Factors influencing performance on the final assessment

Contents

1. Introduction

The Covid-19 pandemic resulted in UTSG closing and teaching occurring through digital media. Teachers' responsibility of "guiding students to gain knowledge independently" has not been fully played in online teaching, resulting in low study efficiency and poor academic achievement for students. High dropout rates and late completion of higher education have become both moral and financial issues. It is a cost to the government and society, a waste of money for the family, and a failure experience for the university student. Early identification of at-risk students allows decision-makers to implement targeted development programmes to reduce dropout and improve academic achievement (Séllei et al., 2021).

This article reports research that applied MULTIPLE LINEAR REGRESSION as a lens to examine the factors that predict student performance on the final STA302 assessment. On running these models against the dataset it was observed that there is a correlation between students' abilities in individual quizzes with the final overall academic performance. It can be determined that the amount of time they spend studying each quiz, the amount of time they spend studying for COVID each week, and the country they live in are significant factors in influencing their total academic success. Students could use this prediction model as a comprehensive guide to carefully plan out the effort they will need to put in in order to score superlative grades in the coming semester.

In the 'Exploratory data analysis' section, our team will illustrate our cleaned data set and give the most important information from data by presenting tables and plots. The 'Model development' section will explain the statistical method the team used to clean and analyze the data set besides the fully interpreted parameters and variables. At the end of this section, a proper model will be established and explained. The 'Conclusion' section will primarily be focused on presenting results from the 'Model development' section as well as some conclusions about the interpretation of these results will be drawn. Finally, our team will end with some discussions about the potential underlying limitations to our approach of the study.

2. Exploratory data analysis

The analysis task is to find the factors which predict student performance on the final STA302 assessment (i.e. quiz 4). The variables available in the data collected for this task are quizzes scores, the number of hours that students spend on studying each quiz, the number of hours that students spend on thinking about covid-19 each week, and student's current stations. Data cleaning is required due to some NA rows founded in the data set. We do so by deleting all NA rows. In addition, an extreme outlier in Covid-19 hours(W1) will be removed. To help expose hidden trends, we disaggregate the data by region. The primary analysis task is approached by fitting a regression model where the Quiz 4 scores are the response variable. However, exploring the data helps us better understand patterns within the data and reveals other interesting features not described by this model.

Histogram of quiz 4 scores (Figure 1) where the bins cover 1 score increment. The distribution of values is skewed left and unimodal. Data skewed to the left is usually a result of a higher boundary in a data set. Most data falls to the right of the graph's peak. On the left-skewed histogram, the mean is smaller than the median. The few smaller values bring the mean down, and the median is minimally affected.

