STA302 Final Project

(Hongcheng Wei, 1006033668, weihongc) & (Wensha Sun, 1006027032, sunwens1)

Aug 22, 2021

Contents

Introduction

Background: Summer of 2021, the world is under a global pandemic caused by the novel coronavirus called COVID-19. Students are ordered to stay home, travelings to campus is restricted, and educations including all the lecturing and assessment are moved online.

Online learning brings up challenges and pressure to everyone. The instructors have to face various technological limitations during lecture. Some students must write the quiz assessment very early in the morning due to time zone differences (those students in Asia).

Under all these difficulties and stress, many questions need to be further investigated. The topic that we are addressing is What are the factors that predict student performance on the final STA302 assessment (i.e. quiz 4)? In particular, we are interested in *How COVID-19 affects student's performance?*.

Result: The effect of COVID-19 isn't the most significant factor to student's performance.

Exploratory Data Analysis

The data we are using is collected through the weekly Quercus quiz. At the end of each week, students are asked the following questions:

- How many hours did you study for STA302 this week?
- How many hours did you think about COVID-19 this week?
- What country are you in? (Only asked in the first quiz)

The weekly assessment quiz is held at the end of each Wednesday lecture. Here are the variables provided in the data set:

- `Country` is the Country that the student is reside in.
- `STA302 hours (W1)`, `STA302 hours (W2)`, `STA302 hours (W3)`, `STA302 hours (W4)` are the time (in hours) that the student put into studying STA302 each week.
- `COVID hours (W1)`, `COVID hours (W2)`, `COVID hours (W3)`, `COVID hours (W4)` are the time (in hours) that the student thinks about COVID-19 each week.
- `Quiz_1_score`, `Quiz_2_score`, `Quiz_3_score`, `Quiz_4_score` are the student's scores on each weekly assessment quiz.

Data Preparation

First, we import the data and do some cleaning:

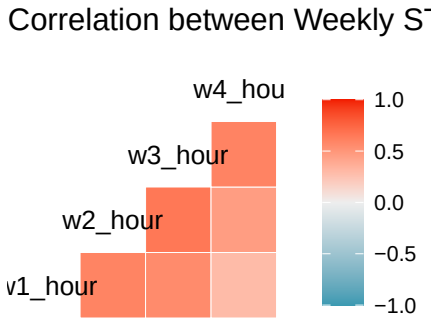
- Read the data, rename the variables to something easy to work with, and unify data types to `numeric`.
- Fill the NA values to the mean of that variable and drop the rows that are missing the response variable `Q4`.

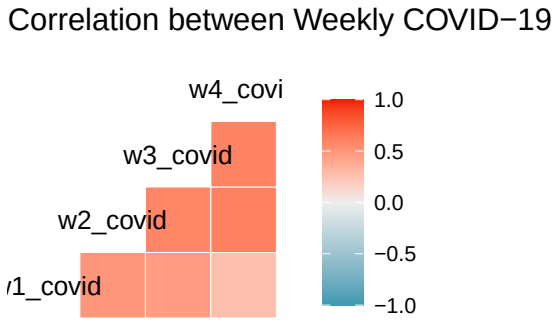
  We choose to replace the NA with mean because the data set is relatively small (around 200 observations). See Limitations of Model for more discussion on this.
- Removes the observation that does not make sense, which includes:
  - The row with 0 hours of studying STA302.
- Create Dummy Variables. Notice that the `Country` variable has the following possible response:

```
##  [1] "Canada"      "UAE"         "China"       "Pakistan"    "India"
##  [6] "South Korea" "Taiwan"      "USA"         "Mongolia"    "china"
## [11] "canada"      "Japan"       "Singapore"   NA
```

We can classify the response into three categories: North America Countries (Canada, USA), Asia Countries (China and Taiwan, Pakistan, India, South Korea, Mongolia, Japan, Singapore), `NA` values. So we can create two dummy variables that represent the two continents: `is_north_america`, `is_asia`.

- Introduce aggregate variables





As we can see from the above plot, both correlation across weekly STA302 hours and weekly COVID-19 hours are pretty high. Therefore it might be a bad idea to include all of them into the model (because it will introduce multicollinearity). Therefore we can create variables that aggregate this information:

  - `mean_hour` is the mean of weekly STA302 hours of each observation.
  - `mean_covid` is the mean of weekly COVID-19 hour of each observation.

- Reorder the variables.

Now we split the data into two part:

- testing data set (the first 20% observation of the original data set)

```
## # A tibble: 6 x 16
##   is_north_america is_asia    q1    q2    q3    q4 w1_hour w2_hour w3_hour
##              <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1                1       0    10  7.8      9    10       3       7       6
## 2                1       0     8  8.2      9     9       4       6       3
## 3                1       0     8  8.2      9     9       5       6       3
## 4                1       0     9  9.4      9     9      10      12      10
## 5                0       1     3  7.42     3     1      10       8       6
## 6                1       0     8  5.8      5     6       6       9       7
## # ... with 7 more variables: w4_hour <dbl>, w1_covid <dbl>, w2_covid <dbl>,
## #   w3_covid <dbl>, w4_covid <dbl>, mean_hour <dbl>, mean_covid <dbl>
```

- training data set (the last 80% observation of the original data set)

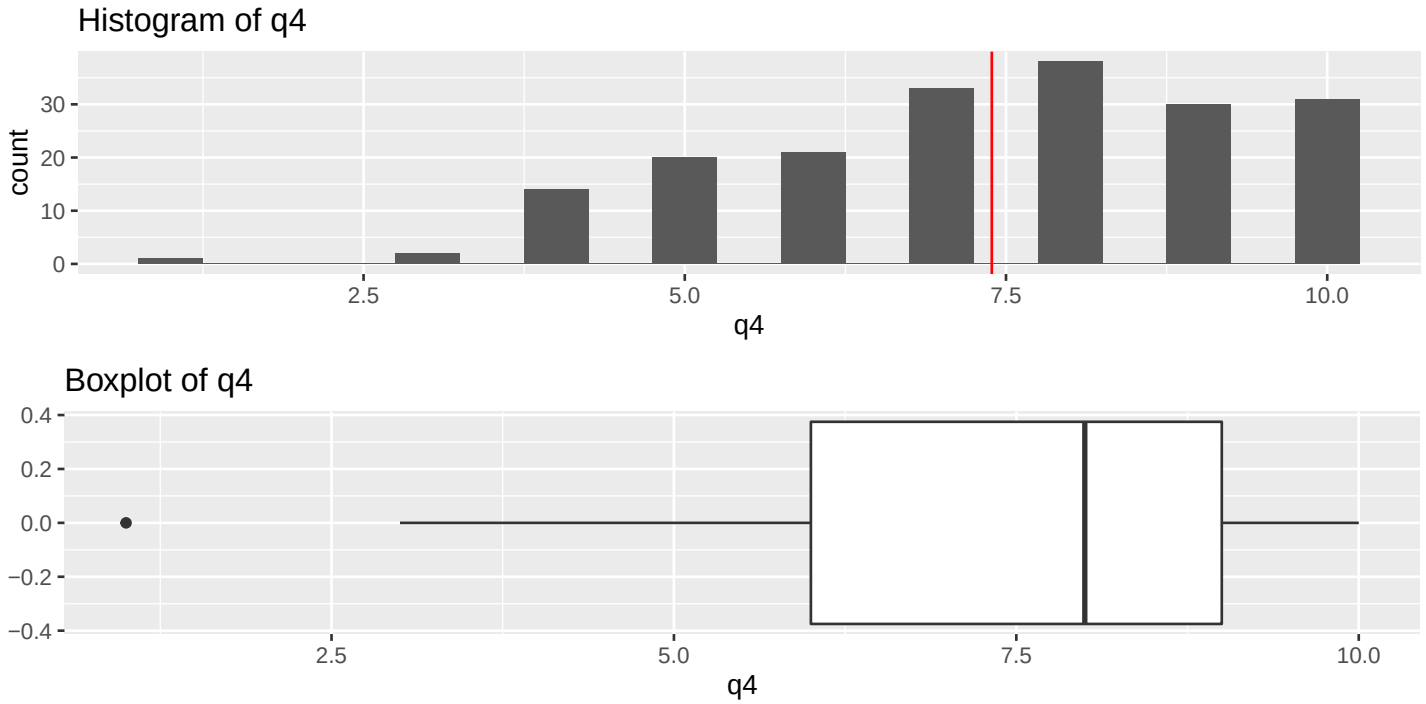
```
## # A tibble: 6 x 16
##   is_north_america is_asia    q1    q2    q3    q4 w1_hour w2_hour w3_hour
##              <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl>   <dbl>   <dbl>   <dbl>
## 1                1       0     8   5.8      5    10       9       9      15
## 2                1       0     6   5.4      5     3       2       3       9.24
## 3                1       0     4   6.6      2     5       6       8       4
## 4                1       0     4   8.2      5     6       8       7.5     9
## 5                0       1     5   1.2      4     5       7       7      12
## 6                1       0    10   5.8      6     4       9      11      10
## # ... with 7 more variables: w4_hour <dbl>, w1_covid <dbl>, w2_covid <dbl>,
## #   w3_covid <dbl>, w4_covid <dbl>, mean_hour <dbl>, mean_covid <dbl>
```

The data set we are using is tiny in terms of observation count, we only have 152 observations in training data and 38 observations in testing data. See the Limitations of Model section for more discussion on this.

