time of a student thinking about covid from week one to week four plus average time of a student spending on diverse things in daily lives from week one to week four.
And mean_nonstudying is a numerical variable.
For the third variable, OH_attendance, it stands for whether a student had attended office hour for at least once from week one to week four.
To be more specific, if a student go to office hour at least once a week, then we define OH_attendance as "Participated".
If a student go to office hour once a week, then we define OH_attendance as "Participated".
If a student go to office hour at least once a week, then we define OH_attendance as "Participated".
If a student never go to office hour, then we define OH_attendance as "Absent".
And OH is a categorical variable.

Histograms

The histograms were used to show the characteristics of students' term test scores, average time spent on studying STA302 from week1 to week 4, and average time spent on thinking about covid and doing miscellaneous from week 1 to week 4. We add a red line to indicate the mean of the distribution, and we add a black line as the median of the distribution.

4) Histogram of distribution of term test

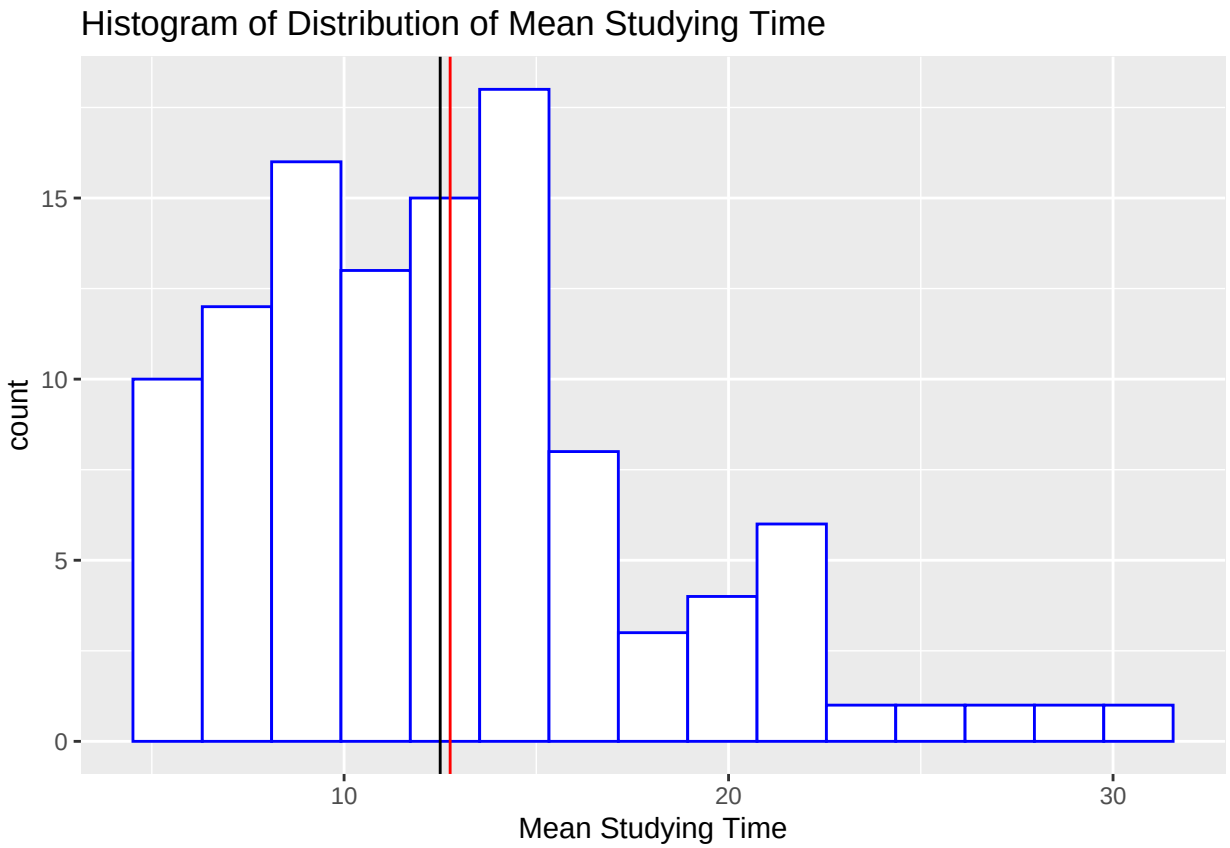



```
## [1] 32.95909
```

```
## [1] 33.25
```

The histogram of the distribution of term test scores is bi-modal, and the median and mean of all student's term test scores are very close(32.96 and 33.25, respectively). And the students' term test scores appear most often around 22-24 and 42-44, so we may consider splitting all students into two groups to analyze the relationship between term test scores and explanatory variables.

5) Histogram of Mean Studying Time

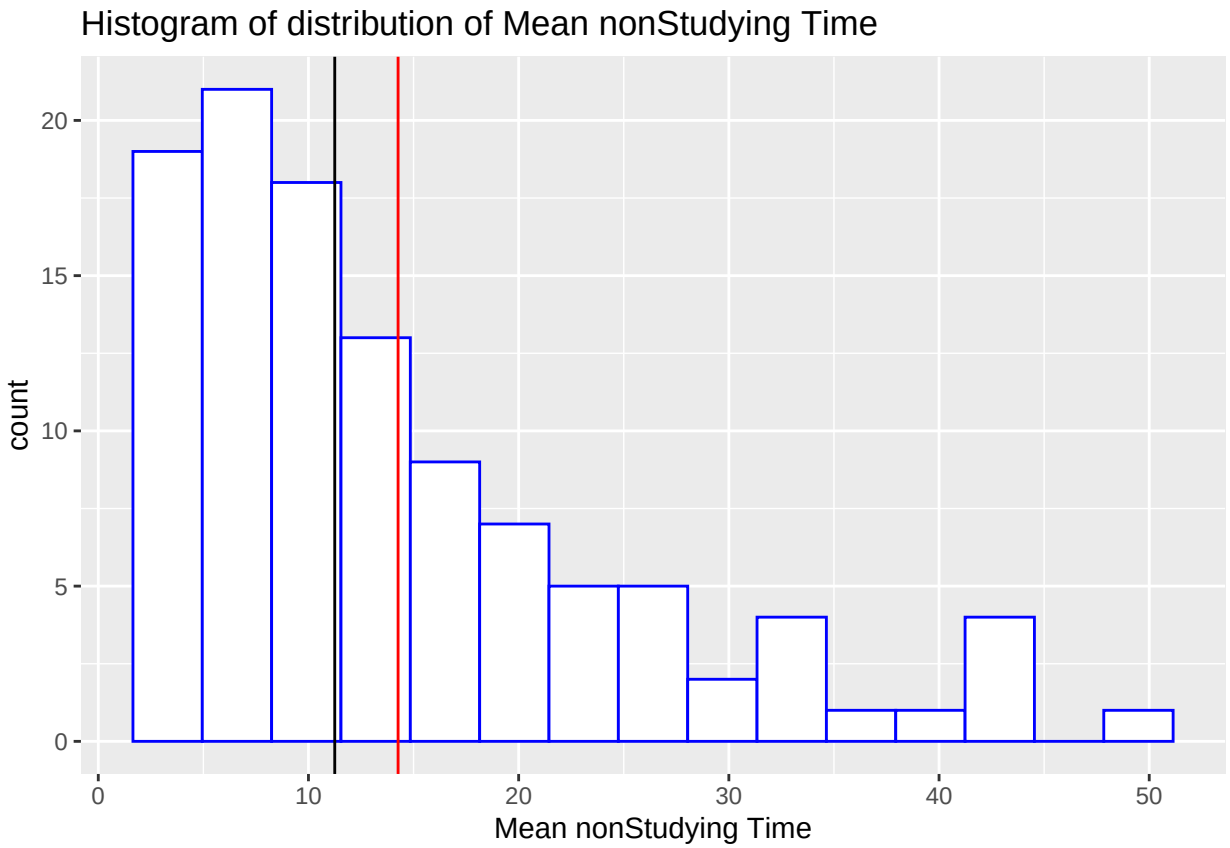



```
## [1] 12.75909
```

```
## [1] 12.5
```

For the histogram of mean studying time, we find that the distribution is right-skewed, which means more students in STA302 have spent less time on studying this course. We added a red vertical line in the graph as the mean of the distribution, which represents the mean of the mean studying time of all students is 12.76 hours. We also added a black vertical line in the graph as the median of the distribution which is 12.5 hours, it represents the median of studying time.