Figure 1:Histogram of STA302 Quiz 4 Scores

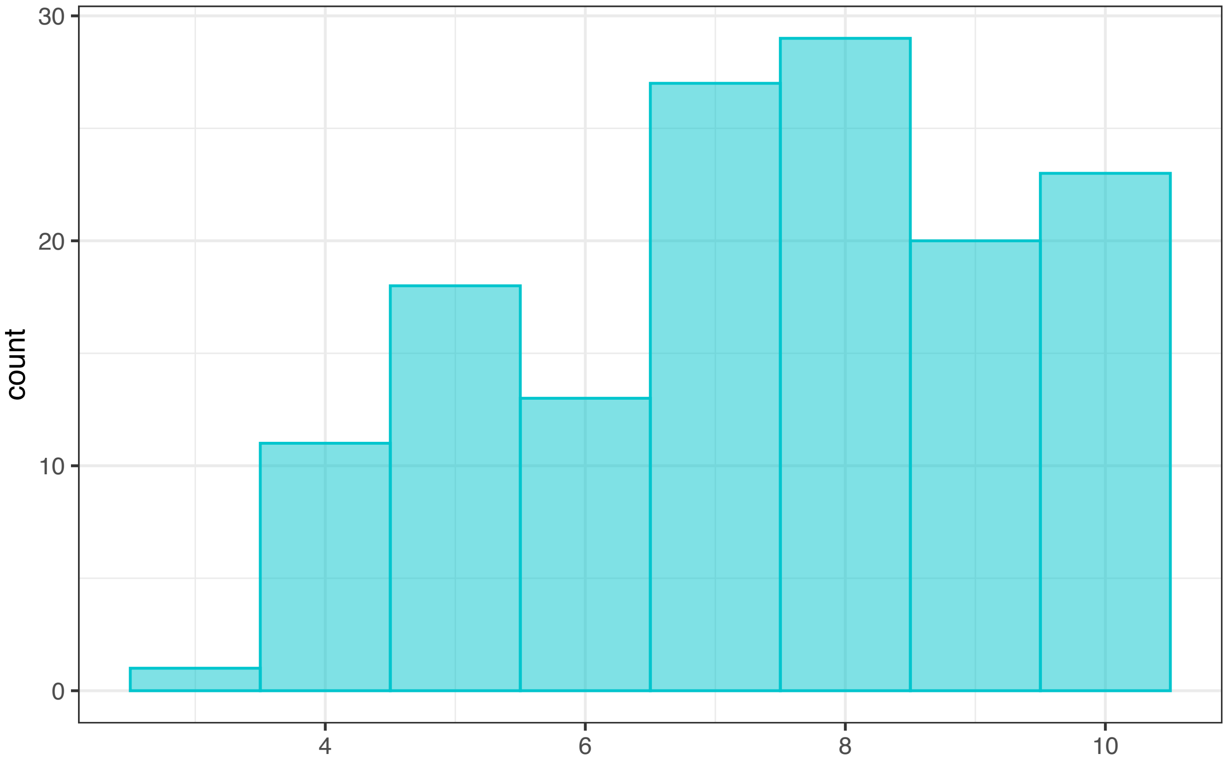


Table 1: STA302 Quiz 4 Score

Min	Max	Mean	Median	SD
3	10	7.4	8	1.9

The following boxplot shows the STA302 students' quiz 4 scores from 5 regions (Figure 2). Boxes overlap with one another. All median lines lie within the overlap among all boxes. Compare the respective medians of each box, the East Asia group has the greatest median while West Asia has the lowest. In addition, North America and Southeast Asia have the same median. The longer box has more dispersed data. The smaller box has less dispersed data. The whiskers show how big a range there is between those two extremes. East Asia group has larger ranges indicate wider distribution that is more scattered data. Also, the box of East Asia is left-skewed where values gather at the upper end, making a short and tight section there. To the left of that crowd, data points spread out, creating a long tail. On the other hand, both boxes of North American and South Asia are right-skewed where values gather at the lower end.

Figure 2:Boxplot of STA302 Quiz 4 Scores by Region

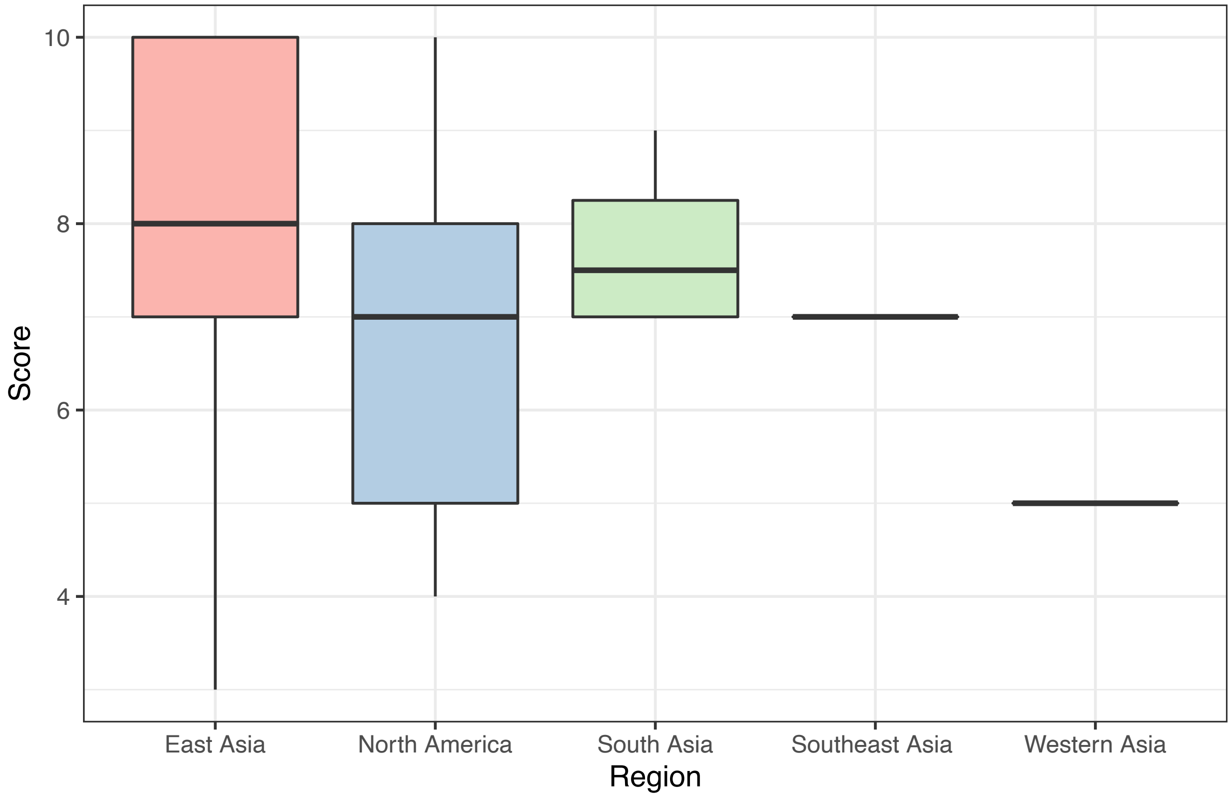
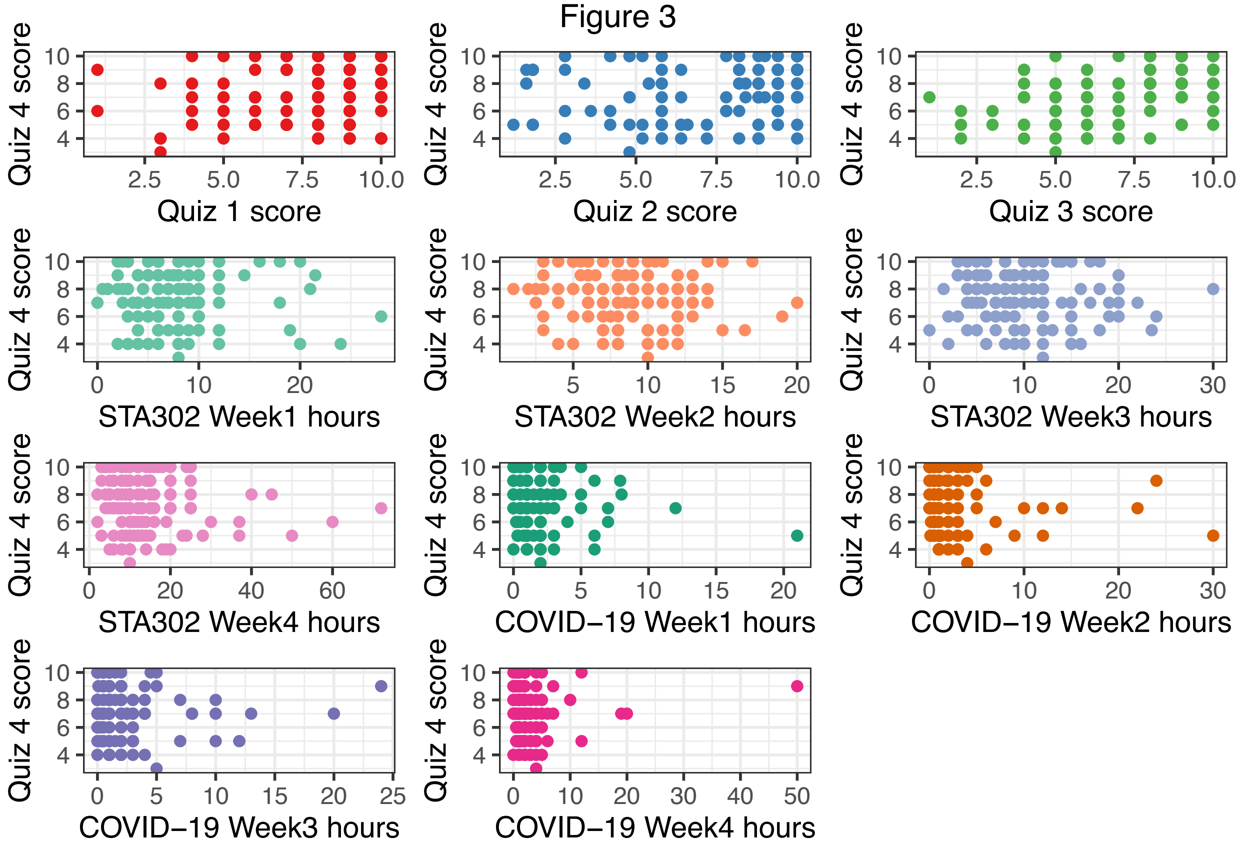