Summary of Important Variables

Quiz 4 score

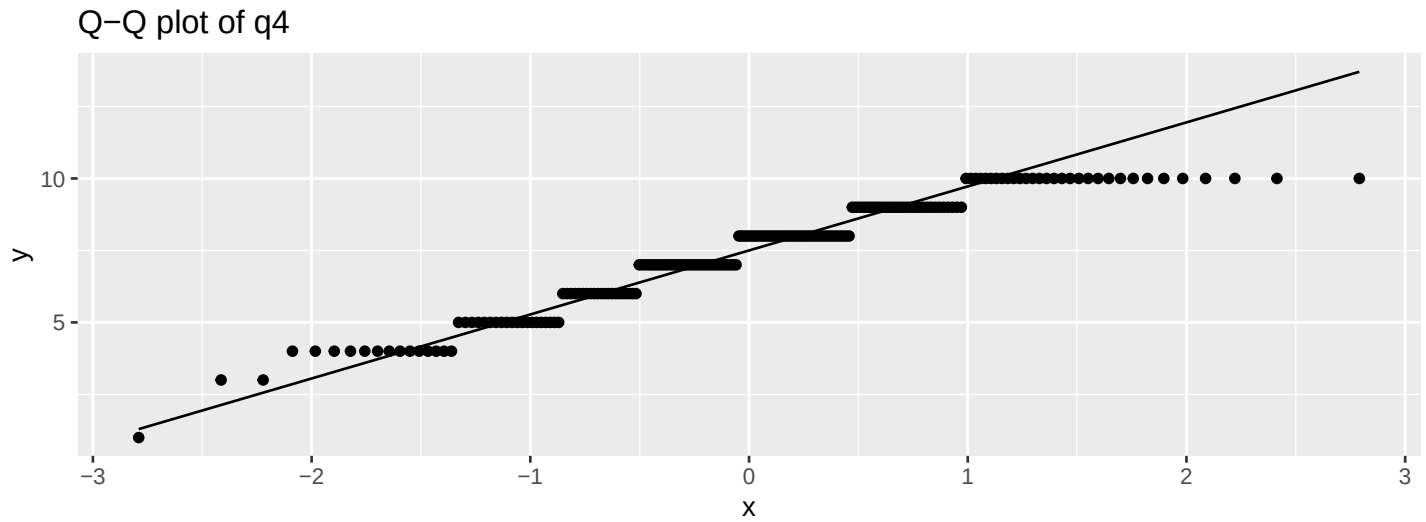
The score of quiz 4 (`q4`) is the response variable of this study.



From the plots:

- The mean of Quiz 4 score is around 7.4
- `q4` is a quiz score, so it ranges between [0, 10], the majority lies between [3.75, 10].
- Quiz 4 score is left-skewed.
- Quiz 4 score is unimodal.
- There is an outlier that has a score of 1.

Notice that q4 might not follow normal distribution due to its skewness.



The Q-Q plot also agrees with our finding; notice that the upper part of the plot does not align with the Q-Q line, indicating the data is heavy on the right side.

Covid Study Ratio

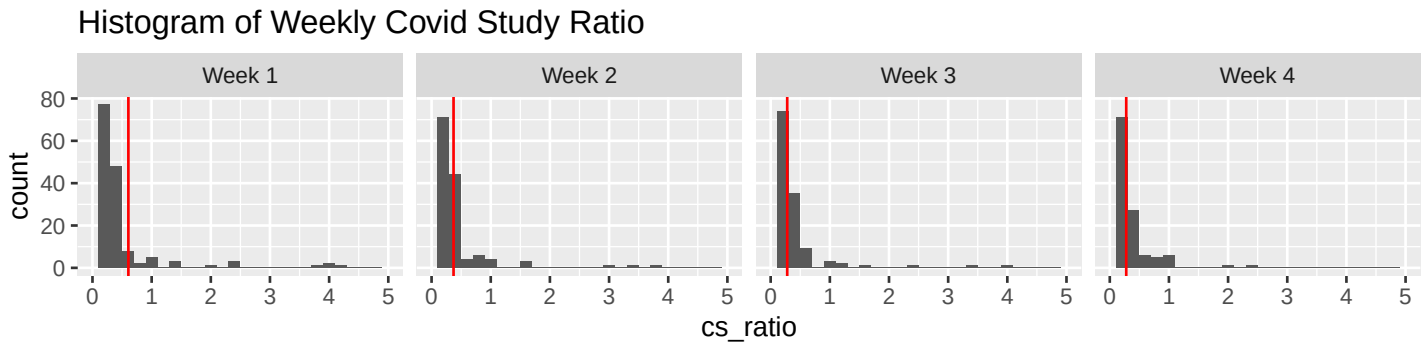
The Covid Study Ratio (or `cs_ratio`) is a variable we plan to use in our model. It is a ratio between hours of thinking about COVID-19 and hours of study in STA302. The `cs_ratio` is calculated as:

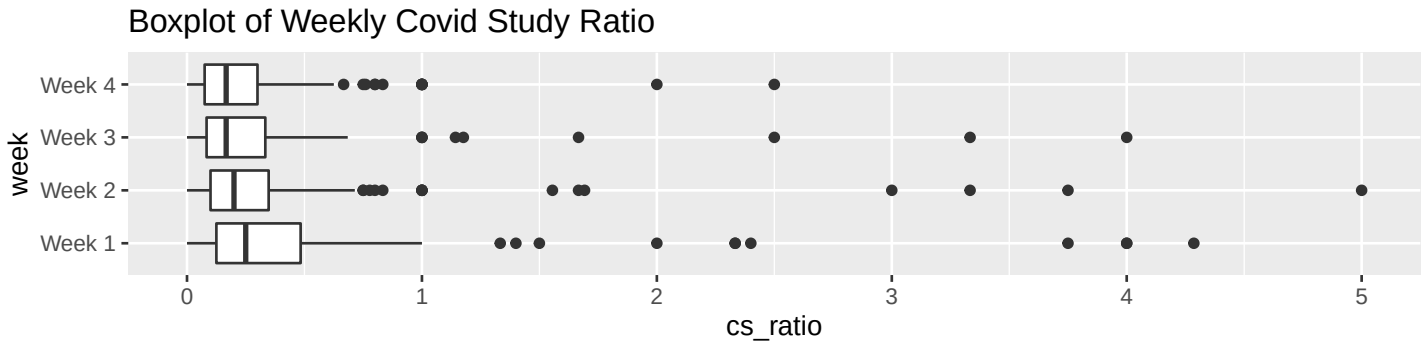
$$\frac{\text{Hour of COVID-19}}{\text{Hour of study STA302}}$$

The purpose of this variable is to measure the effect of the COVID-19 on student's studies.

The Covid Study Ratio reflects how many hours of thinking about COVID-19 per hour of study STA302.

- The `cs_ratio` is high when Hour of COVID-19 is high and Hour of study is low meaning:
  - COVID-19 effects that student a lot, and
  - Student doesn't spend much time on study STA302.
- The `cs_ratio` is low when Hour of COVID-19 is low and Hour of study is high meaning:
  - COVID-19 doesn't effects that student very much, and
  - Student is able to spend a long on study STA302.



