STA302H1 – Final Report

Danny Chen

August 21, 2021

Introduction

The purpose of this report is to study the relationship between a student's country of origin, the time they spent studying for STA302H1 (weeks 1 - 4), the time they spent thinking about COVID-19 (weeks 1 - 4), and their interim STA302H1 quiz scores (quizzes 1 - 3) versus final STA302H1 quiz scores (quiz 4).

Existing studies in pedagogy tend to focus on individual factors that affect course performance, such as the number of hours slept, or the number of hours spent studying for a course. However, this study intends to explore multiple covariates simultaneously to assess their collective effect on final quiz grades, as well as the effects of two covariates on each other.

Experiment Information

The population of interest is a group of students from the online summer 2021 (July - August) STA302H1S cohort, which originally had 227 students at the start of the term, but has 198 students enrolled as of August 13, 2021.

For this study, students were surveyed on Quercus at the end of each week for the first 4 weeks of STA302H1. Each week is specified by a date range below:

- End of Week 1 (July 5 – July 9)
- End of Week 2 (July 12 – July 16)
- End of Week 3 (July 19 – July 23)
- End of Week 4 (July 26 – July 30)

The survey asked about their country of origin (first week only), the number of hours they spent thinking about COVID-19, and the number of hours they spent studying for STA302H1. Quiz scores were also collected and combined into a dataset, where student names were anonymized before the dataset was made available to all STA302H1 students on Quercus for analysis during the STA302H1 final report.

Purpose of Developing Model

The purpose of developing a model is to try to understand the relationship between quiz 4 scores and the other predictor variables (interim STA302H1 quiz scores, study time, COVID contemplation time, country, or some combination of these factors) in explaining the variability among quiz 4 marks.

If the model shows some relationship, it may be possible to use the regression model to determine which of the variables are strong predictors of Quiz 4 grades, and estimate unknown values for quiz 4 scores whenever the values of the independent variables are known.

Developing this model primarily benefits professors and students. Current professors can identify possible weak topics by identifying topics that yield the lowest quiz scores, reflect on things they did/did not help students, and then devote resources to improving lectures or creating carefully curated tutorials that address topics that students find challenging. Teaching stream professors and future STA302H1 professors would inherit these resources so they can establish reasonable STA302H1 learning goals, thoroughly prepare for more formative lectures, and address common student conceptual pitfalls that undermine student quiz scores.

When current STA302H1 students understand the most important factors that contribute to high quiz 4 grades, they can develop informed strategies that can help them effectively learn key material and maximize their grades on formative quizzes. Future students can establish reasonable expectations about workload and develop strategies to maximize their time and success in STA302H1 with available resources.

Plan for Developing Model

The dataset contained a small number of typos, which were cleaned manually rather than programatically. This included removing the word "hours" to safely cast numeric parts of strings as integers, removing non-Unicode characters like "UTF-098", and capitalizing "canada" and "china", so that they would be treated the same as the countries "Canada" and "China." To finish off the data cleaning process, similar columns (i.e., COVID times, study times, and quiz scores) were grouped together to make further data analysis more convenient.

Although some entries in the dataset contained missing (NA) data, missing quiz grades were considered more problematic than rows with only missing number of COVID hours, number of STA302H1 study hours, or even missing countries of origin. To preserve as much of the original dataset as possible, NA countries were categorized as unknown, and NA COVID hours and STA302H1 hours were ignored.

Students who miss 3 or more quizzes were removed from the original dataset, since they may not have quiz 4 scores available. Additionally, students without Quiz 4 scores were also removed.

(TODO: Also remove students without Quiz 4 scores either way?)

Due to the fact that influential outliers tend to have high residuals, they will be removed to prevent them from influencing the magnitude and direction of the regression coefficients, as well as the power of the overall power.

Descriptive statistics such as histograms, boxplots, 5-number summaries, and pairs scatterplots will be created to reveal useful relationships that will help to inform the model selection.

Model diagnostics will be used to verify assumptions of the final model, and address whether or not any variable transformation or variable re-centering is necessary.

Lastly, scholarly research will be consulted to confirm the results and propose ways to improve the final model.

Explanatory Data Analysis

There are a total of 13 variables in the dataset. The response variable is a student's quiz 4 score, and the predictor variables are the remaining 12 variables: a student's country of origin, the time they spent thinking about COVID-19 during weeks 1 - 4, the time they spent studying for STA302H1 during weeks 1 - 4, and their Quiz 1 - 3 scores.

The following table describes each variable, its meaning, and its type:

Variable	Meaning	Type of Variable
`Country`	Student's country of origin	Categorical/nominal
`Quiz_1_Score`	Student's quiz 1 score out of 10	Ordinal numeric
`Quiz_2_Score`	Student's quiz 2 score out of 10	Ordinal numeric
`Quiz_3_Score`	Student's quiz 3 score out of 10	Ordinal numeric
`Quiz_4_Score`	Student's quiz 4 score out of 10	Ordinal numeric
`COVID..hours.W1`	Time student spent thinking about COVID-19 during Week 1 in hours	Continuous numeric
`COVID..hours.W2`	Time student spent thinking about COVID-19 during Week 2 in hours	Continuous numeric
`COVID..hours.W3`	Time student spent thinking about COVID-19 during Week 3 in hours	Continuous numeric
`COVID..hours.W4`	Time student spent thinking about COVID-19 during Week 4 in hours	Continuous numeric
`STA302..hours.W1`	Time student spent studying for STA302H1 during Week 1	Continuous numeric
`STA302..hours.W2`	Time student spent studying for STA302H1 during Week 2	Continuous numeric
`STA302..hours.W3`	Time student spent studying for STA302H1 during Week 3	Continuous numeric
`STA302..hours.W4`	Time student spent studying for STA302H1 during Week 4	Continuous numeric

Note that time spent studying for STA302H1 can include lecture time, review time, quiz time, or assignment time.

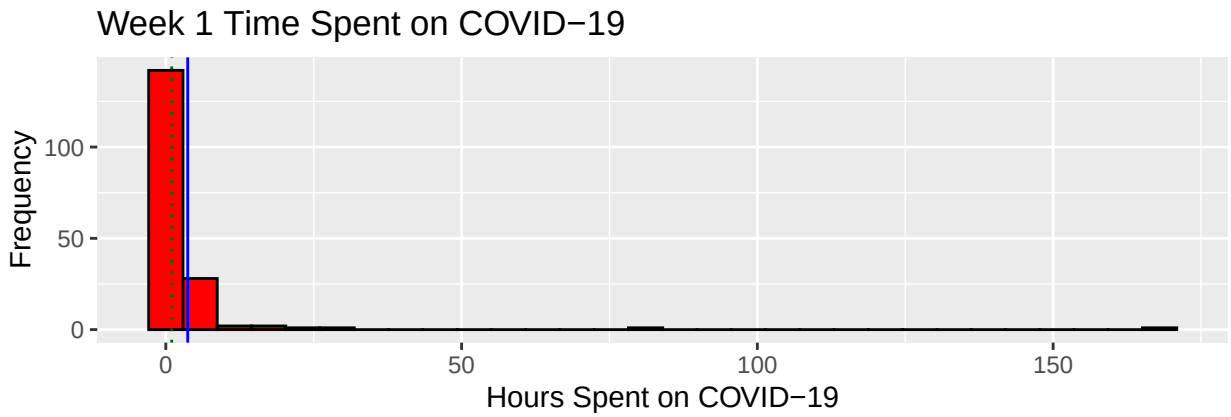
Histograms

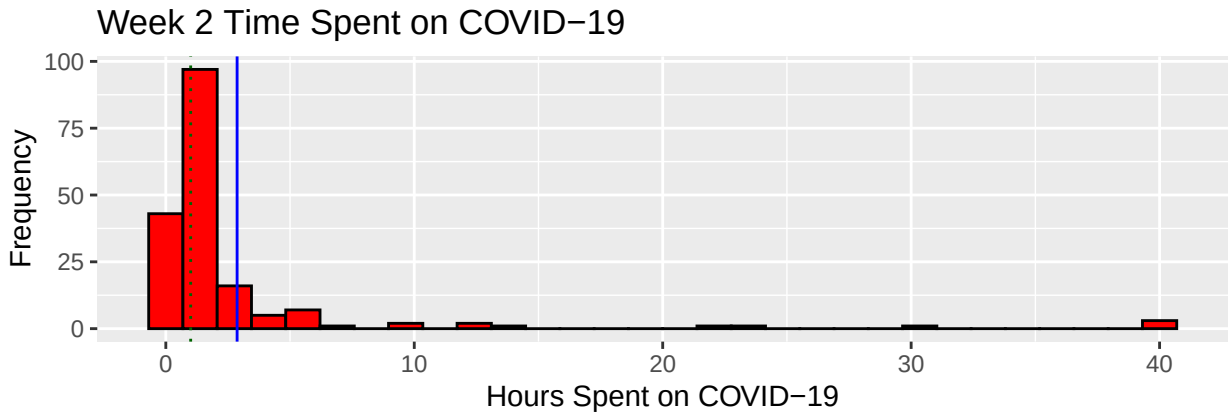
Times spent thinking about COVID-19, study times, and quiz scores among all students were summarized using histograms. The green dotted vertical line represents the mean and the solid blue vertical line represents the median. Seven-number summaries were also included to display meaningful statistics such as the mean, median, minimum, maximum, 1st quartile, 3rd quartile, and the number of NAs.

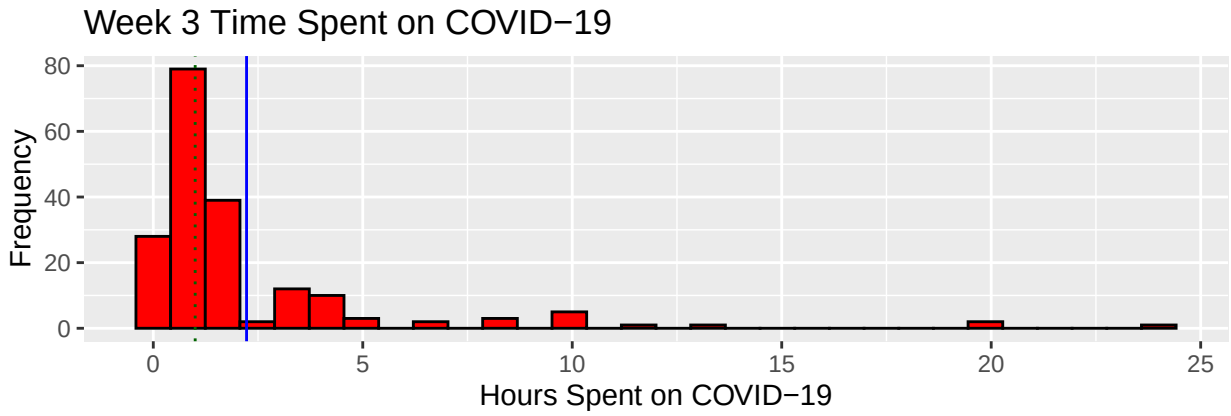
The histograms for COVID hours during weeks 1 - 4 are all right skewed (mean $>$ median), indicating that a few students tend to spend a lot of time thinking about COVID-19 during the week. The median times spent thinking about COVID-19 remained constant at 1 hour/week across all 4 weeks, however the mean time spent thinking about COVID-19 decreased for the first 3 weeks from a maximum of 3.7 hours in week 1 to a minimum of 2.227 hours in week 3, and increased from week 3 to week 4.

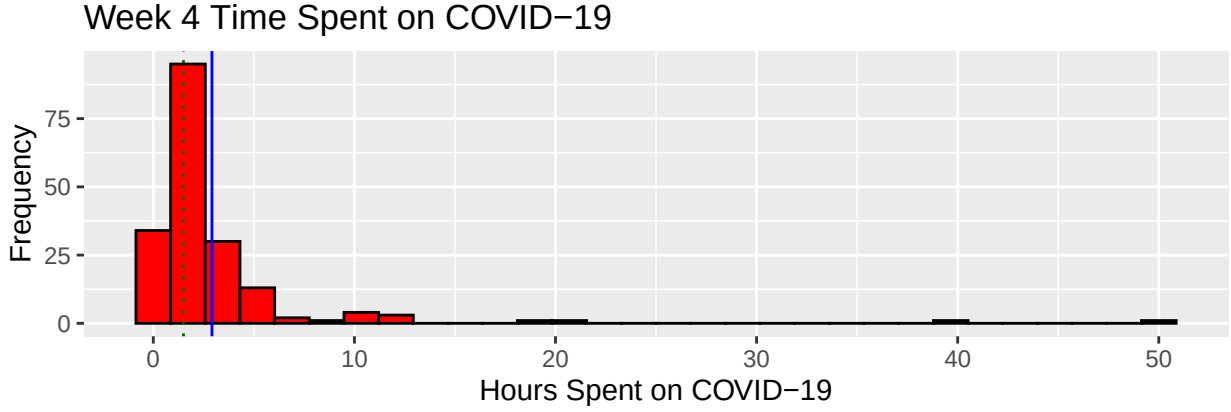
The histograms for study hours during weeks 1 - 3 are approximately normal with a few outliers in the right tail. However, study hours week 4 exhibits a right skewed distribution. In fact, the mean study time increased by approximately 23%, and the median study time increased by approximately 44% from week 3 to week 4.

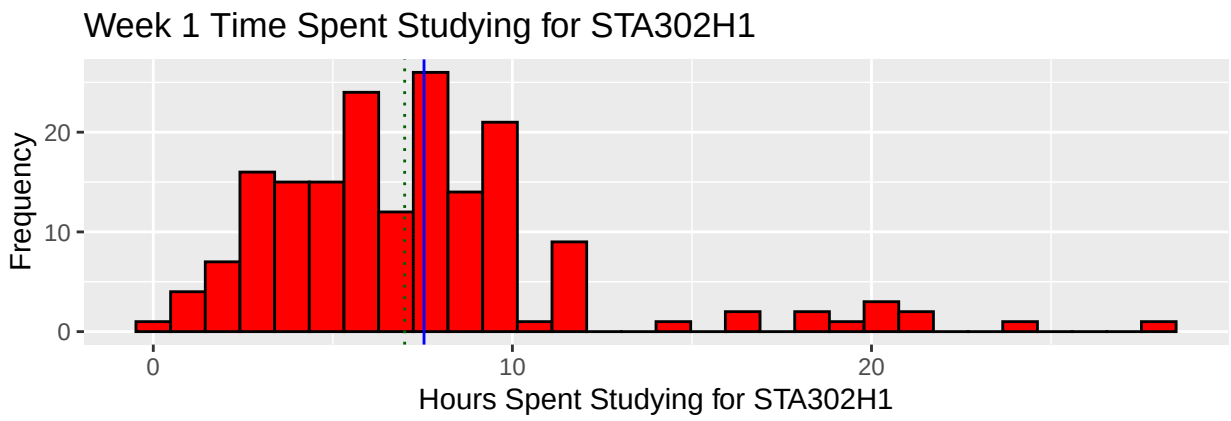
The histograms for quiz scores during weeks 1 - 4 are all left skewed (mean $<$ median), as few students tend to underperform on quizzes. The median for weeks 1, 3, and 4 remain at 8/10 points, except for week 3 which had a median of 8.8/10. Week 1 had the smallest difference between mean and median quiz scores (about 0.26/10 points), while week 2 has the largest divide between mean and median quiz scores (about 1.38/10 points).