6) Histogram of Mean Non-Studying Time

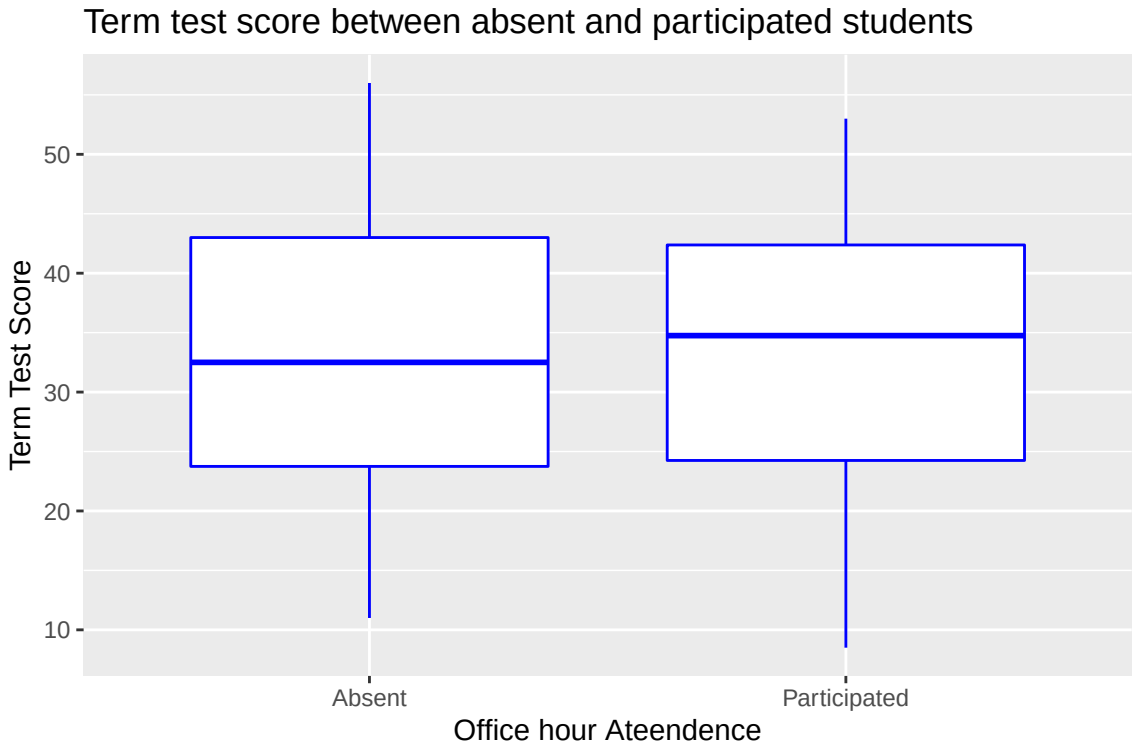



```
## [1] 14.26312
```

```
## [1] 11.25
```

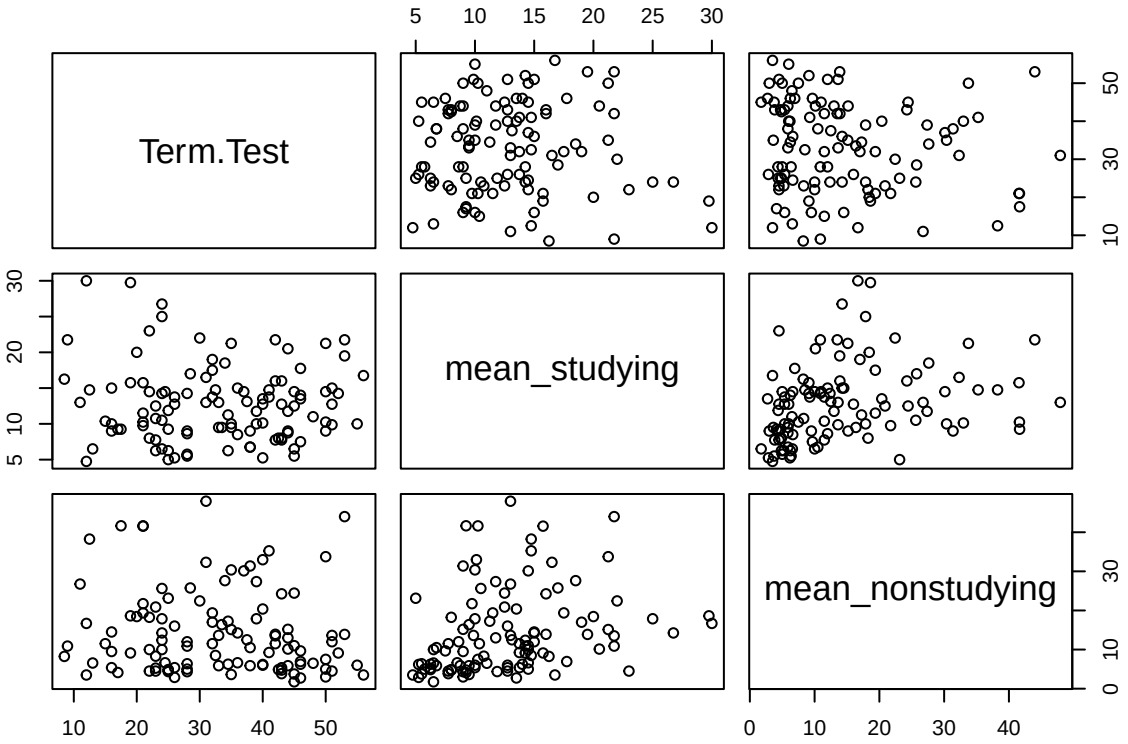
For the histogram of mean non-studying time, we find that the distribution is also right-skewed, which means more students in STA302 will spend less time on considering COVID and miscellaneous in this course offered in the summer. We added a red vertical line in the graph as the mean of the distribution, which represents the average hours on things (excluding studying) for students in STA 302 per week is 14.26 hours. We also added a black vertical line in the graph as the median of the distribution which is 11.25 hours.

7) Boxplot Plot for Categorical Variable OH



This box plot separates STA302 students into two types: students who have participated during the office hour; and students who never show up during office hours. We used the box plot to show the difference in term test scores between students who participated during office hours and those who never showed up. Intuitively, the median of the term test scores of students who at least participated in OH once is 35, slightly higher than that of those who never showed up in OH, which is 33. However, the interquartile range(IQR) of term test scores between these two groups are nearly the same(23-44), meaning the spread of scores of different types of students are very similar.

8) Pairwise scatter plot for numerical variables mean_studying time, mean_nonstuding time and response variable term test score



This pairwise plot between the response variable: term test score, and the two numerical variables, mean studying time and mean no studying time, shows the relationship between every two variables. By looking at this plot, it is evident that there is no potential linear relationship between these variables. The reason of this phenomenon might be that some students do not pay attention to the weekly survey but just answer indiscriminately.

Model Developing

1) Creating Models

In the first step of model developing, we create 7 models to reveal the relation between response variable (Term Test Score) and different predictors.
We create the Model_1 as the model only contains the response variable (Term Test Score);

$$Y_{TermTestScore} \sim No\ Explanatory\ Variables$$

In the Model_2, we use one predictor "mean_studying" (the mean of the time that students spend on studying) to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean studying}}$$

In the Model_3, we use another predictor "mean_nonstudying" (the mean of the time that students spend on considering Covid and miscellaneous) to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean nonstudying}}$$

In the Model_4, we use two predictors "mean_studying" and "mean_nonstudying" to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean studying}} + X_{\text{mean nonstudying}}$$

In the Model_5, we use two predictors "mean_studying" and "OH" (determine whether the student attend any office hours) to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean studying}} + X_{\text{OH}}$$

In the Model_6, we use two predictors "mean_nonstudying" and "OH to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean nonstudying}} + X_{\text{OH}}$$

In the Model_7, we use three predictors"mean_studying", "mean_nonstudying" and "OH_attendance" to estimate the response variable (Term Test Score);

$$Y_{TermTestScore} \sim X_{\text{mean studying}} + X_{\text{mean nonstudying}} + X_{\text{OH}}$$

2) Comparing Models by Empirical Nature

The second step in model developing we decide to compare different models empirically.
According to our experience, high score is proportional to the effort on studying, and we treat 2 predictors (mean_studying, OH_attendance) as the effort on studying. Then we build model_2 and model_5 to reveal the relation between score and effort on studying.
On the other hand, we believe that less effort on studying will cause low score, so we use another predictor "mean_nonstudying" try to find the relation between Term Test Score and time spend on other things (exclude studying).

3) Comparing models

1) by R^2_{adj}

R^2_{adj}	
R^2_{adj1}	0
R^2_{adj2}	-0.006883467
R^2_{adj3}	0.004238174
R^2_{adj4}	-0.004789375
R^2_{adj5}	-0.01516765
R^2_{adj6}	-0.003899393
R^2_{adj7}	-0.01294292

2) by AIC

AIC	
AIC_1	546.5377
AIC_2	548.2784
AIC_3	547.0567
AIC_4	549.0262
AIC_5	550.1565
AIC_6	548.9287
AIC_7	550.8823

For the model selection criterion, we want high R^2_{adj} and low AIC. We know model 3 has the highest R^2_{adj} and model 1 has the lowest AIC; however, The AIC of model 1 and model 3 are very close(546.5377 and 547.0567). Also, we do not need to include many variables due to the high multicollinearity from part 2. Thus, the simple linear regression with the mean_nonstudying variable is the best model.

4) Model Summarization

```
##
## Call:
## lm(formula = Term.Test ~ mean_nonstudying, data = finaldata3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.2373  -9.2001   0.5845   9.3455  23.8892
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       34.8049     1.9020   18.30   <2e-16 ***
## mean_nonstudying  -0.1294     0.1070   -1.21    0.229
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.91 on 108 degrees of freedom
## Multiple R-squared:  0.01337,    Adjusted R-squared:  0.004238
## F-statistic: 1.464 on 1 and 108 DF,  p-value: 0.2289
```

5) Mathematical Interpretation of Selected Model

Since we build this model by selecting $mean_{nonstudying}$ as the predictor, so Y will be our response variable (numerical), which means the score of Term Test, and X will be the predictor variable (numerical), which means the hours spend on things not related to studying.
After we select the model_3, we get β_0=34.8049, which is the y-intercept in our model, in other words, if a student spent 0 hour on considering Covid or miscellaneous, then this student will get 34.8049 in Term Test. The β_1=-0.1294, which is the slope of our fitted line: if a student spent 1 hour on non-studying things, then the student will lose 0.1294 point in Term Test.
By the definition of R^2 we learned in lecture, R^2 means how many variations can be explained by our model, and R^2 is a goodness-of-fit measure for our model. Here we obtain $R^2 = 0.01337$, and note that a low R^2 value indicates that our predictor variable is not explaining much in the variation of our response variable.
We also calculate the p-value of our model, by definition, p-value is the probability of obtaining an effect at least as extreme as the one in your sample data, assuming the truth of the null hypothesis. Here the p-value is 0.2289, which means we may not reject the null hypothesis.

