STA302H1F Final project

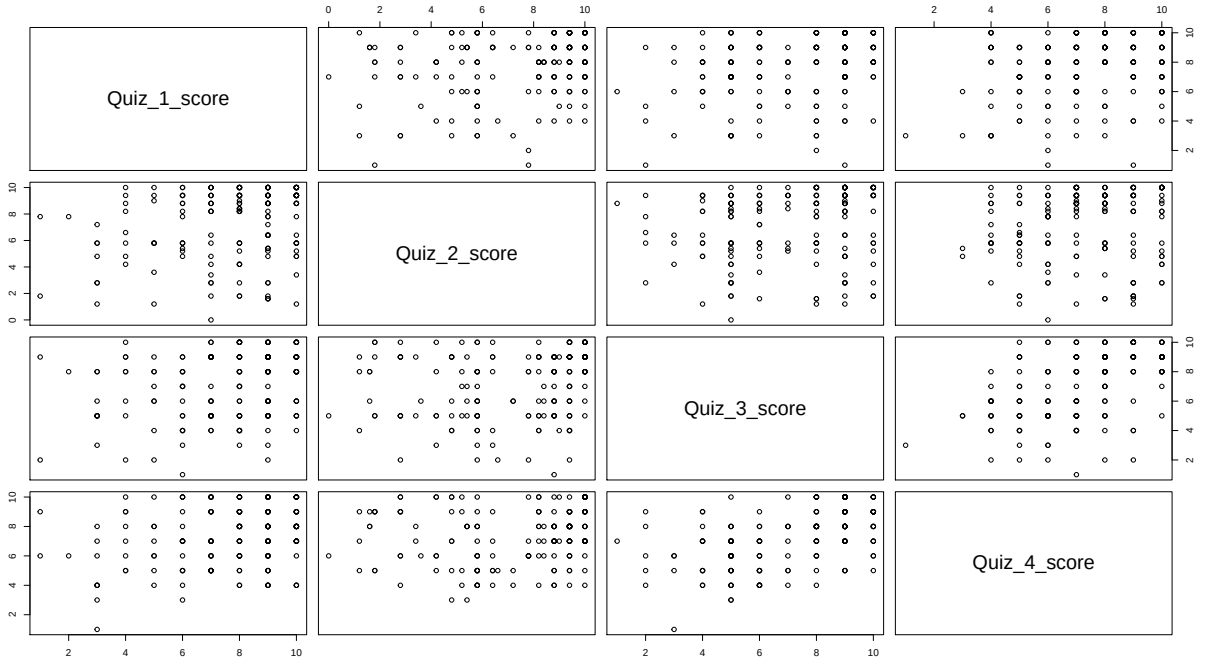
Xinhao Hou 1005426549 Yunkai Fan 1005088584