Table 2: STA302 Quiz 4 Score by Region

Region	Min	Max	Mean	Median	SD
East Asia	3	10	8.0	8.0	1.9
North America	4	10	7.0	7.0	1.8
South Asia	7	9	7.8	7.5	1.0
Southeast Asia	7	7	7.0	7.0	0.0
Western Asia	5	5	5.0	5.0	NA

Scatterplots of Quiz 4 scores vs explanatory variables (Figure 3). The data don't seem to resemble any kind of patterns, thus no relationship exists between Quiz 4 scores and explanatory variables.

Figure 3


Quiz 4 score
4
6
8
10
2.5
5.0
7.5
10.0
Quiz 1 score
Quiz 2 score
Quiz 3 score
0
10
20
STA302 Week1 hours
5
10
15
20
STA302 Week2 hours
0
10
20
30
STA302 Week3 hours
0
20
40
60
STA302 Week4 hours
0
5
10
15
20
COVID−19 Week1 hours
0
10
20
30
COVID−19 Week2 hours
0
5
10
15
20
25
COVID−19 Week3 hours
0
10
20
30
40
50
COVID−19 Week4 hours

3. Model development

In terms of the methods we use, we will be modeling Multiple Linear Regression models and selecting the best fitted one. In practice, the effects on explanatory variables usually exist for two or more explanatory variables. A regression analysis of explanatory variables and multiple explanatory variables that present a linear relationship would be a multiple linear regression.

3.1 Model selection

We will use the forward stepwise selection method as the model selection method. This method does not simply add new independent variables. After each addition of variable, it checks to see if the last independent variable added is still significant in the model. If the variable is no longer significant, it will be removed from the model (P-value > 0.1). Thus, the final model is an optimal combination of independent variables. The forward stepwise method is undoubtedly the robust one and is the most common method for screening independent variables in multiple linear regression. The model finally selected Quiz 3 score as the only predictor.

According to the summaries of the full model and selected model, we discover that the adjusted R-squared of full model is smaller than selected model. This demonstrates the new term improves the model more than would be expected by chance.