6) T-Test for the Selected Model

	value
$\hat{\beta}1$	-0.1294114
standard_error β_1	0.1069579
test_statistic	-1.209929
p_value	0.2289479

Our null hypothesis: there is no linear relation between term test scores and the average time of a student thinking about things not related to studying

$$H_0 : \beta_1 = 0$$

, Our alternate hypothesis: there exists a linear relation between term test scores and the average time of a student thinking about things not related to studying

$$H_a : \beta_1 \neq 0$$

T test of β_1

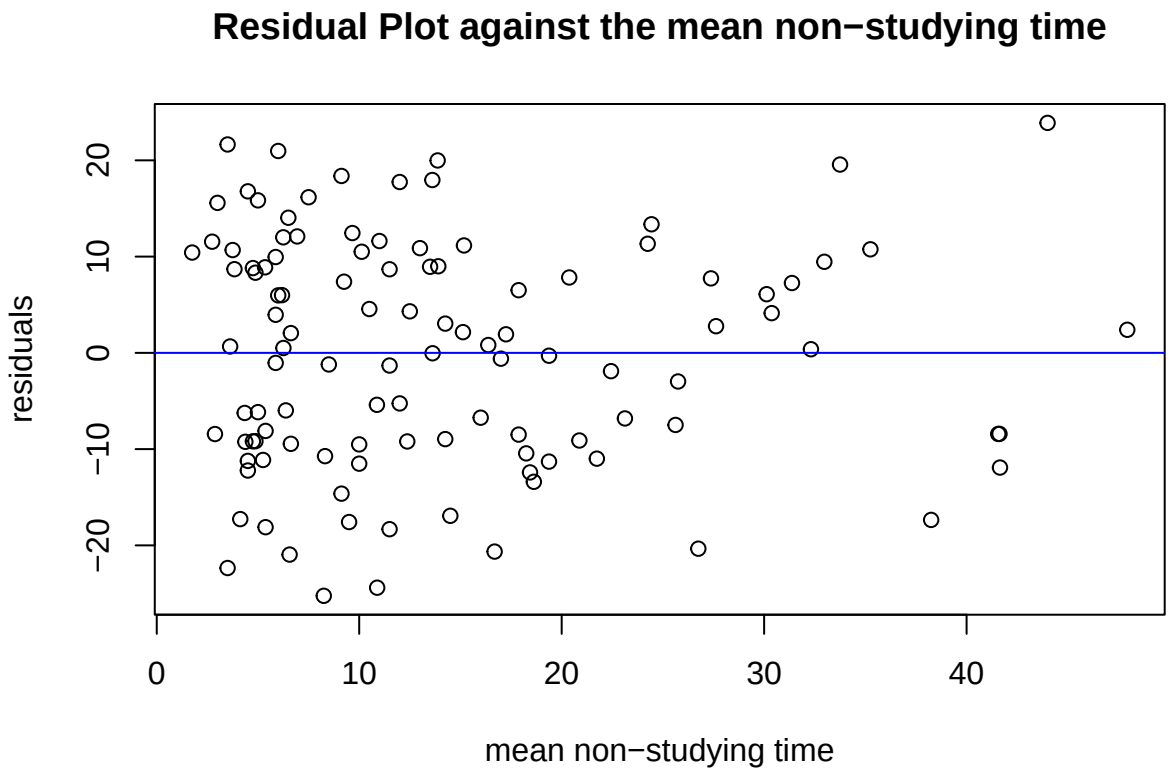
From the summary of our model, we obtain $\hat{\beta}_1$=-0.1294 and standard error=0.107
Then, the test statistic will be -0.1294/0.107=-1.21, and p-value is 0.229, and we should not reject our null hypothesis, which means $\beta_1 = 0$

In this case, we can conclude that there is no linear association between term test scores and the average time of a student spending on diverse things (not related to studying) in daily lives from week one to week four.

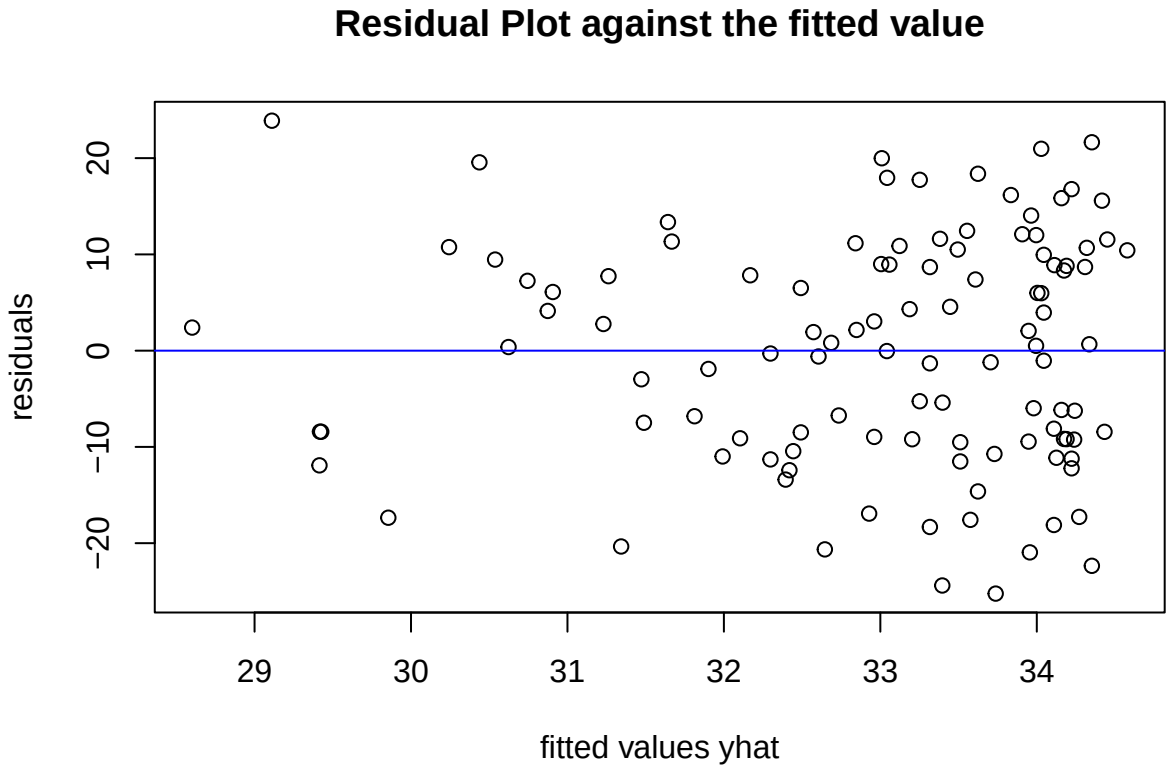
Model Diagnoistic

I will check 4 assumptions.
I will utilize the plot of residuals against the mean non-studying time, the plot of the residuals against the fitted value, and the normal Q_Q plot.

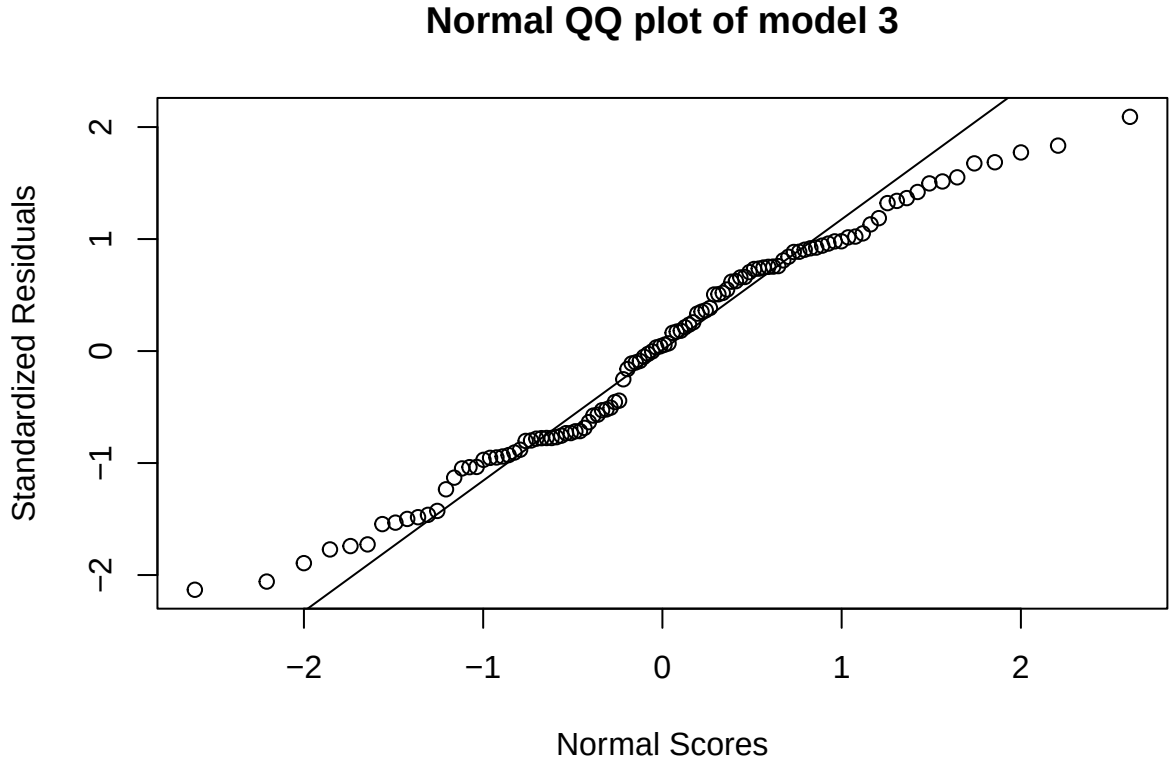
1)Plot of the Residuals against the Mean Non-Studying Time



2) Plot of the Residuals against the Fitted Value



3) Normal Q_Q plot



Assumption 1: Linearity

We know the pattern is not random according to the plot of residuals vs mean non-studying time. So there is a non-linear trend for predictor variable mean_nonstudying and the residuals, indicating a non-linear relationship between mean non-studying time and term test score.