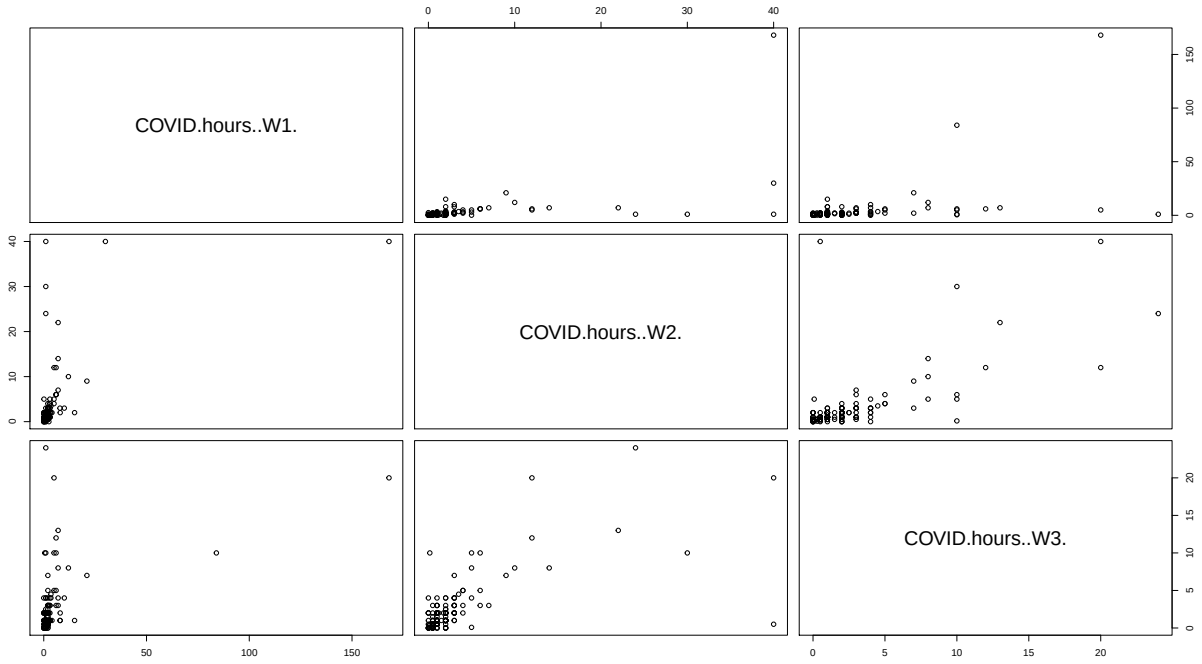
Introduction

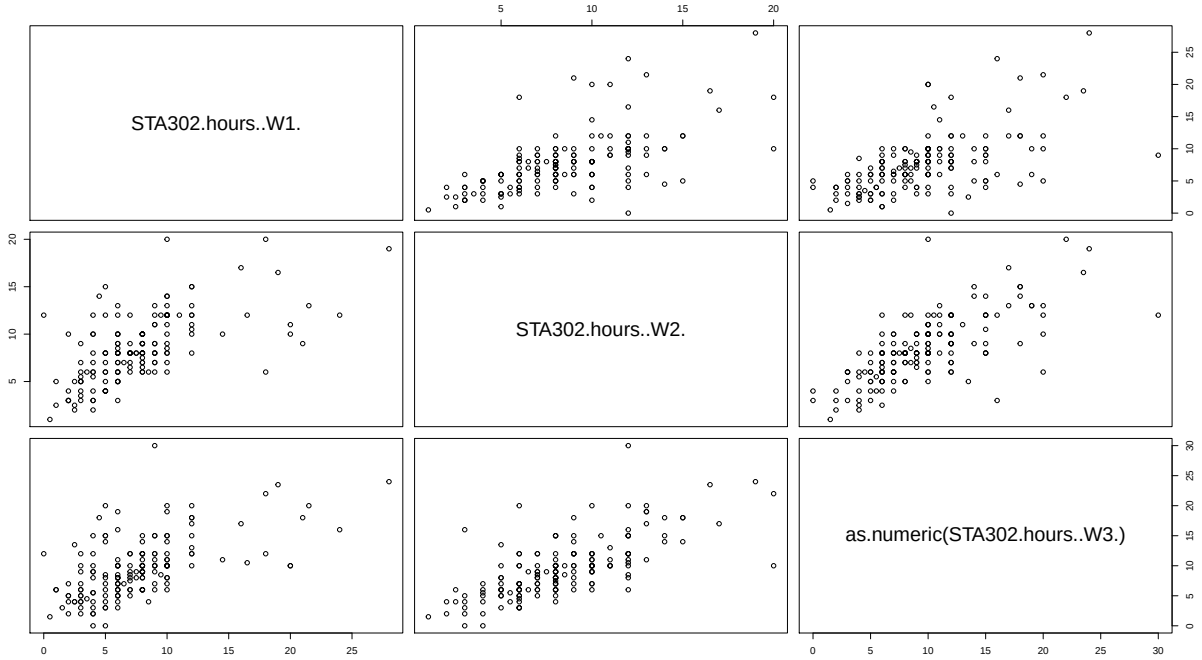
We are trying to predict the Quiz 4 score using data collected from the first 3 weeks of the course. We've collected the amount of hours studied for STA302, the amount of hours each student thought about COVID and the marks of the first 3 quizzes. We're going to build several experimental models and see which data columns would have a reasonably strong relationship with the Quiz 4 score. Then we will build a new linear model with countries isolated to see if the ways each country deal with COVID effect the correlation between our variables.

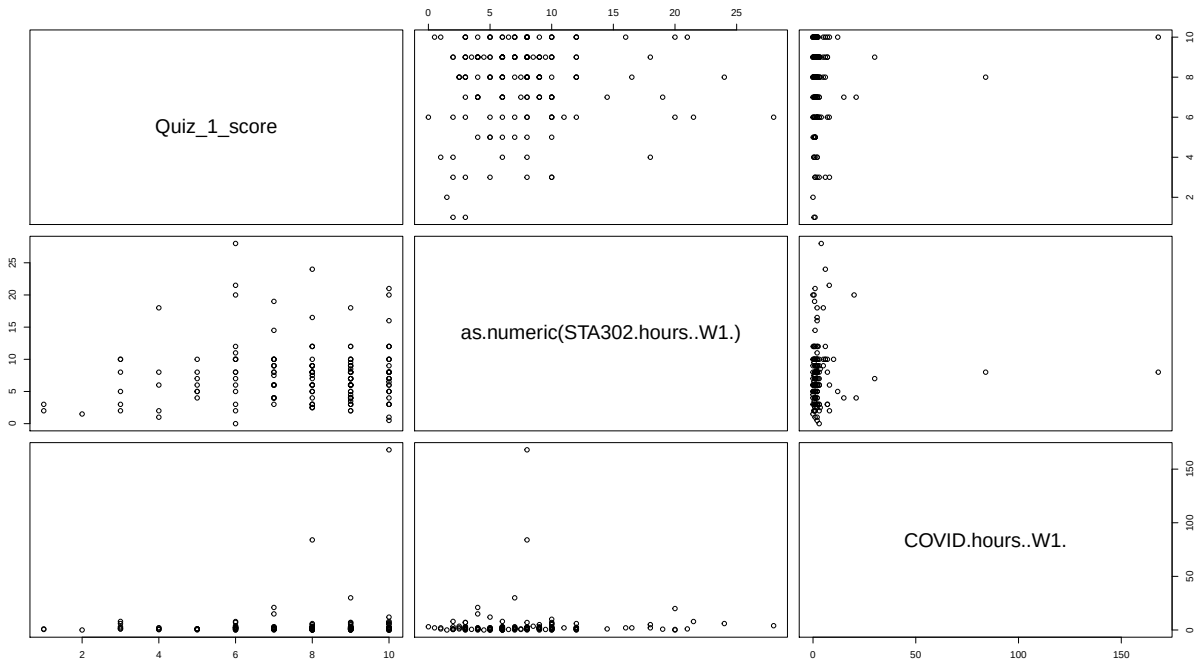
Exploratory data analysis

First we wanted to look at the variables to see if there are any obvious correlation between them.



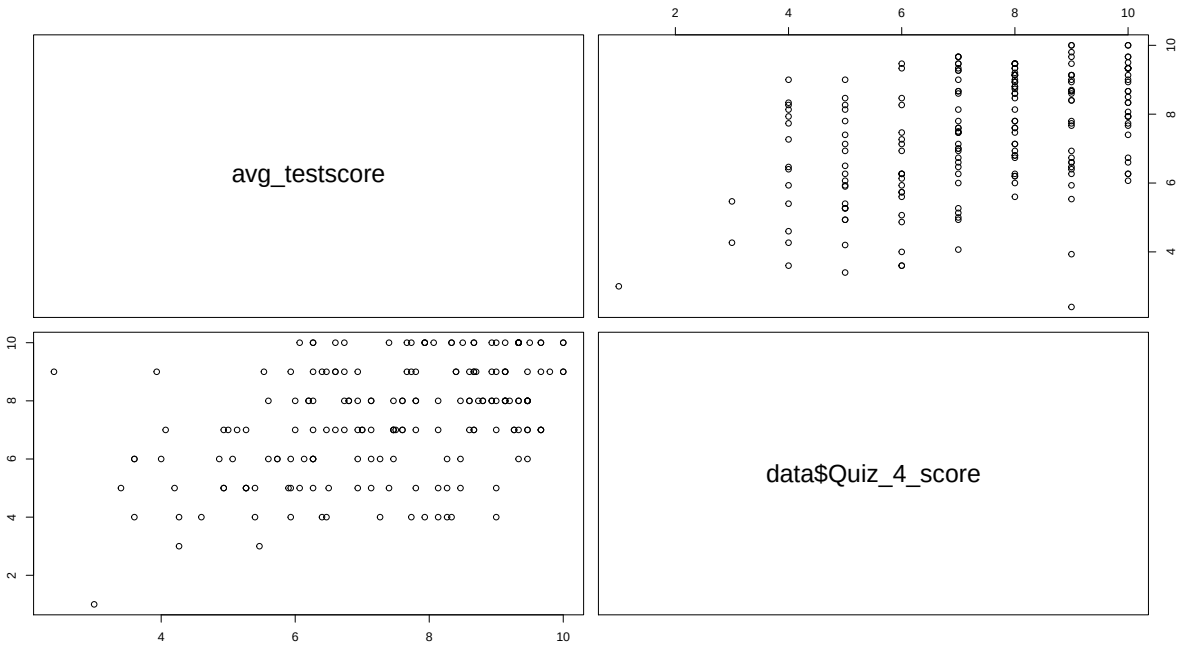

COVID.hours..W1.
COVID.hours..W2.
COVID.hours..W3.