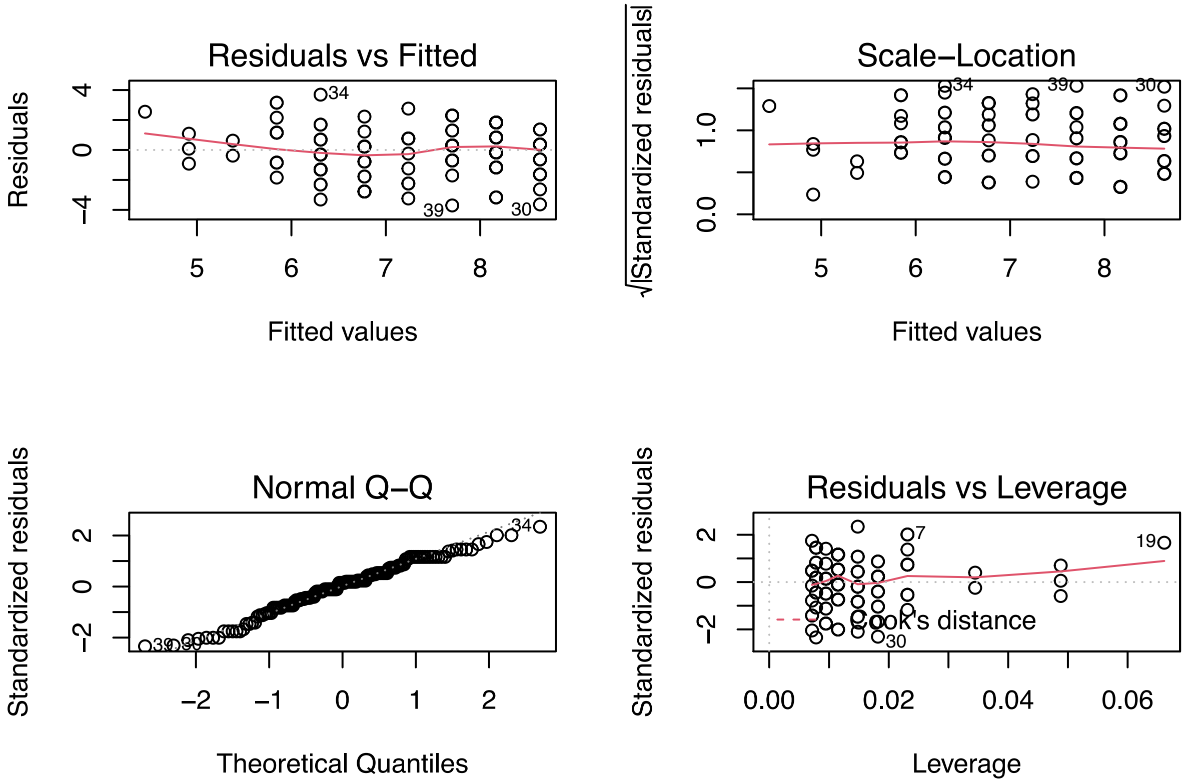
According to the summaries of the full model and selected model, we discover that the adjusted R-squared of full model is smaller than selected model. This demonstrates the new term improves the model more than would be expected by chance.

Table 3: Adjusted R-squared

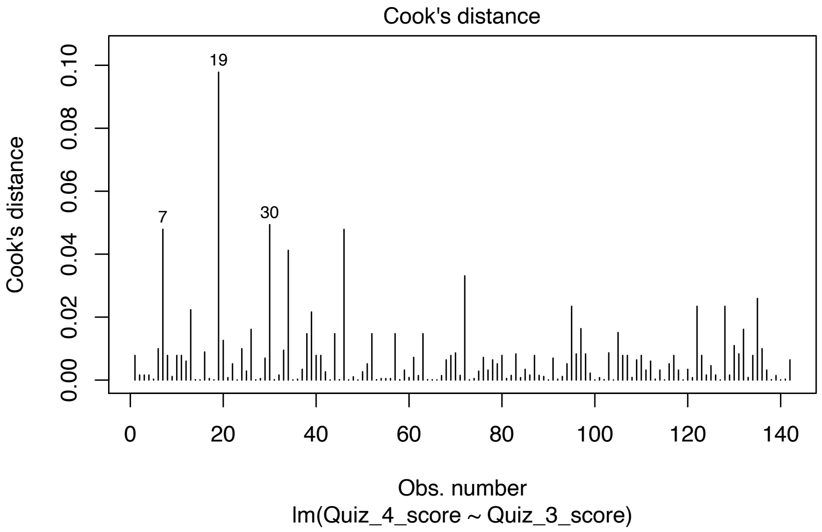
	x
Full_model	0.2645
Selected_model	0.2846

3.2 Model diagnostics

A regression diagnostic is used in statistics to evaluate model assumptions and determine whether or not any observations have a large, undue influence on the analysis (Penrose et al., 1985). The assumptions for linear regression are Linearity, Homoscedasticity, Independence, and Normality. The plot of residuals versus fitted values can be used to examine the assumption of linearity and homoscedasticity. There is no evidence that the linearity assumption is violated if residuals are evenly distributed around a horizontal line without distinct patterns. The Scale-Location plot is useful to check homoscedasticity. If there is a straight line with randomly spread points, there is no indication of heteroscedasticity (constant variance). The Normal Q-Q plot is helpful to evaluate if the errors are normally distributed. There is no indication that the normality assumption is violated if the residuals follow a straight line. Whether there are influential observations can be checked via Residuals versus Leverage plot. When the observations are in the upper right or lower right corner, they have a long Cook's distance and are hence influential.