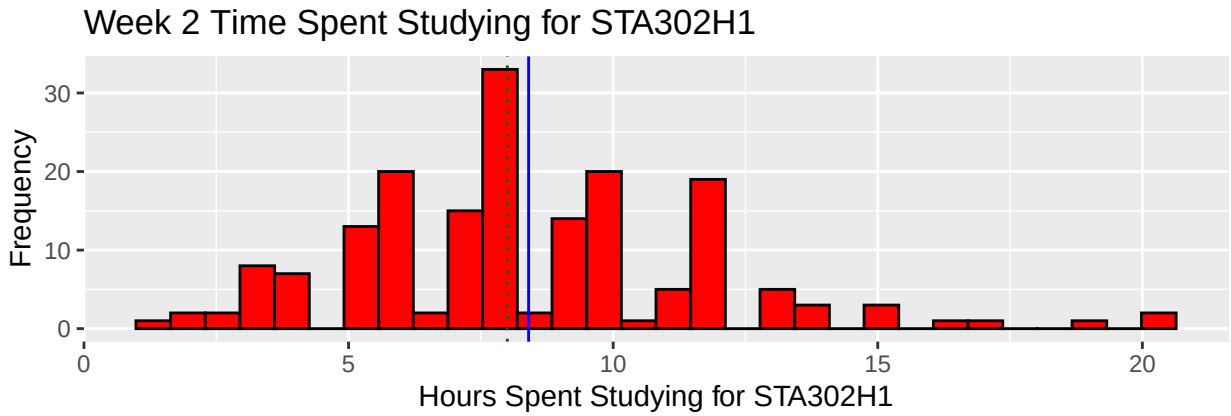

Week 1 Time Spent on COVID−19
Frequency
100
50
0
0
50
100
150
Hours Spent on COVID−19

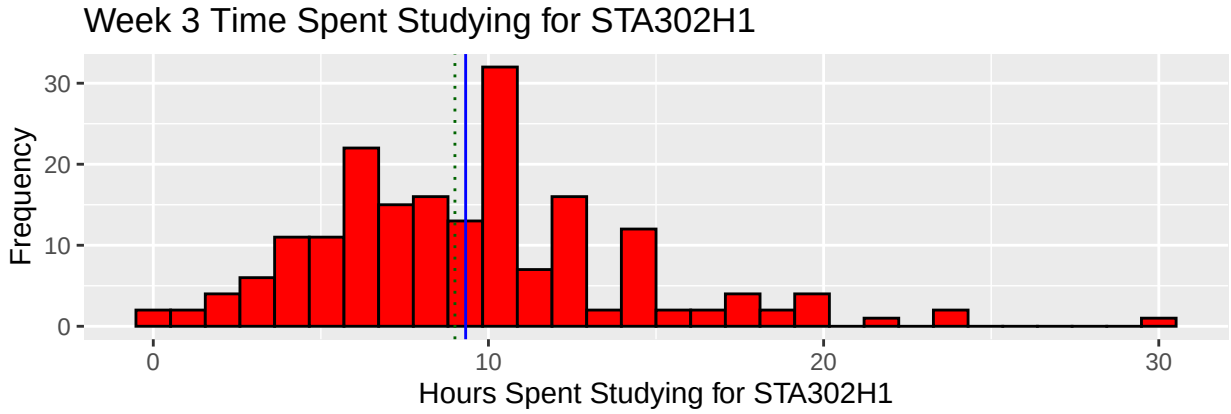

Week 2 Time Spent on COVID−19
Frequency
100
75
50
25
0
0
10
20
30
40
Hours Spent on COVID−19

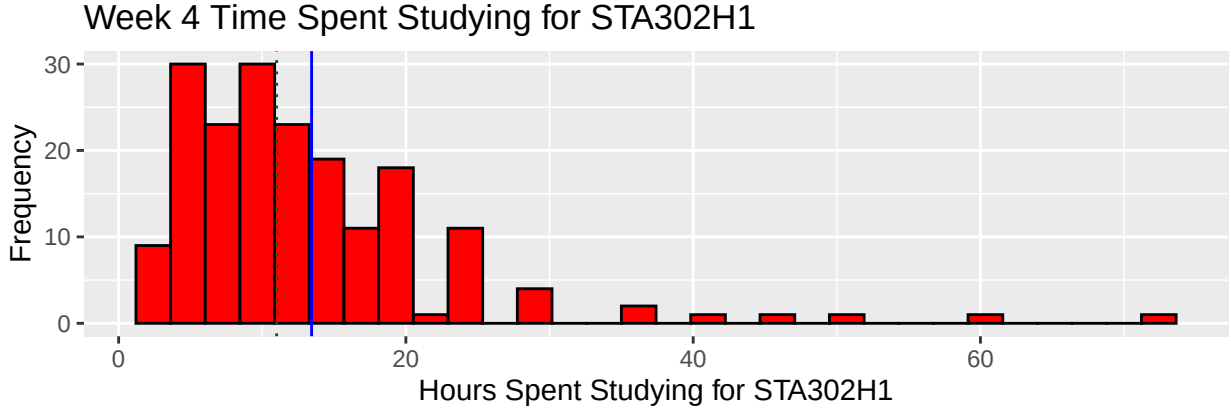

Week 3 Time Spent on COVID−19
Frequency
80
60
40
20
0
0
5
10
15
20
25
Hours Spent on COVID−19

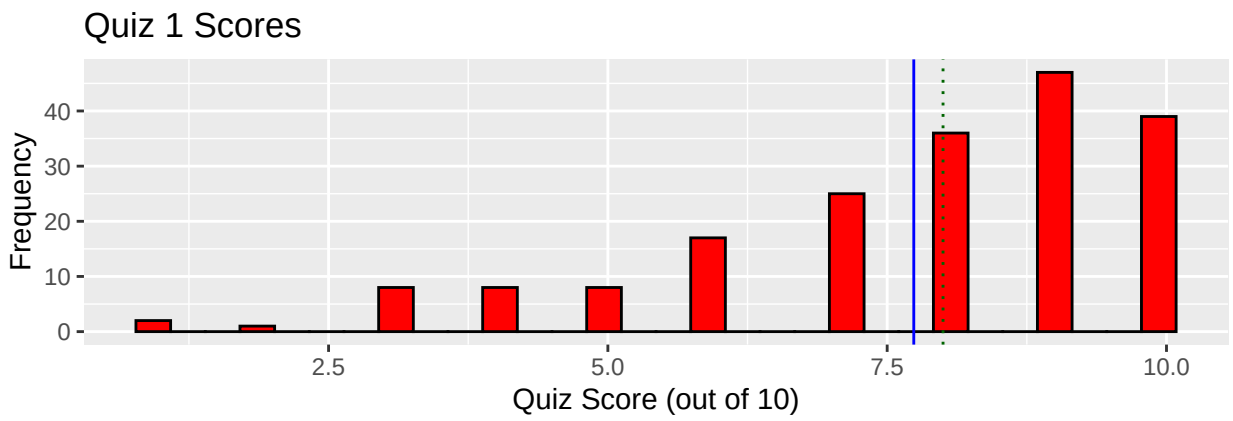

Week 4 Time Spent on COVID−19
Frequency
75
50
25
0
0
10
20
30
40
50
Hours Spent on COVID−19

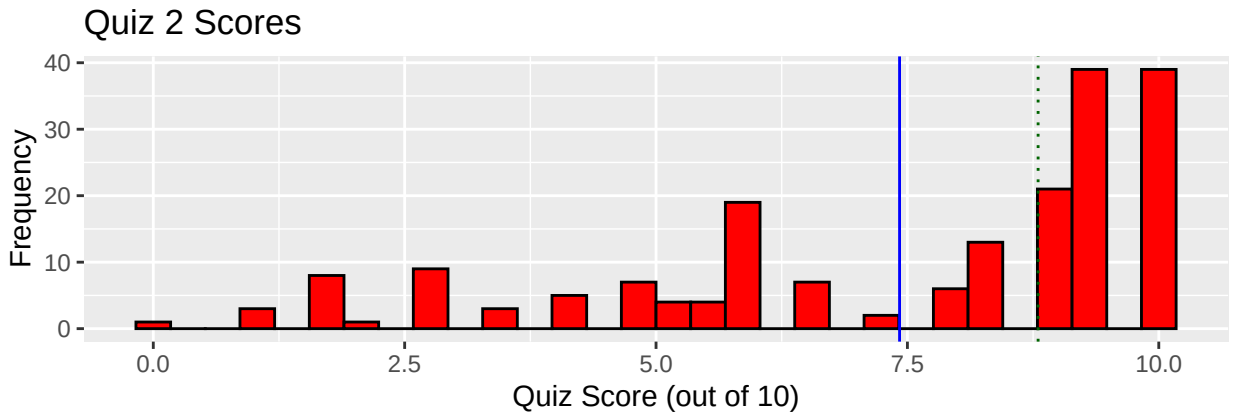

Week 1 Time Spent Studying for STA302H1
Frequency
0
10
20
0
10
20
Hours Spent Studying for STA302H1

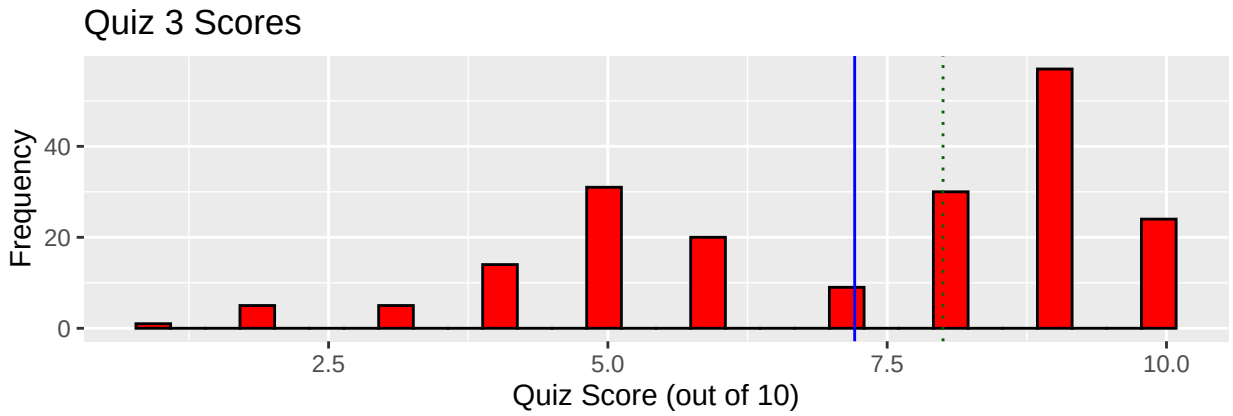

Week 2 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
5
10
15
20
Hours Spent Studying for STA302H1

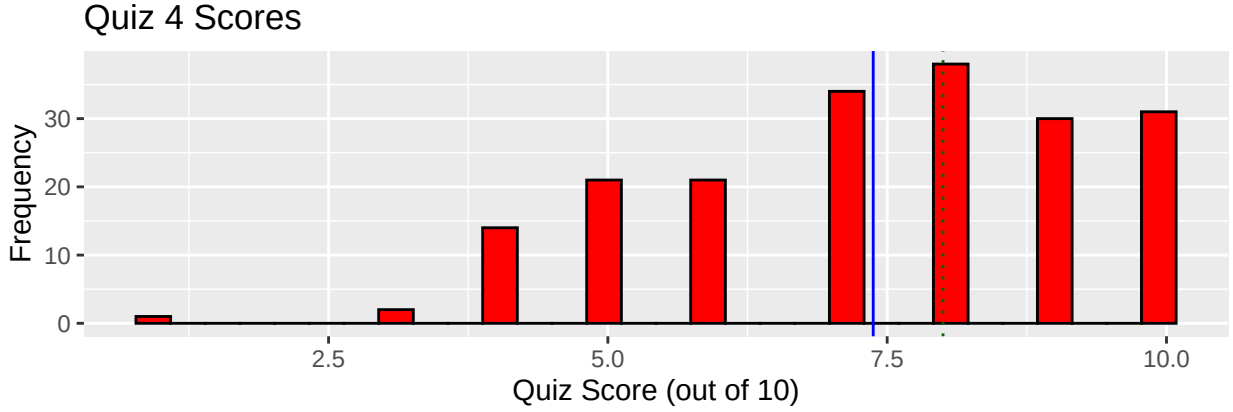

Week 3 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
10
20
30
Hours Spent Studying for STA302H1


Week 4 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
20
40
60
Hours Spent Studying for STA302H1


Quiz 1 Scores
Frequency
0
10
20
30
40
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


Quiz 2 Scores
Frequency
0
10
20
30
40
0.0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


Quiz 3 Scores
Frequency
0
20
40
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


Quiz 4 Scores
Frequency
0
10
20
30
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