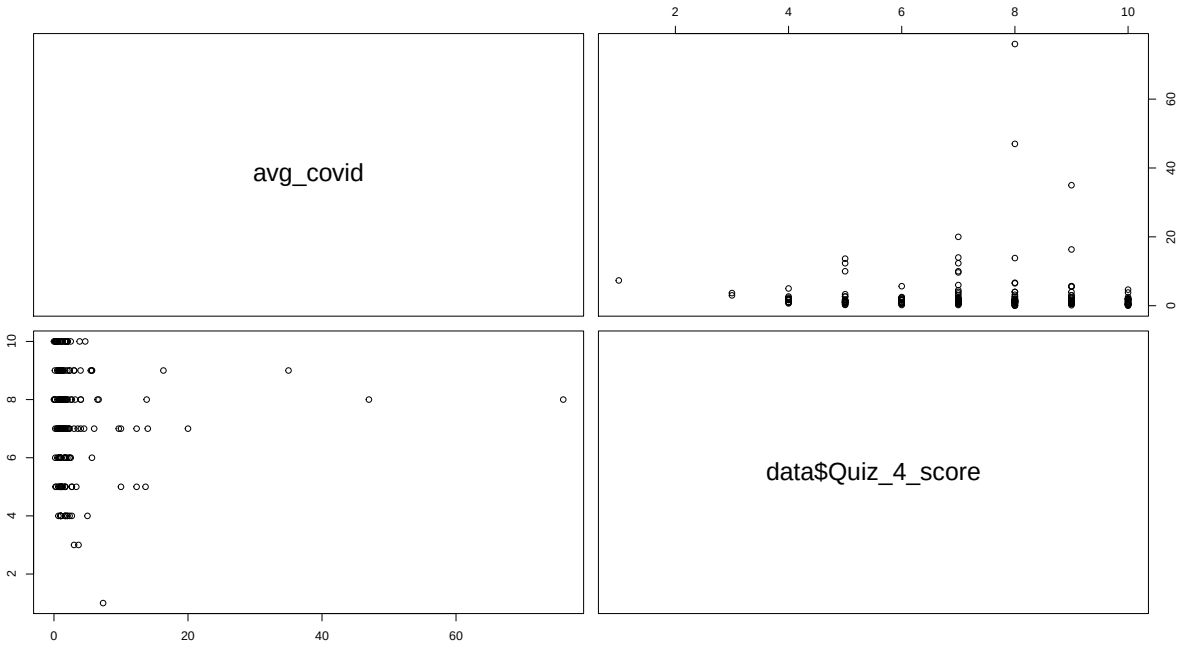

STA302.hours..W1.
STA302.hours..W2.
as.numeric(STA302.hours..W3.)

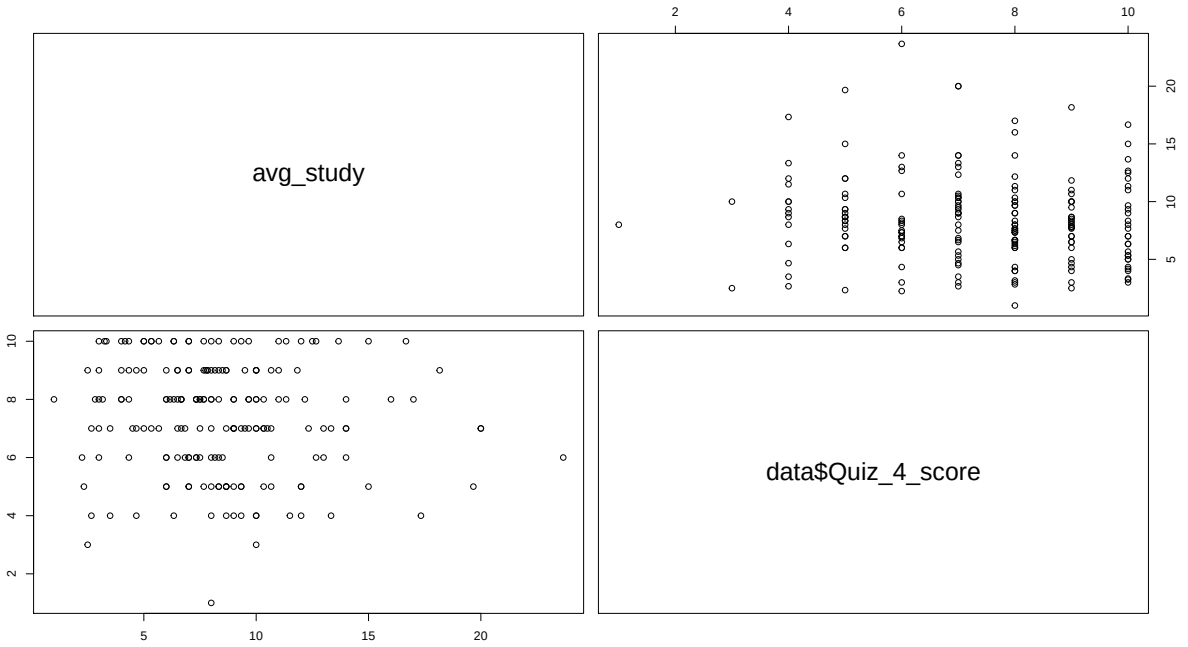


As we can see there no obvious correlation between the quiz scores. There is a positive correlation between the number of hours spend studying STA302 and the number of hours thinking of COVID. This means if a student spend a lot of time thinking about COVID it is likely that student will also spending a lot of time thinking about COVID the next week.

Looking at the correlation graph doesn't show me any obvious distribution between the variables. We decided to calculate an average from the data collected in the first 3 weeks of class to see which variables have a strong correlation with the quiz 4 score.