According to the model diagnostics shown above, we could tell the assumptions are mostly satisfied. However, according to the Residuals versus Leverage plot, there could be outliers and influential points. We, therefore, identify and remove outlying and influential points by applying Studentized deleted residuals. After constructing more statistical tests for outliers and drawing Cook's distance plot, we conclude that there is no existing outlier and the influential points are points 7, 19, 30.



Finally, we remove the influential points and construct a summary for the cleaned model. We discover that after removing the influential points, the adjusted R-squared increases from 0.2846 to 0.3379, which demonstrates we successfully added useful variables.

Table 4: Final adjusted R-squared

	x
Full_model	0.2645
Selected_model	0.2846
Cleaned_model	0.3379

4. Conclusion

The multiple linear regression model allows us to make predictions about the response variable based on the information that is known about the explanatory variables. It helps us to determine which explanatory variables are statistically significant. In our MLR model, the quiz 3 score is statistically significant. It determined that there is a relationship between the response variable and the quiz 3 score caused by something other than chance. We later found a positive linear relationship between them by applying a simple linear regression model.

The Quiz 3 score is the only factor that predicts student performance on the final STA302 assessment (i.e. quiz 4). There is a positive linear relationship between them. We can suggest that for every unit increase in quiz 3 score, there is an associated increase in the estimated mean quiz 4 score by a factor of 0.52457.

Due to some NA rows founded earlier in the data set, we must delete all these rows to obtain a clean data set. This results in a limitation in our model. The absence of data reduces statistical power, which refers to the probability that the test will reject the null hypothesis when it is false. Further, the lost data can cause bias in the estimation of parameters. It can reduce the representativeness of the samples as well (Kang, 2013).

5. Bibliography

1. Séllei, B., Stumphauser, N., & Molontay, R. (2021). *Traits versus Grades—The Incremental Predictive Power of Positive Psychological Factors over Pre-Enrollment Achievement Measures on Academic Performance.* Applied Sciences, *11*(4), 1744. https://doi.org/10.3390/app11041744

2. Penrose, K., Nelson, A., and Fisher, A. (1985). *Generalized Body Composition Prediction Equation for Men Using Simple Measurement Techniques.* Medicine and Science in Sports and Exercise, *17*(2), 189. https://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/R/R5_Correlation-Regression/R5_Correlation-Regression7.html

3. Kang, H. (2013). *The prevention and handling of the missing data. Korean journal of anesthesiology.* Korean J Anesthesiol. *64*(5), 402–406. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3668100/

Appendix

```
alpha <- 0.05
```

```
data <- read.csv("data.csv", header = T)

data <- data %>% na.omit()


data <- data %>% mutate(Region = case_when(Country == "Canada" ~ "North America",
                                           Country == "USA" ~ "North America",

                                           Country == "China" ~ "East Asia",
                                           Country == "South Korea" ~ "East Asia",
                                           Country == "Taiwan" ~ "East Asia",
                                           Country == "Mongolia" ~ "East Asia",

                                           Country == "India" ~ "South Asia",
                                           Country == "Pakistan" ~ "South Asia",

                                           Country == "Singapore" ~ "Southeast Asia ",

                                           Country == "UAE" ~ "Western Asia"))

data <- data %>% filter(COVID.hours..W1. < 119)
```

```
summary_Quiz_4_score <- data %>% summarise(Min = min(Quiz_4_score),
                                           Max = max(Quiz_4_score),

                                           Mean = mean(Quiz_4_score),

                                           Median = median(Quiz_4_score),

                                           SD = sd(Quiz_4_score))


                                           kable(summary_Quiz_4_score,format = "pandoc",
                                                 align=rep(’c’, 5),
                                                 caption = "STA302 Quiz 4 Score",
                                                 digits = 1)
```

Table 1: STA302 Quiz 4 Score

Min	Max	Mean	Median	SD
3	10	7.4	8	1.9

```r
region_score <- data %>% group_by(Region) %>% summarise(Min = min(Quiz_4_score),
                                                        Max = max(Quiz_4_score),

                                                        Mean = mean(Quiz_4_score),

                                                        Median = median(Quiz_4_score),

                                                        SD = sd(Quiz_4_score))


kable(region_score, caption = "STA302 Quiz 4 Score by Region", digits = 1,
      align=rep('c', 5),format = "pandoc")
```