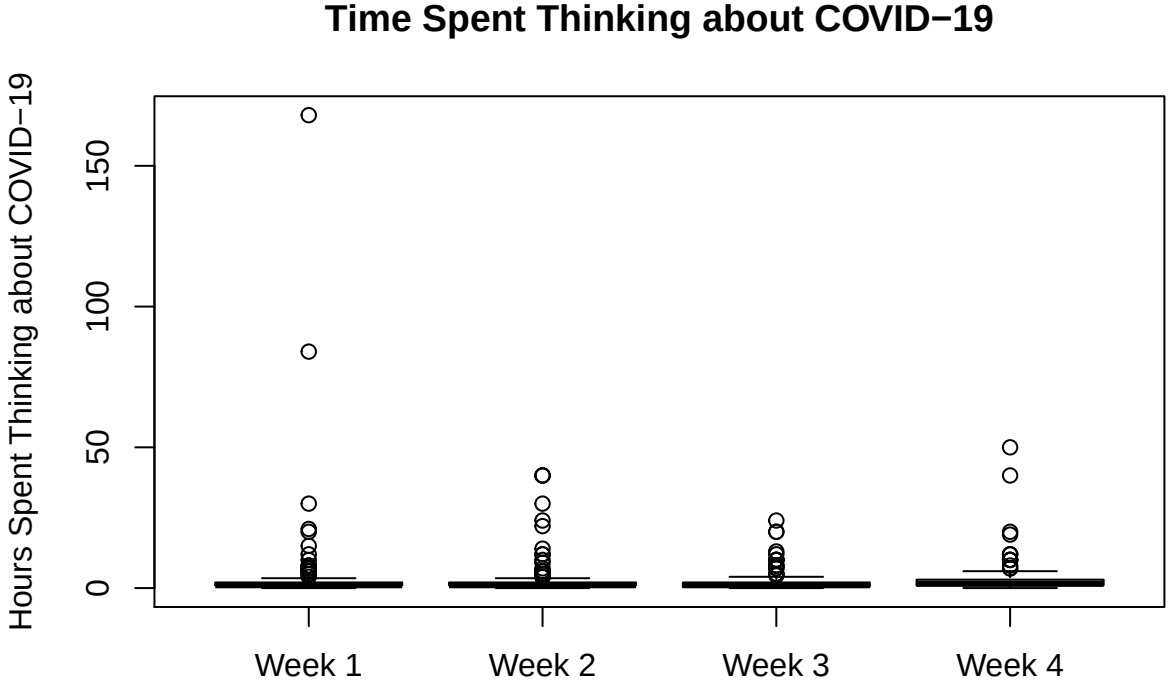
Boxplots

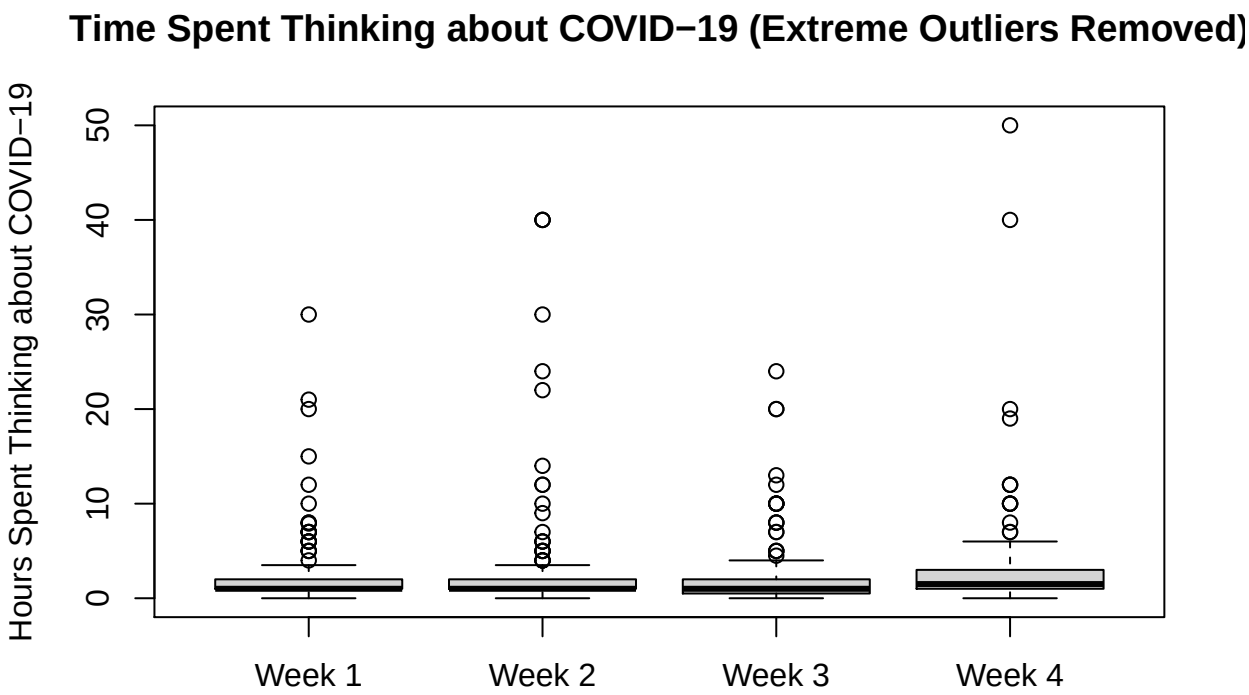
Side-by-side boxplots were used to compare COVID-19 times, study times, and quiz scores among students from all countries.

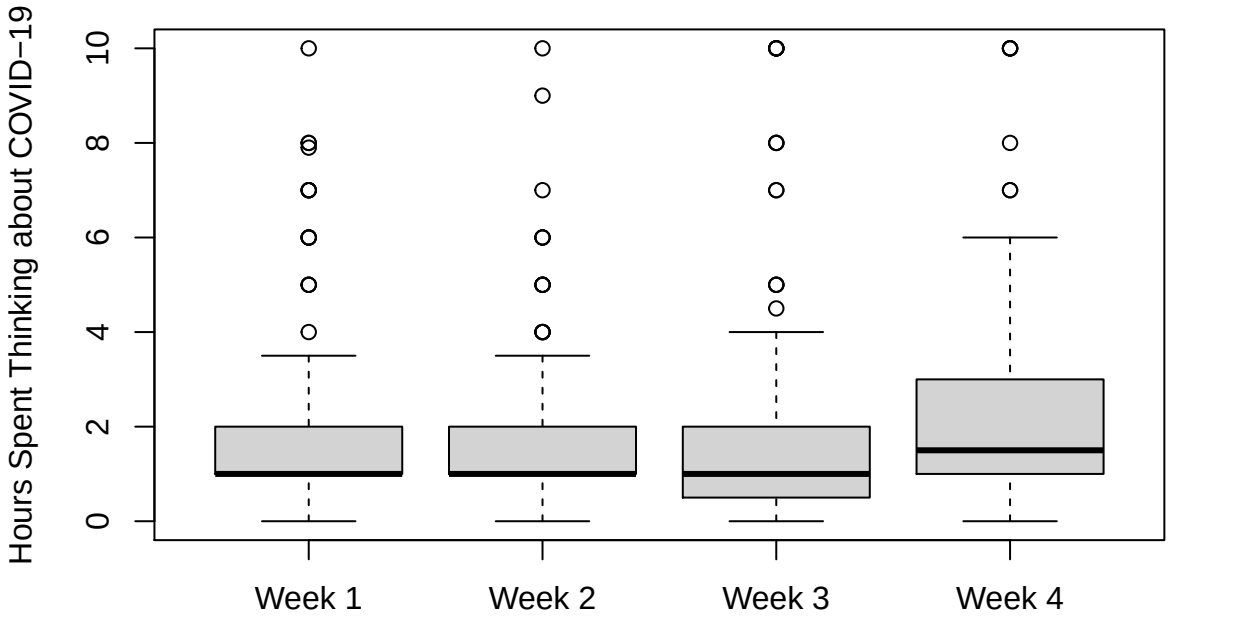
All 4 weeks contain outliers, although week 1 had the largest outliers: two students thought about COVID-19 for 168 and 84 hours. Among all students, weeks 1 - 3 median COVID times were approximately 1 hour with a slightly increased median COVID time of around 1.5 hours from week 3 to week 4. The spread (IQR) of COVID times were both 1 hour for weeks 1 and 3, although there is more spread in week 3 (1.5 hours), and week 4 has the largest spread (2 hours).

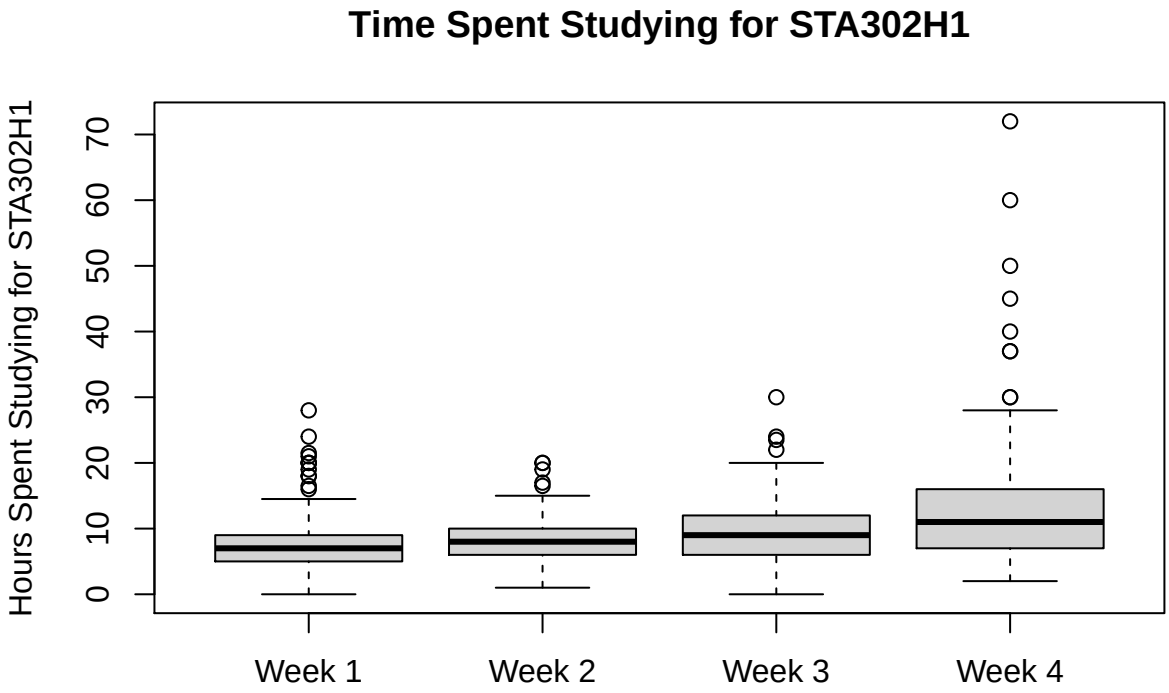
Although all 4 weeks contained outliers, week 4 has outliers ranging from a median study time of 30 - 70 hours. The median amount of study time steadily increased from 7 hours to 9 hours in weeks 1 - 3, with a noticeable increase in median study time from 9 hours to 11 hours from week 3 to week 4. Median study times had similar spreads (IQR) of 4 hours for weeks 1 and 2, but there is more spread in week 3 (6 hours), and even larger spread in week 4 (9 hours).

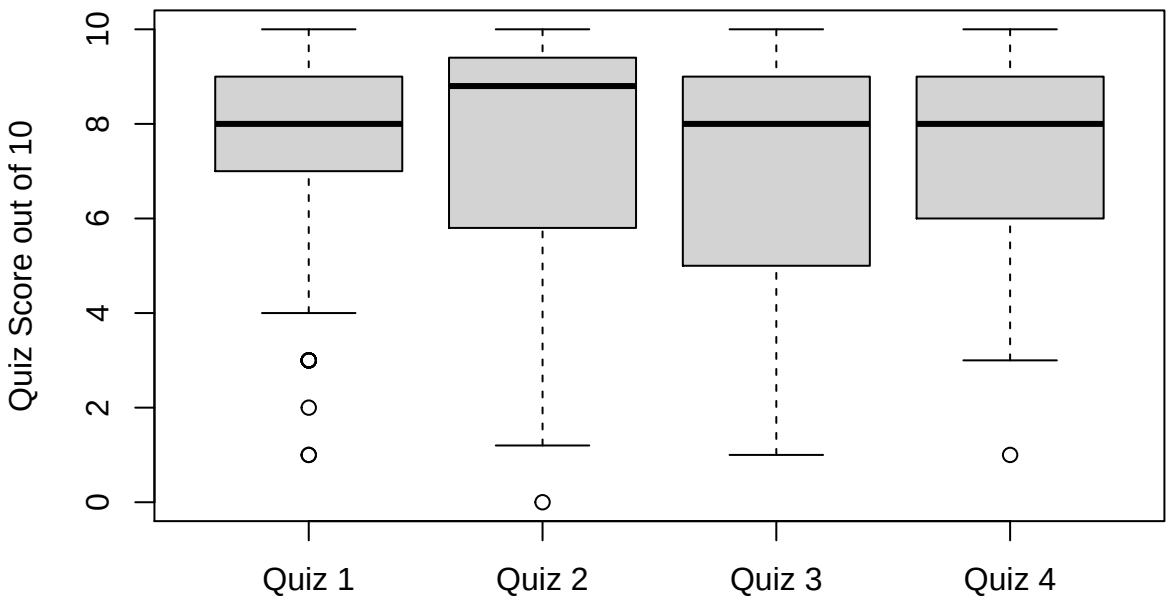
Median quiz 1 scores have the most outliers, although they are milder than the single quiz 2 outlier which displays the lowest quiz 2 score (1/10). Quizzes 1, 3, and 4 has similar median quiz scores of 8/10, with week 2 having the highest median quiz score of 8.8/10. Weeks 2 and 3 have similarly large spreads (IQRs of 3.6 and 4 point gaps), though week 1 has the smallest spread. Week 4 quiz scores are the least skewed.