After calculating the averages we decided to create some experimental linear models to see which variable have a strong correlation.

```
##
## Call:
## lm(formula = data$Quiz_4_score ~ avg_covid)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -6.348 -1.337  0.608  1.662  2.665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.334561   0.152864  47.981   <2e-16 ***
## avg_covid   0.001887   0.019237   0.098    0.922
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.929 on 186 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  5.176e-05,  Adjusted R-squared:  -0.005324
## F-statistic: 0.009627 on 1 and 186 DF,  p-value: 0.9219
```

```
##
## Call:
## lm(formula = data$Quiz_4_score ~ avg_testscore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1769 -1.2184 -0.0561  1.2813  4.2416
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    3.51535    0.55948   6.283 2.21e-09 ***
## avg_testscore  0.51795    0.07321   7.075 2.80e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.719 on 190 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.2085, Adjusted R-squared:  0.2043
## F-statistic: 50.05 on 1 and 190 DF,  p-value: 2.799e-11
```

```
##
## Call:
## lm(formula = data$Quiz_4_score ~ avg_study)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.3607 -1.3975  0.3898  1.5984  3.0648
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.75342    0.34553  22.439   <2e-16 ***
## avg_study   -0.04909    0.03755  -1.307    0.193
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.92 on 186 degrees of freedom
##   (5 observations deleted due to missingness)
## Multiple R-squared:  0.009107,   Adjusted R-squared:  0.00378
## F-statistic: 1.709 on 1 and 186 DF,  p-value: 0.1927
```

Model development

After creating the linear models we found that the time spend studying STA302 and the time spend thinking about COVID have a very weak correlation with quiz 4 score. But we found the average quiz score have a moderately high correlation with quiz 4 score. With this information we decided to see the correlation between quiz 4 score and the scores from week 1 to 3.

```
##
## Call:
## lm(formula = data$Quiz_4_score ~ data$Quiz_1_score + data$Quiz_2_score +
##     data$Quiz_3_score)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8666 -1.0351  0.1492  1.1382  3.6500
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.25380    0.56389   5.770 3.61e-08 ***
## data$Quiz_1_score   0.09121    0.06004   1.519    0.131
## data$Quiz_2_score   0.02976    0.04677   0.636    0.525
## data$Quiz_3_score   0.43879    0.05690   7.711 9.60e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.563 on 172 degrees of freedom
##   (17 observations deleted due to missingness)
## Multiple R-squared:  0.317,  Adjusted R-squared:  0.3051
## F-statistic: 26.61 on 3 and 172 DF,  p-value: 3.444e-14
```

We see the quiz 3 score has the highest correlation and the quiz score of week 1 and 2 has a very low correlation to quiz 4. We decided to build a model with only the quiz 3 score to see if the correlation increases because we think because the correlation between the quiz 1 and quiz 2 score are so low they are essentially adding error to our current model.

```
##
## Call:
## lm(formula = data$Quiz_4_score ~ data$Quiz_3_score)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3775 -0.8904  0.2514  1.1483  4.0967
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.95481    0.40042   9.877  < 2e-16 ***
## data$Quiz_3_score   0.47423    0.05296   8.954 3.32e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.61 on 188 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.2989, Adjusted R-squared:  0.2952
## F-statistic: 80.17 on 1 and 188 DF,  p-value: 3.324e-16
```

As we have suspected removing quiz 1 and quiz 2 score from our linear model slightly decreased our standard error and this shows me that quiz 1 and quiz 2 score wasn't really helping our model.

We wanted to see if other variables from the same week as quiz 3 would help the model but we didn't anything significant. We also checked whether adding a quadratic term would help but it didn't improve the accuracy of the model produced.

We also wanted to see if the hours spend studying and thinking about COVID effected the quiz 4 scores but it didn't show any significant relationship.

Now we decided to isolate the categorical variable of which country the student is in. This is because different countries have different restrictions put in for COVID control.

```
##
##      canada      Canada       china       China       India       Japan
##           2          93           1          59           2           1
##    Mongolia    Pakistan   Singapore South Korea      Taiwan         UAE
##           1           3           2           2           2           2
##         USA
##           1
```

From this table we can see students in our class are mostly from Canada and China so we decided to build 2 models to see if my assumption of quiz 3 score having a strong correlation with quiz 4 score applies to students in both countries.

Both of these models shows a moderately strong correlation between quiz 3 scores. So we decided to perform a t-test with 95% confidence interval to see whether our estimator is significant.

```
qt(0.975, 56)
```

```
## [1] 2.003241
```

```
abststar_china = abs(fit_china$coefficients[2])/summary(fit_china)$coefficients[2,2]
abststar_china > qt(0.975, 56)
```

```
## data_china$Quiz_3_score
##                    TRUE
```