Table 2: STA302 Quiz 4 Score by Region

Region	Min	Max	Mean	Median	SD
East Asia	3	10	8.0	8.0	1.9
North America	4	10	7.0	7.0	1.8
South Asia	7	9	7.8	7.5	1.0
Southeast Asia	7	7	7.0	7.0	0.0
Western Asia	5	5	5.0	5.0	NA
NA	8	8	8.0	8.0	NA

```r
MLR = lm(Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score +
                        STA302.hours..W1. + STA302.hours..W2. + STA302.hours..W3. +
           STA302.hours..W4. + COVID.hours..W1. + COVID.hours..W2. + COVID.hours..W3. +
           COVID.hours..W4. + Region, data=data)

summary(MLR)
```

```
##
## Call:
## lm(formula = Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score +
##     STA302.hours..W1. + STA302.hours..W2. + STA302.hours..W3. +
##     STA302.hours..W4. + COVID.hours..W1. + COVID.hours..W2. +
##     COVID.hours..W3. + COVID.hours..W4. + Region, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.4130 -0.4328  0.0000  0.5295  3.1066
##
## Coefficients: (7 not defined because of singularities)
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -1.23514    2.55458  -0.484 0.630660
## Quiz_1_score                 0.11307    0.11988   0.943 0.349698
## Quiz_2_score                 0.10793    0.08838   1.221 0.227233
## Quiz_3_score                 0.39215    0.10925   3.590 0.000706 ***
## STA302.hours..W1.            0.02907    0.06763   0.430 0.668987
## STA302.hours..W2.            0.03470    0.08989   0.386 0.700943
## STA302.hours..W3.1.5        -0.42679    3.78290  -0.113 0.910584
## STA302.hours..W3.10          1.88745    2.30073   0.820 0.415546
## STA302.hours..W3.11           4.14526    2.47702   1.673 0.099910 .  
## STA302.hours..W3.12           2.76964    2.45368   1.129 0.263895    
## STA302.hours..W3.13           1.00950    3.07991   0.328 0.744330    
## STA302.hours..W3.13.5         7.06421    2.99839   2.356 0.022065 *  
## STA302.hours..W3.14           3.35861    2.67682   1.255 0.214892    
## STA302.hours..W3.15           2.64145    2.56073   1.032 0.306810    
## STA302.hours..W3.16           2.97356    3.16138   0.941 0.351030    
## STA302.hours..W3.17           2.48930    2.76526   0.900 0.371936    
## STA302.hours..W3.18           1.94556    2.66204   0.731 0.467972    
## STA302.hours..W3.19           5.07624    3.95269   1.284 0.204438    
## STA302.hours..W3.2           -1.15791    2.80498  -0.413 0.681353    
## STA302.hours..W3.20           4.17697    3.10135   1.347 0.183561    
## STA302.hours..W3.22           2.32081    3.33320   0.696 0.489191    
## STA302.hours..W3.23.5        -0.45573    3.77237  -0.121 0.904285    
## STA302.hours..W3.24           3.03384    3.26000   0.931 0.356114    
## STA302.hours..W3.3            3.68365    2.21683   1.662 0.102268    
## STA302.hours..W3.30           9.15251    3.42223   2.674 0.009838 ** 
## STA302.hours..W3.4            4.44832    2.42177   1.837 0.071644 .  
## STA302.hours..W3.4.5         -0.30423    2.93119  -0.104 0.917714    
## STA302.hours..W3.5            2.37924    2.45221   0.970 0.336173    
## STA302.hours..W3.5.5          2.94858    3.15746   0.934 0.354467    
## STA302.hours..W3.5.5<U+00A0>  2.81969    2.95501   0.954 0.344155    
## STA302.hours..W3.6            3.51289    2.46281   1.426 0.159412    
## STA302.hours..W3.7            1.84384    2.50145   0.737 0.464189    
## STA302.hours..W3.8            4.00912    2.51949   1.591 0.117288    
## STA302.hours..W3.8.5          3.19732    4.00600   0.798 0.428227    
## STA302.hours..W3.9            2.00773    2.36416   0.849 0.399432    
## STA302.hours..W4.11           0.68536    1.50012   0.457 0.649563    
## STA302.hours..W4.12           2.49836    0.80396   3.108 0.002983 ** 
## STA302.hours..W4.13           1.74574    0.86789   2.011 0.049185 *  
## STA302.hours..W4.14           0.94946    1.04830   0.906 0.369039    
## STA302.hours..W4.14.5         1.03972    3.45029   0.301 0.764290    
## STA302.hours..W4.15           3.10807    0.85903   3.618 0.000646 ***
## STA302.hours..W4.16           0.92721    1.04244   0.889 0.377629    
## STA302.hours..W4.17           1.13292    1.84443   0.614 0.541591    
## STA302.hours..W4.18           0.40745    1.26816   0.321 0.749206    
## STA302.hours..W4.19          -2.26510    3.07847  -0.736 0.464986    
## STA302.hours..W4.2            4.38959    2.49201   1.761 0.083717 .  
## STA302.hours..W4.20           1.32538    0.79550   1.666 0.101382    
## STA302.hours..W4.23           1.31386    1.86382   0.705 0.483829    
## STA302.hours..W4.24           2.90476    2.25651   1.287 0.203387    
## STA302.hours..W4.25           1.79946    1.10903   1.623 0.110403    
## STA302.hours..W4.28           1.05156    2.26740   0.464 0.644641    
## STA302.hours..W4.3            2.32763    1.27863   1.820 0.074139 .  