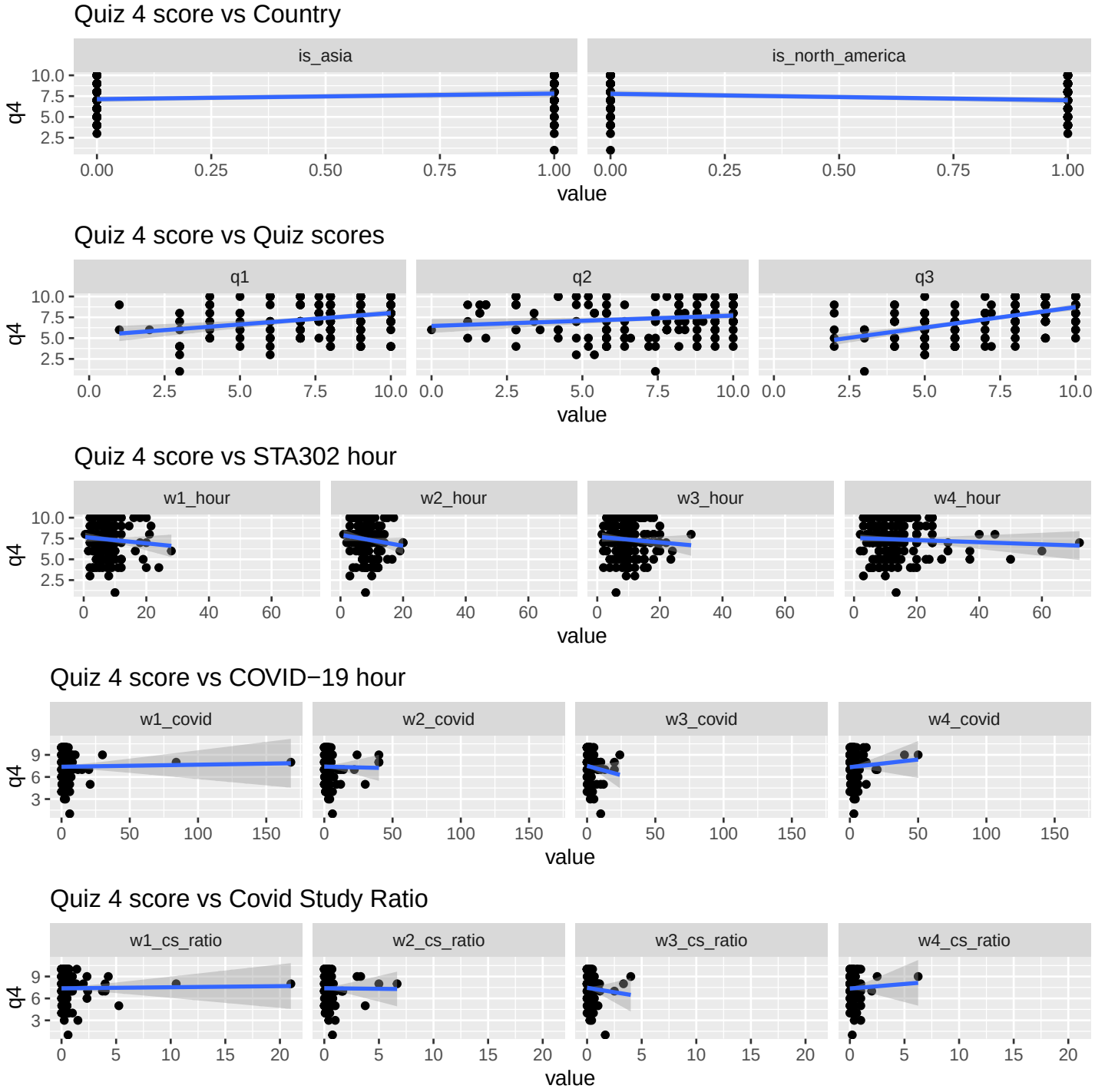


From the plots:

- The mean covid study ratio of each week centers around 0.3, meaning on average students think about 0.3 hours of COVID-19 per hour of study STA302.
- Most of the students have covid study ratio within $[0, 0.3]$.
- Covid Study Ratio is unimodal.
- Some outliers have a very high covid study ratio. These students have a meagre hour of studying STA302.

Correlation between Variables and Quiz 4 scores

Here is how each variable is correlated with the quiz 4 scores.



From the plot:

- The `is_asia` variable and `q4` show a moderate positive correlation, while the `is_north_america` variable and `q4` show a moderate negative correlation.
- Other quiz scores are positively correlated with `q4`.
- The regression line of `STA302 hour`, `COVID-19 hour` and `Covid Study Ratio` to `q4` is nearly flat, indicating that these variables may not be a good explanatory variable.

Model Development

In this section, we will develop the model that best explains the response variable `q4`.

AIC Backward Elimination

We will start with the full model. Notice that we cannot put the aggregate variables `mean_hour` or `mean_covid` in the full model as it will create *perfect collinearity*.

After doing AIC Backward Elimination, here is the model we end up with:

```
##
## Call:
## lm(formula = q4 ~ is_north_america + w3_covid + q1 + q3 + I(w2_covid/w2_hour),
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7244 -0.9720  0.1369  1.1605  3.7957
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           3.52316    0.61418   5.736 5.38e-08 ***
## is_north_america     -0.75064    0.27117  -2.768  0.00637 **
## w3_covid             -0.10799    0.05208  -2.074  0.03988 *
## q1                    0.17514    0.06595   2.656  0.00879 **
## q3                    0.40922    0.06474   6.321 2.98e-09 ***
## I(w2_covid/w2_hour)   0.34728    0.19814   1.753  0.08174 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.552 on 146 degrees of freedom
## Multiple R-squared:  0.3483, Adjusted R-squared:  0.326
## F-statistic: 15.61 on 5 and 146 DF,  p-value: 2.731e-12
```

Manual Models

In this section, we propose some models that we think could be good by common sense.

Manual Model 1: Continent, Previous Quiz Scores and Mean Study Hours

From our own experience:

- Since the quiz always happens on a fixed period each week. The continent that the student resides in will determine the quiz happens at which part of the student's day. Some students will always write the quiz very early in the morning, while others will write the quizzes in the evening. So the location variables `is_north_america` and `is_asia` should be included.
- If students study more on the course material, they should score more marks on quizzes.
- If students were performing well on the previous weeks' quiz, they should also perform well in quiz 4.

```
##
## Call:
## lm(formula = q4 ~ is_asia + is_north_america + q1 + q2 + q3 +
##     mean_hour, data = train_data)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3.6939 -0.9141  0.0161  1.0250  3.8472
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.708698   0.733700   5.055 1.28e-06 ***
## is_asia           0.083119   0.404011   0.206   0.8373
## is_north_america -0.554200   0.389157  -1.424   0.1566
## q1                0.170683   0.068544   2.490   0.0139 *
## q2                0.008926   0.053117   0.168   0.8668
## q3                0.405031   0.066317   6.108 8.83e-09 ***
## mean_hour        -0.042292   0.029839  -1.417   0.1585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.572 on 145 degrees of freedom
## Multiple R-squared:  0.3362, Adjusted R-squared:  0.3088
## F-statistic: 12.24 on 6 and 145 DF,  p-value: 4.073e-11
```

Notice `is_asia`, `q2` and `mean_hour` are not very significant (the p-values are greater than `0.05`), Let us try removing them by doing a partial F-test:

```
## Analysis of Variance Table
##
## Model 1: q4 ~ is_north_america + q1 + q3
## Model 2: q4 ~ is_asia + is_north_america + q1 + q2 + q3 + mean_hour
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    148 363.36
## 2    145 358.36  3    5.0034 0.6748 0.5688
```

The partial F-test yields a p-value of 0.5688, which is much larger than 0.05. We fail to reject the null hypothesis. Therefore the coefficients of the three variables missing in the subset model can be approximate to 0. Hence we can remove them.

```
##
## Call:
## lm(formula = q4 ~ is_north_america + q1 + q3, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7349 -0.9884  0.0255  1.0243  3.8379
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       3.39446    0.61427   5.526 1.44e-07 ***
## is_north_america -0.62052    0.26690  -2.325   0.0214 *
## q1                0.16435    0.06552   2.508   0.0132 *
## q3                0.41468    0.06474   6.405 1.89e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.567 on 148 degrees of freedom
## Multiple R-squared:  0.327,  Adjusted R-squared:  0.3133
## F-statistic: 23.97 on 3 and 148 DF,  p-value: 1.064e-12
```

- After removing the variables, the standard error of `is_north_america` decreases from 0.389157 to 0.26690, indicating that we have removed a term that is correlated with `is_north_america`, and by removing it, we have made `is_north_america` more significant.
- The variable `mean_hour` is removed from the model.