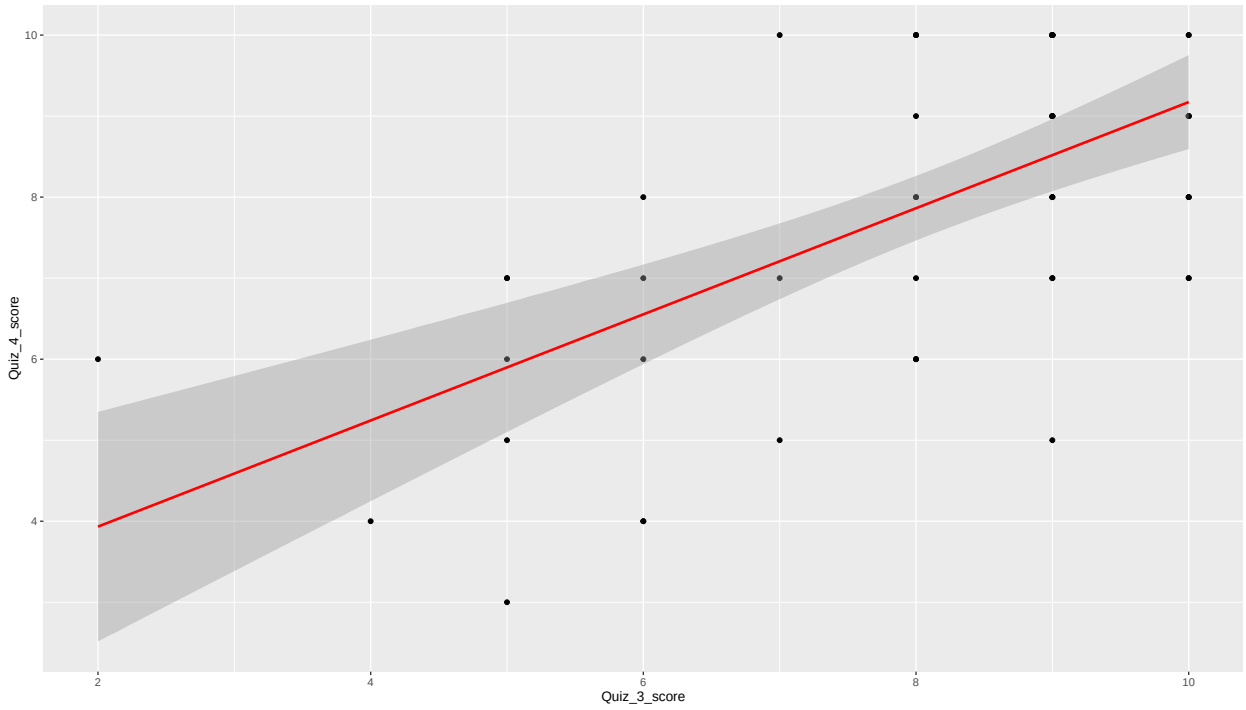
```
qt(0.975, 93)
```

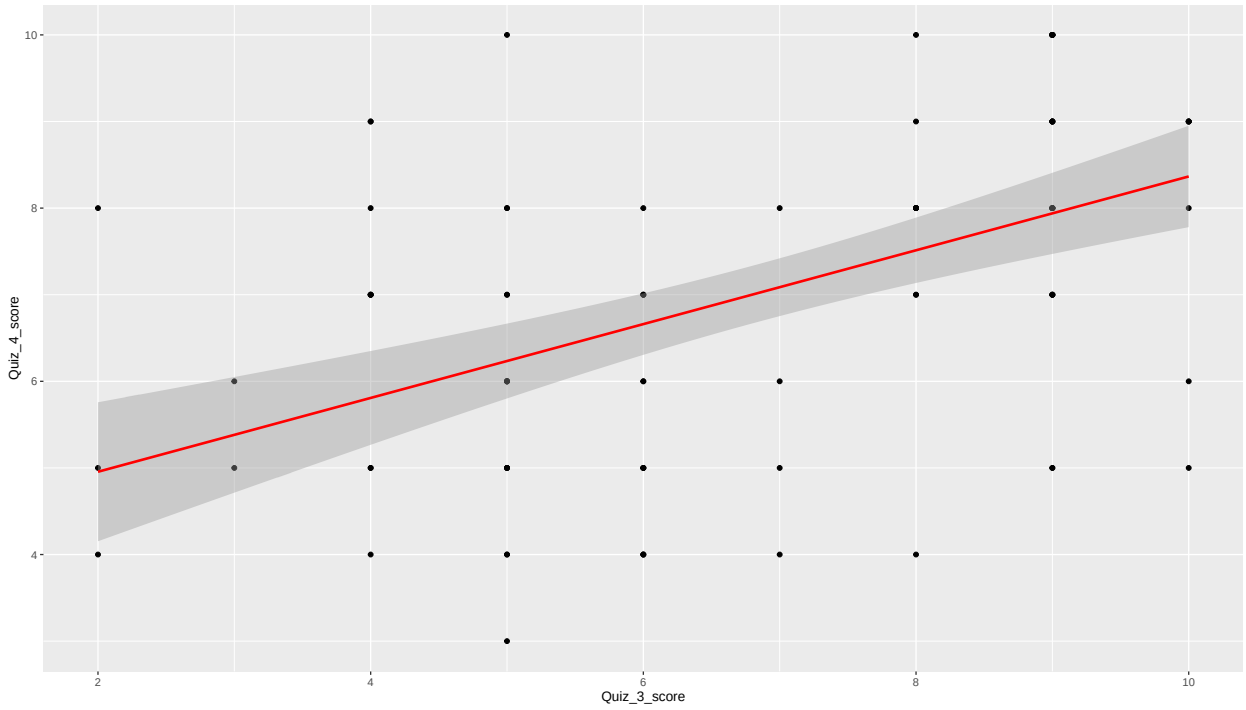
```
## [1] 1.985802
```

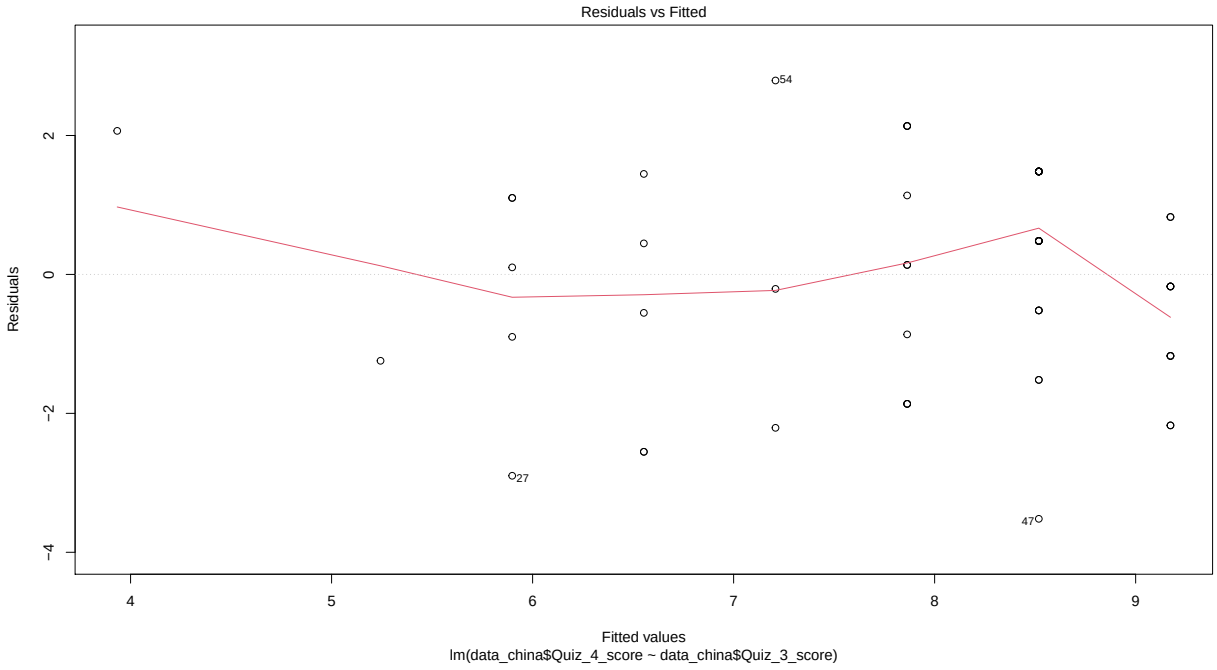
```
abststart_canada = abs(fit_canada$coefficients[2])/summary(fit_canada)$coefficients[2,2]
abststart_canada > qt(0.975, 93)
```