## STA302.hours..W4.30           1.54254    1.94637   0.793 0.431461    
## STA302.hours..W4.37                NA         NA      NA       NA    
## STA302.hours..W4.4            2.28142    1.30819   1.744 0.086755 .  
## STA302.hours..W4.40                NA         NA      NA       NA    
## STA302.hours..W4.45           1.58220    2.32856   0.679 0.499687    
## STA302.hours..W4.5            1.78955    0.93341   1.917 0.060409 .  
## STA302.hours..W4.50           1.30434    4.71282   0.277 0.782998    
## STA302.hours..W4.6            0.79870    0.95234   0.839 0.405283    
## STA302.hours..W4.60          -3.14752    2.34179  -1.344 0.184445    
## STA302.hours..W4.7            1.92984    1.06942   1.805 0.076617 .  
## STA302.hours..W4.7.5 hours    1.35058    1.98036   0.682 0.498112
## STA302.hours..W4.72           0.65718    2.81197   0.234 0.816078
## STA302.hours..W4.8            1.38492    0.74231   1.866 0.067420 .
## STA302.hours..W4.9            2.41726    0.90812   2.662 0.010169 *
## STA302.hours..W4.9.5          1.60004    1.98041   0.808 0.422608
## COVID.hours..W1.             -0.12082    0.14622  -0.826 0.412204
## COVID.hours..W2.              0.06547    0.17558   0.373 0.710646
## COVID.hours..W3.             -0.30086    0.20633  -1.458 0.150475
## COVID.hours..W4.0.083              NA         NA      NA       NA
## COVID.hours..W4.0.333        -0.23070    2.38735  -0.097 0.923368
## COVID.hours..W4.0.5           0.55931    1.16297   0.481 0.632470
## COVID.hours..W4.0.5 hour           NA         NA      NA       NA
## COVID.hours..W4.0.83               NA         NA      NA       NA
## COVID.hours..W4.1             0.50721    1.04073   0.487 0.627936
## COVID.hours..W4.1.5          -0.24670    1.34252  -0.184 0.854880
## COVID.hours..W4.10                 NA         NA      NA       NA
## COVID.hours..W4.12           -0.28763    2.33199  -0.123 0.902288
## COVID.hours..W4.19            0.84115    5.40651   0.156 0.876933
## COVID.hours..W4.2             0.14531    1.05290   0.138 0.890735
## COVID.hours..W4.2.5                NA         NA      NA       NA
## COVID.hours..W4.20            1.88350    2.62245   0.718 0.475660
## COVID.hours..W4.3            -0.01677    1.22986  -0.014 0.989172
## COVID.hours..W4.4             0.56983    1.26563   0.450 0.654309
## COVID.hours..W4.5             1.04672    1.24004   0.844 0.402271
## COVID.hours..W4.50            7.08316    4.38999   1.613 0.112364
## COVID.hours..W4.6             1.22139    3.64717   0.335 0.738984
## COVID.hours..W4.7             0.13821    2.56269   0.054 0.957186
## RegionNorth America          -0.65146    0.50056  -1.301 0.198525
## RegionSouth Asia             -0.38802    1.26667  -0.306 0.760506
## RegionSoutheast Asia          1.85569    1.76966   1.049 0.298942
## RegionWestern Asia           -0.49525    2.04161  -0.243 0.809235
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.516 on 55 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.7455, Adjusted R-squared:  0.3521
## F-statistic: 1.895 on 85 and 55 DF,  p-value: 0.006071
```

```
step = stepAIC(MLR, direction = "both")
```

```
## Start:  AIC=156.58
## Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score + STA302.hours..W1. +
##     STA302.hours..W2. + STA302.hours..W3. + STA302.hours..W4. +
##     COVID.hours..W1. + COVID.hours..W2. + COVID.hours..W3. +
##     COVID.hours..W4. + Region
##
##                      Df Sum of Sq    RSS    AIC
## - COVID.hours..W4.  14    19.182 145.57 148.50
## - COVID.hours..W2.   1     0.320 126.71 154.94
## - STA302.hours..W2.  1     0.343 126.73 154.96
## - STA302.hours..W1.  1     0.425 126.82 155.05
## - COVID.hours..W1.   1     1.569 127.96 156.32
## <none>                           126.39 156.58
## - Quiz_1_score       1     2.044 128.44 156.84
## - Region             4     8.995 135.39 158.27
## - Quiz_2_score       1     3.427 129.82 158.35
## - COVID.hours..W3.   1     4.886 131.28 159.93
## - STA302.hours..W4. 27    71.056 197.45 165.48
## - STA302.hours..W3. 24    82.274 208.67 179.27
## - Quiz_3_score       1    29.611 156.00 184.26
##
## Step:  AIC=148.5
## Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score + STA302.hours..W1. +
##     STA302.hours..W2. + STA302.hours..W3. + STA302.hours..W4. +
##     COVID.hours..W1. + COVID.hours..W2. + COVID.hours..W3. +
##     Region
##
##                     Df Sum of Sq    RSS    AIC
## - STA302.hours..W1.  1     0.006 145.58 146.51
## - STA302.hours..W2.  1     0.091 145.66 146.59
## - Quiz_2_score       1     1.211 146.78 147.67
## <none>                           145.57 148.50
## - Quiz_1_score       1     4.200 149.77 150.51
## - STA302.hours..W4. 30    84.498 230.07 153.04
## - Region             4    13.833 159.41 153.30
## - COVID.hours..W2.   1     7.736 153.31 153.80
## - COVID.hours..W3.   1    10.186 155.76 156.04
## + COVID.hours..W4.  14    19.182 126.39 156.58
## - COVID.hours..W1.   1    13.837 159.41 159.30
## - STA302.hours..W3. 27    96.901 242.48 166.44
## - Quiz_3_score       1    33.919 179.49 176.03
##
## Step:  AIC=146.51
## Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score + STA302.hours..W2. +