Manual Model 2 : Most Recent

The idea of this model is that: quiz 4 happens at week 4, therefore, it must be related to the student's status in the most recent week (week 4) and the student's performance in the most recent quiz (quiz 3), so we use these variables to build the model.

From manual model 1, we know that we `is_asia` and `is_north_america` are highly correlated, so we should only include one of them in the model.

```
##
## Call:
## lm(formula = q4 ~ is_north_america + w4_hour + w4_covid + I(w4_covid/w4_hour) +
##     q3, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6222 -0.8914  0.1444  0.9708  3.8247
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4.77028    0.58737   8.121 1.77e-13 ***
## is_north_america     -0.44056    0.26407  -1.668   0.0974 .
## w4_hour              -0.02968    0.01641  -1.809   0.0725 .
## w4_covid              0.11242    0.07526   1.494   0.1374
## I(w4_covid/w4_hour) -1.66303    1.02565  -1.621   0.1071
## q3                    0.45540    0.06332   7.192 3.07e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.59 on 146 degrees of freedom
## Multiple R-squared:  0.316,  Adjusted R-squared:  0.2926
## F-statistic: 13.49 on 5 and 146 DF,  p-value: 8.133e-11
```

Manual Model 3 : COVID-19 Model

The idea of this model is that: under the pressure of the global pandemic, COVID-19 should be the biggest factor in determining the student's performance. So we use `mean_hour`, `mean_covid` and `mean_sc_ratio` as explanatory variables.

```
##
## Call:
## lm(formula = q4 ~ mean_covid + mean_hour + I(mean_covid/mean_hour),
##     data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4814 -1.4172  0.3556  1.5585  3.1816
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7.97536    0.52193  15.281   <2e-16 ***
## mean_covid                0.06459    0.12628   0.511    0.610
## mean_hour                -0.06362    0.04671  -1.362    0.175
## I(mean_covid/mean_hour) -0.61272    1.28462  -0.477    0.634
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.896 on 148 degrees of freedom
## Multiple R-squared:  0.01431,    Adjusted R-squared:  -0.005671
## F-statistic: 0.7162 on 3 and 148 DF,  p-value: 0.5438
```

Notice that none of the variables are significant, and the R^2_{adj} is negative, which indicates this model may not be predictive at all.

Model Selection

```
## # A tibble: 4 x 6
##   Model             p_prime SSRes   R_squared_adj C_p    AIC
##   <chr>             <chr>   <chr>   <chr>         <chr>  <chr>
## 1 AIC Reduced Model 6       351.836 0.326         1.149  572.928
## 2 Manual Model 1    4       363.361 0.313         1.773  573.828
## 3 Manual Model 2    6       369.287 0.293         8.15   580.287
## 4 Manual Model 3    4       532.163 -0.006        69.493 631.824
```

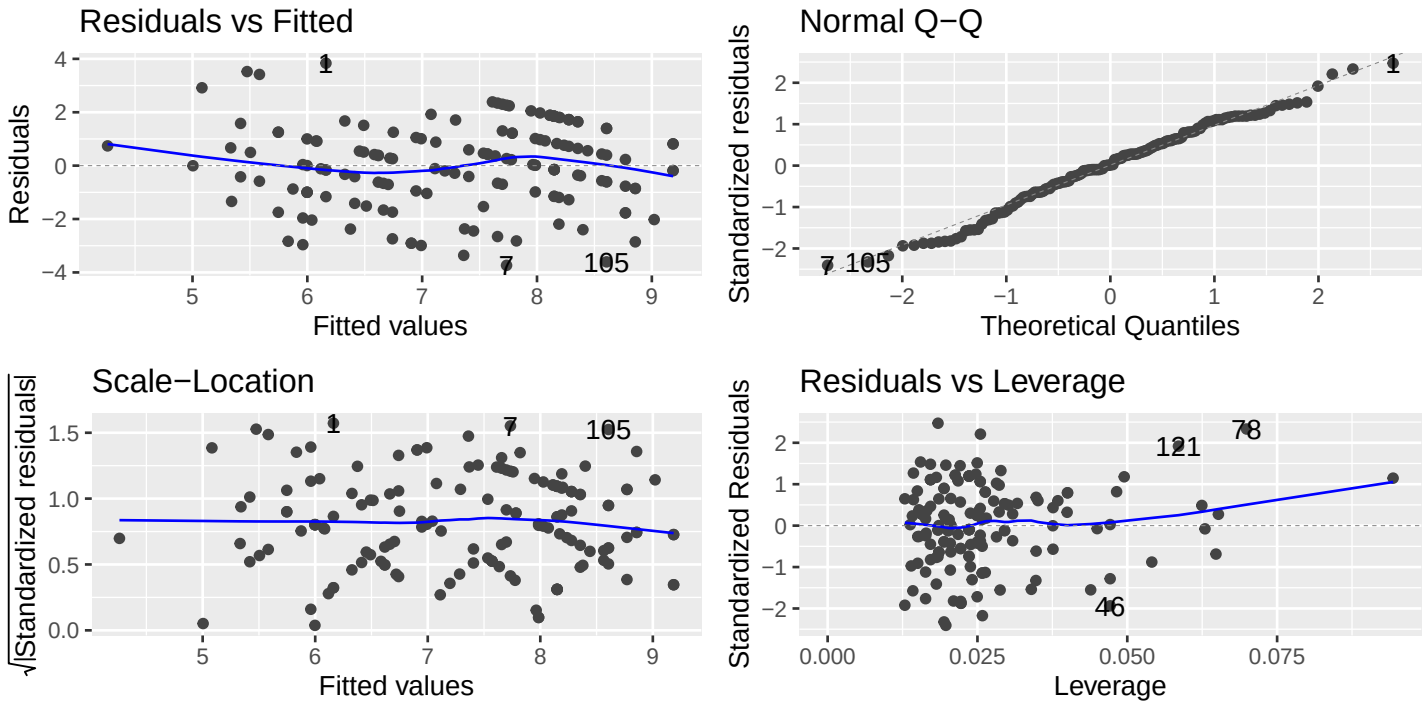
- The AIC reduced model has
  - the smallest $SSRes$,
  - the highest R^2_{adj},
  - the smallest C_p (but C_p is not close to p'),
  - the smallest AIC.

  However, this model is more challenging to interpret due to its algorithm-generated models' nature, but it is the best model among the four proposed models.
- The manual model 1 is the simplest model (with the lowest number of variables). Compare to the AIC reduced model, the manual model 1 has
  - a slightly higher $SSRes$,
  - a slightly lower R^2_{adj},
  - a slightly higher C_p (the C_p is also not close to to p'),
  - nearly the same AIC.

  Despite being slightly worst than the AIC reduced model, the manual model 1 is straightforward to interpret.
- The manual model 2 is no better model than manual model 1. It has:
  - higher $SSRes$,
  - lower R^2_{adj},
  - higher C_p (the C_p is higher than p'),
  - higher AIC.
- The manual model 3 is worst model. It has:
  - very high $SSRes$,
  - negative R^2_{adj},
  - very high C_p (the C_p is a lot higher than p'),
  - very high AIC.

Conclusion: The best model is the AIC reduced model, but it is also the hardest to interpret. Since we wish to describe "the factor that predicts the student's performance", we will select the manual model 1, which is slightly worst, but highly interpretable as our model to further investigate.

Model Diagnostic



1. Outlying and Influential Observations

- Outlying and influential X observations, we are checking the leverage value for outlying observations and are checking the cooks distance for influential observation.

```
## # A tibble: 6 x 7
##   row_index     Y Y_hat leverage cooks_distance is_outlier is_influential
##       <dbl> <dbl> <dbl>    <dbl>          <dbl> <lgl>      <lgl>
## 1         1    10  6.16   0.0184         0.0287  FALSE      FALSE
## 2         2     3  5.83   0.0222         0.0190  FALSE      FALSE
## 3         3     5  4.26   0.0624         0.00395 FALSE      FALSE
## 4         4     6  5.50   0.0400         0.00108 FALSE      FALSE
## 5         5     5  5.87   0.0376         0.00317 FALSE      FALSE
## 6         6     4  6.91   0.0222         0.0200  FALSE      FALSE
```

We filter the observation that is both influential and outlier.

```
## # A tibble: 0 x 7
## # ... with 7 variables: row_index <dbl>, Y <dbl>, Y_hat <dbl>, leverage <dbl>,
## #   cooks_distance <dbl>, is_outlier <lgl>, is_influential <lgl>
```

Since there are no observations that are both outlying and influential, we do not need to remove any observations.