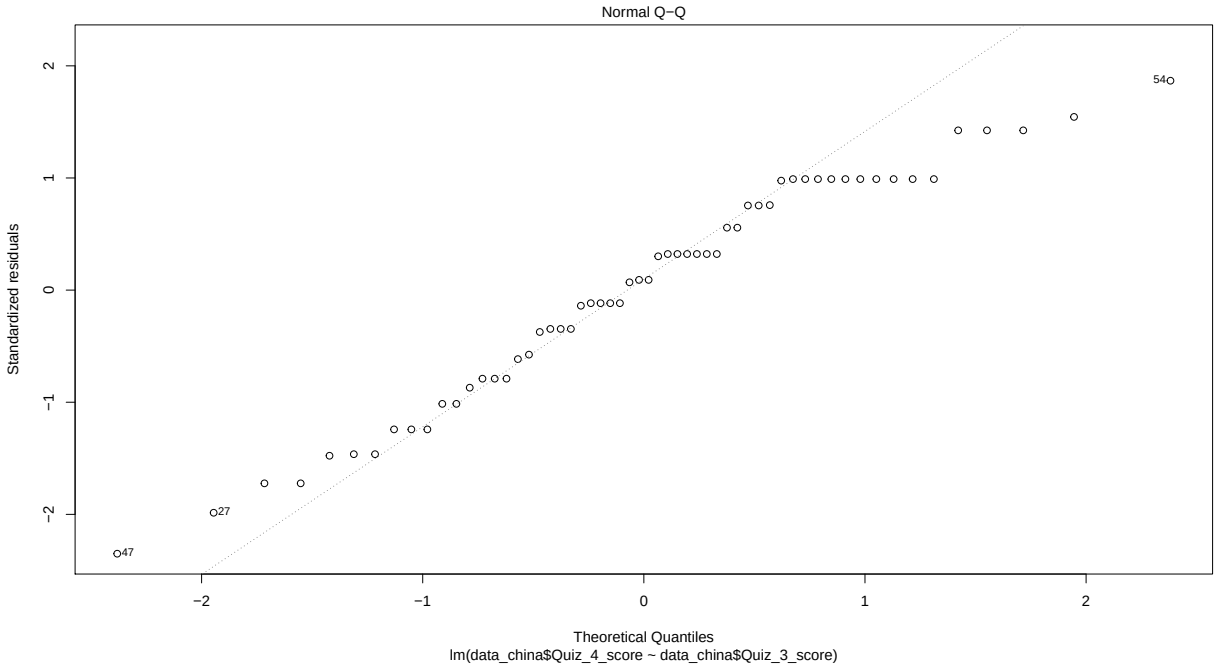
```
## data_canada$Quiz_3_score
##                     TRUE
```

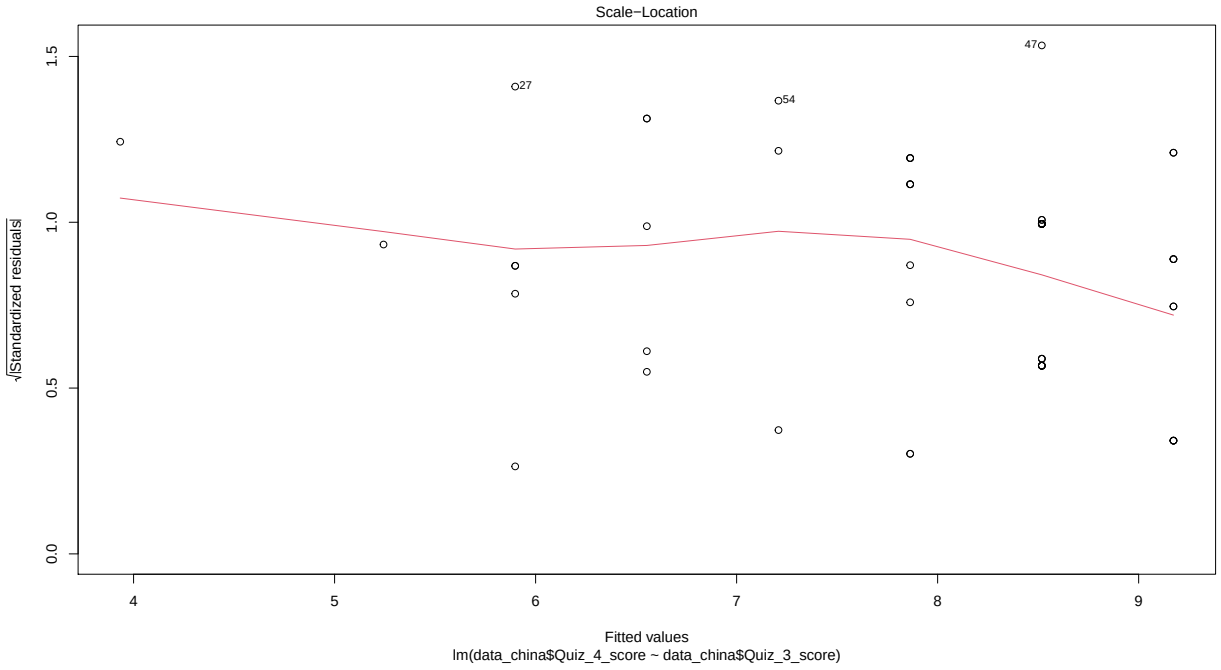
Both our models are proven to be statistically significant via the t-test. We decided to plot the data and see if there are any outliers effecting our model.