Assumption 2: Independece of Errors

According to the residual vs fitted values plot of model3, there is no cluster of residuals. And we know the data are randomly collected. Thus, the assumption that there is no relationship between the residuals and the variable is not violated.

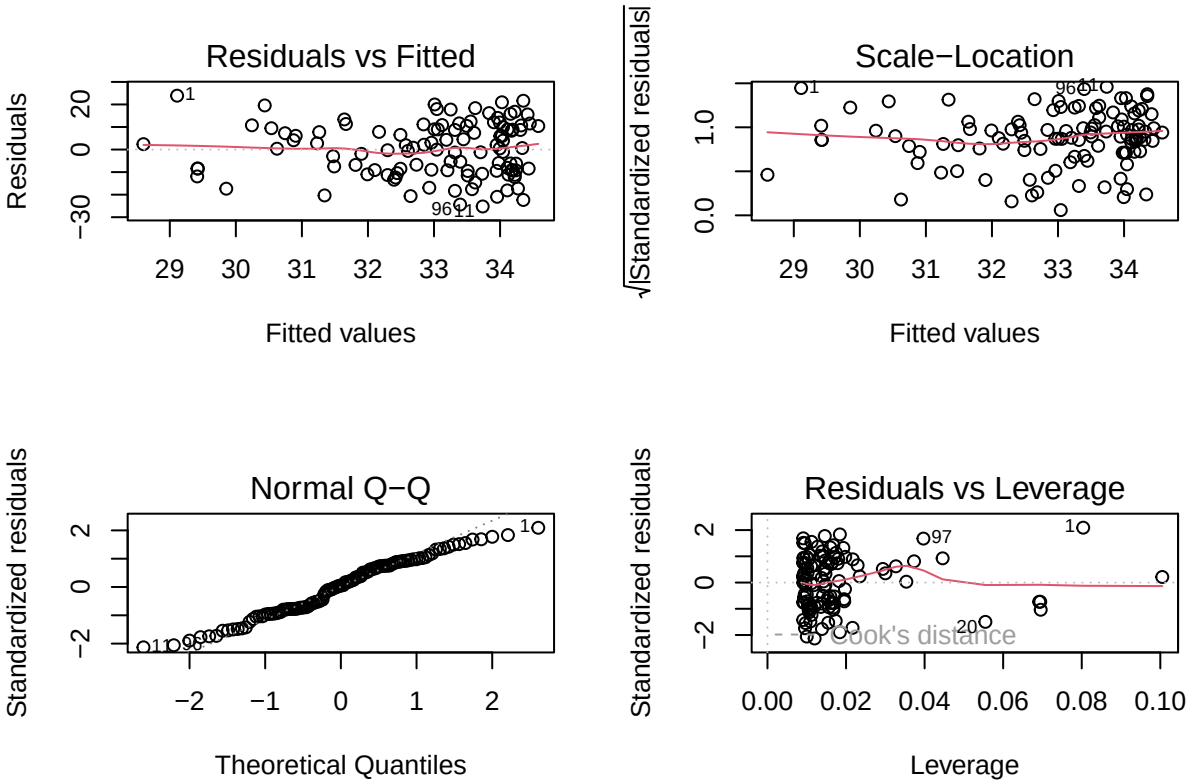
Assumption 3: Constant Variance of Errors(Homoscedasticity)

Since there is no fanning pattern in the residual vs fitted values plot of model3, and as the fitted values become larger, the variance does not increase. Therefore, the residuals have a constant variance.

Assumption 4: Normality of Errors

Based on the Normal Q-Q plot, we observed both heavy left and right tails with some observations that have large residuals. However, the distribution still looks symmetric, and thus the assumption of normality of errors is satisfied.

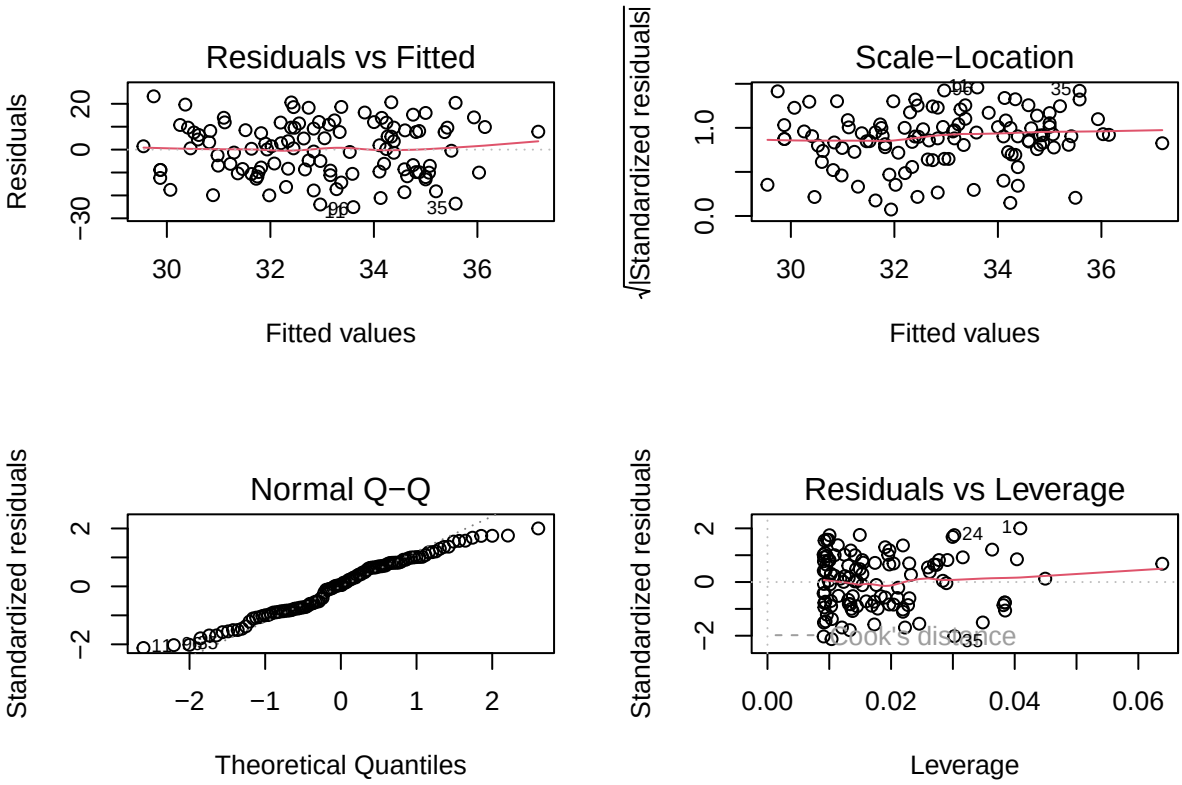
4)Checking for Influential Obeservations and Outliers



Based on the residuals vs. leverage plot, we can see that observations 1, 20, and 97 are leverage points(outlier on X). Moreover, since there are no points out of the Cook's distance boundary, indicating there are no influential points exist under the final model with just mean non-studying time predictor.

5) Any Transformation Necessary

Because of the assumption that errors' independence, homoscedasticity and normality are satisfied, while the linearity assumption is violated, the transformation on X will be considered to be adopted.



After transformation on X, the new model with log(mean_nonstudying) variable would improve the linearity problem according to the residuals vs fitted plot.

```
Rsq_adj_transformation <- summary(model_transformation)$adj.r.squared
Rsq_adj_transformation
```

```
## [1] 0.01182917
```

Here we obtain the R^2_{adj} is 0.0118 for the model after transformation, which is much higher than the R^2_{adj} we obtained for our original model (0.0042). So here are 1.18% of variations can be explained by our new model.

6) T-test after Transformation

	value
$\hat{\beta}1$	-2.304398
standard_error β_1	1.517888
test_statistic	-1.518162
p_value	0.1318956

Our null hypothesis: there is no linear relation between term test scores and the average time of a student thinking about things not related to studying

$$H_0 : \beta_1 = 0,$$

Our alternate hypothesis: there exists a linear relation between term test scores and the average time of a student thinking about things not related to studying

$$H_a : \beta_1 \neq 0$$

T test of β_1
After we apply the transformation to our model, we obtain the p-value is 0.132, so we still fail to reject our null hypothesis, which means there is no linear association between term test scores and the average time of a student spending on diverse things (not related to studying) in daily lives from week one to week four.

