# STA302H1 – Final Project Descriptive Statistics

Danny Chen

August 10, 2021

## Import STA302H1 Study Time and COVID Contemplation Time vs. Quiz Performance Dataset

## Data Cleaning

First, I'll clean my data.

```
rearranged_data <- sta302_performance_data %>%
    # Create a new "country" column, which is just "Country" but whose entries are factors.
    mutate(country = as.factor(Country)) %>%

    # TODO: Replace quiz grades, covid hours, and sta302h1 study hours with their
    # TODO: median counterparts.

    # Remove the "X" column: it's simply the row number, which isn't very useful.
    # Remove the "Country" column: column "country" already exists
    dplyr::select(-X, -Country) %>%

    # Rearrange similar columns side-by-side.
    relocate(country,
             COVID.hours..W1., COVID.hours..W2., COVID.hours..W3., COVID.hours..W4.,
             STA302.hours..W1., STA302.hours..W2., STA302.hours..W3., STA302.hours..W4.,
             Quiz_1_score, Quiz_2_score, Quiz_3_score, Quiz_4_score)
```

## Helper Functions

```
num_column_NAs = function(predictor_variable) {
  sum(is.na(predictor_variable))
}
```

```
row_nums_of_NA_columns = function(data, predictor_variable) {
  which(is.na(predictor_variable))
}
```

```
rows_with_num_NAs = function(data, num_NAs) {
  return (rowSums(is.na(data)) == num_NAs)
}
```

```
row_nums_of_NA_rows = function(data, num_NAs) {
  return (which(rows_with_num_NAs(data, num_NAs)))
}
```

```
display_histogram <- function(data, predictor_variable, histogram_title, x_axis_label) {
  ggplot(data = tibble(data), mapping = aes(x = predictor_variable)) +
    geom_histogram(col = "black", fill = "red", bins = 30) +
    labs(title = histogram_title, y = "Frequency", x = x_axis_label) +
    geom_vline(mapping = aes(xintercept = mean(predictor_variable, na.rm = TRUE)),
               color = "blue", linetype = "solid") +
    geom_vline(mapping = aes(xintercept = median(predictor_variable, na.rm = TRUE)),
               color = "dark green", linetype = "dotted")
}
```

```
display_boxplot <- function(data, predictor_variable, boxplot_title, y_axis_label) {
  ggplot(mapping = aes(x = Country, y = predictor_variable)) +
    geom_boxplot(mapping = aes(x = Country, y = predictor_variable)) +
    labs(title = boxplot_title, x = "Country", y = y_axis_label)
}
```

```
get_row_nums_to_exclude <- function(data) {
  row_nums_with_3_NAs = which(rows_with_num_NAs(data, 3))
  row_nums_with_4_NAs = which(rows_with_num_NAs(data, 4))
  row_nums_to_exclude <- union(row_nums_with_3_NAs,
                               row_nums_with_4_NAs)
  return (row_nums_to_exclude)
}
```

```
display_correlation_by_country <- function(country_data) {
  colnames(country_data) <- c("W1COV", "W2COV", "W3COV", "W4COV",
                              "W1302", "W2302", "W3302", "W4302",
                              "Q1", "Q2", "Q3", "Q4")
  round(cor(country_data, use = "pairwise.complete.obs", method = "pearson"), 2)
}
```

```r
display_residual_plot <- function(data, model, predictor_variable, predictor_variable_name) {
  fit = fitted(model)
  residuals = resid(model)
  ggplot(data = data, aes(x = predictor_variable, y = residuals)) +
    geom_point() +
    geom_hline(yintercept = 0) +
    labs(title = paste0("Residual Plot for Variable ", predictor_variable_name),
         x = predictor_variable_name, y = paste0("Residuals of ", predictor_variable_name))
}
```

# Special Tables

## Rows With At Least One NA

Rows with at least one NA deserve closer examination.
Some of the rows might only have 1 - 2 NAs and are therefore salvageable, which is OK.
Other rows may contain 3 or more NAs, and might indicate students who have dropped STA302H1. We'd like to exclude them from our analysis.

Here are the number of rows with 0 - 4 NAs.

```
##   nrows_0_NAs nrows_1_NAs nrows_2_NAs nrows_3_NAs nrows_4_NAs
## 1         143           9          16          19           1
```

## Columns with NAs

```
##   week1_covid week2_covid week3_covid week4_covid
## 1          26          22          21          40
```

```
##   week1_sta302 week2_sta302 week3_sta302 week4_sta302
## 1           26           22           20           40
```

```
##   quiz1_score quiz2_score quiz3_score quiz4_score
## 1          13          36          31          34
```

## Number of Missed Quizzes

```
##   miss_0_quizzes miss_1_quizzes miss_2_quizzes miss_3_quizzes miss_4_quizzes
## 1            176             20              3             24              4
```

## Who to Exclude from the Dataset?

Identify rows with at least 3 missing quiz marks. These indicate students who have dropped STA302H1, and who should be excluded from the final data.

Notice that we didn't check the number of NAs for country of origin, COVID hours, and STA302H1 hours, since some students either forgot or abstained. So there's no reason to exclude these students from our final dataset.

```
row_nums_to_exclude <- get_row_nums_to_exclude(quiz_grades)
remaining_data = rearranged_data[-row_nums_to_exclude,]
```

### Rows with Mistyped Columns

Rows whose columns are mis-typed may need to be corrected via imputation.

```
rows_with_mistyped_columms = remaining_data[c(38, 83, 84, 117),]
# row 83: Country -> "canada" -- DONE
# row 84: Country -> "canada" -- DONE

# row 117: COVID.hours..W4. -> 0.5 hours -- DONE

# row 38:  STA302.hours..W3. -> 5.5<U+00A0> -- DONE
# row 117: STA302.hours..W4. -> 7.5 hours -- DONE
```

```
# library(janitor)
# use it to clean up data.
```

### Rows Without Country Entry

Taking out the country column can come in handy for functions like cor() where factors aren't allowed.

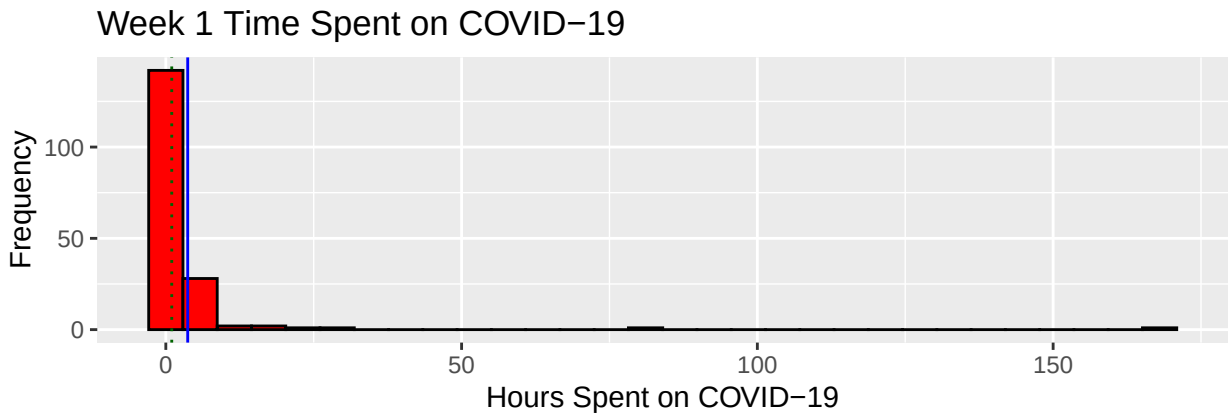
```
rows_with_no_country = remaining_data %>%
  dplyr::select(-country)
```

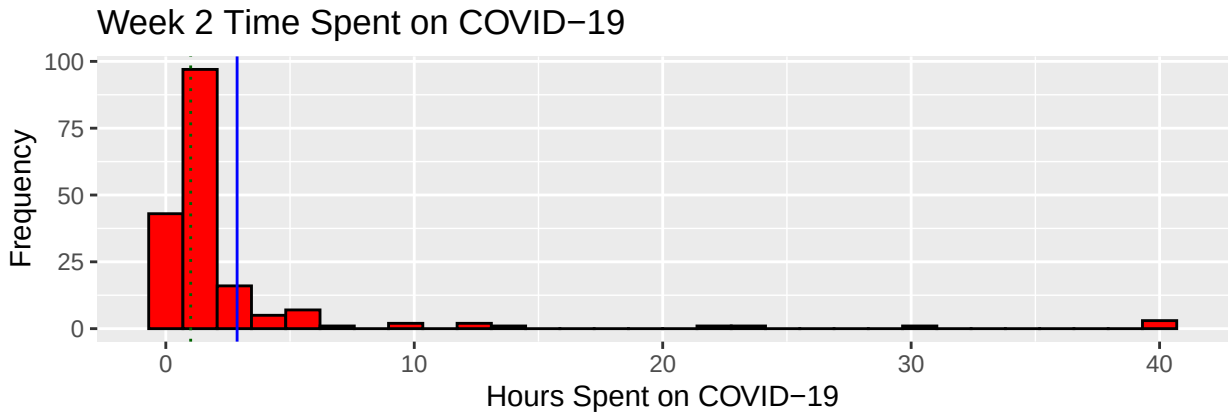
### Rows Filtered by Country

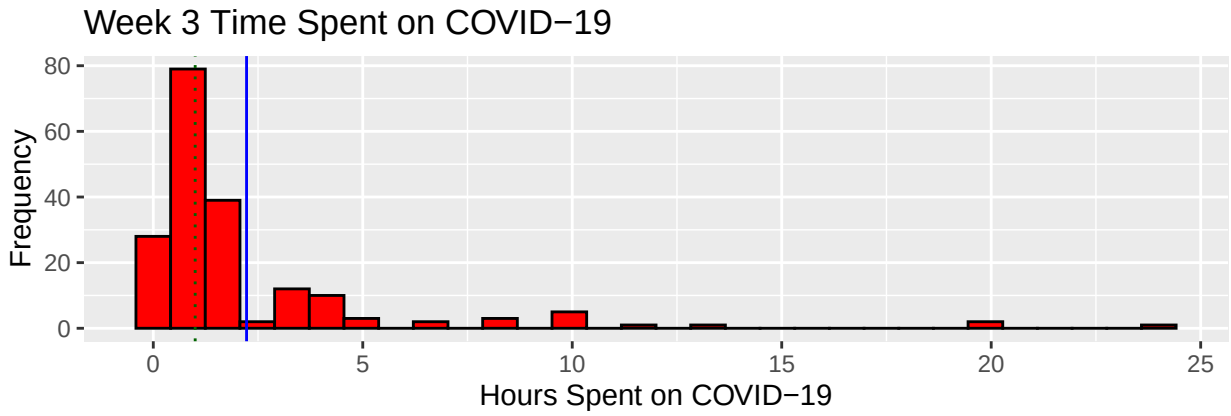
This is useful if we want data for individual countries.
Only the first and last code snippets are shown.