Time Spent Thinking about COVID−19 (Extreme Outliers Removed)
Hours Spent Thinking about COVID−19
0
10
20
30
40
50
Week 1
Week 2
Week 3
Week 4

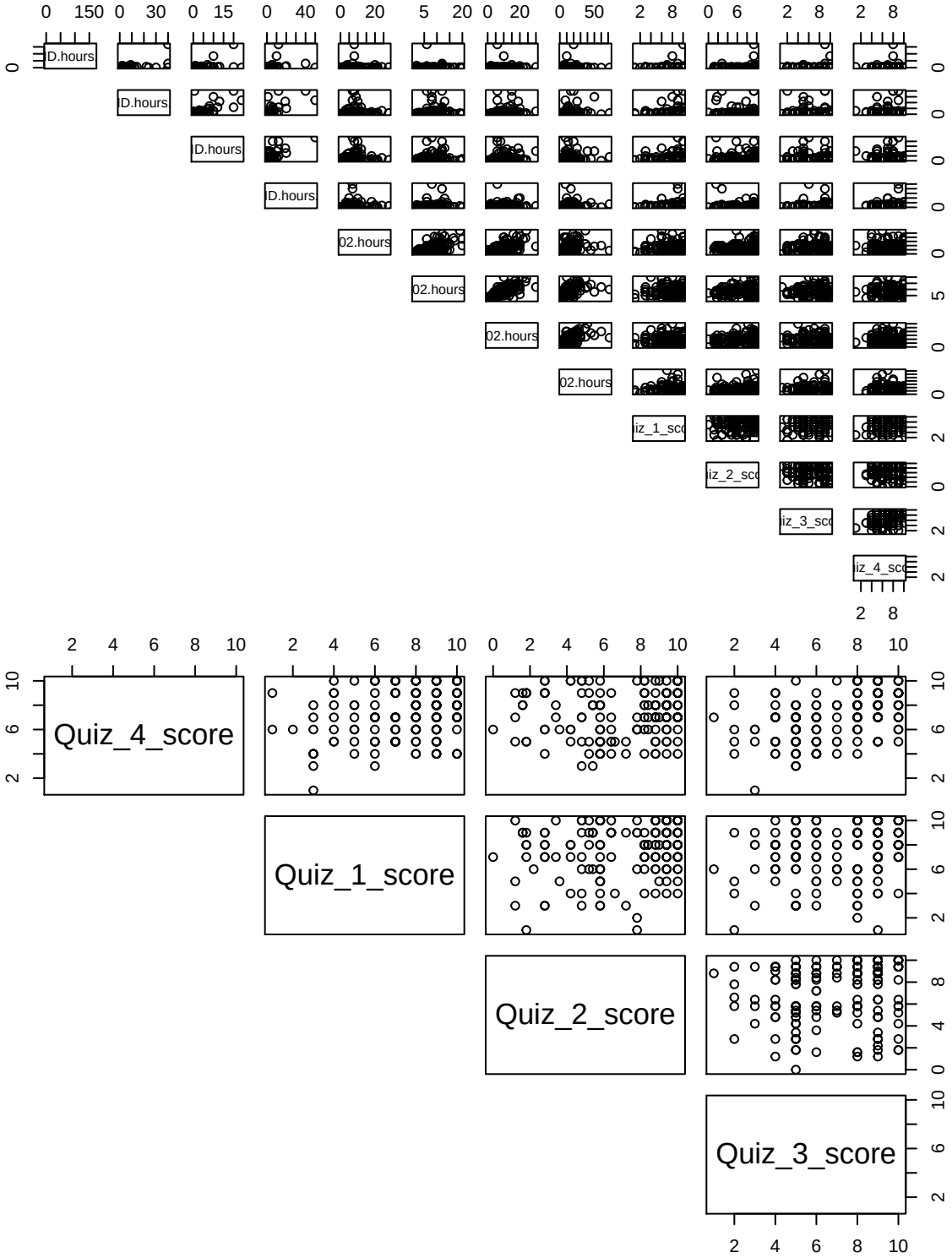

Time Spent Thinking about COVID−19 (Moderate Outliers Removed
Hours Spent Thinking about COVID−19
0
2
4
6
8
10
Week 1
Week 2
Week 3
Week 4

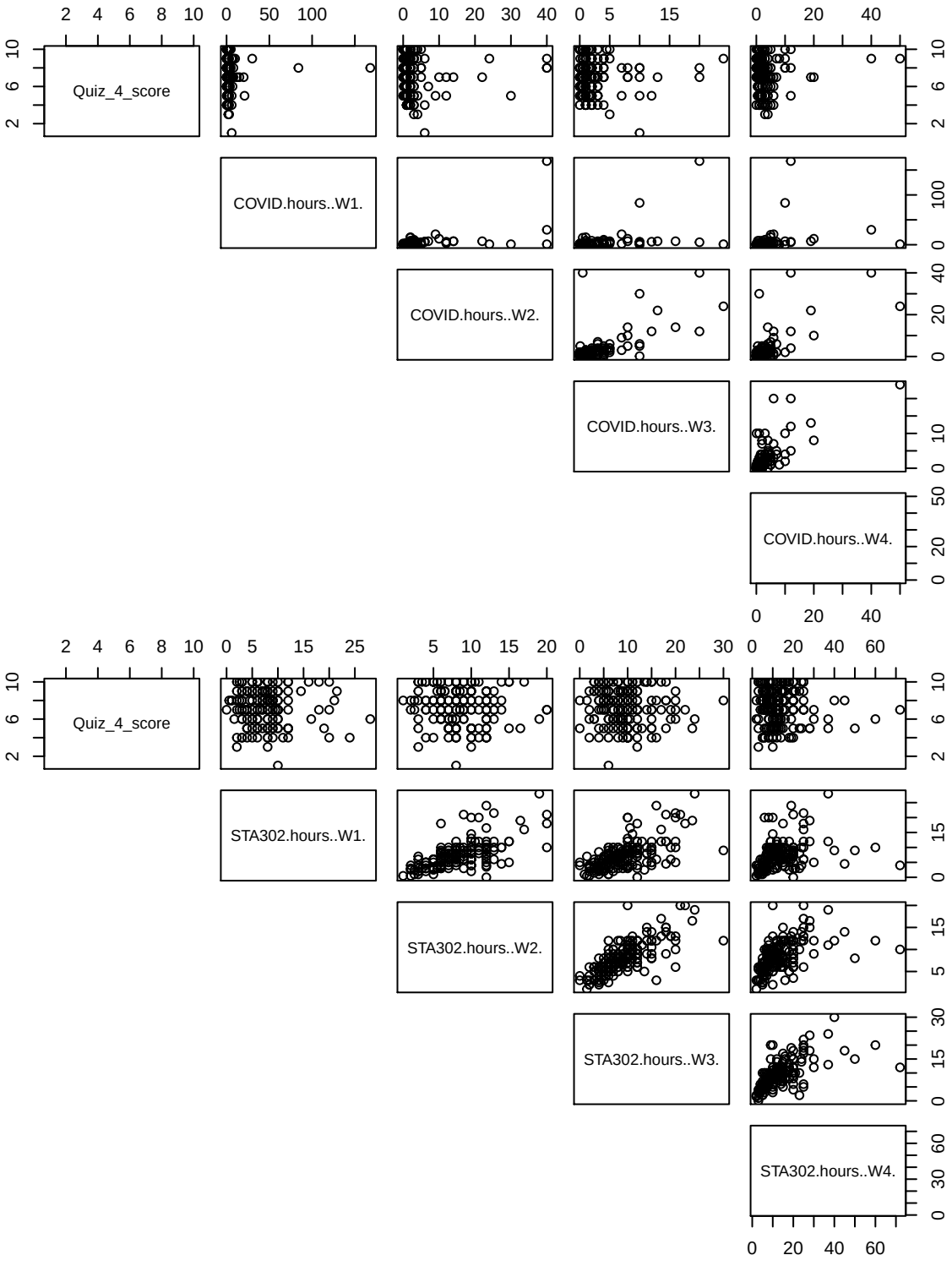

Time Spent Studying for STA302H1
Hours Spent Studying for STA302H1
0
10
20
30
40
50
60
70
Week 1
Week 2
Week 3
Week 4


Quiz Scores
Quiz Score out of 10
0
2
4
6
8
10
Quiz 1
Quiz 2
Quiz 3
Quiz 4

Pairwise Scatterplots

Pairwise scatterplots were created to observe the relationships between all combinations of variables, especially between the response variable and the explanatory variables. Since the pairs scatterplot were symmetric along its main diagonal, the bottom half of the pairs scatterplot could be safely omitted.





Correlation Matrix

A correlation matrix was constructed to determine the correlations of each pair of variables in the dataset. Since the correlation matrices for underrepresented countries (i.e. every country except Canada, China, and NA) contained a lot of NA or ± 1 entries, it made more sense to use the correlation matrix for all countries. A correlation between 0.0 - 0.3 is considered low, 0.3 - 0.5 is moderate, and 0.5 - 1.0 is high.

Model Development

Proposing an Initial Model

Simpler (e.g., linear, quadratic) relationships were prioritized over more complex relationships (e.g., higher order polynomial, logarithmic, square root) for the sake of straightforward analysis and intuitive interpretation. The pairs scatterplot and the correlation matrix were primarily used to derive most of the terms for the initial model.

The non-interactive covariate terms were determined by examining the pairs scatterplots for quiz 4 scores regressed on quiz 1 - 3 scores, weeks 1 - 4 COVID-19 times, and weeks 1 - 4 STA302H1 study times to hypothesize a relationship between quiz 4 scores and each predictor variable.

The weeks 1 - 4 COVID times vs. quiz 4 scatterplots show a moderate quadratic relationship between quiz 4 scores and weeks 1 - 4 COVID times (except for week 3 COVID, which seems to have a stronger linear relationship than a quadratic relationship). Therefore, both the linear and quadratic terms for weeks 1, 2, and 4 COVID times (i.e., `covid1`, `covid1 ** 2`, `covid2`, `covid2 ** 2`, `covid4`, and `covid4 ** 2` – along with `covid3` only) were included in the initial model.

Further analysis indicates that there is a possibly moderate linear relationship between quiz 4 scores and weeks 1 - 4 study times. Therefore, only the linear terms for weeks 1 - 4 study times (i.e., `study1`, `study2`, `study3`, and `study4`) were included in the model.

Similarly, there might be a moderate linear relationship between quiz 4 scores and quiz 1 - 3 scores. Therefore, only the linear terms for quiz 1 - 3 scores (i.e., `quiz1`, `quiz2`, and `quiz3`) were included in the model.

(TODO: Show scatterplot of covid1 vs. quiz4.)
(TODO: Display summary statistics for covid1 vs. quiz4 in appendix.)
(TODO: Show scatterplot of covid2 vs. quiz4.)
(TODO: Display summary statistics for covid2 vs. quiz4 in appendix.)
(TODO: Show scatterplot of covid3 vs. quiz4.)
(TODO: Display summary statistics for covid3 vs. quiz4 in appendix.)
(TODO: Show scatterplot of covid4 vs. quiz4.)
(TODO: Display summary statistics for covid4 vs. quiz4 in appendix.)

(TODO: Show scatterplot of study1 vs. quiz4.)
(TODO: Display summary statistics for study1 vs. quiz4 in appendix.)
(TODO: Show scatterplot of study2 vs. quiz4.)
(TODO: Display summary statistics for study2 vs. quiz4 in appendix.)
(TODO: Show scatterplot of study3 vs. quiz4.)
(TODO: Display summary statistics for study3 vs. quiz4 in appendix.)
(TODO: Show scatterplot of study4 vs. quiz4.)
(TODO: Display summary statistics for study4 vs. quiz4 in appendix.)

(TODO: Show scatterplot of quiz1 vs. quiz4.)
(TODO: Display summary statistics for quiz1 vs. quiz4 in appendix.)
(TODO: Show scatterplot of quiz2 vs. quiz4.)
(TODO: Display summary statistics for quiz2 vs. quiz4 in appendix.)
(TODO: Show scatterplot of quiz3 vs. quiz4.)
(TODO: Display summary statistics for quiz3 vs. quiz4 in appendix.)

Pairs of covariates, as well as response variable-covariate combinations from the correlation matrix with high correlation were used as a heuristic for determining potentially significant interaction terms in the initial model.

The initial model was fitted using `quiz4` as the response variable; and `quiz1`, `quiz2`, `quiz3`, `covid1`, `covid2`, `covid3`, `covid4`, (including 3 quadratic terms 8 interaction terms), and `country` as predictor variables. However, only the predictor variables `quiz3`, `I(covid1 ** 2)`, `I(covid2 ** 2)`, `I(covid1 * covid2)`,

and `I(study1 * study2)` were significant at the 5% significance level. The global F-statistic value is 3.098, and the global p-value is approximately 1.436×10^{-5}. The residual standard error is 1.582. The multiple R^2 value is 0.4211 and the adjusted R^2 value of 0.2851.

Characteristic	**Beta**	**95% CI**	**p-value**
quiz1	0.03	-0.13, 0.20	0.7
quiz2	0.05	-0.07, 0.17	0.4
quiz3	0.48	0.32, 0.63	<0.001
covid1	0.18	-0.07, 0.43	0.2
I(covid1^2)	0.02	0.00, 0.03	0.029
covid2	0.29	-0.09, 0.67	0.13
I(covid2^2)	-0.02	-0.05, 0.00	0.046
covid3	-0.05	-0.30, 0.20	0.7
covid4	-0.25	-0.55, 0.06	0.11
I(covid4^2)	0.02	-0.01, 0.05	0.2
I(covid1 * covid2)	-0.07	-0.14, -0.01	0.029
I(covid2 * covid3)	0.05	-0.01, 0.11	0.12
I(covid2 * covid4)	0.04	-0.01, 0.09	0.093
I(covid3 * covid4)	-0.08	-0.18, 0.02	0.13
I(study1 * study2)	-0.02	-0.03, 0.00	0.018
I(study1 * study3)	0.01	0.00, 0.02	0.14
I(study2 * study3)	0.01	0.00, 0.02	0.095
I(study3 * study4)	0.00	0.00, 0.00	0.14
country			
Canada			
China	0.59	-0.10, 1.3	0.092
India	0.87	-1.5, 3.2	0.5
Mongolia	-13	-51, 26	0.5
Pakistan	-0.15	-3.3, 3.0	>0.9
Singapore	1.2	-2.1, 4.5	0.5
South Korea	-0.02	-2.3, 2.3	>0.9
Taiwan	-1.2	-3.5, 1.1	0.3
UAE	-0.63	-3.9, 2.6	0.7
USA	1.5	-2.0, 5.0	0.4

Improving the Original Model

To refine the original model, backwards selection was used to remove insignificant terms (terms whose p-values were > 0.05) and find the subset model with the lowest AIC value.

The final model has `quiz4` as the response variable; `quiz3`, the quadratic term `I(covid1 ** 2)`, and the interaction terms `I(covid1 * covid2)`, `I(covid2 * covid3)`, `study1 * study2`, `study1 * study3`, `study2 * study3`, `study3 * study4` as the predictor variables. Out of all of the predictors variables in the final model, only the predictor variables `quiz3`, `I(covid1 ** 2`, `I(covid1 * covid2)`, `I(study1 * study2)`, `I(study2 * study3)`, and `I(study3 * study4)` are significant at the 5% significance level. The global F-statistic value increased to 8.764, the global p-value increased to 1.374×10^{-9}. The residual standard error decreased marginally to 1.561. Even though the multiple R^2 value dropped significantly to 0.3435, the adjusted R^2 value increase slightly to 0.3043 – there is a smaller difference between the R^2 value and the adjusted R^2 value in the final model than the original model. The coefficient of 0.499110 for `quiz3` suggests that for every 1 point increase in quiz 3, there is an increase in quiz 4 grades by about 0.50/10 points. There also exists a quadratic relationship between variables `covid1` and `quiz4`, as well as an interaction effect between `covid1` and `covid2`, `study1` and `study2`, `study2` and `study3`, and `study3` and `study4`. All further analyses will be performed on the final model.