Conclusion

1) Purpose of Final Model

We want to have a model that potentially reveals the pattern between term test grade and mean of studying time, mean of non-studying time, and attendance of office hours. If possible, the model needs to both fit and predict term test grades based on the subset of the predictor variables. What is more, if we find an appropriate model and understand the relationship between variables like mean studying time and term test score, we will be able to find the recommended mean studying time each week, which will result in a high score.

2) Interpretation of Selected Model and Remaining Limitations

In this study, we select the model about the relation between term test scores and the time spend on things not related to studying, and we adjust this model to fix the linearity after transformation. But unfortunately, we do not believe the model we build is meaningful: because we have no evidence to show that there is a linear relation between time spend on things not related to studying and term test scores by t-test.

The limitation of this model starts from the fundamental issues with how the experiment was conducted. Recall that the surveys are worth credits in the course, and therefore, it is likely that students simply finish the surveys only for the sake of finishing learning task. Also, the predictor variables selected are not as much traceable as the instructor wish them to be.

There should be more variables that are reasonable and traceable, such as setting up a course website for learning using the slides and course notes with timer attached on it. The survey should also be done in a daily basis which will not be a big issue considering the workload for such a small task. Also, the amount of time possible to be designated to this project is rather limited due to the schedule of final exam. If more time allowed, a more generalized linear model and non-linear model will probably yield a better reasonable model for both fitting and predicting.

What is more, we think office hour is not a good explanatory variable, because firstly, our studying styles are changed due to COvid, which means students tend to working independently at home; secondly, Piazza is a good platform for every student solving their inquiries, thus, students can improve their learning without attending office hour. Thirdly, people who receive high score are definitely good at this course. Under this circumstance, there might be no need for them to go to the office hour.

3) Potential Improvements

One potential way to improve our model is considering using variables more wisely. Firstly, we do not include a variable “Familiarity” (How comfortable did a student feel about the course material?), because we do not think this variable has some relation with term test scores. Also, maybe we are supposed to use interaction between different variables to build more models as candidate models. Therefore we may find a better model to estimate the term test score by some predictors. Secondly, we should consider re-arranging categorical variables in different way. For example, there are 4 responses for a categorical variable “OH” (How often did a student attend OH per week?), and we re-arrange this variable as “attend” (have joined the office hour) and “absent” (never join the office hour), now I think this rearrangement may not be very appropriate because the times of attending office hour is related to understanding the course material and we should only describe them as “attend” or “absent”.

Another potential improvement for our model is considering different regression model. Now we are only familiar with linear regression model, but what if the real relation we want to find is a non-linear regression model such as a polynomial regression model? Therefore, if we can use a different type of regression model (non-linear) to fit our data, then we may obtain a more precise and meaningful model to estimate the term test scores more accurate.

Appendix

```
finaldata1 <- finaldata %>%
  mutate(mean_studying = (Studying + Studying2 + Studying3 + Studying4)/4,
                         mean_covid = (COVID + COVID2 + COVID3 + COVID4)/4,
                         mean_miscellaneous = (Miscellaneous + Miscellaneous2 +
                                                 Miscellaneous3 + Misceallenous4)/4)
finaldata2 <- finaldata1 %>% mutate(mean_nonstudying =
                                      (mean_covid + mean_miscellaneous)/2 )
finaldata3 <- finaldata2 %>% mutate(OH_attendence =
                         case_when(OH == 'At least once a week' ~ 'Participated',
                         OH == 'Once a week' ~ 'Participated',
                OH == 'Less than one time a week (on average)' ~  'Participated',
                                                        OH == 'Never' ~ 'Absent'))
```

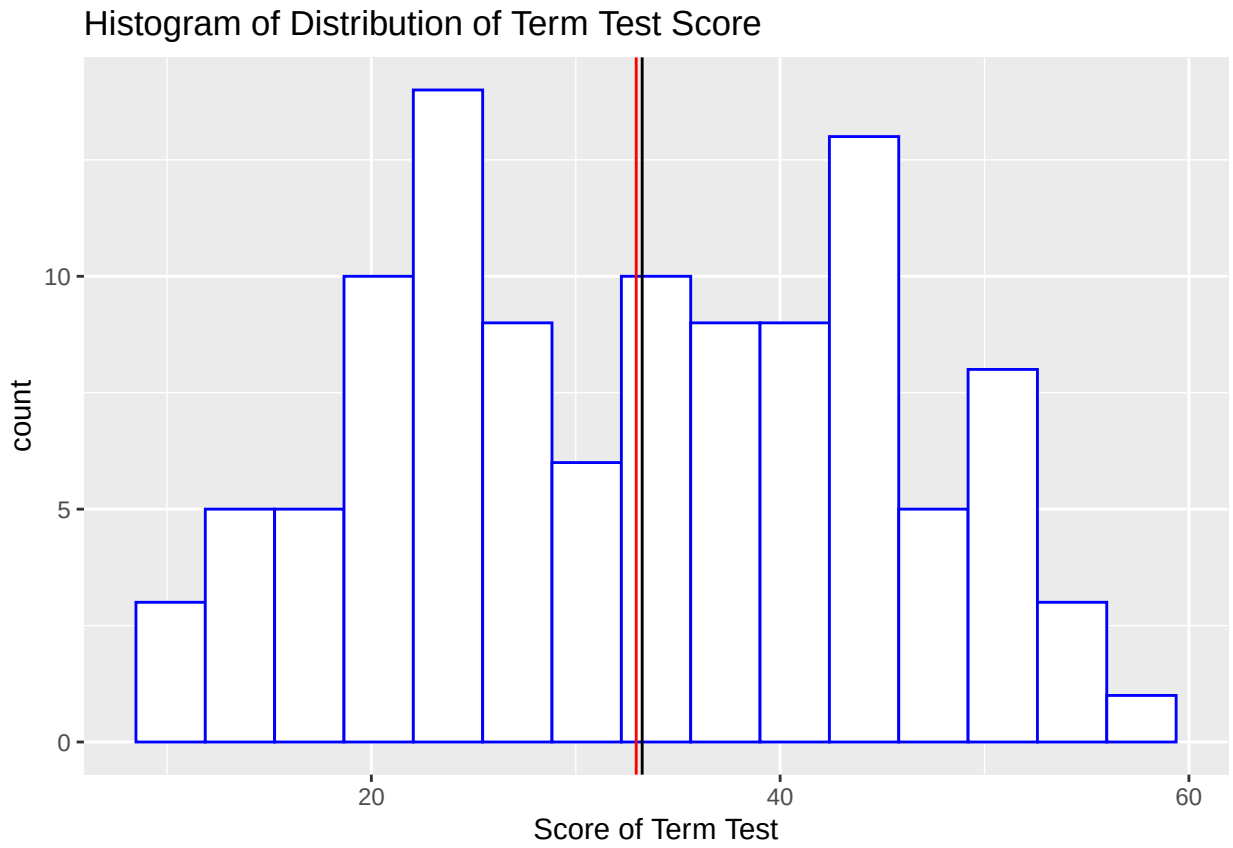
```
r1<-cor(finaldata3$mean_studying,finaldata3$mean_nonstudying)
r2<-cor(finaldata3$mean_nonstudying,finaldata3$mean_studying)
VIF1 <- 1/(1-r1^2)
VIF2 <- 1/(1-r2^2)
VIF_mean <- mean(VIF1, VIF2)
```

```
VIF_mean
```

```
## [1] 1.086629
```

```
finaldata3 %>% ggplot(aes(x = Term.Test))+
  geom_histogram(fill = 'white',
                 color = 'blue',
                 bins = 15)+
  geom_vline(xintercept=mean(finaldata3$Term.Test), color="red")+
  geom_vline(xintercept=median(finaldata3$Term.Test), color="black")+

  labs(x = "Score of Term Test",
       y = "count",
       title = "Histogram of Distribution of Term Test Score")
```




```
mean(finaldata3$Term.Test)
```