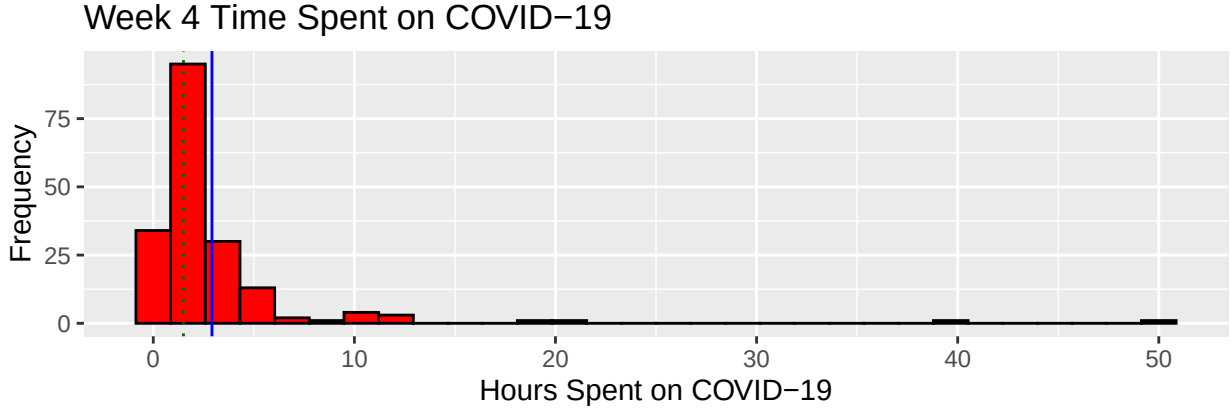
```
canada <- remaining_data %>%
  filter(as.character(country) == "Canada") %>%
  dplyr::select(-country)

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

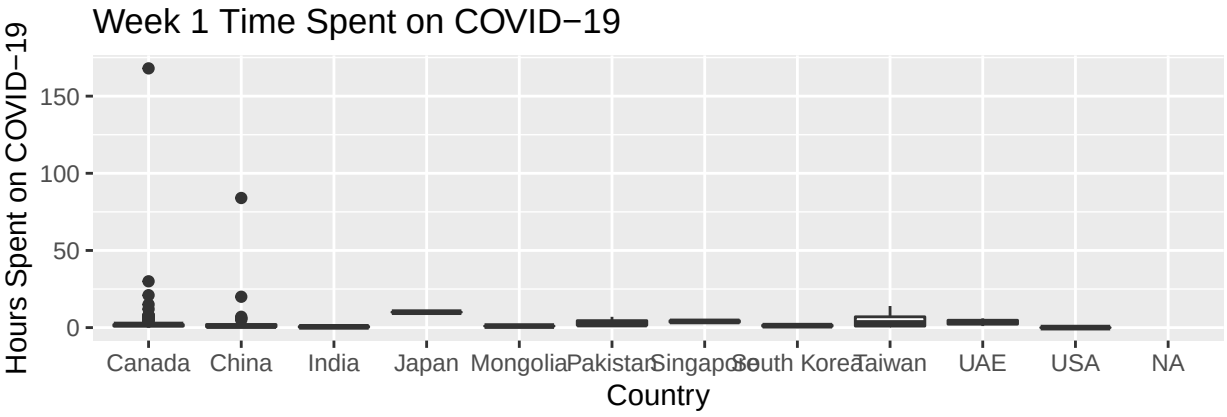
```
##             Country
## Canada           97
## China            63
## India             2
## Japan             1
## Mongolia          1
## Pakistan          3
## Singapore         2
## South_Korea       2
## Taiwan            3
## UAE               2
## USA               2
## Unknown          21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

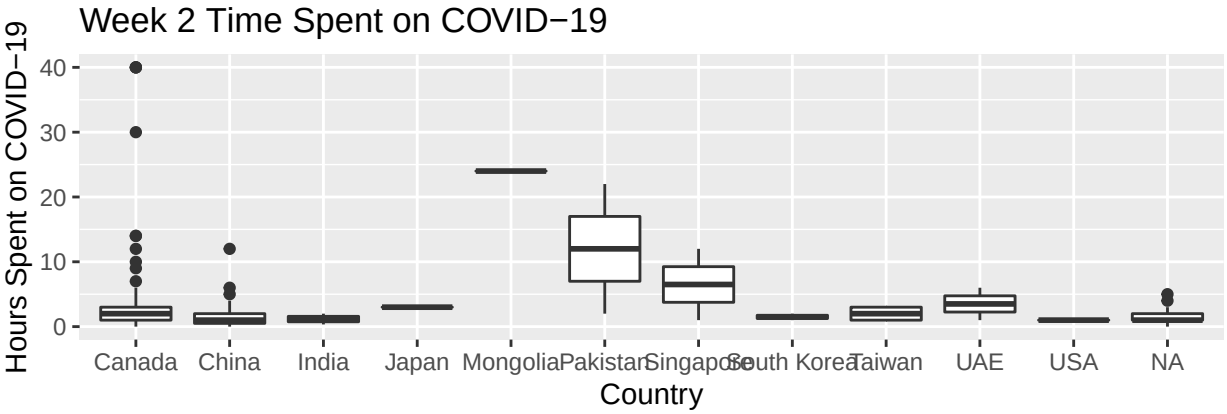

Week 1 Time Spent on COVID−19
Frequency
100
50
0
0
50
100
150
Hours Spent on COVID−19

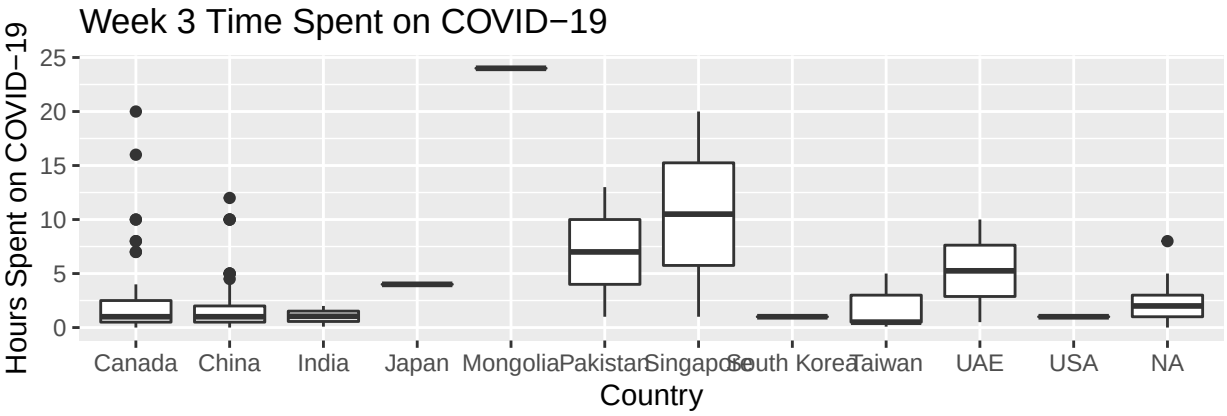

Week 2 Time Spent on COVID−19
Frequency
100
75
50
25
0
0
10
20
30
40
Hours Spent on COVID−19

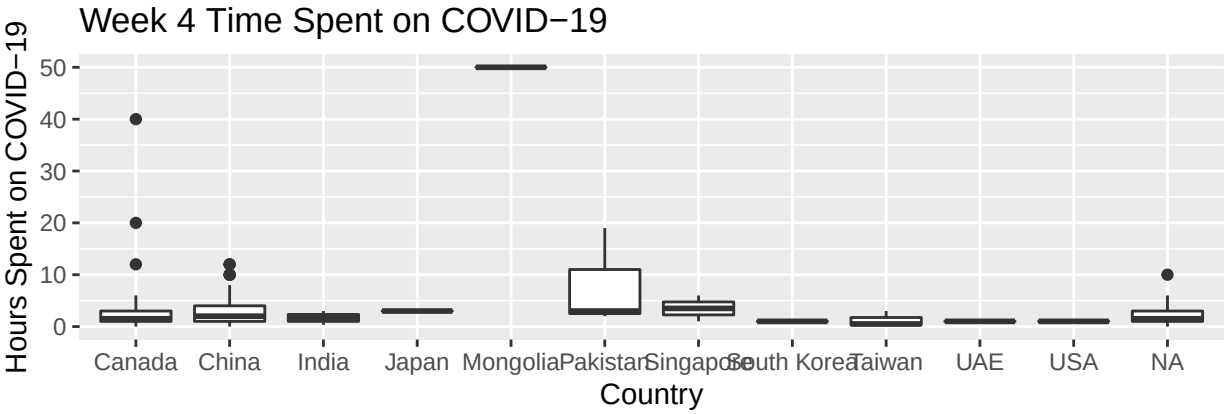

Week 3 Time Spent on COVID−19
Frequency
80
60
40
20
0
0
5
10
15
20
25
Hours Spent on COVID−19

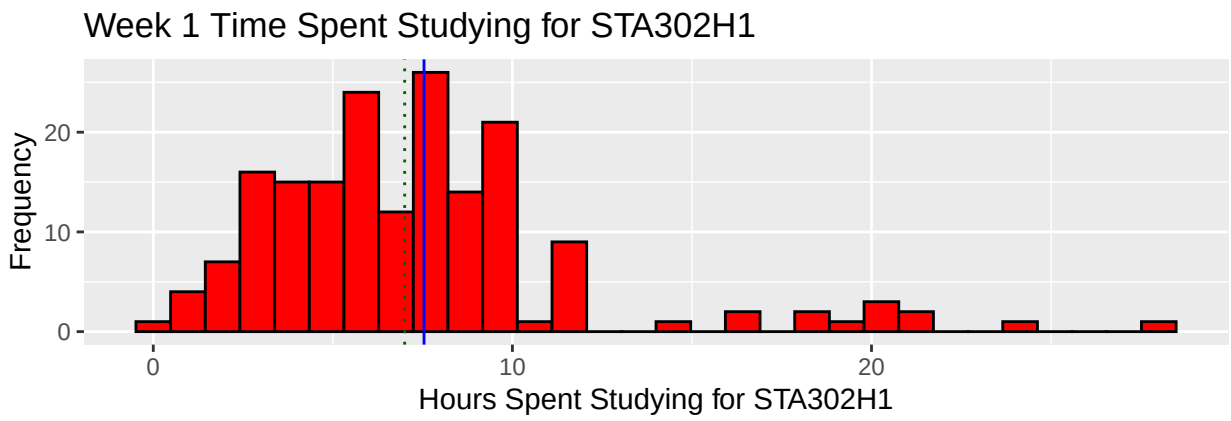

Week 4 Time Spent on COVID−19
Frequency
75
50
25
0
0
10
20
30
40
50
Hours Spent on COVID−19

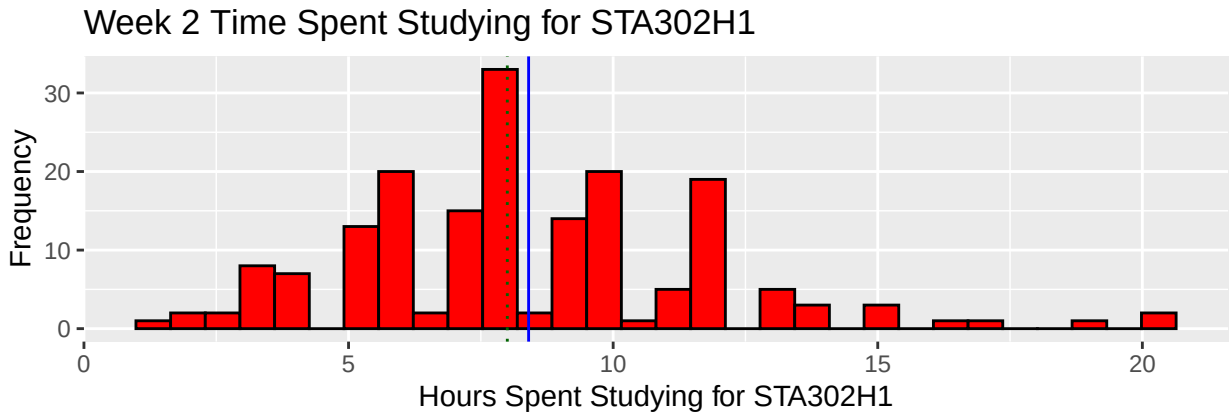

Week 1 Time Spent on COVID−19
Hours Spent on COVID−19
150
100
50
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