Characteristic	**Beta**	**95% CI**	**p-value**
quiz3	0.50	0.38, 0.62	<0.001
I(covid1^2)	0.00	0.00, 0.01	0.046
I(covid1 * covid2)	-0.02	-0.04, 0.00	0.043
I(covid2 * covid3)	0.00	0.00, 0.01	0.066
I(study1 * study2)	-0.02	-0.03, 0.00	0.014
I(study1 * study3)	0.01	0.00, 0.02	0.2
I(study2 * study3)	0.01	0.00, 0.02	0.029
I(study3 * study4)	0.00	0.00, 0.00	0.035

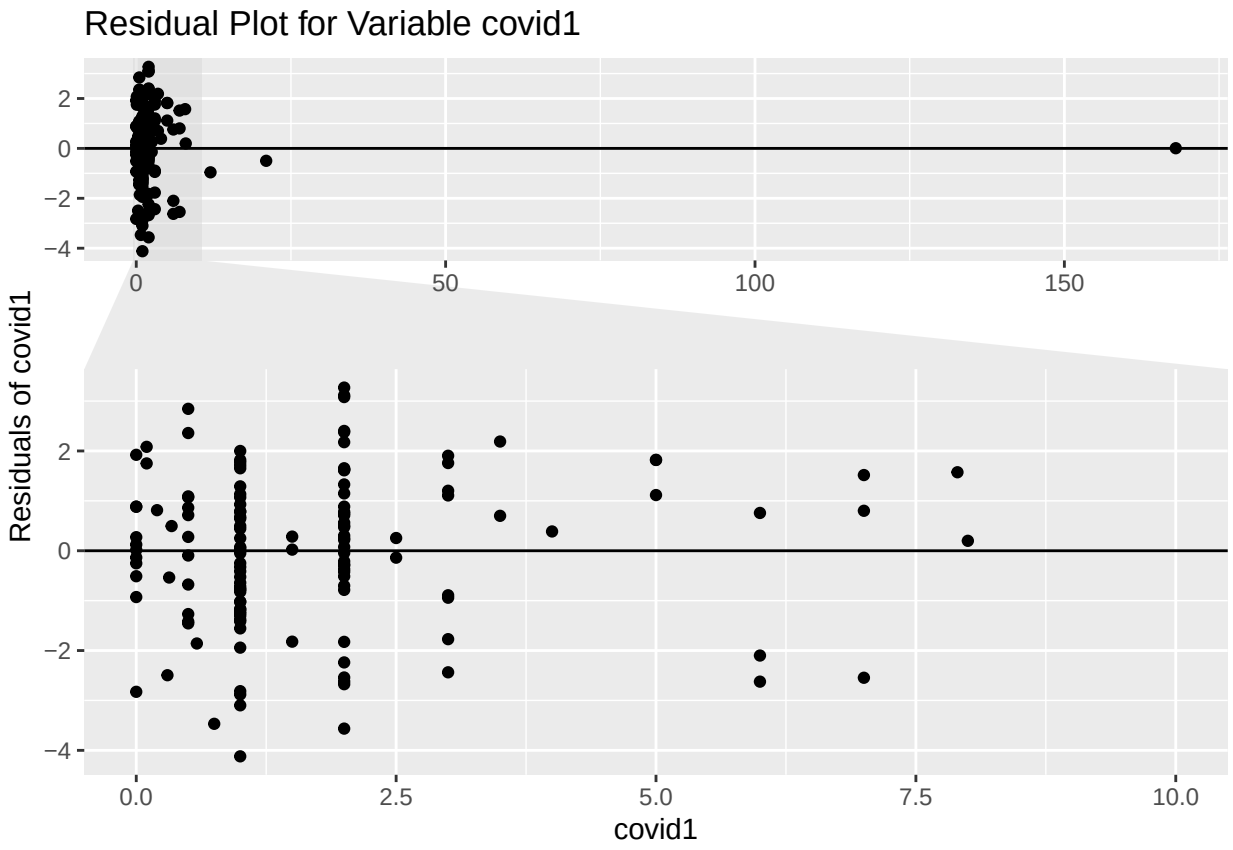
Model Diagnostics

Model Validity

To show that a quadratic relationship for the final model is valid, the following assumptions must hold.

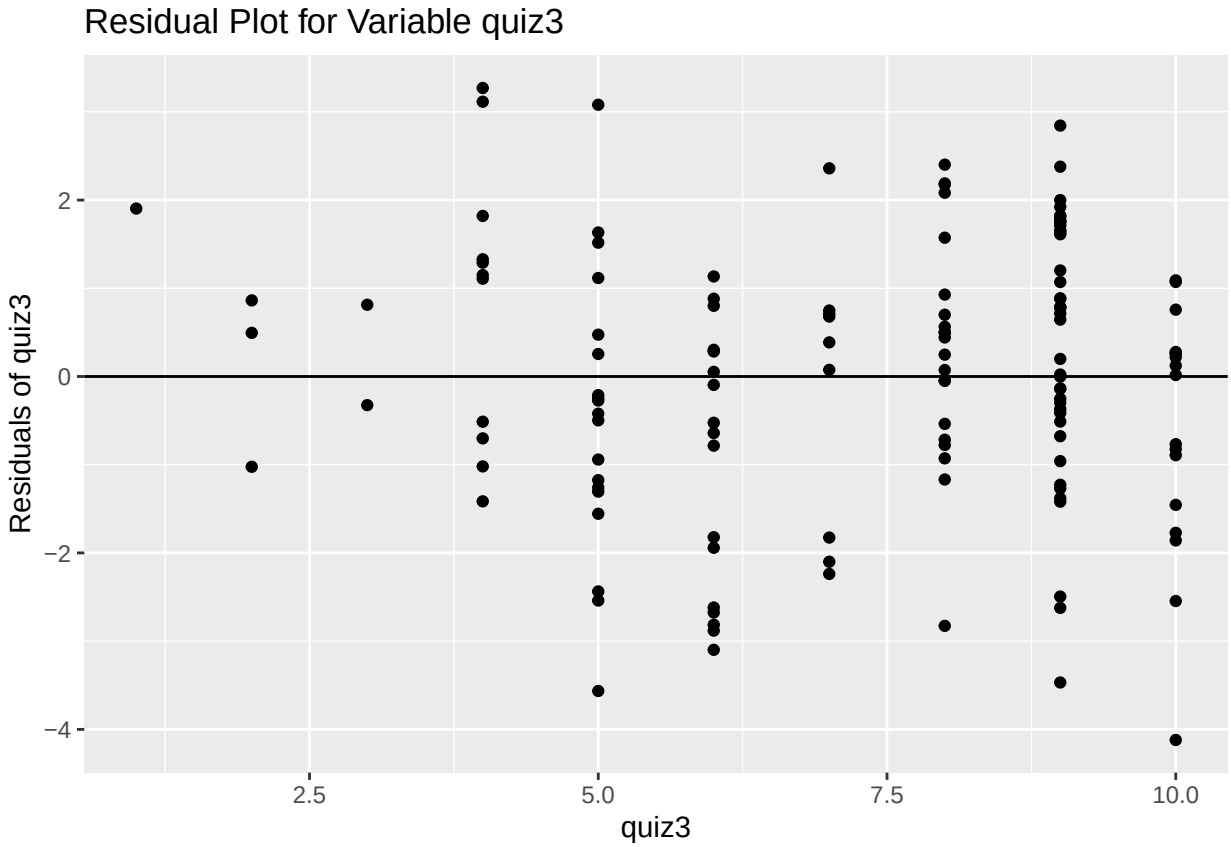
Assumption 1. Quadratic Functional Form

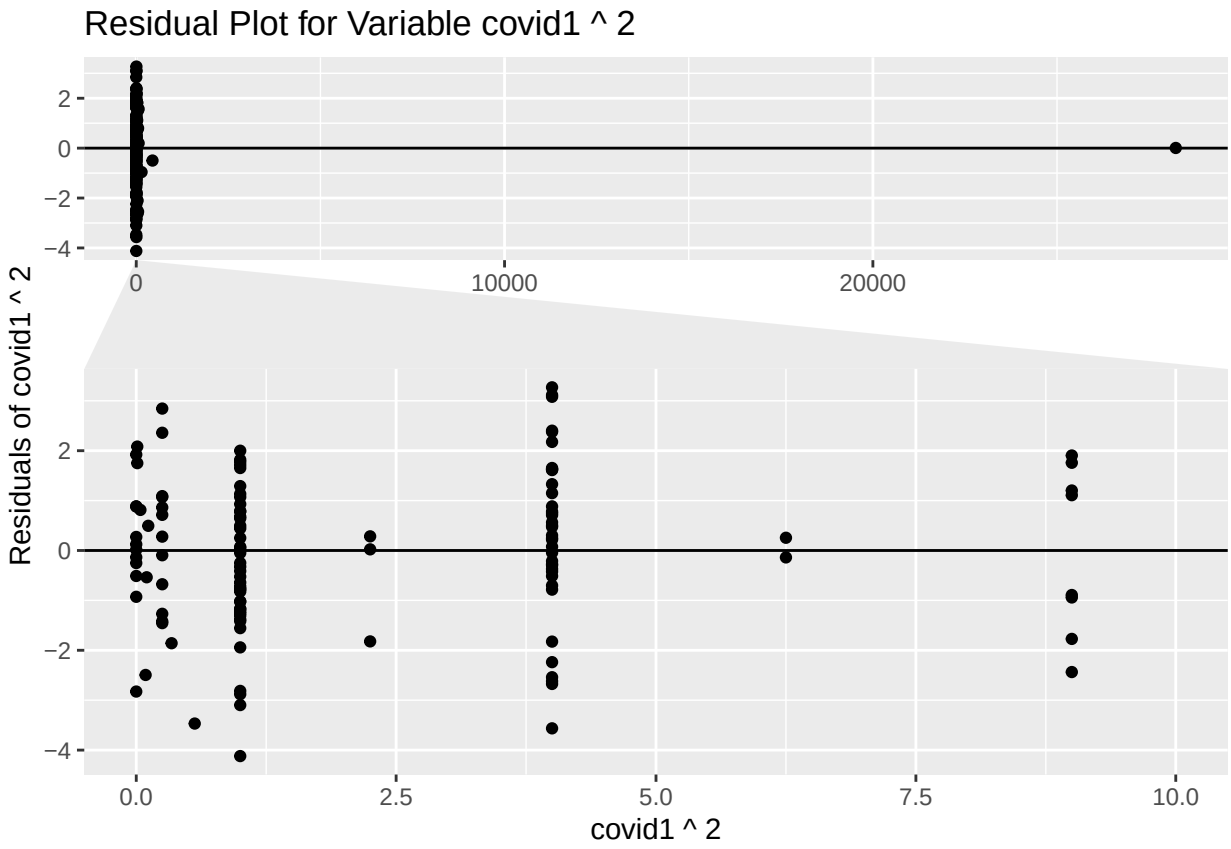
Since the quiz 4 vs. covid1 scatterplot in figure X indicates a moderate quadratic relationship between the variable, and the residual vs. covid1 plot for the fitted model is random, the final model has a quadratic relationship.