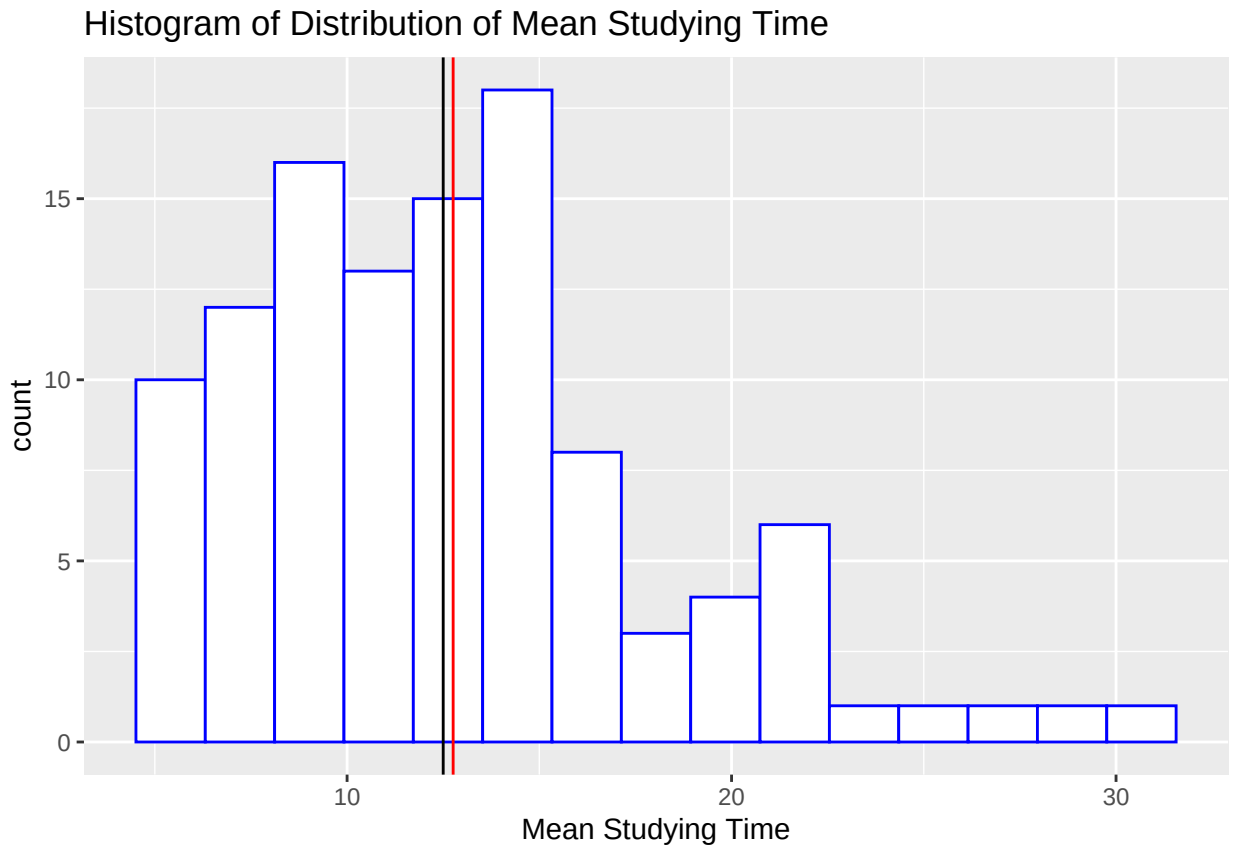
```
## [1] 32.95909
```

```
median(finaldata3$Term.Test)
```

```
## [1] 33.25
```

```
finaldata3 %>% ggplot(aes(x = mean_studying))+
  geom_histogram(fill = 'white',
                 color = 'blue',
                 bins = 15)+
  geom_vline(xintercept=mean(finaldata3$mean_studying), color="red")+
  geom_vline(xintercept=median(finaldata3$mean_studying), color="black")+

  labs(x = "Mean Studying Time",
       y = "count",
       title = "Histogram of Distribution of Mean Studying Time")
```



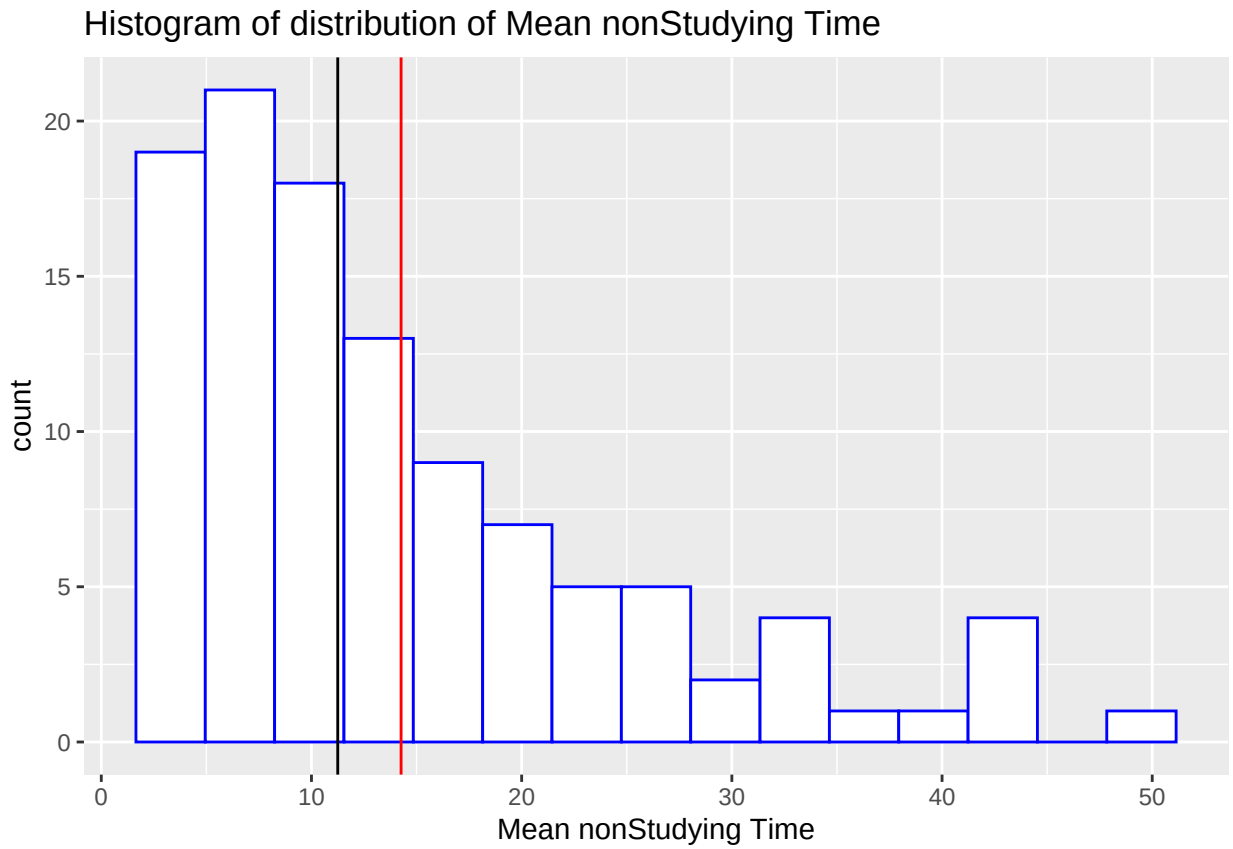

```
mean(finaldata3$mean_studying)
```

```
## [1] 12.75909
```

```
median(finaldata3$mean_studying)
```

```
## [1] 12.5
```

```
finaldata3 %>% ggplot(aes(x = mean_nonstudying))+
  geom_histogram(fill = 'white',
                 color = 'blue',
                 bins = 15)+
    geom_vline(xintercept=mean(finaldata3$mean_nonstudying), color="red")+
  geom_vline(xintercept=median(finaldata3$mean_nonstudying), color="black")+
  labs(x = "Mean nonStudying Time",
       y = "count",
       title = "Histogram of distribution of Mean nonStudying Time")
```



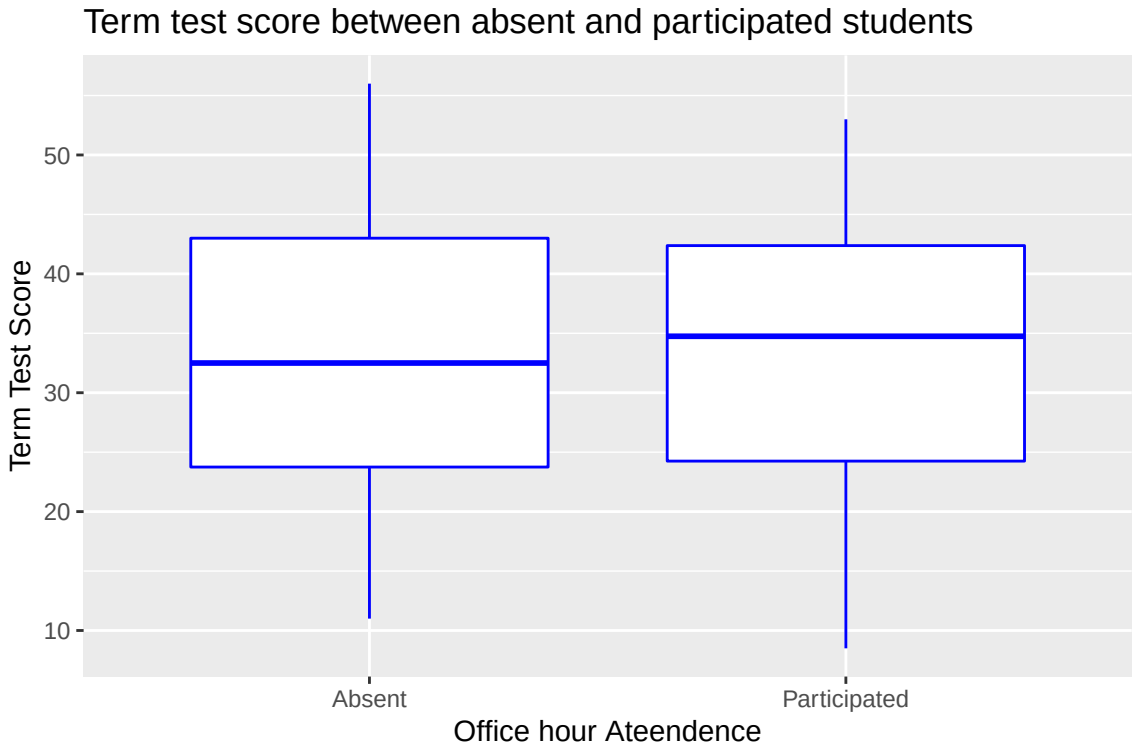

```
mean(finaldata3$mean_nonstudying)
```

```
## [1] 14.26312
```