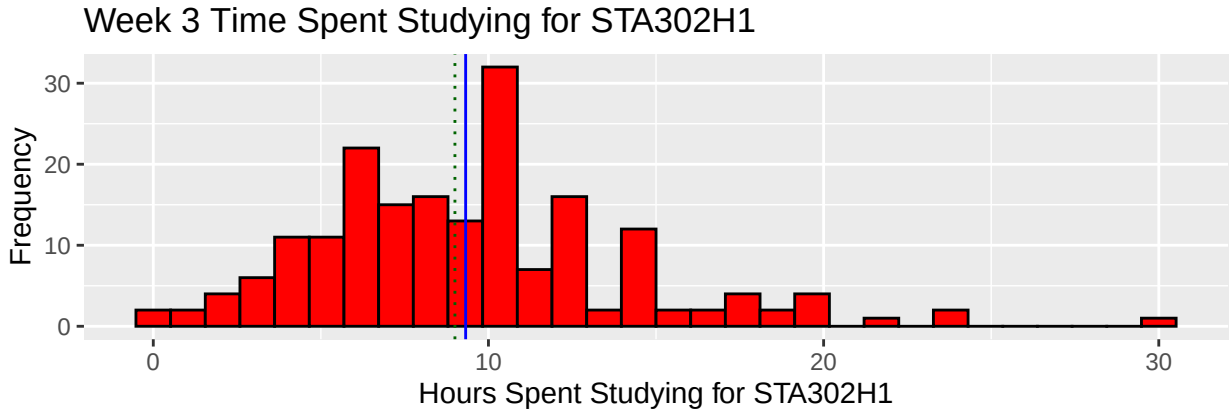

Week 2 Time Spent on COVID−19
Hours Spent on COVID−19
40
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

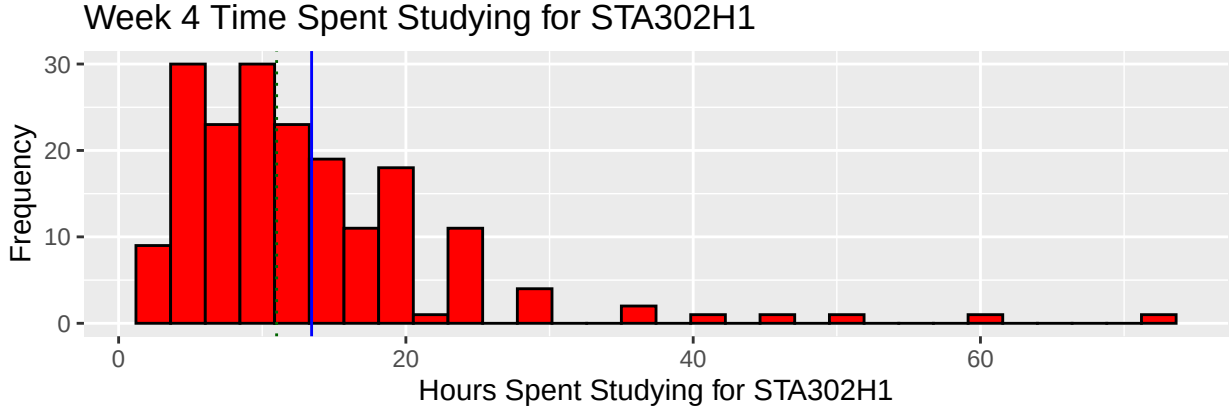

Week 3 Time Spent on COVID−19
Hours Spent on COVID−19
25
20
15
10
5
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

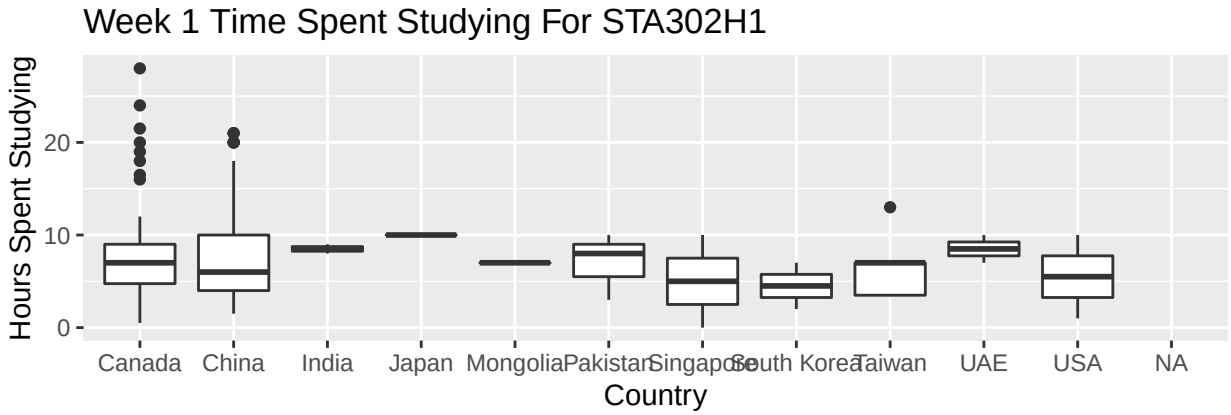

Week 4 Time Spent on COVID−19
Hours Spent on COVID−19
50
40
30
20
10
0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

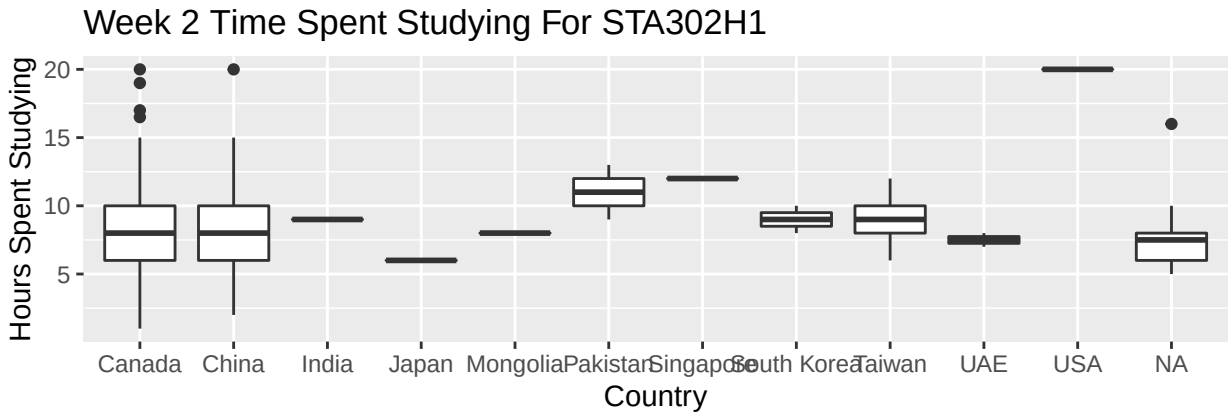

Week 1 Time Spent Studying for STA302H1
Frequency
0
10
20
0
10
20
Hours Spent Studying for STA302H1

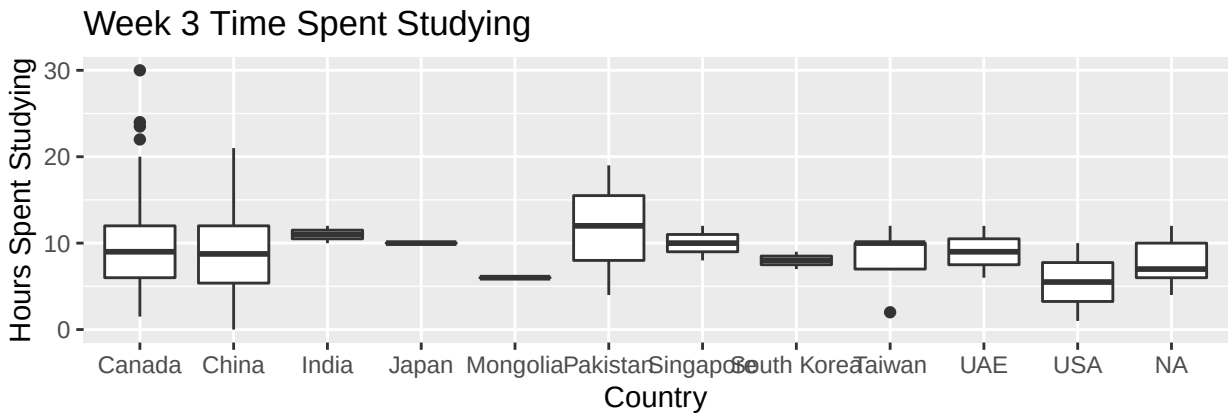

Week 2 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
5
10
15
20
Hours Spent Studying for STA302H1

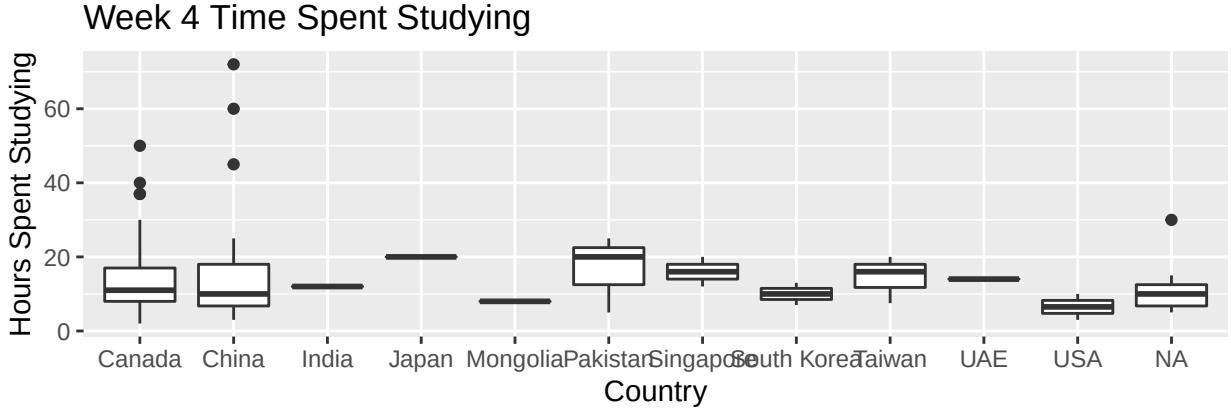

Week 3 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
10
20
30
Hours Spent Studying for STA302H1

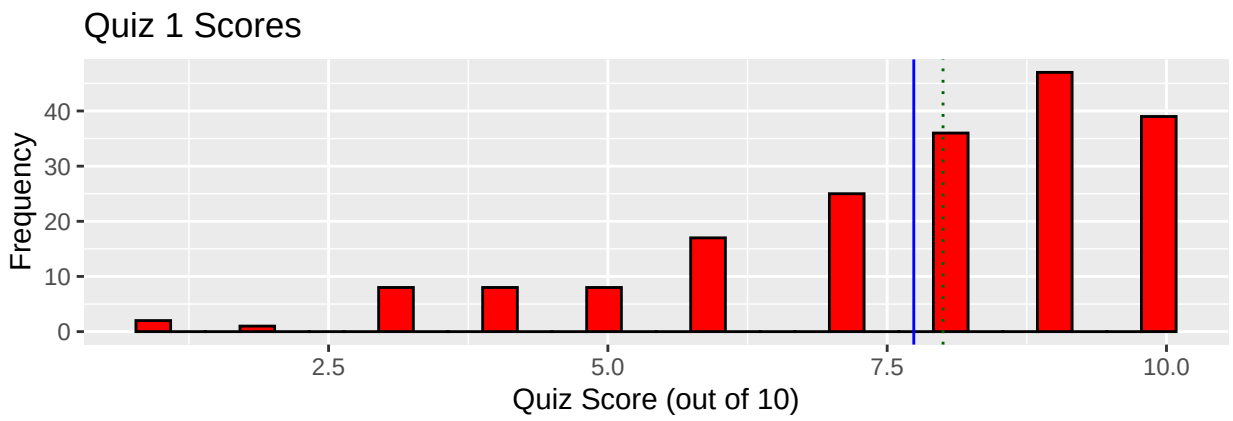

Week 4 Time Spent Studying for STA302H1
Frequency
0
10
20
30
0
20
40
60
Hours Spent Studying for STA302H1