- Outlying and influential Y observation, we are testing the studentized deleted residuals on an significant level of 0.05 for outlying observation and checking the cooks distance for influential observation.

```
## # A tibble: 6 x 7
##   row_index     Y Y_hat      t cooks_distance is_outlier is_influential
##       <dbl> <dbl> <dbl>  <dbl>          <dbl> <lgl>      <lgl>
## 1         1    10  6.16  2.52         0.0287  FALSE      FALSE
## 2         2     3  5.83 -1.84         0.0190  FALSE      FALSE
## 3         3     5  4.26  0.486        0.00395 FALSE      FALSE
## 4         4     6  5.50  0.322        0.00108 FALSE      FALSE
## 5         5     5  5.87 -0.568        0.00317 FALSE      FALSE
## 6         6     4  6.91 -1.89         0.0200  FALSE      FALSE
```

We filter the observation that is both influential and outlier.

```
## # A tibble: 0 x 7
## # ... with 7 variables: row_index <dbl>, Y <dbl>, Y_hat <dbl>, t <dbl>,
## #   cooks_distance <dbl>, is_outlier <lgl>, is_influential <lgl>
```

Since there are no observations that are both outlying and influential, we do not need to remove any observations.

2. Linearity Assumption

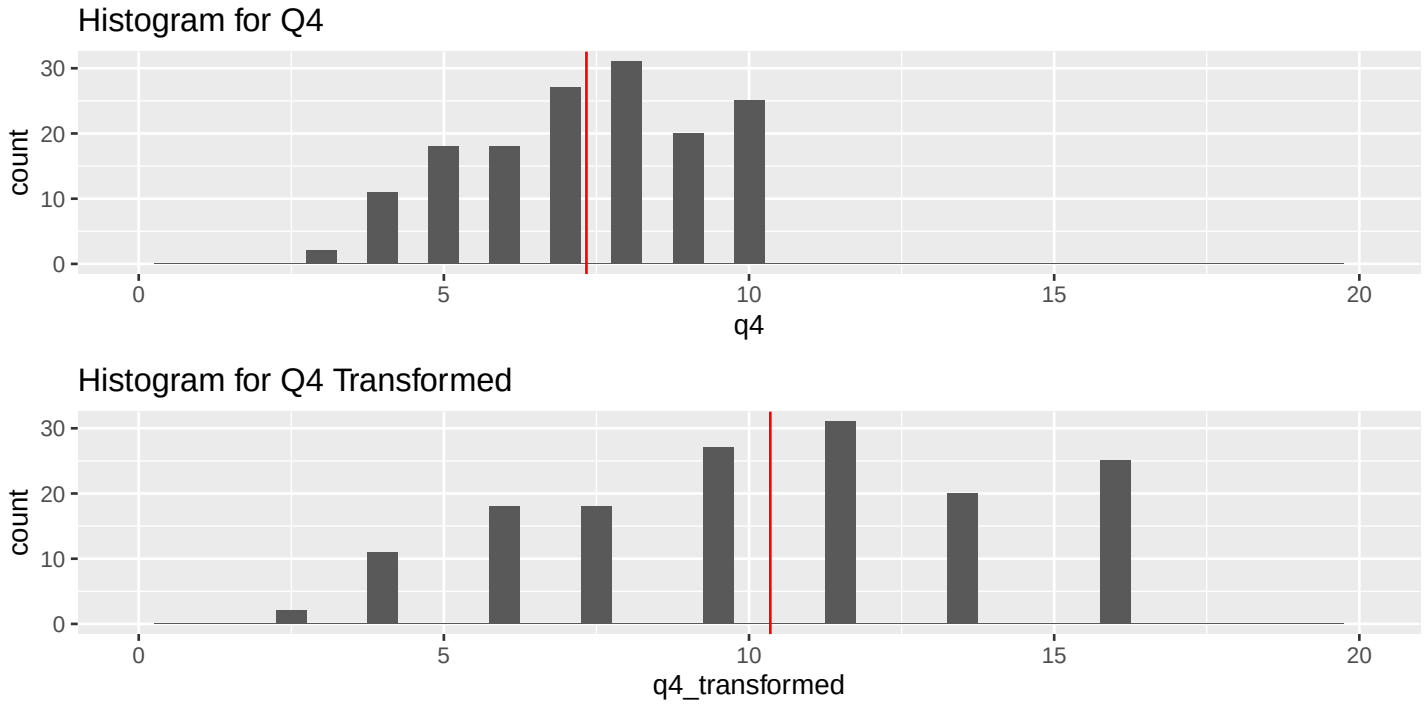
There are no obvious non-random patterns in the Residual vs Fitted value plot, which suggests that the linearity assumption is satisfied.

There are some strange diagonal lines in the graph. These diagonal lines are possibly due to q4 being not truly continuous, i.e. there is no quiz 4 scores of 9.9 points. See the Limitations of Model section for more discussion on this.

3. Normality Assumption

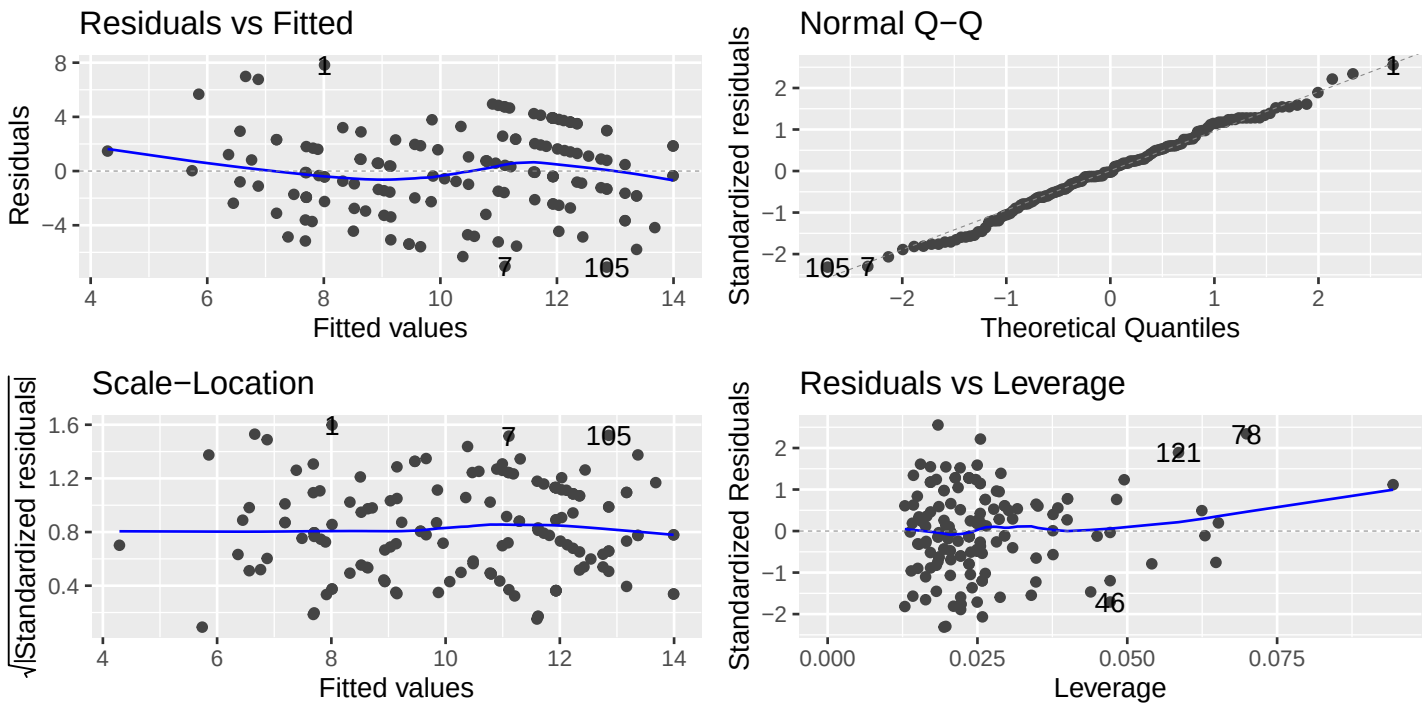
Before checking for homoscedasticity, we should make sure the normality assumption is satisfied. Looking at the Normal Q-Q plot, the lower portion of the plot is under the Q-Q line, meaning we have a left-skewed distribution. This result matches up with the conclusion from the above section at the summary of Quiz 4 score.