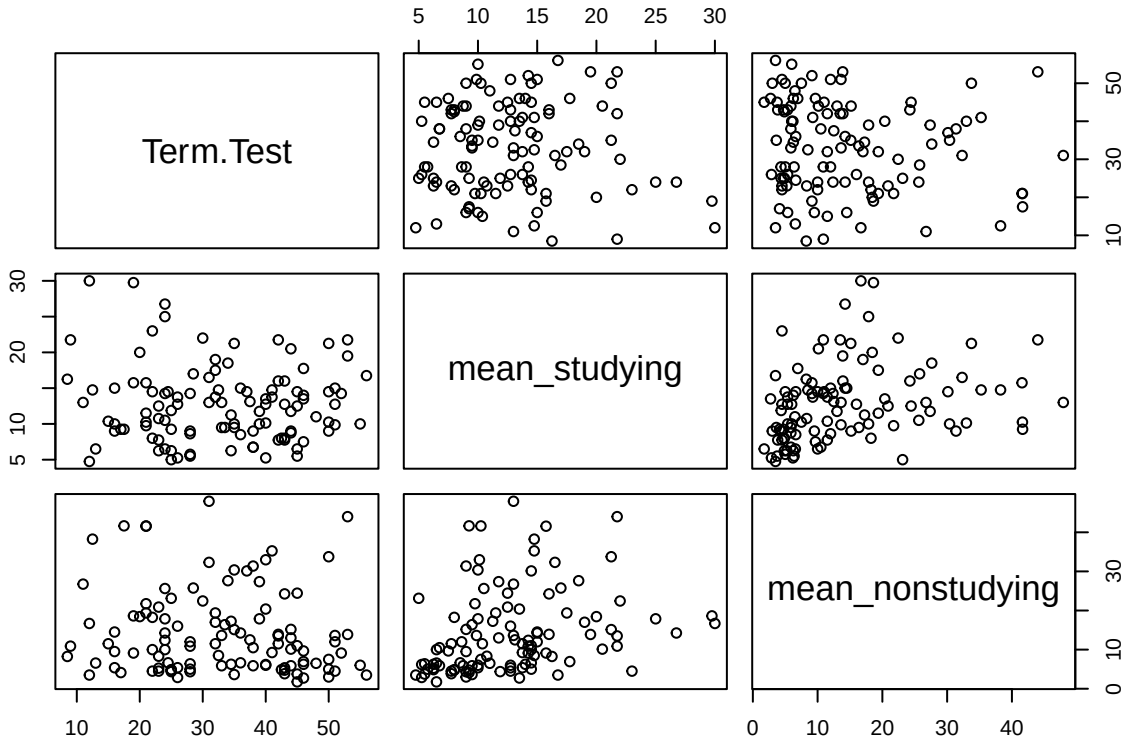
```
    median(finaldata3$mean_nonstudying)
```

```
## [1] 11.25
```

```
finaldata3 %>% ggplot(aes(x = OH_attendence, y = Term.Test))+
  geom_boxplot(fill = 'white',
                color = 'blue',
                )+
  labs(x = "Office hour Ateendence",
       y = "Term Test Score",
       title = "Term test score between absent and participated students")
```




```
pairs(~Term.Test + mean_studying + mean_nonstudying, data = finaldata3)
```




```
model1 <- lm(Term.Test ~ 1, data = finaldata3)
model2 <- lm(Term.Test ~ mean_studying, data = finaldata3)
model3 <- lm(Term.Test ~ mean_nonstudying, data = finaldata3)
model4 <- lm(Term.Test ~ mean_studying +
               mean_nonstudying, data = finaldata3)
```

```r
model5 <- lm(Term.Test ~ mean_studying + OH_attendence, data = finaldata3)
model6 <- lm(Term.Test ~ mean_nonstudying +
               OH_attendence, data = finaldata3)
model7 <- lm(Term.Test ~ mean_studying +
               mean_nonstudying + OH_attendence, data = finaldata3)
```

```r
Rsq_adj1 <- summary(model1)$adj.r.squared
Rsq_adj2 <- summary(model2)$adj.r.squared
Rsq_adj3 <- summary(model3)$adj.r.squared
Rsq_adj4 <- summary(model4)$adj.r.squared
Rsq_adj5 <- summary(model5)$adj.r.squared
Rsq_adj6 <- summary(model6)$adj.r.squared
Rsq_adj7 <- summary(model7)$adj.r.squared
```

```r
n <- 110
p1 <- length(model1$coefficients) - 1
p2 <- length(model2$coefficients) - 1
p3 <- length(model3$coefficients) - 1
p4 <- length(model4$coefficients) - 1
p5 <- length(model5$coefficients) - 1
p6 <- length(model6$coefficients) - 1
p7 <- length(model7$coefficients) - 1
SSres1<-sum(model1$residuals^2)
SSres2<-sum(model2$residuals^2)
SSres3<-sum(model3$residuals^2)
SSres4<-sum(model4$residuals^2)
SSres5<-sum(model5$residuals^2)
SSres6<-sum(model6$residuals^2)
SSres7<-sum(model7$residuals^2)
AIC1<-n*log(SSres1)-n*log(n)+2*(p1+1)
AIC2<-n*log(SSres2)-n*log(n)+2*(p2+1)
AIC3<-n*log(SSres3)-n*log(n)+2*(p3+1)
AIC4<-n*log(SSres4)-n*log(n)+2*(p4+1)
AIC5<-n*log(SSres5)-n*log(n)+2*(p5+1)
AIC6<-n*log(SSres6)-n*log(n)+2*(p6+1)
AIC7<-n*log(SSres7)-n*log(n)+2*(p7+1)
```

```r
standard_error_beta_1=summary(model3)$coefficients[2,2]
beta_1_hat <- model3$coefficients[2]
p_value <- summary(model3)$coefficients[2,4]
test_statistic = summary(model3)$coefficients[2,3]
beta_1_hat
```

```
## mean_nonstudying
##       -0.1294114
```

```r
standard_error_beta_1
```

```
## [1] 0.1069579
```

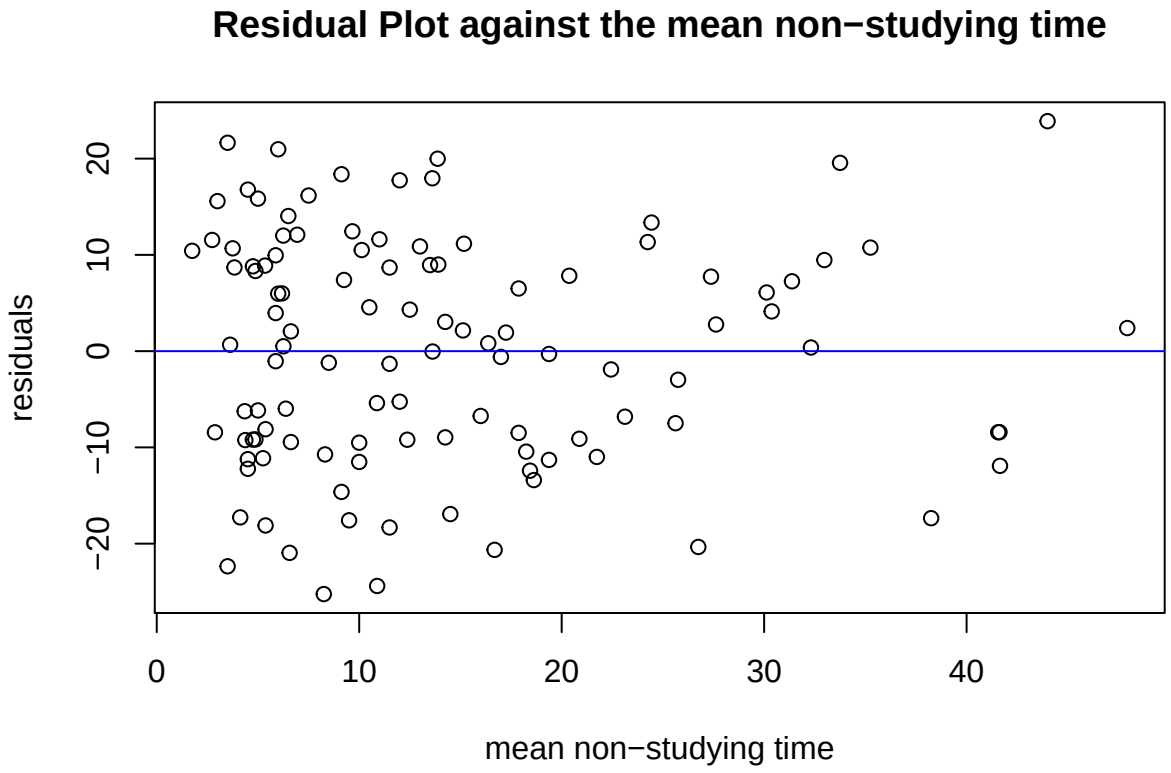
```r
test_statistic
```

```
## [1] -1.209929
```