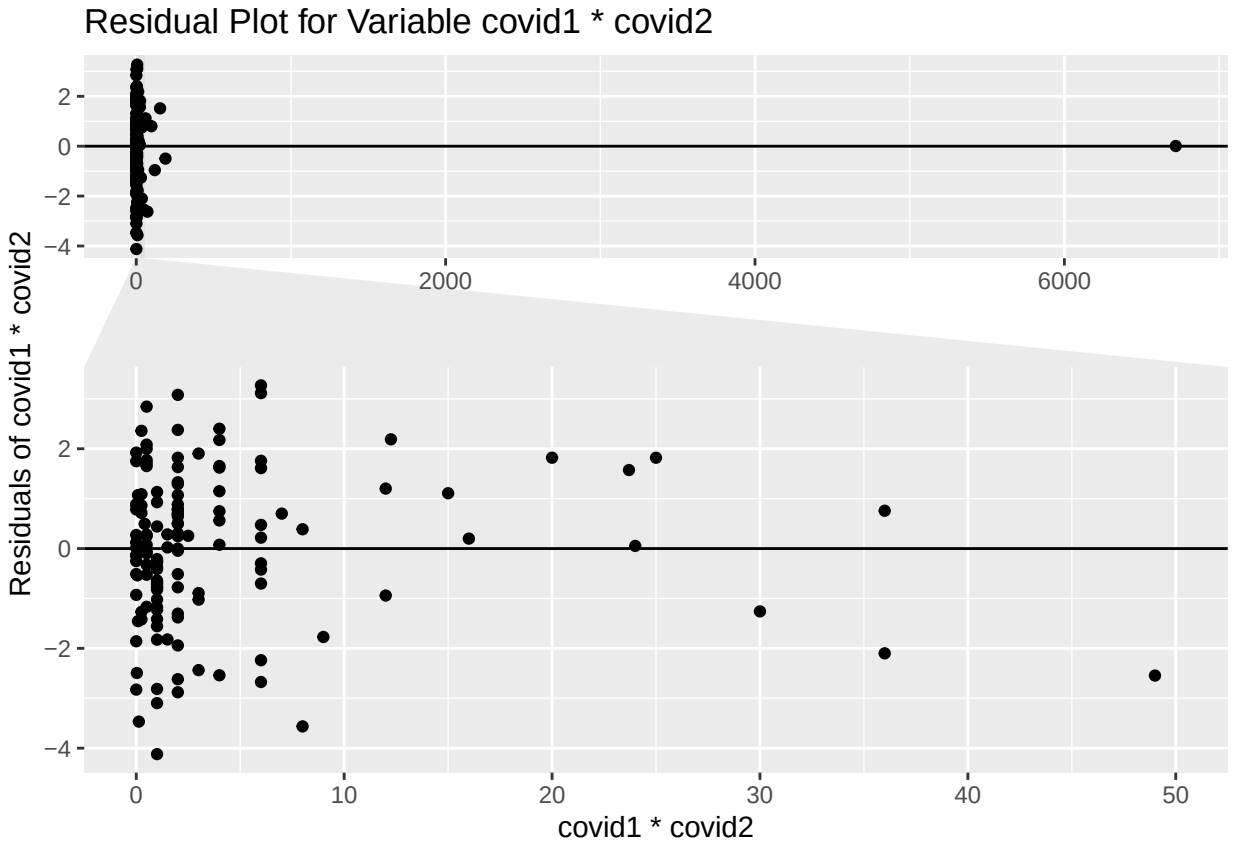


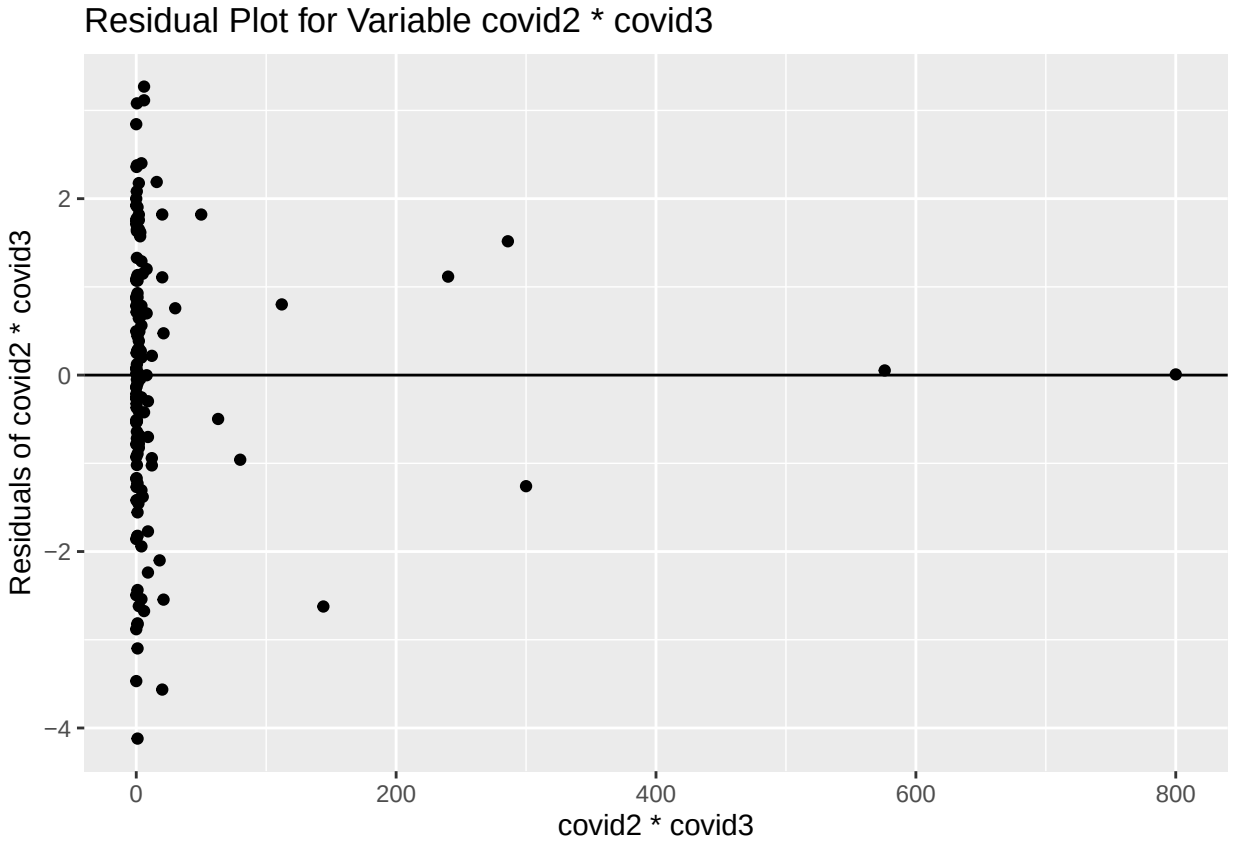
Assumption 2. Independence of Errors

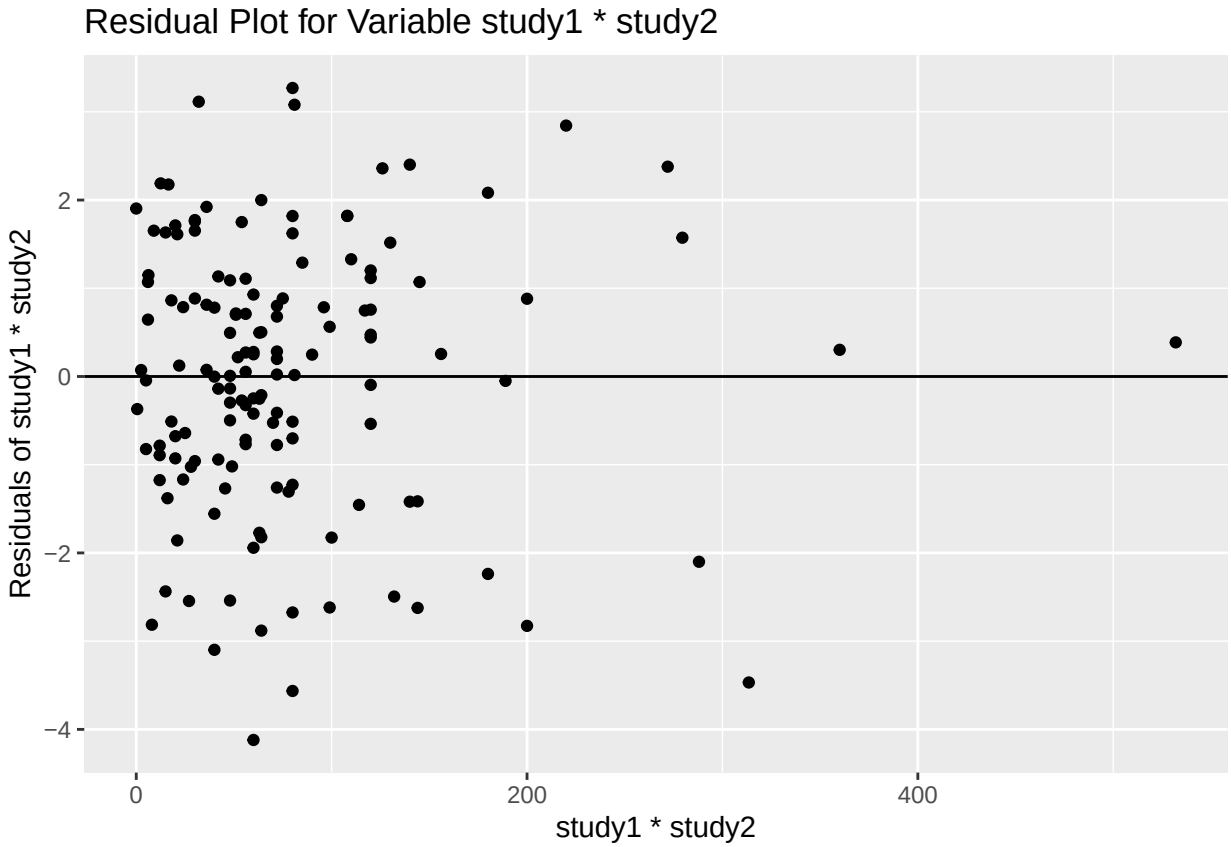
From above, the residual vs. covid1 plot for the fitted model show no discernible relationship, and neither do the residual plots for the other predictor variables (ingoring all outliers). The dataset comes from a random sample, so the error terms must be independent.

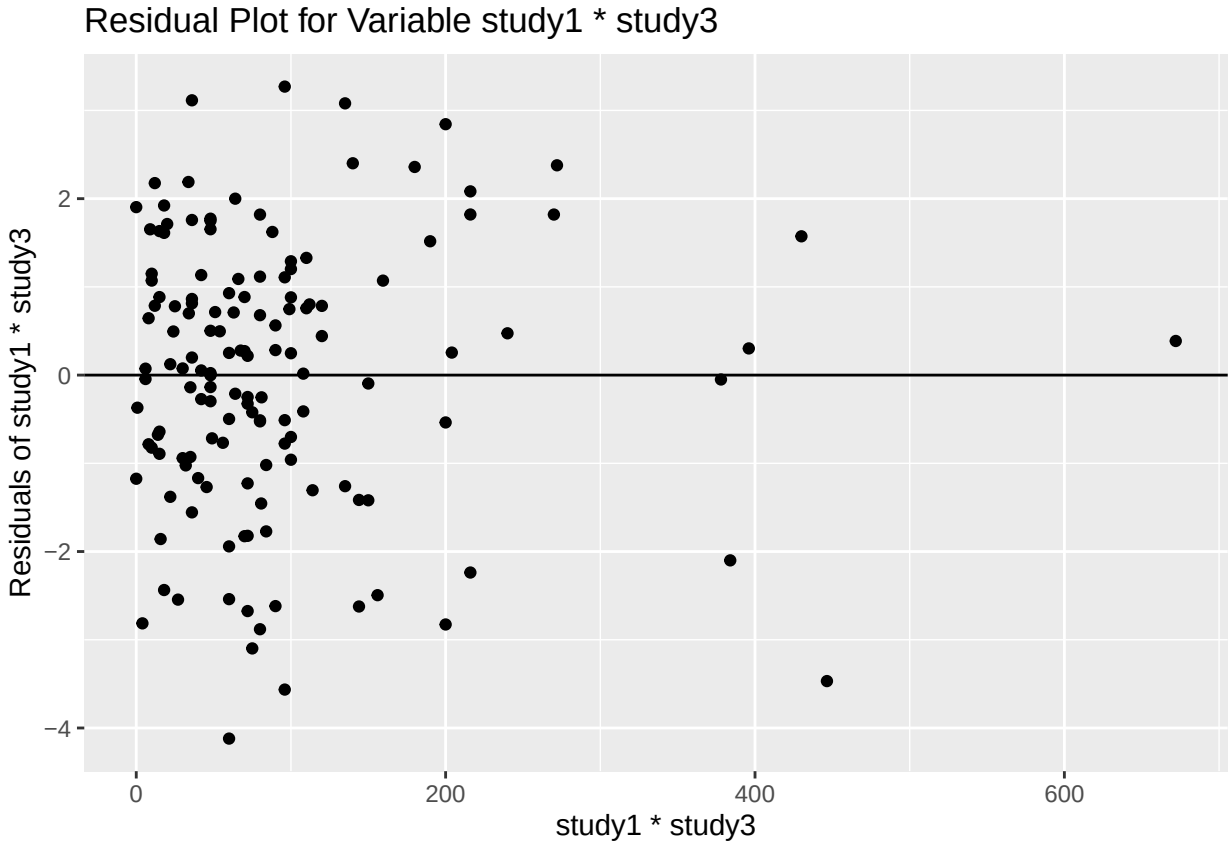

Residual Plot for Variable quiz3
Residuals of quiz3
2
0
−2
−4
2.5
5.0
7.5
10.0
quiz3

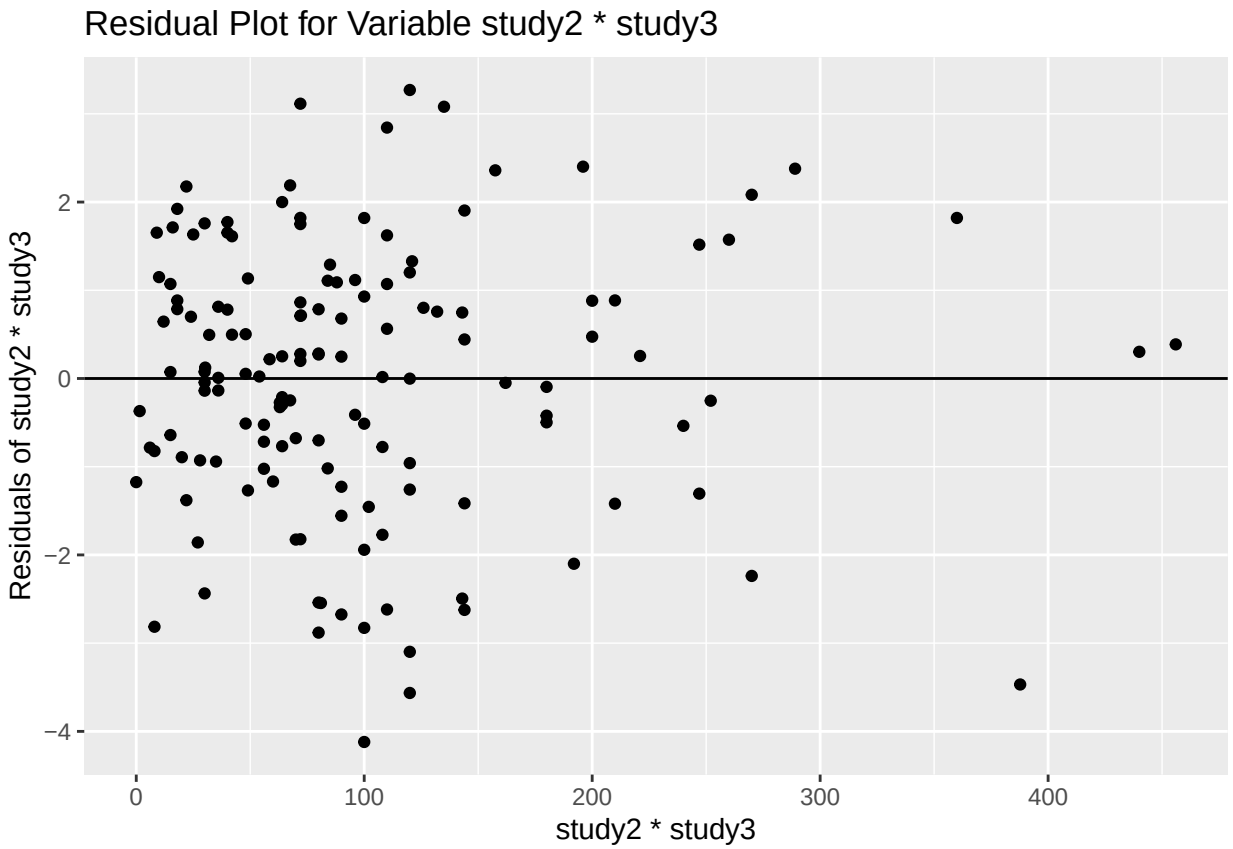

Residual Plot for Variable covid1 ^ 2
Residuals of covid1 ^ 2
2
0
−2
−4
0
10000
20000
0.0
2.5
5.0
7.5
10.0
covid1 ^ 2