Box-cox suggest a transformation with $\lambda = 1.35$.



Notice that the distribution of `q4_transformed` looks more normal, but it is still left-skewed. See the Limitations of Model section for more discussion on this.

Then we refit the model, using the transformed quiz 4 scores. Here is the diagnostic plot for the refitted model:



4. Homoscedasticity assumption

The Scale-Location plot for the refitted model shows that the line is approximately straight, and the point appears random, suggesting that the constant variance assumption is satisfied.

5. Independence assumption

Since the data is independently collected, we can assume the independence assumption hold.

6. Multicollinearity

We can test multicollinearity by looking at the variance inflation factor of the refitted model.

```
## is_north_america               q1               q3
##         1.101795         1.117164         1.127380
```

None of the VIF is greater than 10, The mean VIF is:

```
## [1] 1.115446
```

which is not very big. Therefore we can assume there is no severe multicollinearity.

Final Model

The final model we propose is as follow:

$$\text{Quiz 4 Score} = \left(1.35 \times \left(\begin{bmatrix}1 & \text{is\_north\_america} & \text{q1} & \text{q3}\end{bmatrix} \cdot \begin{bmatrix}2.6279 \\ -1.2363 \\ 0.3129 \\ 0.8235\end{bmatrix}\right) + 1\right)^{\frac{1}{1.35}}$$

Where:

- `is_north_america` is a dummy variable that has value 1 if the student resides in North American (Canada or USA).
- `q1` is the score of the student in Quiz 1.
- `q3` is the score of the student in Quiz 3.

```
##
## Call:
## lm(formula = q4_transformed ~ is_north_america + q1 + q3, data = train_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0910 -1.9348 -0.0803  1.9817  7.8300
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.6279     1.2131   2.166   0.0319 *
## is_north_america  -1.2363     0.5271  -2.345   0.0203 *
## q1                 0.3129     0.1294   2.418   0.0168 *
## q3                 0.8235     0.1279   6.441 1.57e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.094 on 148 degrees of freedom
## Multiple R-squared:  0.3271, Adjusted R-squared:  0.3134
## F-statistic: 23.98 on 3 and 148 DF,  p-value: 1.054e-12
```

Model Validation

We validate our model against the testing data set. The $MSPE$ of the model on the testing data is:

```
## [1] 2.785388
```

The $MSRes$ of the final model on the training data is:

```
## [1] 9.575827
```

Since the difference of $MSPE$ and $MSRes$ is relatively big, this indicates our model may suffer on predictive ability. See the Limitations of Model section for more discussion on this.

Conclusion & Discussion

Model Interpretation

The proposed final model is in the form:

$$\text{Quiz 4 Score} = (1.35 \times (\beta_0 + \beta_1 \times \text{is\_north\_america} + \beta_2 \times \text{q1} + \beta_3 \times \text{q3}) + 1)^{\frac{1}{1.35}}$$

Where β_0, β_1, β_2, β_3 are coefficient calculated using least square method for linear models.

- $\beta_0 = 2.6279$ is the coefficient of the intercept term. It means that the true mean of Quiz 4 score of students that *does not reside in North America*, and *scores* 0 *on quiz 1 and quiz 3* is $(1.35 * 2.6279 + 1)^{\frac{1}{1.35}} = 3.070794$.
- $\beta_1 = -1.2363$ is the coefficient of a dummy variable `is_north_america`. It means that the regression function for students reside in North America is $(1.35 * 1.2363 + 1)^{\frac{1}{1.35}} = 2.069257$ lower than the regression function for students reside in Asia or the students who don't say which continent they reside in.
- $\beta_2 = 0.3129$ is the coefficient of the variable `q1`. It means that the mean increase of Quiz 4 score is $(1.35*0.3129+1)^{\frac{1}{1.35}} = 1.298234$ per 1 point increase of Quiz 1 score when other variables are held constant.
- $\beta_3 = 0.8235$ is the coefficient of the variable `q3`. It means that the mean increase of Quiz 4 score is $(1.35*0.8235+1)^{\frac{1}{1.35}} = 1.739691$ per 1 point increase of Quiz 3 score when other variables are held constant.

Advantages of Model

The final model is very simple and straightforward:

- It only contains three variables,
- It is computationally simple,
- Each of the variables is intuitive to understand.

Limitations of Model

- Data limitations:
  - The size of the data set is tiny. The data set only has 228 observations, which is already quite small to start with. After removed the "useless" observations, we are only left with 190 observations. 20% of the data set is used for validation, which is only left with 152 observations for training.
  - Response bias during data collection, some data points are clearly out of range. For example, you can't think about COVID-19 for 168 hours in a week. Also, the survey questions are somewhat unclear. In my opinion, "the hour of thinking about COVID-19" isn't a good question because how "thinking about COVID-19" is defined differently for different individuals. Therefore it will be hard to measure this variable fairly.
  - The data set has quite a few missing values. All methods of handling NA have drawbacks. In our analysis, we choose to fill the NA values with the mean of that variable. This method of handling NA preserves data size but sacrifices the accuracy of the data.
  - Not many of the variables are predictive. Some of the variables collected are not predictive against Quiz 4 Score. We should consider including other sets of variables in further studies.
- Model limitations:
  - The quiz score may not be genuinely continuous. Since the marking of the quiz is done by summing the partial marks of each question, therefore there will be some score that can't be reached. For example, no one will ever score 9.9 points if there are no partial marks of 0.1. This is called an ordinal data (link to website).
  - The quiz score may not follows a normal distribution. This is the nature of assessment scores because if the quiz score truly does follow a normal distribution, it will imply that half of the class is failed.
  - The model we are proposing is possibly not the best. As mentioned in the Model Validation section, the difference between $MSPE$ in the testing data set and $MSRes$ in the training data set is relatively big, meaning our model doesn't generalize well into other samples from the population. One possible cause is that the data set is too small.

Conclusion

The proposed final model tells us:

- The score of the final assessment of STA302 (i.e. Quiz 4) can be predicted using previous quiz scores of the student (Quiz 1 and Quiz 3 scores) and which Country the student reside in.
- The students in North America (Canada or the USA) have a negative impact on the Quiz 4 score. This is possibly due to the epidemic in North America was more severe during the two months of studying STA302.
- Notice *all the variables/models related to COVID-19 are eliminated.* This tells us *the effect of COVID-19 isn't the most significant factor to student's performance.* Otherwise, if COVID-19 truly is the most significant factor, models with COVID-19 related variables should show a big increase in predictive power towards student performance.

Group Work Division

This final project report for STA302 is coauthored by Hongcheng Wei and Wensha Sun. The discussions is done though zoom meetings, and collaboration is done through Github.

The contribution of both of us is the same. Here is a detailed list of the millstones each of us achieved.

Project Millstones	Wensha Sun Contribution	Hongcheng Wei Contribution
Introduction	Discussion of all parts.	Writing of all parts.
Data Cleaning	Writing of all parts, including choosing aggregate variables, and come up with method to fix NA.	Discussion of all parts.
Data Exploratory	Summary of quiz 4 score, discussion of all parts.	Propose covid study ratio and writing of relationship between other variables, discussion of all parts.
Model Development	Propose manual model 1 and manual model 3, and discussion of all parts.	Writing of AIC reduced model, propose manual model 2, and discussion of all parts.
Model Selection	Discussion of all parts.	Writing of all parts.
Model Diagnostic	Discussion of all parts, writing of Outlying & Influential, Linearity Assumption, Multicollinearity.	Discussion of all parts, writing of Normality Assumption, Homoscedasticity Assumption, Independence Assumption.
Model Validation	Discussion of all parts.	Writing of all parts.
Model Interpretation	Writing of all parts.	Discussion of all parts.
Model Advantage and Limitations	Discussion on all parts.	Writing of all parts.
Conclusion	Writing of all parts.	Discussion on all parts.

Appendix

Reference

- Plot one variables against many others: https://www.datanovia.com/en/blog/how-to-plot-one-variable-against-multiple-others/. The idea of the plot in Correlation between Variables and Quiz 4 scores is borrowed from the above website.
- Box-cox Transformation: https://www.statisticshowto.com/box-cox-transformation/. The website provides details on Box-cox transformation, which is used in addressing the violation of Normality Assumption.
- Ordinal Data: https://www.formpl.us/blog/ordinal-data. This website introduces ordinal data, which is referenced in Limitations of Model.