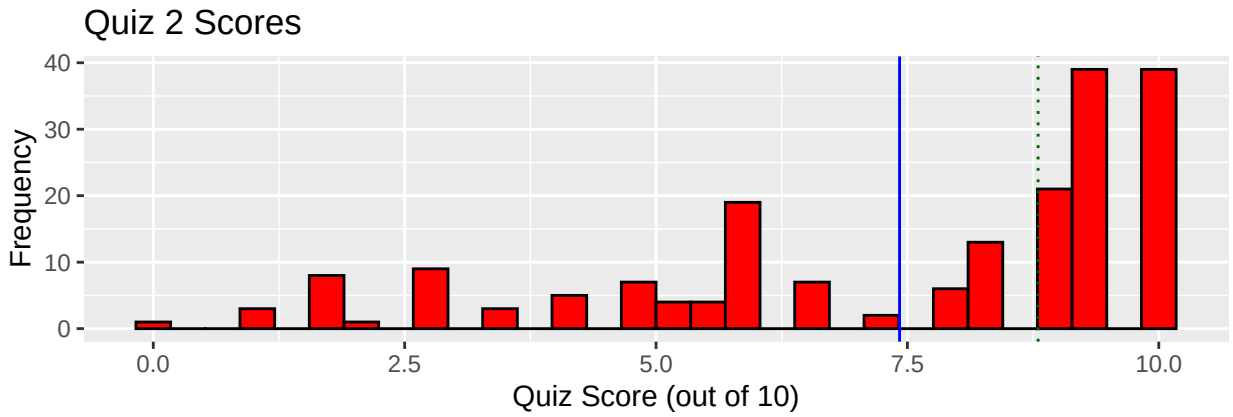

Week 1 Time Spent Studying For STA302H1
Hours Spent Studying
0
10
20
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

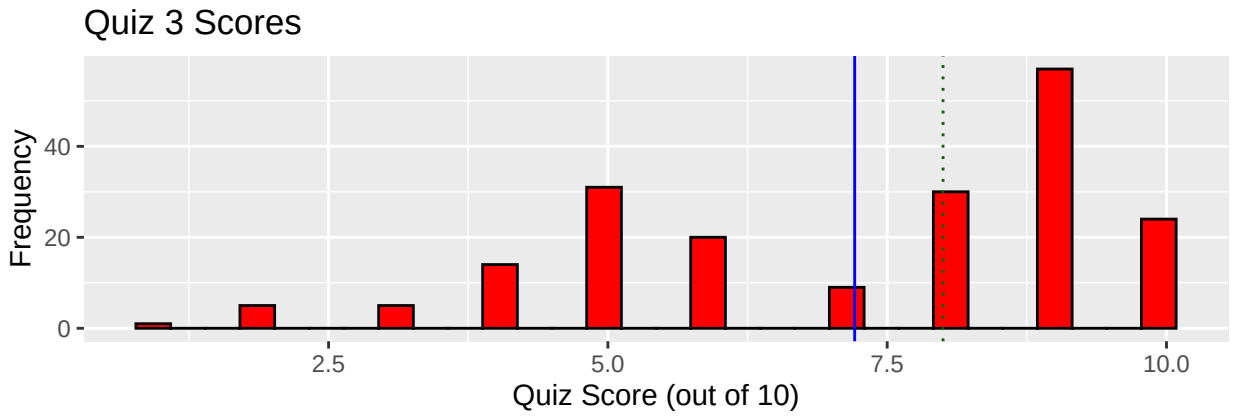

Week 2 Time Spent Studying For STA302H1
Hours Spent Studying
5
10
15
20
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

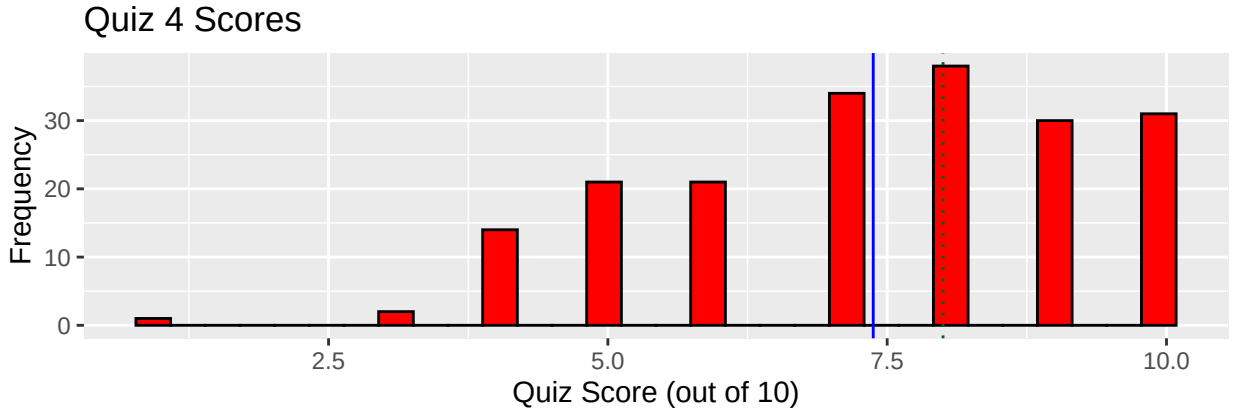

Week 3 Time Spent Studying
Hours Spent Studying
0
10
20
30
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

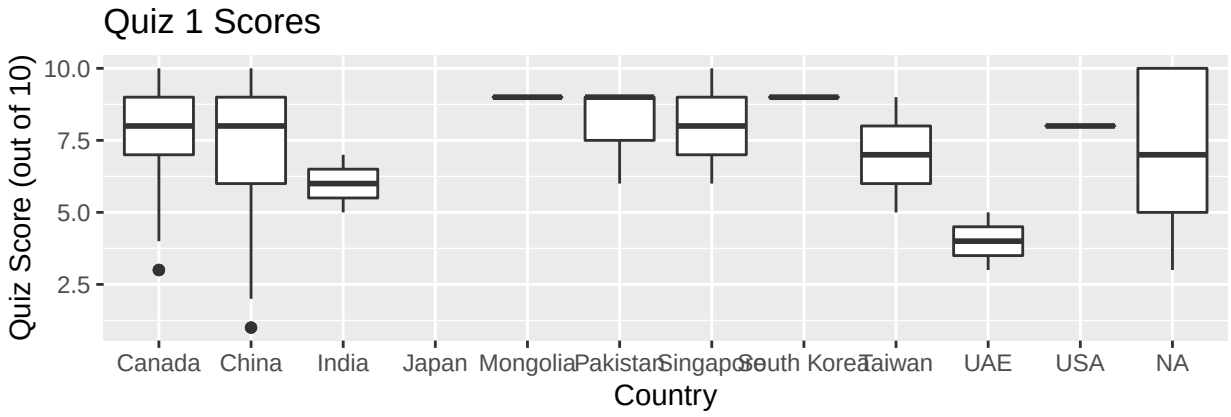

Week 4 Time Spent Studying
Hours Spent Studying
0
20
40
60
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

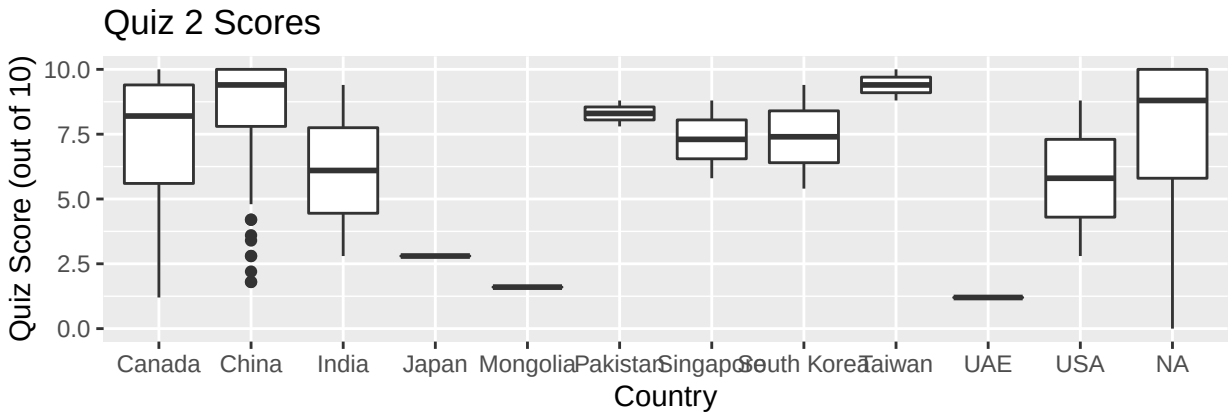

Quiz 1 Scores
Frequency
0
10
20
30
40
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

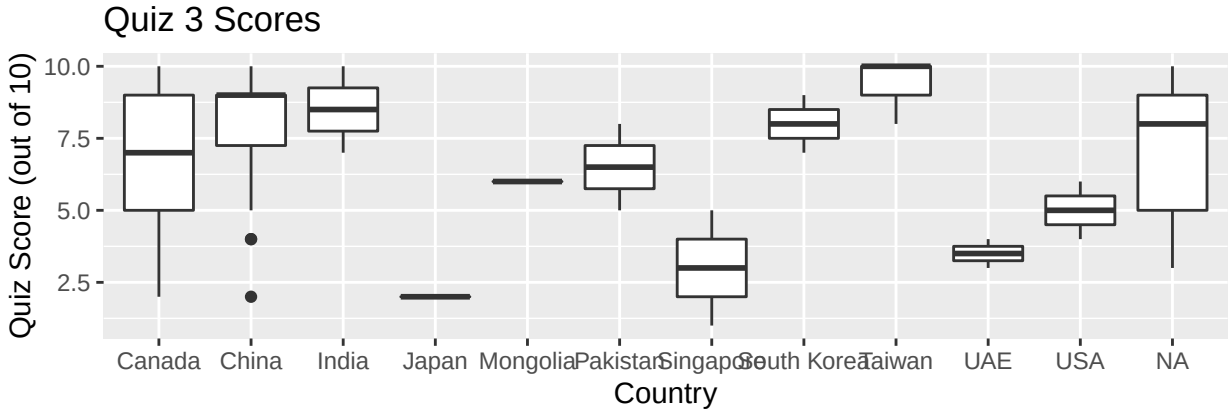

Quiz 2 Scores
Frequency
0
10
20
30
40
0.0
2.5
5.0
7.5
10.0
Quiz Score (out of 10)

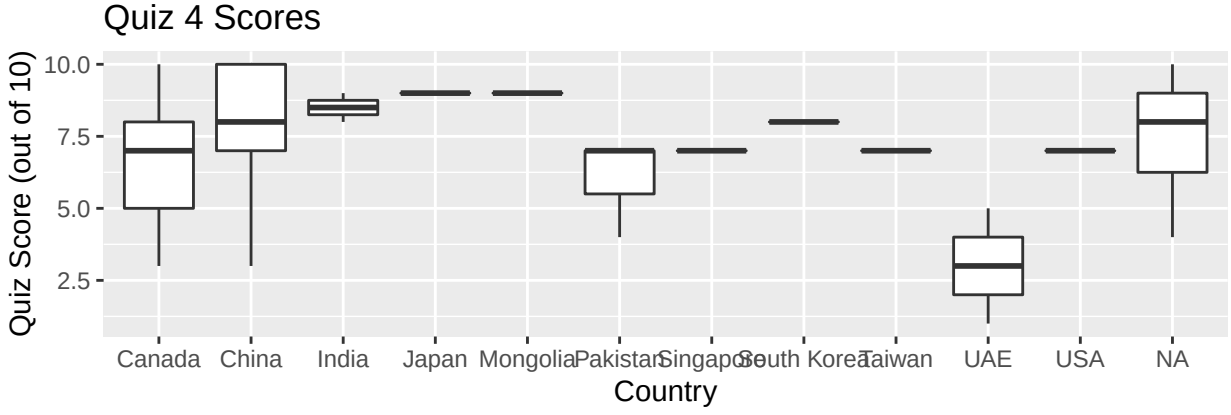

Quiz 3 Scores
Frequency
0
20
40
2.5
5.0
7.5
10.0
Quiz Score (out of 10)


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


Quiz 1 Scores
Quiz Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country


Quiz 2 Scores
Quiz Score (out of 10)
10.0
7.5
5.0
2.5
0.0
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country


Quiz 3 Scores
Quiz Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country


Quiz 4 Scores
Quiz Score (out of 10)
10.0
7.5
5.0
2.5
Canada
China
India
Japan
Mongolia
Pakistan
Singapore
South Korea
Taiwan
UAE
USA
NA
Country