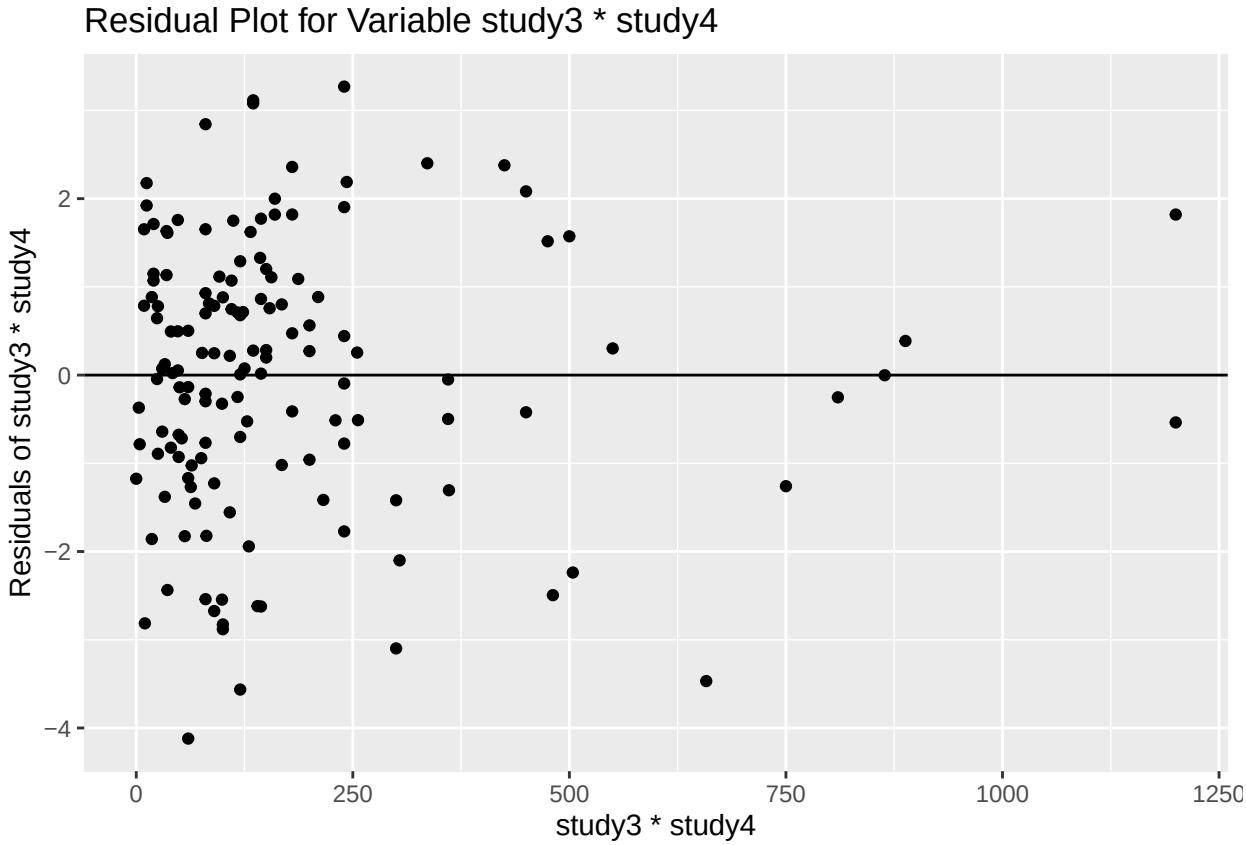

Residual Plot for Variable covid1 * covid2
Residuals of covid1 * covid2
covid1 * covid2
2
0
−2
−4
0
2000
4000
6000
10
20
30
40
50


Residual Plot for Variable covid2 * covid3
Residuals of covid2 * covid3
covid2 * covid3
2
0
−2
−4
0
200
400
600
800


Residual Plot for Variable study1 * study2
Residuals of study1 * study2
2
0
−2
−4
0
200
400
study1 * study2


Residual Plot for Variable study1 * study3
Residuals of study1 * study3
2
0
−2
−4
0
200
400
600
study1 * study3


Residual Plot for Variable study2 * study3
Residuals of study2 * study3
2
0
−2
−4
0
100
200
300
400
study2 * study3


Residual Plot for Variable study3 * study4
Residuals of study3 * study4
2
0
−2
−4
0
250
500
750
1000
1250
study3 * study4

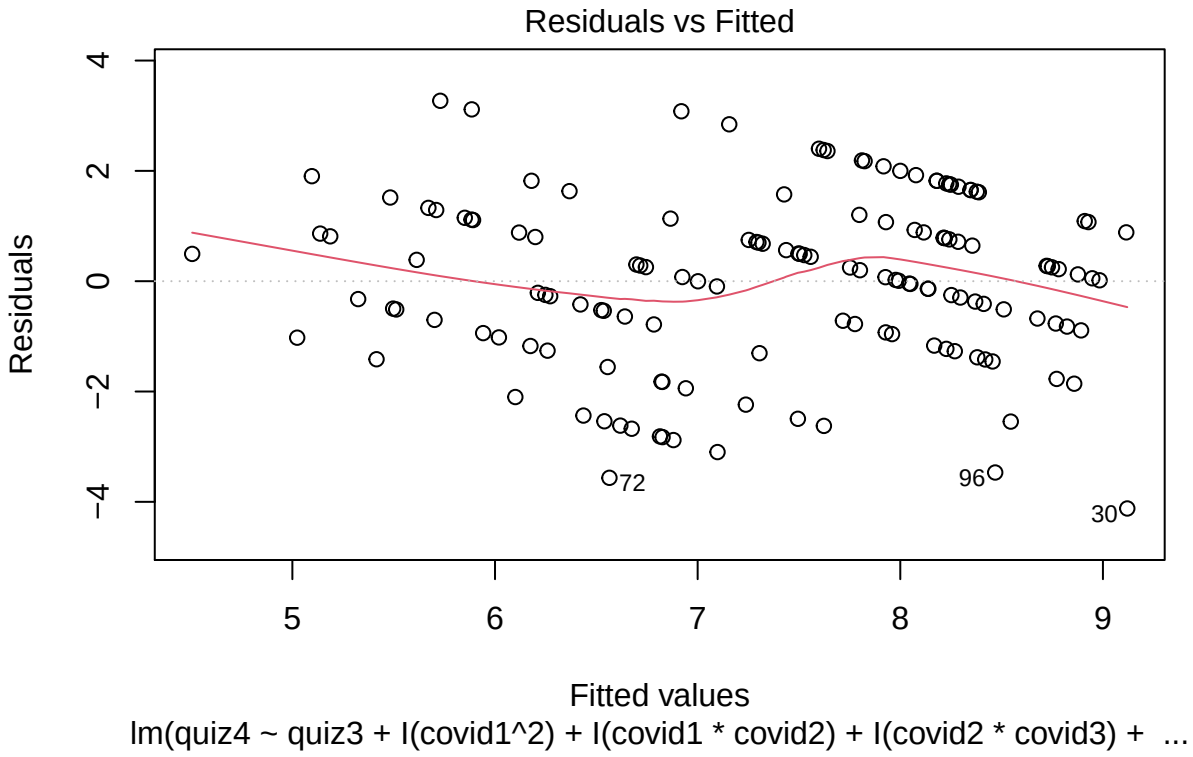
Assumption 3. Homoscedasticity (constant variance)

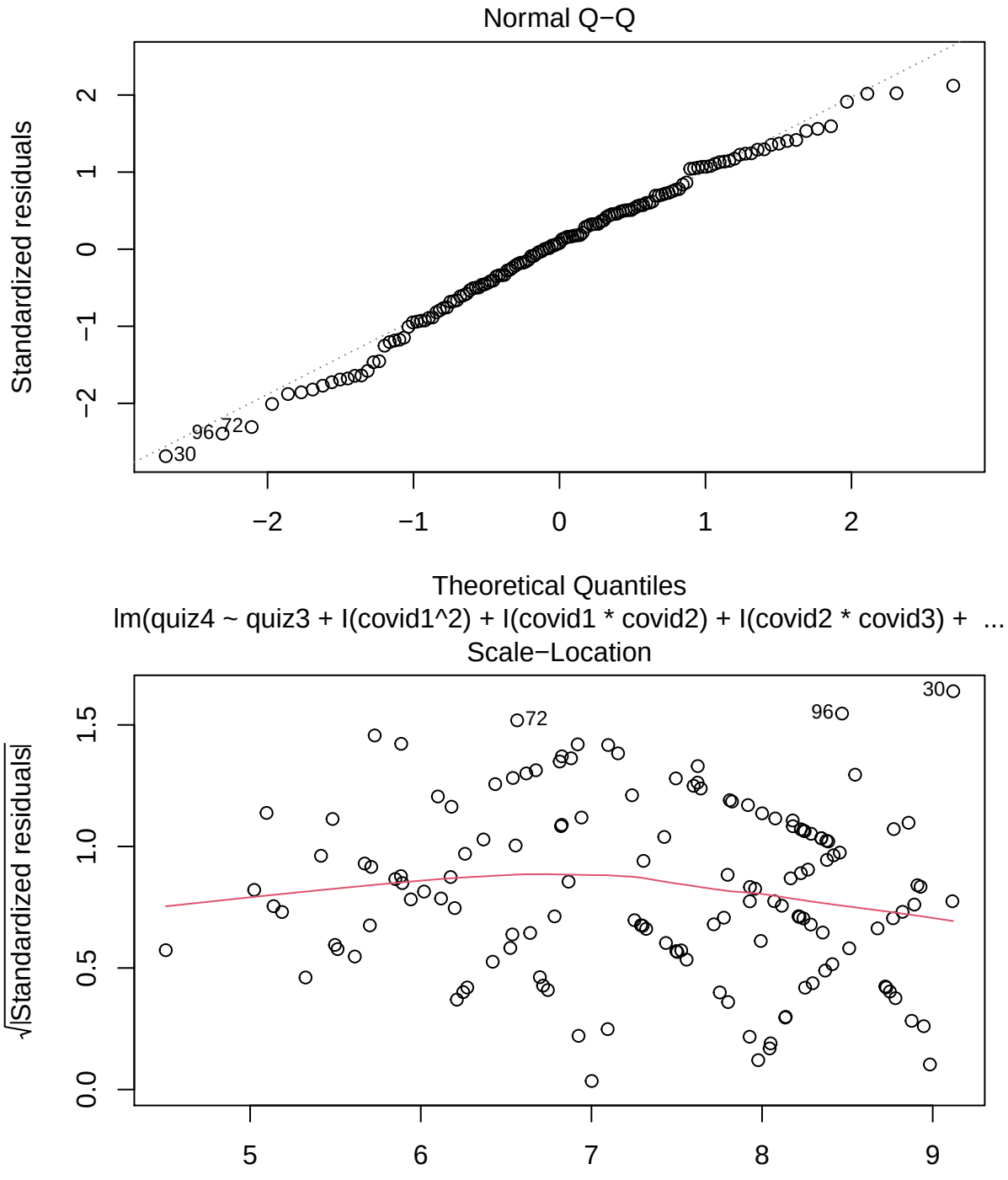
The scale-location model appears to have a straight horizontal line with randomly spread points, and the residual vs. fitted model show equally spread residuals around the horizontal line. Additionally, the residual plots for all predictor variables do not show any trend as fits increase. Therefore, the errors terms have constant variance.

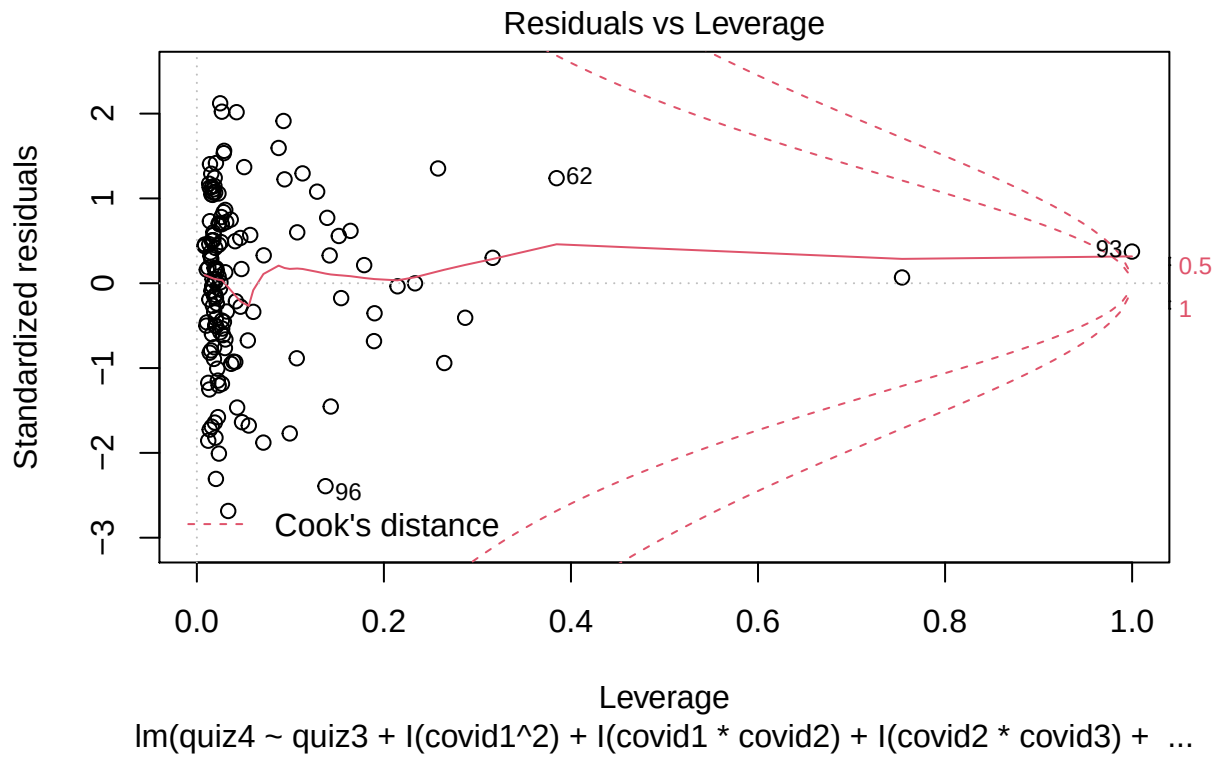
Assumption 4. Normality of Error

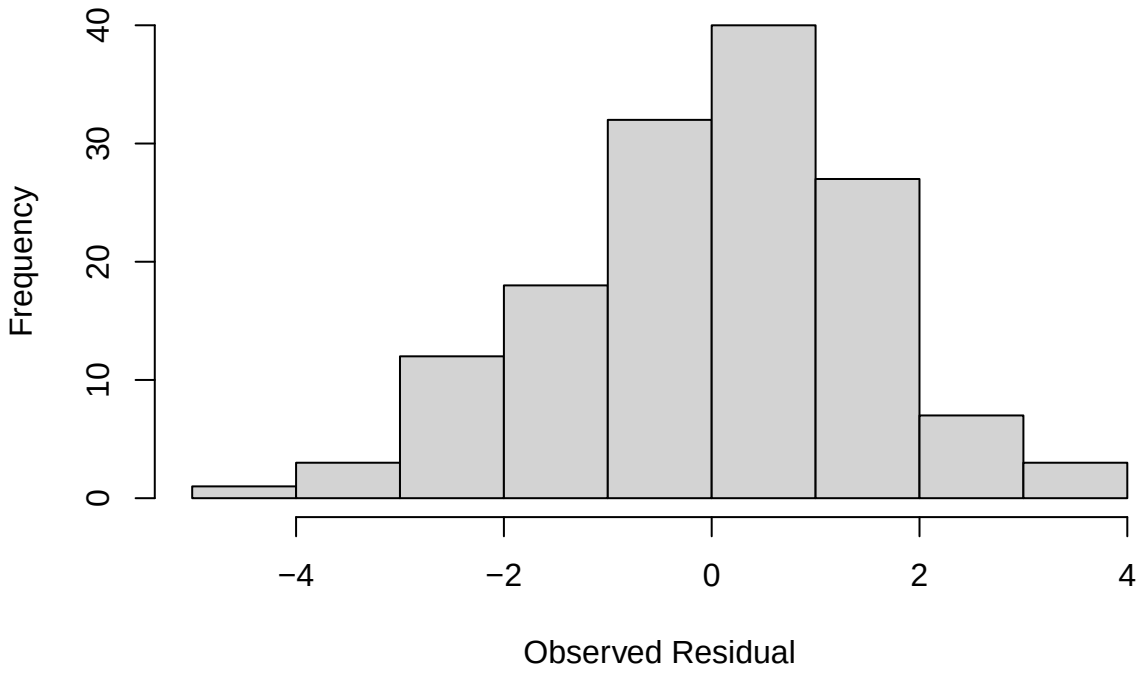
Almost all of the points in the middle follow the QQplot line very closely, except for the last 2 points in the right-tail which is only a little bit right-skewed. By inspection, the histograms of residuals for the final model looks approximately normal. Therefore, the error terms are approximately normal.