Code

All the code used to process the data and generate the figures and result in this report.

```
# setup
knitr::opts_chunk$set(echo = FALSE, warning = FALSE, fig.height = 2, fig.width = 8, dpi = 300)
library(tidyverse)
library(GGally)
library(ggfortify)
library(MASS)
library(car)
library(gridExtra)
library(plyr)
library(lmtest)
library(pracma)
library(olsrr)
# read data, rename, unify data types
data <- read_csv("data.csv") %>%
    dplyr::rename(w1_hour = "STA302 hours (W1)") %>%
    mutate(w1_hour = as.numeric(w1_hour)) %>%
    dplyr::rename(w1_covid = "COVID hours (W1)") %>%
    mutate(w1_covid = as.numeric(w1_covid)) %>%
    dplyr::rename(q1 = "Quiz_1_score") %>%
    mutate(q1 = as.numeric(q1)) %>%
    dplyr::rename(w2_hour = "STA302 hours (W2)") %>%
    mutate(w2_hour = as.numeric(w2_hour)) %>%
    dplyr::rename(w2_covid = "COVID hours (W2)") %>%
    mutate(w2_covid = as.numeric(w2_covid)) %>%
    dplyr::rename(q2 = "Quiz_2_score") %>%
    mutate(q2 = as.numeric(q2)) %>%
    dplyr::rename(w3_hour = "STA302 hours (W3)") %>%
    mutate(w3_hour = as.numeric(w3_hour)) %>%
    dplyr::rename(w3_covid = "COVID hours (W3)") %>%
    mutate(w3_covid = as.numeric(w3_covid)) %>%
    dplyr::rename(q3 = "Quiz_3_score") %>%
    mutate(q3 = as.numeric(q3)) %>%
    dplyr::rename(w4_hour = "STA302 hours (W4)") %>%
    mutate(w4_hour = as.numeric(w4_hour)) %>%
    dplyr::rename(w4_covid = "COVID hours (W4)") %>%
    mutate(w4_covid = as.numeric(w4_covid)) %>%
    dplyr::rename(q4 = "Quiz_4_score") %>%
    mutate(q4 = as.numeric(q4))
# handle NA
data <- data %>% replace_na(list(
    w1_hour = mean(data$w1_hour, na.rm = TRUE),
```

```
    w2_hour = mean(data$w2_hour, na.rm = TRUE),
    w3_hour = mean(data$w3_hour, na.rm = TRUE),
    w4_hour = mean(data$w4_hour, na.rm = TRUE),
    w1_covid = mean(data$w1_covid, na.rm = TRUE),
    w2_covid = mean(data$w2_covid, na.rm = TRUE),
    w3_covid = mean(data$w3_covid, na.rm = TRUE),
    w4_covid = mean(data$w4_covid, na.rm = TRUE),
    q1 = mean(data$q1, na.rm = TRUE),
    q2 = mean(data$q2, na.rm = TRUE),
    q3 = mean(data$q3, na.rm = TRUE)
)) %>% drop_na(q4)
# removes the rows with 0 hours of study STA302.
data <- data %>%
    dplyr::filter(w1_hour > 0 & w2_hour > 0 & w3_hour > 0 & w4_hour > 0)
# the possible values to Country
unique(data$Country)
# dummy variables for country
data <- data %>% mutate(is_north_america = ifelse(
        !is.na(Country) & (Country == "Canada" | Country == "canada" |  Country == "USA"), 1, 0
    )) %>%
    mutate(is_asia = ifelse(
        !is.na(Country) & (
            Country == "China" | Country == "china" | Country == "Taiwan" | Country == "Pakistan" |
            Country == "Singapore" | Country == "South Korea" | Country == "UAE" | Country == "Mongolia" |
            Country == "Japan" | Country == "India"
        ), 1, 0
    ))
# corrolation maps between weekly hour and weekly COVID
corr_hours <- ggcorr(cbind(
  w1_hour = data$w1_hour,
  w2_hour = data$w2_hour,
  w3_hour = data$w3_hour,
  w4_hour = data$w4_hour
)) + ggtitle("Correlation between Weekly STA302 Hour")
corr_covid <- ggcorr(cbind(
    w1_covid = data$w1_covid,
    w2_covid = data$w2_covid,
    w3_covid = data$w3_covid,
    w4_covid = data$w4_covid
)) + ggtitle("Correlation between Weekly COVID-19 Hour")
grid.arrange(corr_hours, corr_covid, nrow = 1)
# add aggregate variables
data <- data %>%
  mutate(mean_hour = rowMeans(cbind(w4_hour, w3_hour, w2_hour, w1_hour))) %>%
  mutate(mean_covid = rowMeans(cbind(w4_covid, w3_covid, w2_covid, w1_covid)))
# Reorder variables
data <- data %>% dplyr::select(
    is_north_america, is_asia,
    q1, q2, q3, q4,
    w1_hour, w2_hour, w3_hour, w4_hour,
    w1_covid, w2_covid, w3_covid, w4_covid,
    mean_hour, mean_covid
  )
# split data into training and testing
split <- floor(nrow(data) * 0.2)
test_data <- data[1:split, ]
train_data <- data[(split+1):nrow(data), ]
head(test_data)
head(train_data)
# histogram and boxplot for q4
```

```
q4_histogram <- ggplot(data, aes(x = q4)) +
    geom_histogram(binwidth = 0.5) +
    geom_vline(xintercept = mean(data$q4), color = "red") +
    ggtitle("Histogram of q4")
q4_boxplot <- ggplot(data, aes(x = q4)) +
    geom_boxplot() +
    ggtitle("Boxplot of q4")
grid.arrange(q4_histogram, q4_boxplot, ncol = 1)
# QQ plot of q4
ggplot(data, aes(sample = q4)) +
    stat_qq() + stat_qq_line() +
    ggtitle("Q-Q plot of q4")
# sc_score of each week
w1_cs_ratio <- data.frame(week="Week 1", cs_ratio=data$w1_covid / data$w1_hour)
w2_cs_ratio <- data.frame(week="Week 2", cs_ratio=data$w2_covid / data$w2_hour)
w3_cs_ratio <- data.frame(week="Week 3", cs_ratio=data$w3_covid / data$w3_hour)
w4_cs_ratio <- data.frame(week="Week 4", cs_ratio=data$w4_covid / data$w4_hour)
cs_ratio <- rbind(w1_cs_ratio, w2_cs_ratio, w3_cs_ratio, w4_cs_ratio)
# the plots for cs_ratio
cs_ratio_histogram <- ggplot(cs_ratio, aes(x = cs_ratio)) +
    geom_histogram(binwidth = 0.2) + xlim(0, 5) + facet_wrap(~week, nrow = 1) +
    geom_vline(data=cs_ratio %>% filter(week == "Week 1"), aes(xintercept = mean(cs_ratio)), color="red") +
    geom_vline(data=cs_ratio %>% filter(week == "Week 2"), aes(xintercept = mean(cs_ratio)), color="red") +
    geom_vline(data=cs_ratio %>% filter(week == "Week 3"), aes(xintercept = mean(cs_ratio)), color="red") +
    geom_vline(data=cs_ratio %>% filter(week == "Week 4"), aes(xintercept = mean(cs_ratio)), color="red") +
    ggtitle("Histogram of Weekly Covid Study Ratio")
cs_ratio_boxplot <- ggplot(cs_ratio, aes(x = cs_ratio, y = week)) +
    geom_boxplot() + xlim(0, 5) +
    ggtitle("Boxplot of Weekly Covid Study Ratio")
grid.arrange(cs_ratio_histogram, cs_ratio_boxplot, ncol = 1)
# construct the gathered data
country_gathered <- data %>%
  dplyr::select(c(is_north_america, is_asia, q4)) %>%
  gather(key = "variables", value = "value", -q4)
quiz_gathered <- data %>%
  dplyr::select(c(q1, q2, q3, q4)) %>%
  gather(key = "variables", value = "value", -q4)
hour_gathered <- data %>%
  dplyr::select(c(w1_hour, w2_hour, w3_hour, w4_hour, q4)) %>%
  gather(key = "variables", value = "value", -q4)
covid_gathered <- data %>%
  dplyr::select(c(w1_covid, w2_covid, w3_covid, w4_covid, q4)) %>%
  gather(key = "variables", value = "value", -q4)
cs_ratio_gathered <- cbind(q4 = data$q4,
             w1_cs_ratio = w1_cs_ratio$cs_ratio,
             w2_cs_ratio = w2_cs_ratio$cs_ratio,
             w3_cs_ratio = w3_cs_ratio$cs_ratio,
             w4_cs_ratio = w4_cs_ratio$cs_ratio) %>%
  as_tibble() %>%