##     STA302.hours..W3. + STA302.hours..W4. + COVID.hours..W1. +
##     COVID.hours..W2. + COVID.hours..W3. + Region
##
##                     Df Sum of Sq    RSS    AIC
## - STA302.hours..W2.  1     0.103 145.68 144.61
## - Quiz_2_score       1     1.251 146.83 145.71
## <none>                           145.58 146.51
## + STA302.hours..W1.  1     0.006 145.57 148.50
## - Quiz_1_score       1     4.199 149.78 148.52
## - STA302.hours..W4. 30    84.897 230.48 151.29
## - Region             4    14.303 159.88 151.72
## - COVID.hours..W2.   1     7.777 153.36 151.84
## - COVID.hours..W3.   1    10.411 155.99 154.25
## + COVID.hours..W4.  14    18.763 126.82 155.05
## - COVID.hours..W1.   1    13.931 159.51 157.39
## - STA302.hours..W3. 27    96.898 242.48 164.44
## - Quiz_3_score       1    35.103 180.68 174.97
##
## Step:  AIC=144.61
## Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score + STA302.hours..W3. +
##     STA302.hours..W4. + COVID.hours..W1. + COVID.hours..W2. +
##     COVID.hours..W3. + Region
##
```

```
##                      Df Sum of Sq    RSS    AIC
## - Quiz_2_score       1     1.277 146.96 143.84
## <none>                           145.68 144.61
## + STA302.hours..W2.  1     0.103 145.58 146.51
## - Quiz_1_score       1     4.159 149.84 146.57
## + STA302.hours..W1.  1     0.018 145.66 146.59
## - Region             4    14.243 159.93 149.76
## - COVID.hours..W2.   1     7.777 153.46 149.94
## - STA302.hours..W4. 30    87.136 232.82 150.71
## - COVID.hours..W3.   1    10.309 155.99 152.25
## + COVID.hours..W4.  14    18.524 127.16 153.43
## - COVID.hours..W1.   1    14.723 160.41 156.18
## - STA302.hours..W3. 27    97.642 243.32 162.94
## - Quiz_3_score       1    36.505 182.19 174.13
##
## Step:  AIC=143.84
## Quiz_4_score ~ Quiz_1_score + Quiz_3_score + STA302.hours..W3. +
##     STA302.hours..W4. + COVID.hours..W1. + COVID.hours..W2. +
##     COVID.hours..W3. + Region
##
##                      Df Sum of Sq    RSS    AIC
## <none>                           146.96 143.84
## + Quiz_2_score       1     1.277 145.68 144.61
## + STA302.hours..W2.  1     0.129 146.83 145.71
## + STA302.hours..W1.  1     0.077 146.88 145.76
## - Quiz_1_score       1     5.665 152.62 147.17
## - COVID.hours..W2.   1     9.212 156.17 150.41
## - STA302.hours..W4. 30    89.735 236.69 151.04
## - Region             4    17.638 164.60 151.82
## - COVID.hours..W3.   1    13.210 160.17 153.97
## - COVID.hours..W1.   1    14.718 161.68 155.29
## + COVID.hours..W4.  14    16.098 130.86 155.48
## - STA302.hours..W3. 27    97.671 244.63 161.69
## - Quiz_3_score       1    36.070 183.03 172.78
```

```
step$anova
```

```
## Stepwise Model Path
## Analysis of Deviance Table
##
## Initial Model:
## Quiz_4_score ~ Quiz_1_score + Quiz_2_score + Quiz_3_score + STA302.hours..W1. +
##     STA302.hours..W2. + STA302.hours..W3. + STA302.hours..W4. +
##     COVID.hours..W1. + COVID.hours..W2. + COVID.hours..W3. +
##     COVID.hours..W4. + Region
##
## Final Model:
## Quiz_4_score ~ Quiz_1_score + Quiz_3_score + STA302.hours..W3. +
##     STA302.hours..W4. + COVID.hours..W1. + COVID.hours..W2. +
##     COVID.hours..W3. + Region
##
##
##                  Step Df     Deviance Resid. Df Resid. Dev      AIC
## 1                                            55   126.3924 156.5790
## 2  - COVID.hours..W4. 14 19.181637954        69   145.5740 148.5014
## 3 - STA302.hours..W1.  1  0.005679191        70   145.5797 146.5069
## 4 - STA302.hours..W2.  1  0.102957857        71   145.6827 144.6066
## 5      - Quiz_2_score  1  1.277195879        72   146.9599 143.8373
```

```r
MLR1 <- lm(Quiz_4_score ~ Quiz_3_score, data = data)
summary(MLR1)
```

```
##
## Call:
## lm(formula = Quiz_4_score ~ Quiz_3_score, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7034 -1.1047  0.2263  1.1560  3.6912
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   3.98464    0.46692   8.534 2.10e-14 ***
## Quiz_3_score  0.46484    0.06152   7.556 4.91e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.588 on 140 degrees of freedom
## Multiple R-squared:  0.2897, Adjusted R-squared:  0.2846
## F-statistic:  57.1 on 1 and 140 DF,  p-value: 4.906e-12
```

```r
t <- rstudent(MLR1)

Pii <- hatvalues(MLR1)

n <- length(data$Quiz_4_score)


p_pri = length(coef(MLR1))

t_cr <- qt(1-alpha/(2*n),n-(p_pri+1))

which(abs(t) > t_cr)
```

```
## named integer(0)
```

```r
Data1 <- data[-c(7,19,30),]

MLR2 <- lm(Quiz_4_score ~ Quiz_3_score, data = Data1)
```