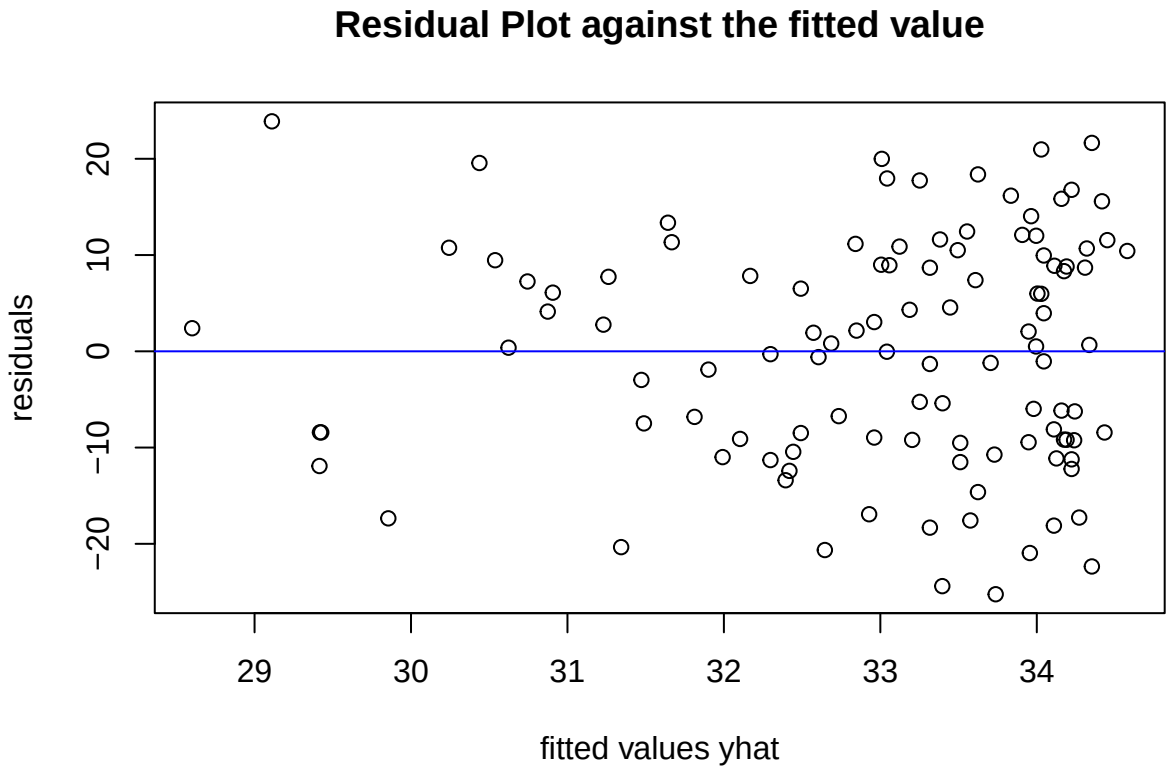
```r
p_value
```

```
## [1] 0.2289479
```

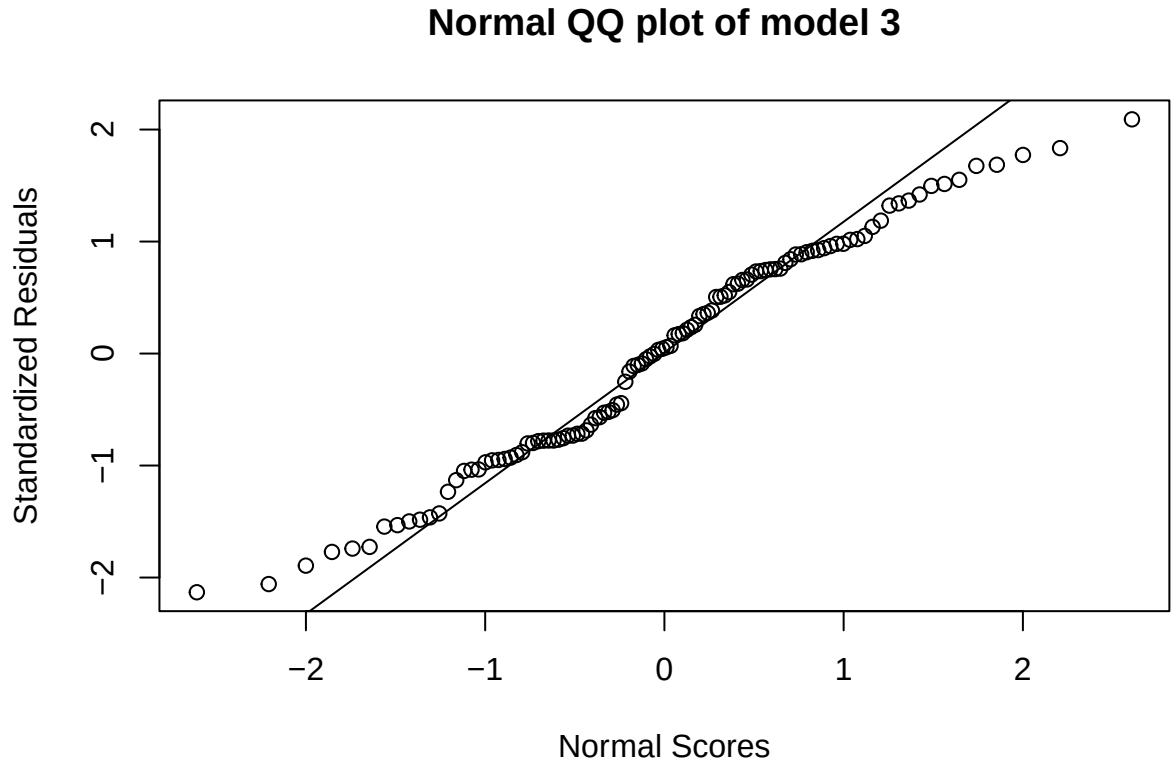
```
plot(finaldata3$mean_nonstudying,resid(model3),
     xlab = 'mean non-studying time',
     ylab = 'residuals',
     main = 'Residual Plot against the mean non-studying time')
abline( h = 0, col = 'blue')
```



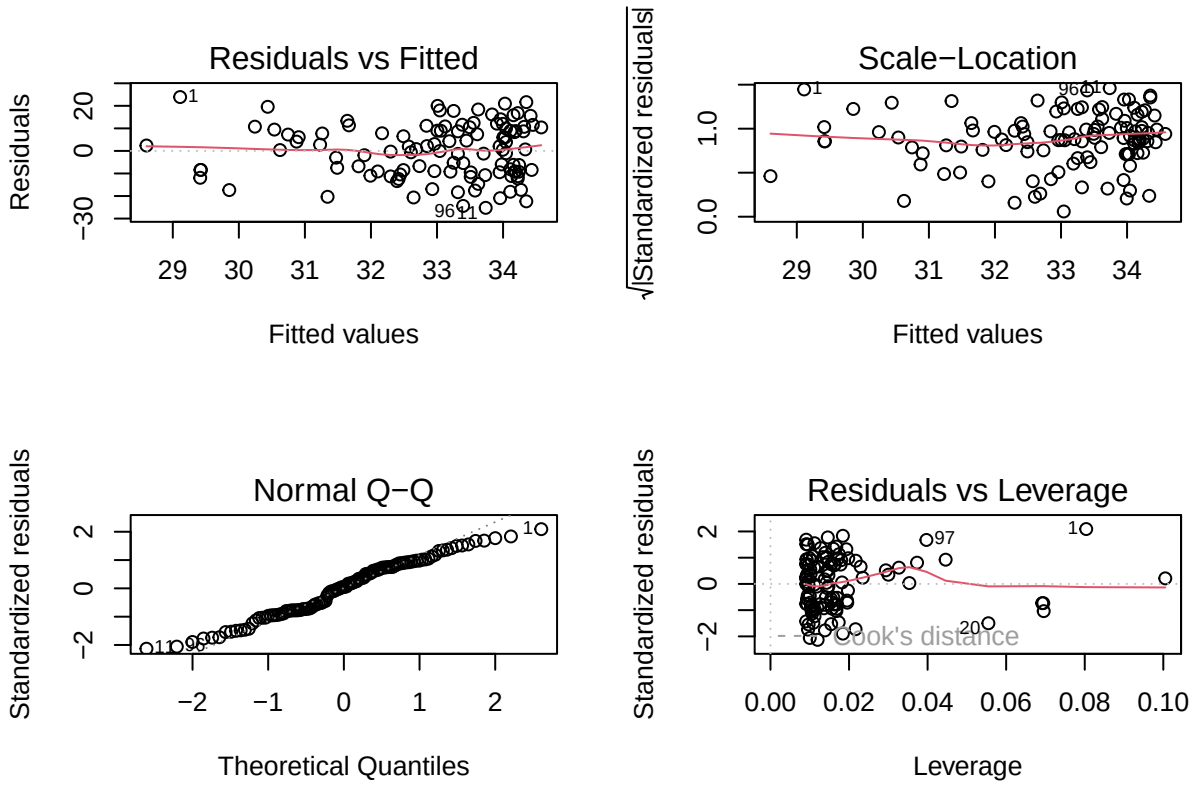

```
plot(fitted(model3),resid(model3),
     xlab = 'fitted values yhat',
     ylab = 'residuals',
     main = 'Residual Plot against the fitted value')
abline( h = 0, col = 'blue')
```




```
model3_check = rstandard(model3)
qqnorm(model3_check,
       xlab = "Normal Scores",
       ylab = "Standardized Residuals",
       main = "Normal QQ plot of model 3")
qqline(model3_check)
```




```
layout(matrix(c(1,2,3,4),2,2)) # yields 4 graphs/page
plot(model3)
```




```
Rsq_adj_transformation <- summary(model_transformation)$adj.r.squared
Rsq_adj_transformation
```

```
## [1] 0.01182917
```

```
standard_error_beta_1=summary(model_transformation)$coefficients[2,2]
beta_1_hat <- model_transformation$coefficients[2]
p_value <- summary(model_transformation)$coefficients[2,4]
test_statistic = summary(model_transformation)$coefficients[2,3]
beta_1_hat
```

```
## log((mean_nonstudying))
##               -2.304398
```

```
standard_error_beta_1
```

```
## [1] 1.517888
```

```
test_statistic
```

```
## [1] -1.518162
```

```
p_value
```

```
## [1] 0.1318956
```