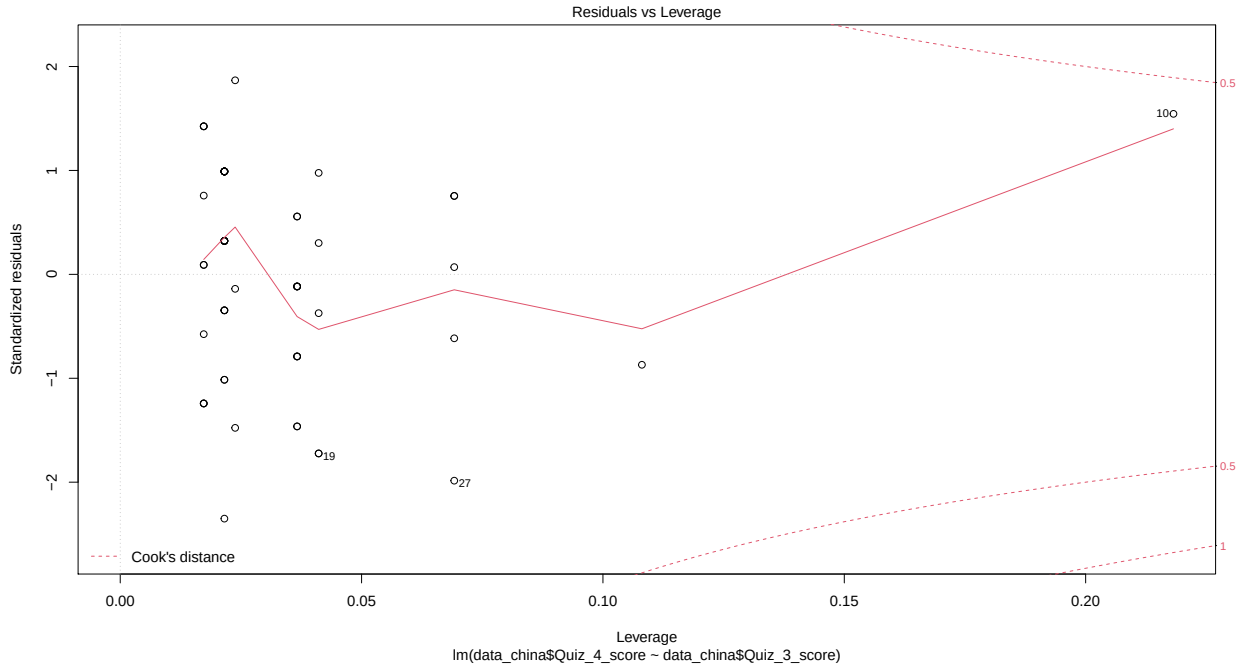

Quiz_4_score
10
8
6
4
2
4
6
8
10
Quiz_3_score

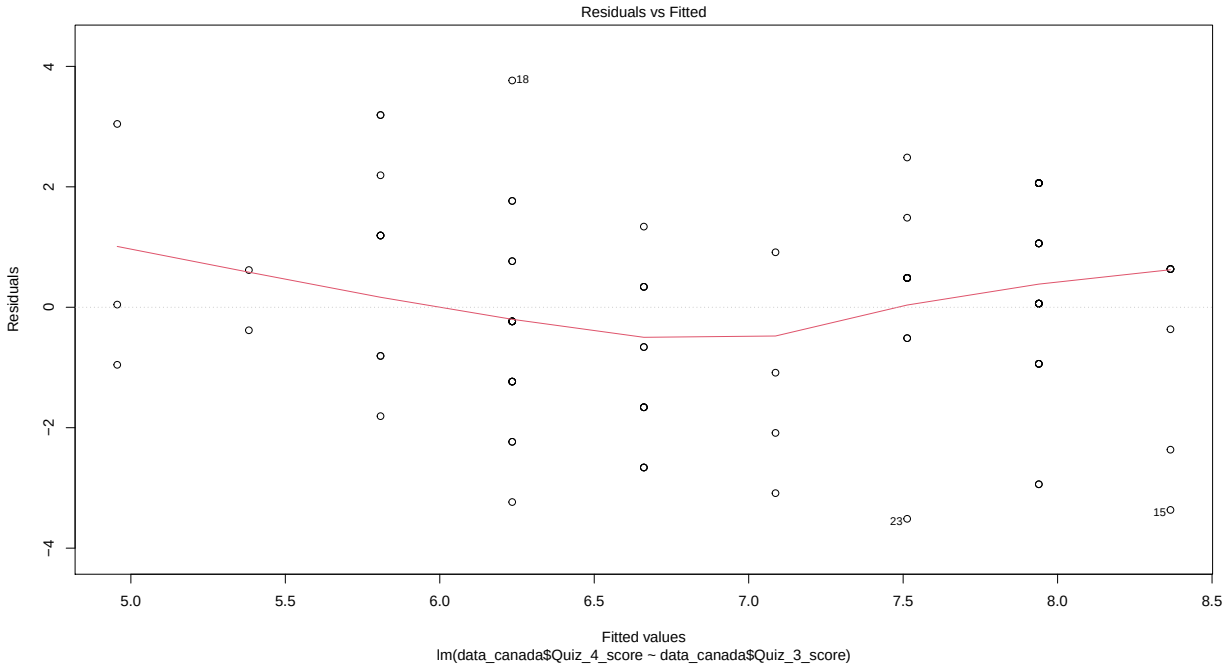

Quiz_4_score
10
8
6
4
2
4
6
8
10
Quiz_3_score

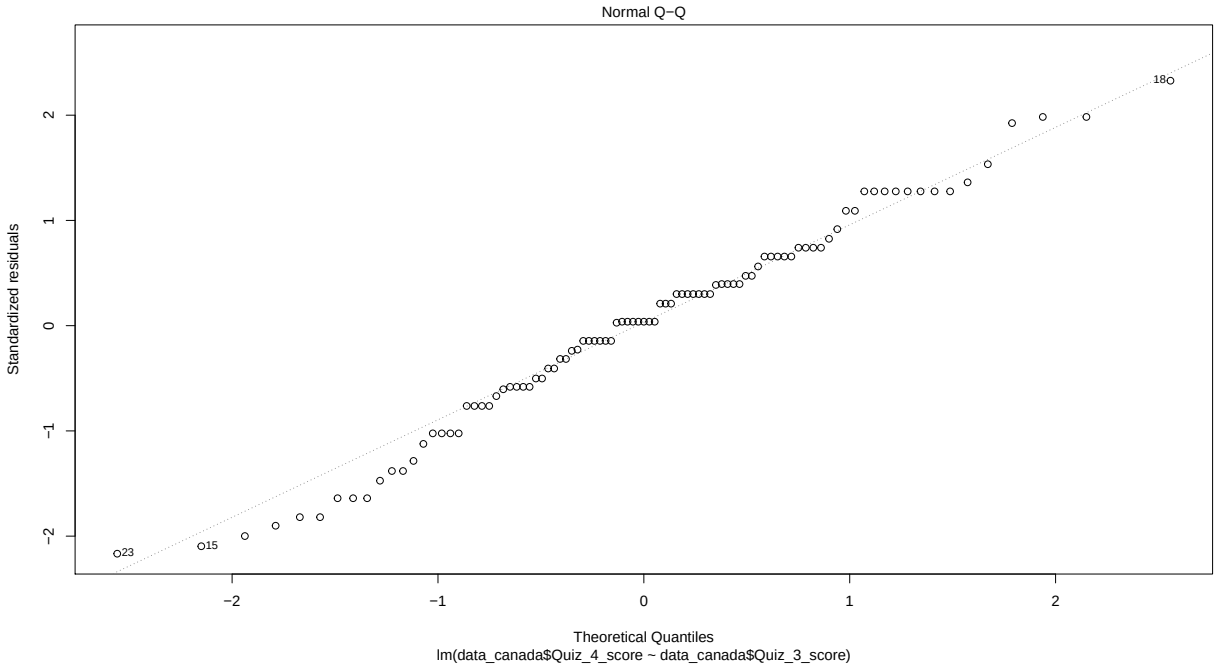

Residuals vs Fitted
Residuals
Fitted values
lm(data_china$Quiz_4_score ~ data_china$Quiz_3_score)
54
27
47