## 5-Number Summary Statistics

```
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
```

### Full Model

```
# single variable per term = additive model
first_model = lm(
  quiz4 ~
    quiz1  # scatterplot seems to have no relationship
  + quiz2  # scatterplot seems to have no relationship
  + quiz3  # scatterplot looks more linear
  + covid1 # must add this linear term b/c i have a quadratic term
  + I(covid1 ^ 2)  # scatterplot looks more quadratic
  + covid2 # must add this linear term b/c i have a quadratic term
  + I(covid2 ^ 2)  # scatterplot looks more quadratic
  + covid3
  # + I(covid3 ^ 2) # scatterplot looks less quadratic
  + covid4  # must add this linear term b/c i have a quadratic term
  + I(covid4 ^ 2) # scatterplot looks more quadratic
  + I(covid1 * covid2) # first impressions from correlation matrix
  + I(covid2 * covid3) # correlation = 0.67
  + I(covid2 * covid4) # discard: correlation = 0.71
  + I(covid3 * covid4) # correlation = 0.72
  + I(study1 * study2)  # correlation = 0.61
  + I(study1 * study3)  # correlation = 0.58
  + I(study2 * study3)  # correlation = 0.70
  + I(study3 * study4)  # correlation = 0.62
  + country  # for simplicity, but backwards process shows this term is not significant
)
summary(first_model)
```

```
stepAIC(first_model, direction = "forward")$anova
```

```
stepAIC(first_model, direction = "backward")$anova
```

```
stepAIC(first_model, direction = "both")$anova
```

### Final Model

```
final_model = lm(
  quiz4 ~ quiz3
  + I(covid1 ^ 2)   # don't remove, else all other terms become insignificant
  + I(covid1 * covid2)
  + I(covid2 * covid3)  # don't remove, else all other terms become insignificant
  + I(study1 * study2)
  + I(study1 * study3)  # maybe don't remove?
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(final_model)
```

```
stepAIC(final_model, direction = "forward")$anova
```

```
stepAIC(final_model, direction = "backward")$anova
```

```
stepAIC(final_model, direction = "both")$anova
```

### Final Model with Some Terms I Pruned Myself

```
# I decide to remove more terms for simplicity.
third_model = lm(
quiz4 ~ quiz3
  # + I(covid1 ^ 2)  # this lone quadratic term add a lot of complexity for negligible change in R^2 an
  # + I(covid1 * covid2) + I(covid2 * covid3) # these terms alone add complexity -- harder to interpret
  + I(study1 * study2)
# + I(study1 * study3)  # make weeks consecutive: "want to see correlation from week to week", rather t
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(third_model)
```

```
# Doing stepAIC on a well-fitted model produces the same model.
# The model is already in a "steady state."
stepAIC(third_model, direction = "both")$anova
```

```
stepAIC(third_model, direction = "forward")$anova
```

```
stepAIC(third_model, direction = "backward")$anova
```

### Simplistic Model

```
fourth_model = lm(quiz4 ~ quiz3)
summary(fourth_model)
```

```
stepAIC(fourth_model, direction = "forward")$anova
```

```
stepAIC(fourth_model, direction = "backward")$anova
```

```
stepAIC(fourth_model, direction = "both")$anova
```

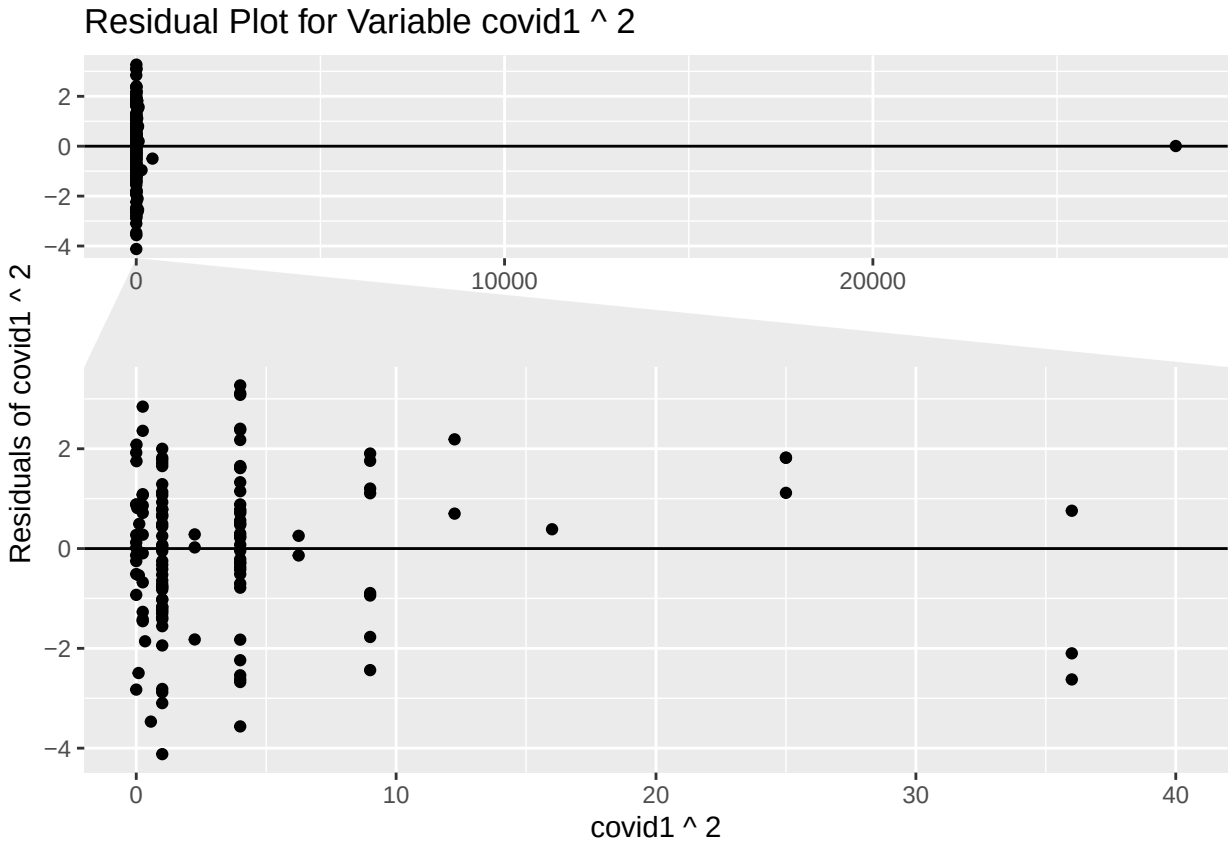
### Linear Model Only

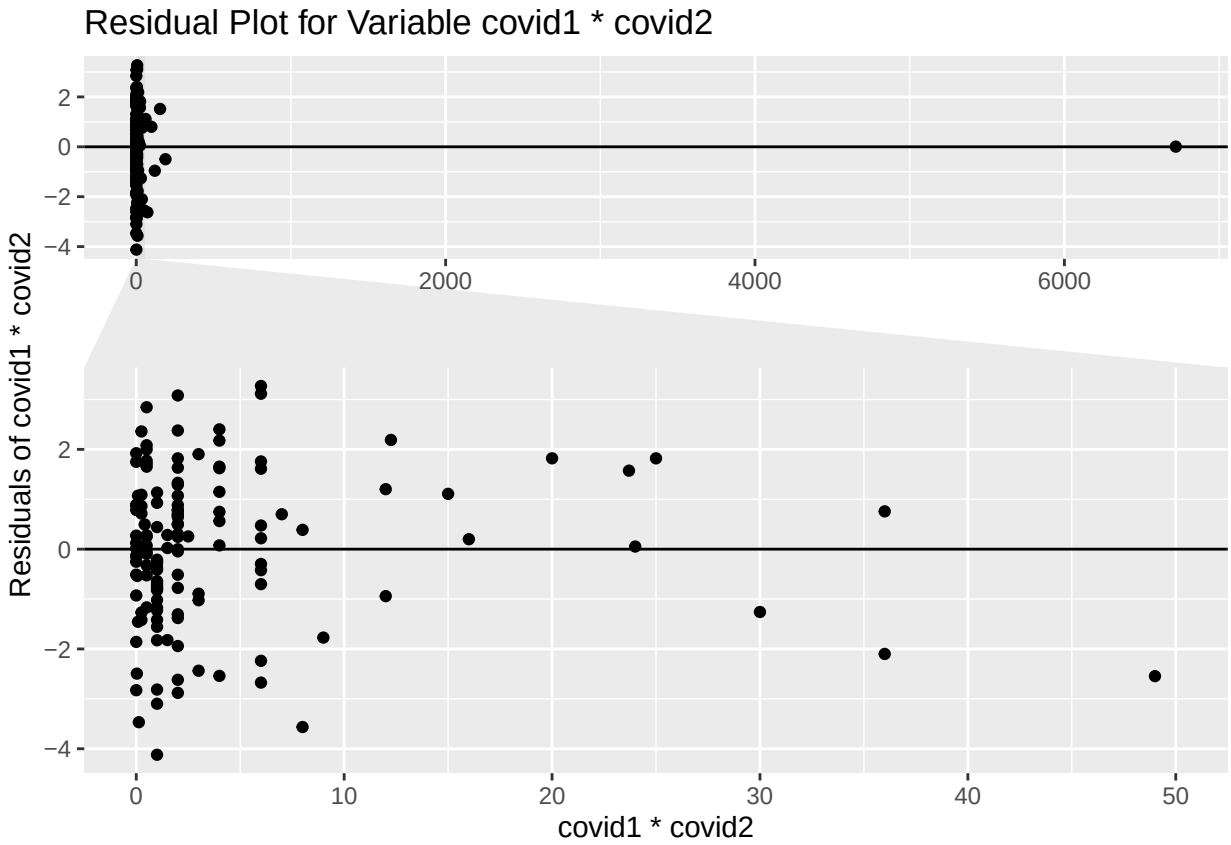
```
additive_model = lm(
  quiz4 ~ quiz1 + quiz2 + quiz3
  + covid1 + covid2 + covid3 + covid4
  + study1 + study2 + study3 + study4
  + country
)
summary(additive_model)
```

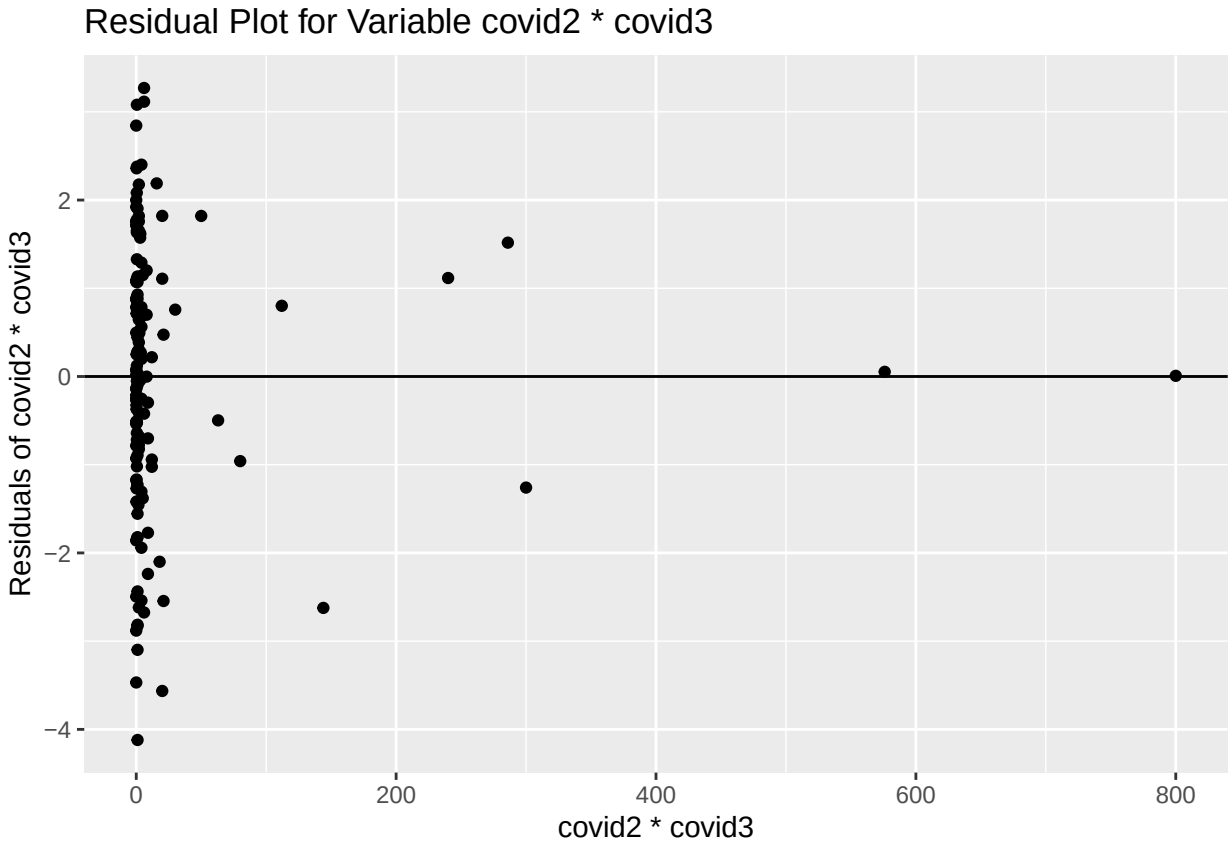
```
stepAIC(additive_model, direction = "forward")$anova
```