  gather(key = "variables", value = "value", -q4)
# gather plot
q4_country_plot <- ggplot(country_gathered, aes(x = value, y = q4)) +
    geom_point() +
    geom_smooth(method=lm) +
    facet_wrap(~variables, nrow = 1) +
    ggtitle("Quiz 4 score vs Country")
q4_quiz_plot <- ggplot(quiz_gathered, aes(x = value, y = q4)) +
    geom_point() +
    geom_smooth(method=lm) +
    facet_wrap(~variables, nrow = 1) +
```

```
    ggtitle("Quiz 4 score vs Quiz scores")
q4_hour_plot <- ggplot(hour_gathered, aes(x = value, y = q4)) +
    geom_point() +
    geom_smooth(method=lm) +
    facet_wrap(~variables, nrow = 1) +
    ggtitle("Quiz 4 score vs STA302 hour")
q4_covid_plot <- ggplot(covid_gathered, aes(x = value, y = q4)) +
    geom_point() +
    geom_smooth(method=lm) +
    facet_wrap(~variables, nrow = 1) +
    ggtitle("Quiz 4 score vs COVID-19 hour")
q4_cs_ratio_plot <- ggplot(cs_ratio_gathered, aes(x = value, y = q4)) +
    geom_point() +
    geom_smooth(method=lm) +
    facet_wrap(~variables, nrow = 1) +
    ggtitle("Quiz 4 score vs Covid Study Ratio")
grid.arrange(
  q4_country_plot,
  q4_quiz_plot,
  q4_hour_plot,
  q4_covid_plot,
  q4_cs_ratio_plot, ncol = 1)
full_model <- lm(
  q4 ~ is_asia + is_north_america +
    w1_hour + w2_hour + w3_hour + w4_hour +
    w1_covid + w2_covid + w3_covid + w4_covid +
    q1 + q2 + q3 +
    I(w1_covid / w1_hour) + I(w2_covid / w2_hour) +
    I(w3_covid / w3_hour) + I(w4_covid / w4_hour), data = train_data)
aic_reduced_model <- step(full_model, trace = 0, direction = "backward")
aic_reduced_model_summary <- summary(aic_reduced_model)
aic_reduced_model_summary
# manual model 1
manual_model_1 <- lm(q4 ~ is_asia + is_north_america + q1 + q2 + q3 + mean_hour, data = train_data)
summary(manual_model_1)
# manual model 1 reduced
manual_model_1_reduced <- lm(q4 ~ is_north_america + q1 + q3, data = train_data)
anova(manual_model_1_reduced, manual_model_1)
manual_model_1_fixed <- manual_model_1_reduced
manual_model_1_fixed_summary <- summary(manual_model_1_fixed)
manual_model_1_fixed_summary
# manual model 2
manual_model_2 <- lm(q4 ~ is_north_america + w4_hour + w4_covid +
                       I(w4_covid / w4_hour) + q3, data = train_data)
manual_model_2_summary <- summary(manual_model_2)
manual_model_2_summary
# manual model 3
manual_model_3 <- lm(q4 ~ mean_covid + mean_hour + I(mean_covid / mean_hour), data = train_data)
manual_model_3_summary <- summary(manual_model_3)
manual_model_3_summary
# SSRes
models_SSRes <- c(
  sum(aic_reduced_model$residuals ^ 2),
  sum(manual_model_1_fixed_summary$residuals ^ 2),
  sum(manual_model_2_summary$residuals ^ 2),
  sum(manual_model_3_summary$residuals ^ 2)
)
# r^2_adj
models_r_squared_adj <- c(
  aic_reduced_model_summary$adj.r.squared,
```

```
  manual_model_1_fixed_summary$adj.r.squared,
  manual_model_2_summary$adj.r.squared,
  manual_model_3_summary$adj.r.squared
)
# mallow's C_p
models_mallow_C_p <- c(
  ols_mallows_cp(aic_reduced_model, full_model),
  ols_mallows_cp(manual_model_1_fixed, full_model),
  ols_mallows_cp(manual_model_2, full_model),
  ols_mallows_cp(manual_model_3, full_model)
)
# AIC
models_AIC <- c(
  AIC(aic_reduced_model),
  AIC(manual_model_1_fixed),
  AIC(manual_model_2),
  AIC(manual_model_3)
)
# models report
models_report <- cbind(
  Model = c("AIC Reduced Model", "Manual Model 1", "Manual Model 2", "Manual Model 3"),
  p_prime = c(
    length(coef(aic_reduced_model)),
    length(coef(manual_model_1_fixed)),
    length(coef(manual_model_2)),
    length(coef(manual_model_3))
  ),
  SSRes = round(models_SSRes, 3),
  R_squared_adj = round(models_r_squared_adj, 3),
  C_p = round(models_mallow_C_p, 3),
  AIC = round(models_AIC, 3)
  ) %>% as_tibble() %>%
  mutate()
models_report
# set the selected models
selected_model <- manual_model_1_fixed
# diagnostic plots
autoplot(selected_model)
# Outlying and influential for X
n <- nrow(train_data)
p_prime <- length(coef(selected_model))
# calculate leverage value
model_leverage <- hatvalues(selected_model)
# calculate cooks distance
model_cooks_distance <- cooks.distance(selected_model)
# construct report
model_outlying_influential_x <- cbind(
    row_index = seq(1:n),
    Y = train_data$q4,
    Y_hat = fitted(selected_model),
    leverage = model_leverage,
    cooks_distance = model_cooks_distance
  ) %>%
  as_tibble() %>%
  mutate(is_outlier = leverage > 0.5) %>%
  mutate(is_influential = cooks_distance > qf(0.5, p_prime, n - p_prime))
head(model_outlying_influential_x)
model_outlying_influential_x %>% filter(is_outlier & is_influential)
# Outlying and influential for Y
# calculate studentized deleted residual
```

```
model_studentized_deleted_residual <- rstudent(selected_model)
# alpha value
alpha <- 0.05
# construct report
model_outlying_influential_y <- cbind(
    row_index = seq(1:n),
    Y = train_data$q4,
    Y_hat = fitted(selected_model),
    t = model_studentized_deleted_residual,
    cooks_distance = model_cooks_distance
  ) %>%
  as_tibble() %>%
  mutate(is_outlier = abs(t) > qt(1 - alpha/(2*n), n - p_prime - 1)) %>%
  mutate(is_influential = cooks_distance > qf(0.5, p_prime, n - p_prime))
head(model_outlying_influential_y)
model_outlying_influential_y %>% filter(is_outlier & is_influential)
# apply the transformation
lambda <- 1.35
train_data <- train_data %>% mutate(q4_transformed = (q4 ^ lambda - 1) / lambda)
q4_before_hist <- ggplot(train_data, aes(x = q4)) +
  geom_histogram(binwidth = 0.5) +
  geom_vline(xintercept = mean(train_data$q4), color = "red") +
  xlim(0, 20) +
  ggtitle("Histogram for Q4")
q4_transformed_hist <- ggplot(train_data, aes(x = q4_transformed)) +
  geom_histogram(binwidth = 0.5) +
  geom_vline(xintercept = mean(train_data$q4_transformed), color = "red") +
  xlim(0, 20) +
  ggtitle("Histogram for Q4 Transformed")
grid.arrange(q4_before_hist, q4_transformed_hist, ncol = 1)
# the refitted model
model_refitted <- lm(formula = q4_transformed ~ is_north_america + q1 + q3, data = train_data)
# diagnostic plots for refitted model
autoplot(model_refitted)
# VIF of the refitted models
model_vif <- vif(model_refitted)
model_vif
# mean VIF
mean(model_vif)
# summary of final model
final_model <- model_refitted
summary(final_model)
# MSPE
model_predicted_q4 <- (lambda * predict(final_model, test_data) + 1) ^ (1 / lambda)
model_mspe <- mean((test_data$q4 - model_predicted_q4) ^ 2)
model_mspe
# MSRes
n <- nrow(train_data)
p_prime <- length(coef(final_model))
model_msres <- sum(final_model$residuals ^ 2) / (n - p_prime)
model_msres
```