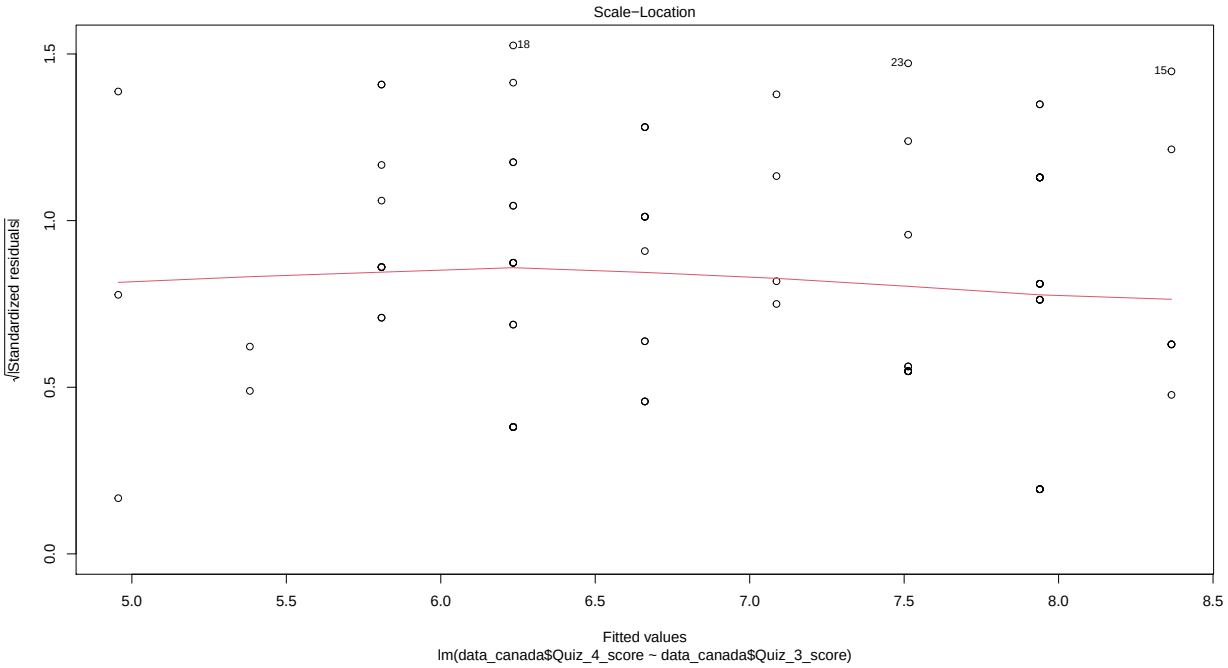

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(data_china$Quiz_4_score ~ data_china$Quiz_3_score)
54
27
47

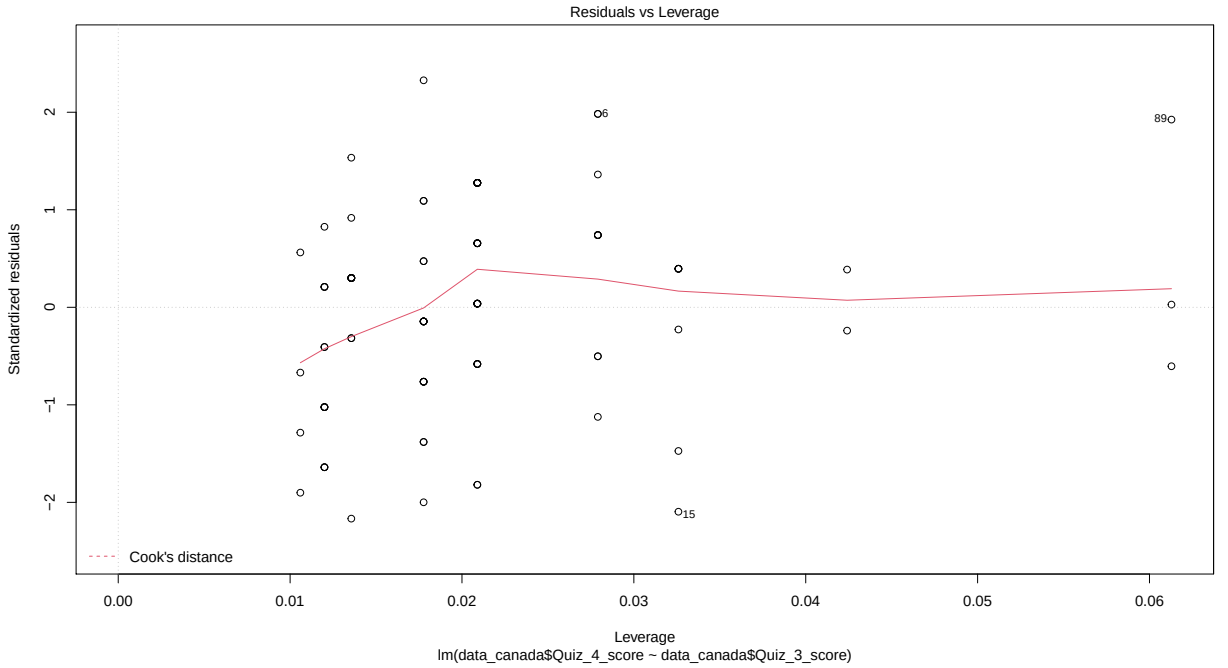

Scale−Location
√|Standardized residuals|
Fitted values
lm(data_china$Quiz_4_score ~ data_china$Quiz_3_score)
27
54
47


Residuals vs Leverage
Standardized residuals
Leverage
lm(data_china$Quiz_4_score ~ data_china$Quiz_3_score)
Cook's distance
0.5
1
10
19
27


Residuals vs Fitted
Residuals
Fitted values
lm(data_canada$Quiz_4_score ~ data_canada$Quiz_3_score)
18
23
15


Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(data_canada$Quiz_4_score ~ data_canada$Quiz_3_score)
18
23
15





After plotting the linear models, we can observe that both plots consists of linear relationships in the Residuals vs Fitted plots as the residuals are equally spreaded around the horizontal line. In the normal Q-Q plots, both plots show the residuals are normally distributed as the residuals follow in a straight line, thus there is no indication of violation in normality assumption. In the Scale-Location plots, the residuals can be observed randomly spreaded and are staying in equal distancing among the horizon lines, thus there's no indicaiton of homoscedasticity. In the Residuals vs Leverage plots, all observations are within the Cook's distance, meaning none of the observations are influential and can impact our models in a significant way, therefore removing outliers is not necessary.

Project Distribution

This report is written by Xinhao Hou and Yunkai Fan. In this project, Xinhao completes the majority of coding in R Markdown, including data cleaning, model building, and plotting. Yunkai contributes to the written part of the report, including management of report sections and management of images. In this project, we have both focus on the distributed sections in the majority of our time, however we also provide assistance to each other on their sections as well.

Conclusion

Our model provides a reasonable prediction of quiz 4 score using the score from the previous quiz using a simple linear model. In the real world this can help the instructor to predict a very rough estimate of how the class will do on the next quiz and whether adjusting the next assignments is necessary. The model we built showed an estimated correlation of 50% between the quiz 3 and quiz 4 score. This means for every mark a student would get on quiz 3 the same student would likely get half a mark quiz 4 in addition to the base grade/intercept term. The limitation of this model would be this model is built in a time where the world is in a pandemic and may not be accurate in predicting quiz scores after the world returns to normal. Also this model only provides a very rough estimate of quiz 4 marks with residual standard error up to 1.6 predicting a number from 0 to 10. In the real world this kind of error could lead to course average to be off a letter grade.