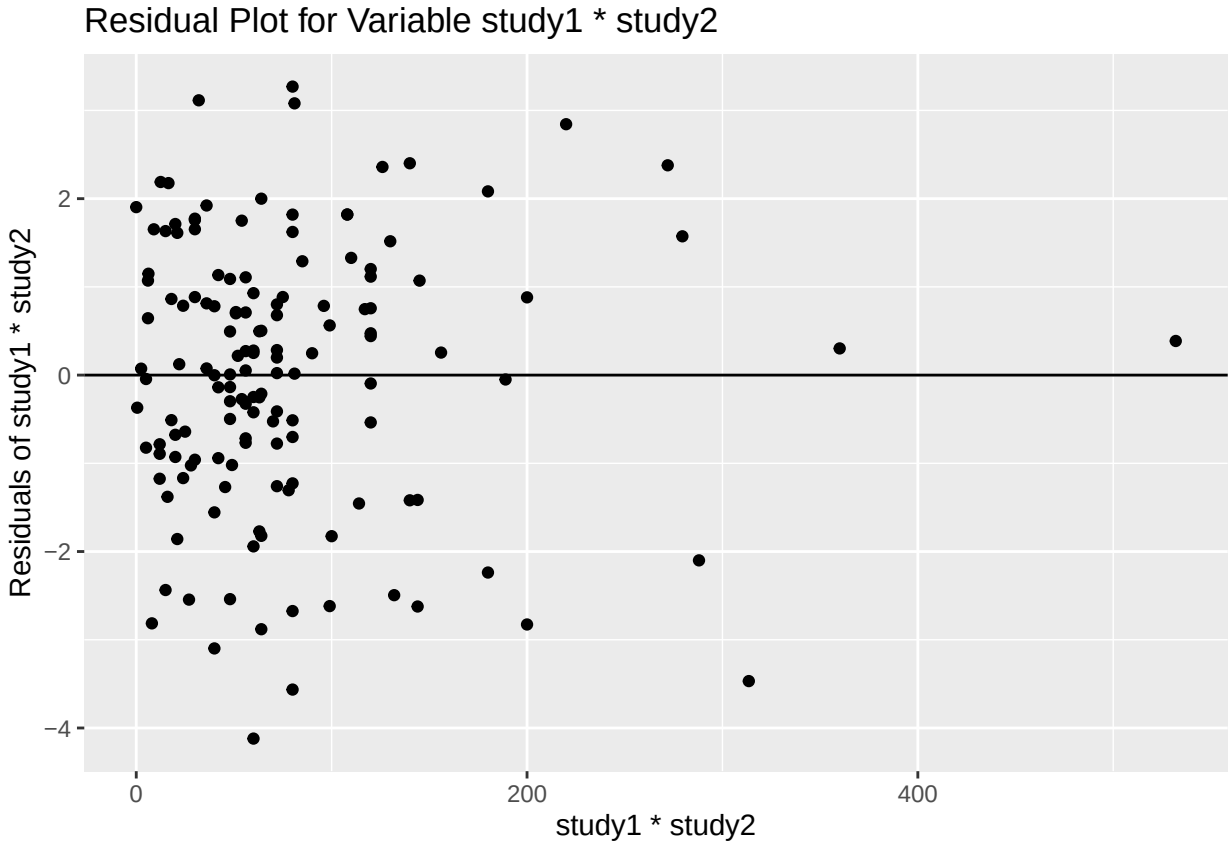
```
stepAIC(additive_model, direction = "backward")$anova
```

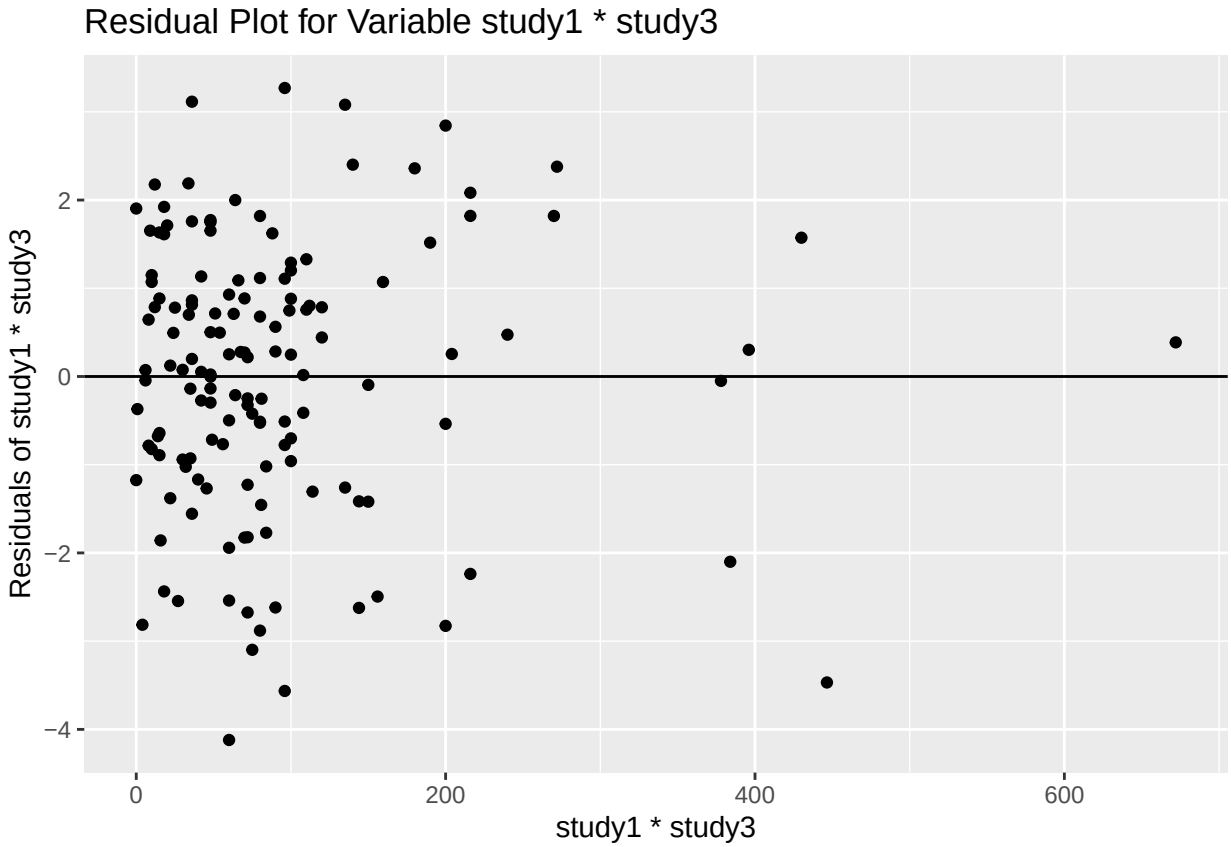
```
stepAIC(additive_model, direction = "both")$anova
```

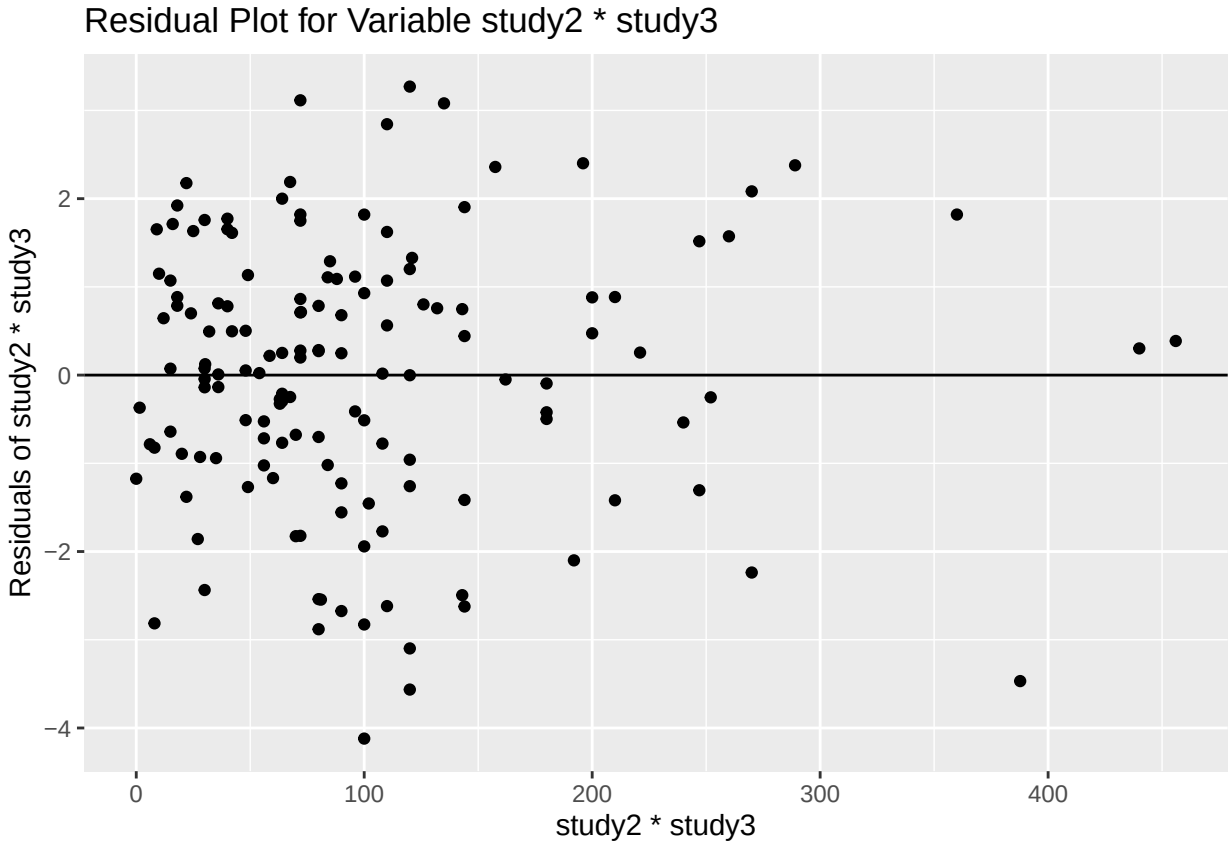

Residual Plot for Variable covid1 ^ 2
Residuals of covid1 ^ 2
covid1 ^ 2
2
0
−2
−4
0
10000
20000
0
10
20
30
40