Hence, our model satisfies the assumptions for a quadratic model.




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 * covid3) + ...
Scale−Location
√|Standardized residuals|
Fitted values
lm(quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 * covid3) + ...


Residuals vs Leverage
Standardized residuals
Leverage
lm(quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 * covid3) + ...
Cook's distance
62
93
96
0.5
1


Histogram of Residuals
Frequency
Observed Residual

55/45 Training/Testing Split

Obtaining an independent dataset is infeasible, and the existing dataset has sufficiently many data points ($n = 143$ after removing all NAs). Therefore, it is possible to use 55% of the dataset to develop a regression model and the remaining 45% of the dataset to validate the trained model.


```
mean(predicted_values - actual_values)
```

```
## [1] -0.7223123
```

```
median(predicted_values - actual_values)
```

```
## [1] -0.1126939
```

Moreover, the histogram of residuals looks approximately symmetric, with the mean of this new distribution being -0.2102 (see figure X in appendix). To be sure, since there are $n = 142$ points, the central limit theorem states that the sample mean is approximately normally distributed.

T-test for significance

For our t-test, we hypothesized that the mean of residuals was identically equal to 0. More formally,

- $H_0 : \mu_{residuals} = 0$
- $H_1 : \mu_{residuals} \neq 0$

(TODO: Add t-test results in a nice table.)

The t-test results showed that the p-value is 0.7938, which means that we fail to reject H_0, meaning that $\mu_{residuals} = 0$ holds. The 95% confidence interval for the mean of residuals is $(-0.4510668, 0.3462948)$, and since $0 \in (-0.4510668, 0.3462948)$, we again fail to reject H_0 and conclude that $\mu_{residuals} = 0$.
(TODO: See figure X in appendix for t-test output.)

Therefore, we've shown that our linear model is a reasonable model.

Checking for Influential Outliers

The residuals vs. leverage plot was used to check for the presence of influential data points. There are no points in the upper right and lower right quadrants of the residuals vs. leverage plot. Even though there is a point sitting directly on the Cook's distance boundary but not too far away from the Cook's distance boundary, no influential outliers exist that could undermine the predictability of the final model.

Any Variable Transformation?

No variable transformations were performed on the final model because the error terms are independent, homoscedastic, and approximately normal.

Any or Recentering Necessary?

No variable re-centering was required since only a few entries in the correlation matrix have high correlation. Most of the entries in the correlation matrix had low to moderate correlation, so re-centering variables likely has negligible effect on the correlation matrix.

(TODO: Correlation matrix of final model.) (TODO: VIF for multicollinearity.)

Conclusion

Purpose of Final Model

Recall that the purpose of the model is to see whether previous quiz scores, time spent thinking about COVID, and study time can predict future quiz scores.

Interpretation of Final Model

The final model suggests a strong evidence of a positive relationship between quiz 3 scores and quiz 4 scores, as well as study times between consecutive weeks (i.e., (STA302W1, STA302W2), (STA302W2, STA302W3), (STA302W3, STA302W4)), keeping all other variables constant.

Quiz 3 is much closer in difficulty to Quiz 4 because students are used to the online Quercus quiz format. Students better understood how many decimal places they should round their final answers to from Quizzes 1 and 2, and the style of Quiz 3 questions are very similar to Quiz 4. Students also ramped up their study efforts as they anticipated more challenging quizzes as the semester progressed.

Although STA302H1 studying increased throughout the semester, the final model suggests that students who made a consistent effort to start studying during the 1st week of classes tend to score higher than students who started studying during the 2nd or 3rd week of classes. (TODO: Insert author here) shows that studying many hours last minute (mass learning) is less effective than studying a few hours a day throughout the term (spaced learning). By studying frequently throughout the semester, students have more opportunities to review STA302H1 material and their STA302H1 knowledge has more time to "compound" throughout the semester.

(TODO: Insert scholarly source about how far students start studying for quizzes in advance vs. test scores.)
(TODO: Insert scholarly source about how far students start studying for quizzes in advance improves material retention.)

Generalizability of Model

This model is only generalizable to online courses rather than in-person courses. Students may enroll in online courses from various time zones, as opposed to a standardized time zone for in person courses. STA302H1 is also a 3rd year course, so naturally 3rd year students tend to study more for their courses and score higher on average on quizzes than 1st year students.

Remaining Limitations and Problems with Model

This model fails to capture variables such as weekly anxiety levels, the weekly number of hours students sleep, and the weekly number of hours students participate in physical activity.

(TODO: Studies show high anxiety impairs course performance)
(TODO: Studies show high anxiety and more COVID think time)
(TODO: Studies show moderate sleep improves course performance)
(TODO: Studies show moderate exercise improves course performance)

The final dataset excluded 28 STA302H1 dropped students, and some blank entries remained for missing survey responses, and missing quiz scores remained due to some students skipping quizzes.

Some students may have underreported the number of study hours under the assumption that it only involves the number of hours spent attending lectures and the number of hours spend reviewing lecture notes, when in fact some students included hours spent doing assignments, writing quizzes, and even hours spent attending office hours – overall the number of hours spent doing any kind of coursework for STA302H1.

Despite these limitations, further research may help inform us on benefits of study and COVID think times vs. quiz scores.

Proposed Improvements with Model

One way to improve the model is to introduce composite variables in the model, such as a student happiness index as a function of a student's anxiety levels, COVID, and physical activity.

Another improvement is to use empirical research to propose some more new terms to improve model.

Another improvement to this model is to take the median of a student's 1 - 3 quizzes that they wrote, the median of a student's 1 - 4 COVID hours they report, and the median of a student's 1 - 4 study hours they report. Using the median as opposed to the mean makes these values less prone to skewness, and even smaller residuals since medians are more reliable statistic than mean.

Appendix

```
summary(remaining_data$COVID.hours..W1.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##     0.0     1.0     1.0     3.7     2.0   168.0      21
```

```
summary(remaining_data$COVID.hours..W2.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   1.000   2.869   2.000  40.000      19
```

```
summary(remaining_data$COVID.hours..W3.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   0.500   1.000   2.227   2.000  24.000      11
```

```
summary(remaining_data$COVID.hours..W4.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   1.000   1.500   2.917   3.000  50.000      13
```

```
summary(remaining_data$STA302.hours..W1.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.000   7.000   7.539   9.000  28.000      21
```

```
summary(remaining_data$STA302.hours..W2.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   8.000   8.403  10.000  20.000      19
```

```
summary(remaining_data$STA302.hours..W3.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    0.00    6.00    9.00    9.32   12.00   30.00      10
```

```
summary(remaining_data$STA302.hours..W4.)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    2.00    7.00   11.00   13.44   16.00   72.00      13
```

```r
summary(remaining_data$Quiz_1_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   7.000   8.000   7.738   9.000  10.000       8
```

```r
summary(remaining_data$Quiz_2_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   0.000   5.800   8.800   7.422   9.400  10.000       8
```

```r
summary(remaining_data$Quiz_3_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   5.000   8.000   7.209   9.000  10.000       3
```

```r
summary(remaining_data$Quiz_4_score)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   1.000   6.000   8.000   7.375   9.000  10.000       7
```

```r
canada <- remaining_data %>%
  filter(as.character(country) == "Canada") %>%
  dplyr::select(-country)

unknown <- remaining_data %>%
  filter(is.na(as.character(country))) %>%
  dplyr::select(-country)
```

```
##             Country
## Canada           97
## China            63
## India             2
## Japan             1
## Mongolia          1
## Pakistan          3
## Singapore         2
## South_Korea       2
## Taiwan            3
## UAE               2
## USA               2
## Unknown          21
```

```r
display_correlation_matrix(all_countries)
```

```
##       W1COV W2COV W3COV W4COV W1302 W2302 W3302 W4302    Q1    Q2    Q3    Q4
## W1COV  1.00  0.56  0.48  0.27  0.04 -0.03 -0.01  0.04  0.08  0.06  0.07  0.02
## W2COV  0.56  1.00  0.67  0.71  0.05  0.08  0.17  0.19  0.13 -0.10 -0.12 -0.01
## W3COV  0.48  0.67  1.00  0.72  0.08  0.08  0.14  0.13  0.09 -0.07 -0.11 -0.09
## W4COV  0.27  0.71  0.72  1.00  0.02  0.07  0.09  0.07  0.12 -0.10  0.02  0.06
```

```
## W1302  0.04  0.05  0.08  0.02  1.00  0.61  0.58  0.30  0.05  0.13 -0.04 -0.08
## W2302 -0.03  0.08  0.08  0.07  0.61  1.00  0.70  0.48  0.00  0.06 -0.05 -0.11
## W3302 -0.01  0.17  0.14  0.09  0.58  0.70  1.00  0.62 -0.01  0.08 -0.12 -0.08
## W4302  0.04  0.19  0.13  0.07  0.30  0.48  0.62  1.00 -0.01  0.04 -0.05 -0.06
## Q1     0.08  0.13  0.09  0.12  0.05  0.00 -0.01 -0.01  1.00  0.25  0.29  0.29
## Q2     0.06 -0.10 -0.07 -0.10  0.13  0.06  0.08  0.04  0.25  1.00  0.23  0.19
## Q3     0.07 -0.12 -0.11  0.02 -0.04 -0.05 -0.12 -0.05  0.29  0.23  1.00  0.55
## Q4     0.02 -0.01 -0.09  0.06 -0.08 -0.11 -0.08 -0.06  0.29  0.19  0.55  1.00
```

```
summary(first_model)
```

```
##
## Call:
## lm(formula = quiz4 ~ quiz1 + quiz2 + quiz3 + covid1 + I(covid1^2) +
##     covid2 + I(covid2^2) + covid3 + covid4 + I(covid4^2) + I(covid1 *
##     covid2) + I(covid2 * covid3) + I(covid2 * covid4) + I(covid3 *
##     covid4) + I(study1 * study2) + I(study1 * study3) + I(study2 *
##     study3) + I(study3 * study4) + country)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5884 -0.8610  0.1800  0.8824  3.2815
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.184400   0.797065   3.995 0.000114 ***
## quiz1                 0.034421   0.081102   0.424 0.672054
## quiz2                 0.047852   0.061074   0.784 0.434941
## quiz3                 0.477087   0.079290   6.017 2.16e-08 ***
## covid1                0.178659   0.126969   1.407 0.162094
## I(covid1^2)           0.016115   0.007279   2.214 0.028818 *
## covid2                0.289324   0.192110   1.506 0.134802
## I(covid2^2)          -0.023657   0.011719  -2.019 0.045850 *
## covid3               -0.053594   0.125976  -0.425 0.671317
## covid4               -0.248941   0.154339  -1.613 0.109497
## I(covid4^2)           0.020698   0.014617   1.416 0.159476
## I(covid1 * covid2)   -0.074201   0.033661  -2.204 0.029489 *
## I(covid2 * covid3)    0.050008   0.031997   1.563 0.120826
## I(covid2 * covid4)    0.040835   0.024083   1.696 0.092671 .
## I(covid3 * covid4)   -0.076459   0.050768  -1.506 0.134798
## I(study1 * study2)   -0.016578   0.006879  -2.410 0.017537 *
## I(study1 * study3)    0.007613   0.005076   1.500 0.136424
## I(study2 * study3)    0.007761   0.004604   1.686 0.094568 .
## I(study3 * study4)   -0.001958   0.001328  -1.474 0.143221
## countryChina          0.585571   0.344768   1.698 0.092127 .
## countryIndia          0.873927   1.174061   0.744 0.458175
## countryMongolia     -12.901734  19.426608  -0.664 0.507938
## countryPakistan      -0.148747   1.593692  -0.093 0.925800
## countrySingapore      1.191079   1.651695   0.721 0.472296
## countrySouth Korea   -0.015750   1.146622  -0.014 0.989064
## countryTaiwan        -1.213168   1.161154  -1.045 0.298309
## countryUAE           -0.631273   1.649231  -0.383 0.702598
## countryUSA            1.456298   1.765878   0.825 0.411256
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.582 on 115 degrees of freedom
## Multiple R-squared:  0.4211, Adjusted R-squared:  0.2851
## F-statistic: 3.098 on 27 and 115 DF,  p-value: 1.436e-05
```

```
summary(final_model)
```

```
##
## Call:
## lm(formula = quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) +
##     I(covid2 * covid3) + I(study1 * study2) + I(study1 * study3) +
##     I(study2 * study3) + I(study3 * study4))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1202 -0.9349  0.0755  1.0711  3.2701
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         3.874214   0.518252   7.476 8.98e-12 ***
## quiz3               0.499110   0.062645   7.967 6.22e-13 ***
## I(covid1^2)         0.004241   0.002104   2.016   0.0458 *
## I(covid1 * covid2) -0.018384   0.009007  -2.041   0.0432 *
## I(covid2 * covid3)  0.004796   0.002588   1.853   0.0661 .
## I(study1 * study2) -0.015069   0.006079  -2.479   0.0144 *
## I(study1 * study3)  0.006443   0.004472   1.441   0.1520
## I(study2 * study3)  0.009345   0.004246   2.201   0.0295 *
## I(study3 * study4) -0.002544   0.001195  -2.130   0.0350 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.561 on 134 degrees of freedom
## Multiple R-squared:  0.3435, Adjusted R-squared:  0.3043
## F-statistic: 8.764 on 8 and 134 DF,  p-value: 1.374e-09
```

```
##
##  One Sample t-test
##
## data:  predicted_values - actual_values
## t = -1.263, df = 65, p-value = 0.2111
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -1.8644405  0.4198159
## sample estimates:
##  mean of x
## -0.7223123
```