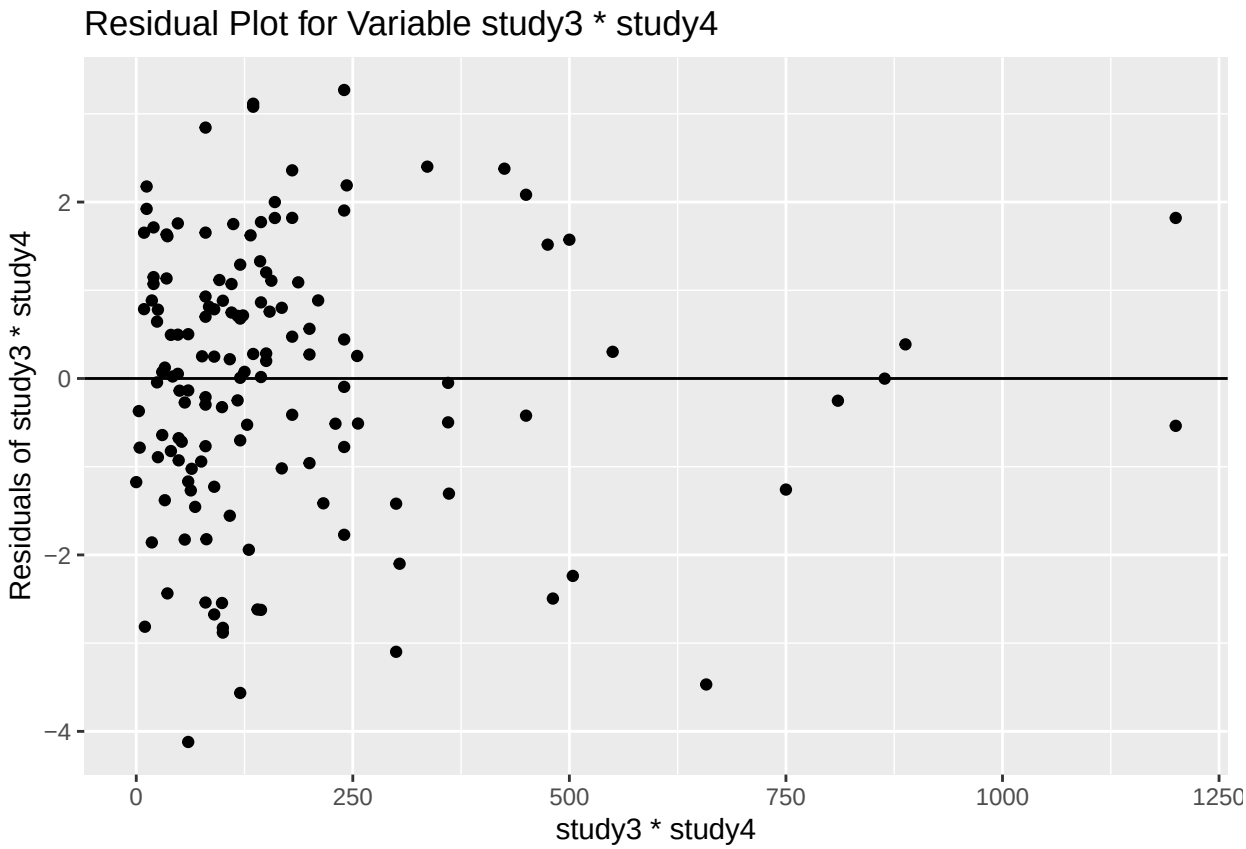

Residual Plot for Variable covid1 * covid2
Residuals of covid1 * covid2
covid1 * covid2
2
0
−2
−4
0
2000
4000
6000
0
10
20
30
40
50

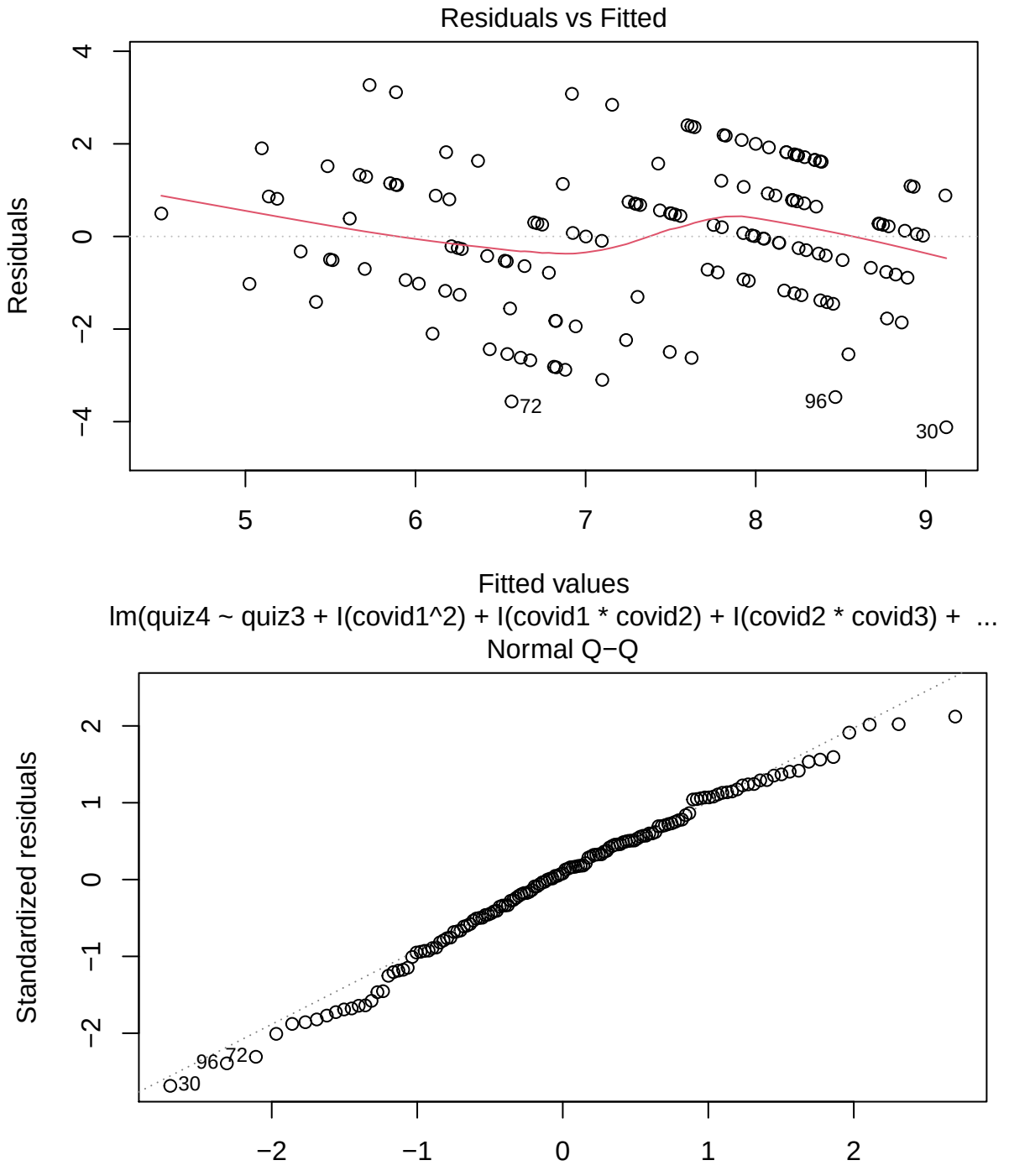

Residual Plot for Variable covid2 * covid3
Residuals of covid2 * covid3
2
0
−2
−4
0
200
400
600
800
covid2 * covid3


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

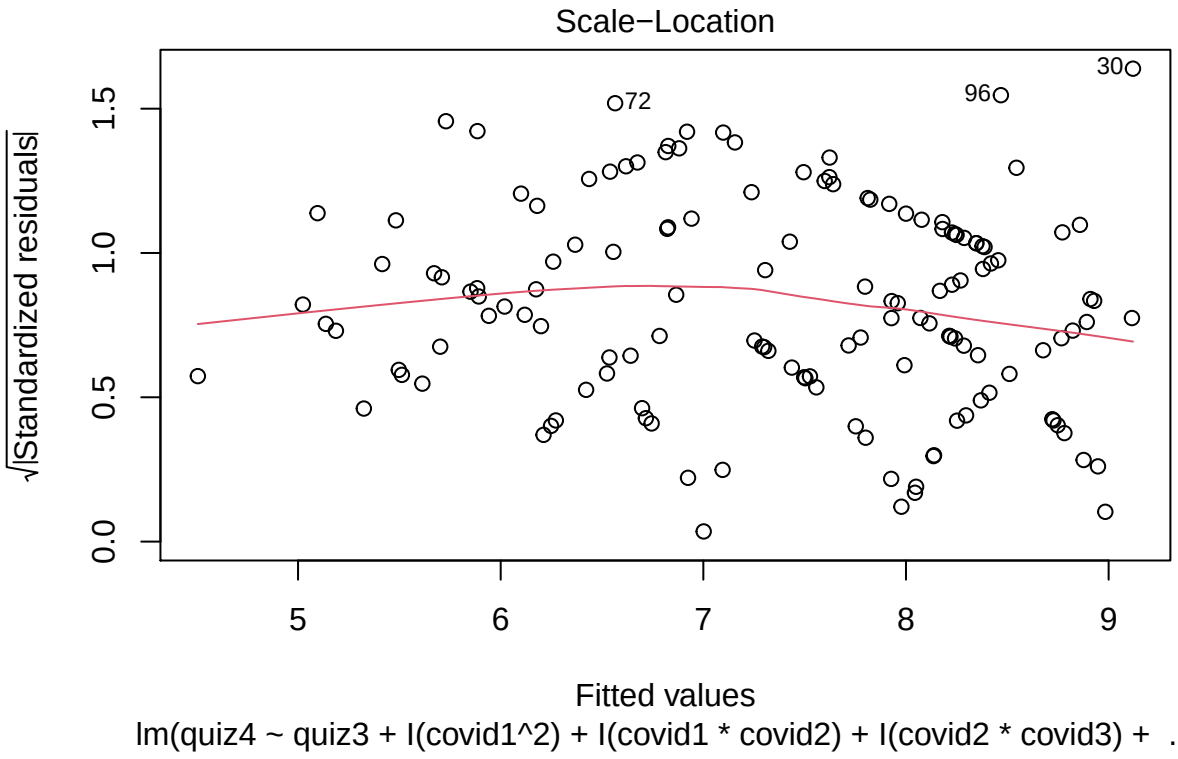

Residual Plot for Variable study1 * study3
Residuals of study1 * study3
2
0
−2
−4
0
200
400
600
study1 * study3


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

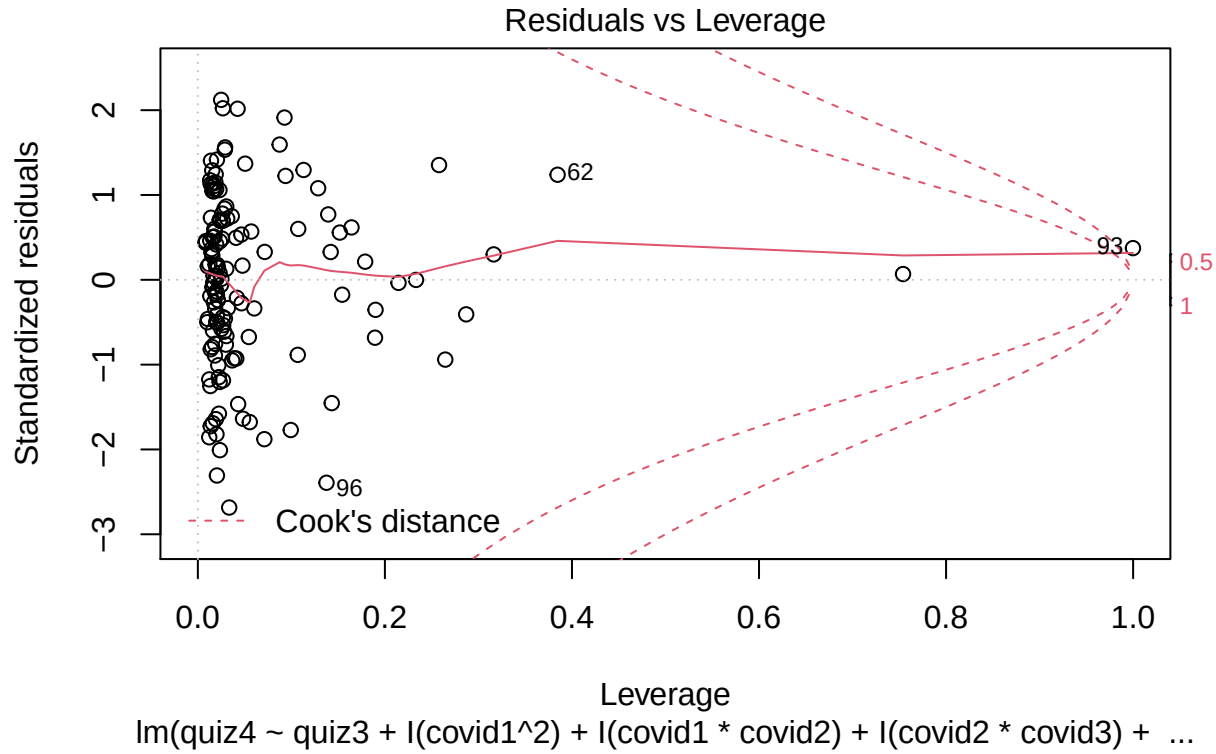

Residual Plot for Variable study3 * study4
Residuals of study3 * study4
2
0
−2
−4
0
250
500
750
1000
1250
study3 * study4


Residuals vs Fitted
Residuals
Fitted values
lm(quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 * covid3) + ...
72
96
30

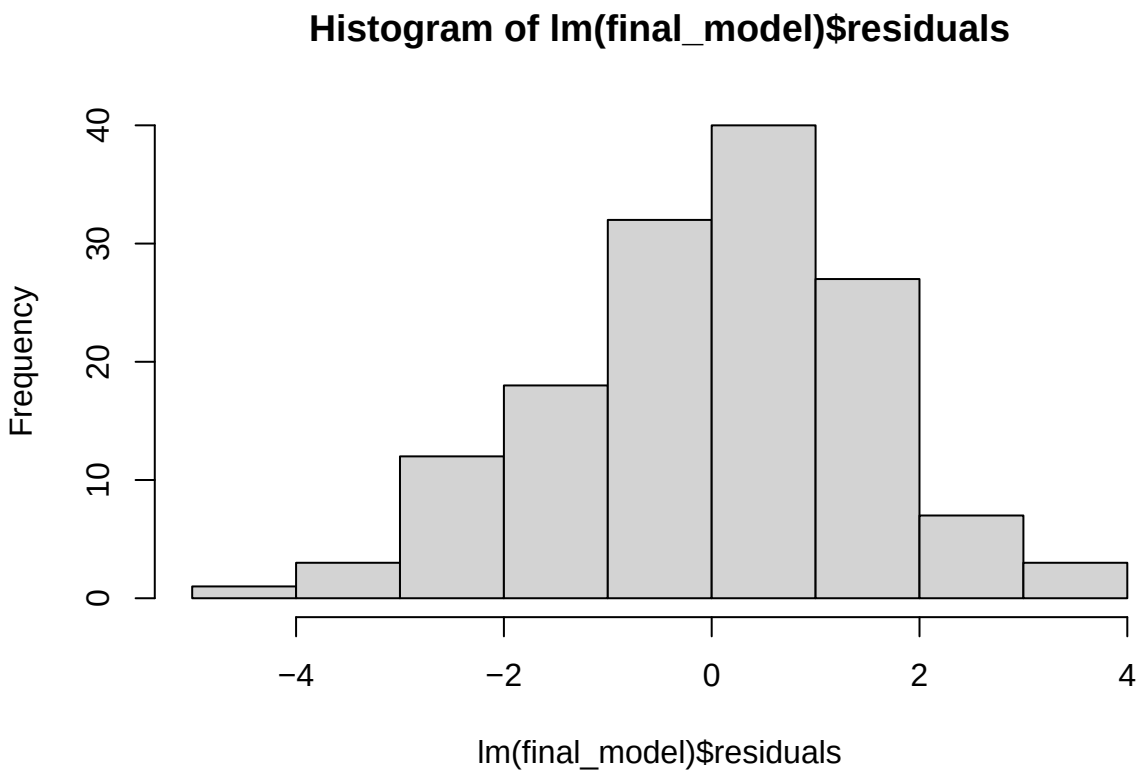

Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(quiz4 ~ quiz3 + I(covid1^2) + I(covid1 * covid2) + I(covid2 * covid3) + ...
96
72
30




```
## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced

## Warning in sqrt(crit * p * (1 - hh)/hh): NaNs produced
```



### Goodness of Current Model




```
mean(lm(final_model)$residuals)
```

```
## [1] -1.651627e-17
```

```
median(lm(final_model)$residuals)
```

```
## [1] 0.07546203
```

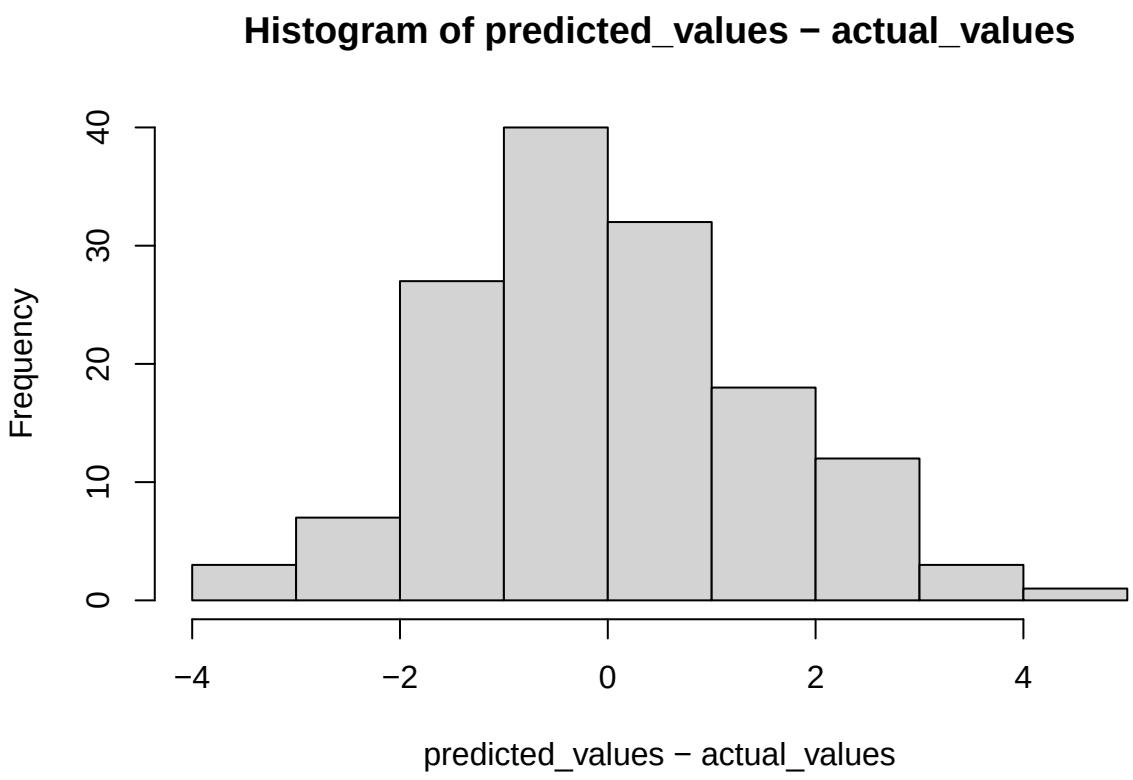
### Try Predicting on the Fitted Values

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

```
## [1] 5.664449e-15
```

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

```
## [1] -0.07546203
```




```
t.test(predicted_values - actual_values)
```

```
##
##  One Sample t-test
##
## data:  predicted_values - actual_values
## t = 4.467e-14, df = 142, p-value = 1
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  -0.250671  0.250671
## sample estimates:
##    mean of x
## 5.664449e-15
```

## 50/50 Training/Testing

### Partitioning Phase

Source: https://stackoverflow.com/questions/17200114/how-to-split-data-into-training-testing-sets-using-sample-function

```
library(caTools)  # for sample.split
set.seed(888)
sample = sample.split(remaining_data_no_NAs, SplitRatio = 0.55)
training_data = subset(remaining_data_no_NAs, sample == TRUE)
testing_data  = subset(remaining_data_no_NAs, sample == FALSE)
```

### Training Phase

```
final_model = lm(
  quiz4 ~ quiz3
  + I(covid1 ^ 2)   # don't remove, else all other terms become insignificant
  + I(covid1 * covid2)
  + I(covid2 * covid3)  # don't remove, else all other terms become insignificant
  + I(study1 * study2)
  + I(study1 * study3)  # maybe don't remove?
  + I(study2 * study3)
  + I(study3 * study4)
)
summary(final_model)
```

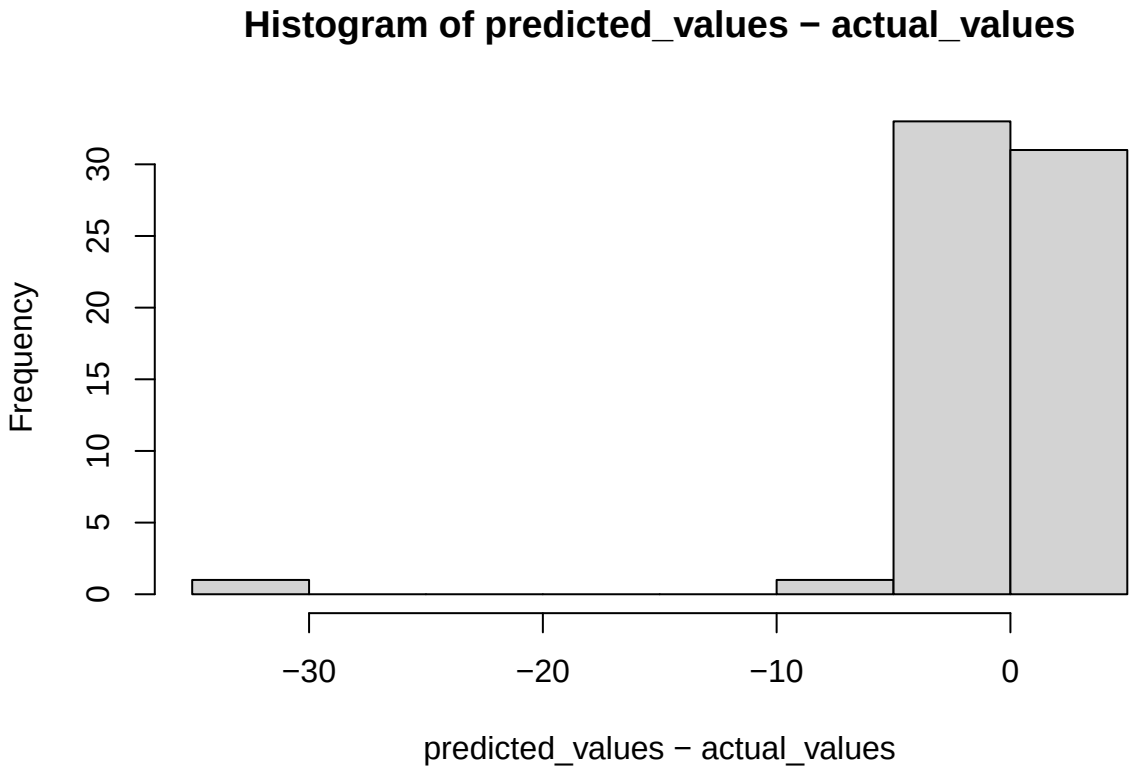
### Testing Phase

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

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

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

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




## One sample t-test on Mean

- $H_0 : \mu_{residuals} = 0$
- $H_1 : \mu_{residuals} \neq 0$
- is -0.427222 statistically different from 0?
- the p-value should be small.
- n = 77, so by CLT sample mean is approximately normal

```
t.test(predicted_values - actual_values)
```

```
##
##  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
```

- p-value = 0.9693
- t-value = -0.038679