Group 21

CASE STUDY: WHAT FACTORS AFFECT THE CGPA OF STUDENTS AT UTSC

Author(s):
Ankit Jhurani (1003697065)

Responsible for formulating the Introduction, Hypothesis, and Conclusion. Additionally, performed initial the data analysis to plan out the strategy for the research.

Pranav Sethi (1004911205)

Responsible for data cleaning and organization of dataset to fit the parameter description. Also performed splitting of the dataset into training and validation.

Aaisha Eid ()

Responsible for finding the suitable final model. Successfully transformed ideas from the entire team into code to find the best model that answers our hypothesis

John Jarlos (1004916781)

Responsible for model validation as well as remedial measures needed to make the final model more suitable.

UNIVERSITY OF TORONTO SCARBOROUGH, STAC67 WINTER 2021

Case Study: What Affects Cumulative Grade Point Average of Students at the University of Toronto, Scarborough

Table of Contents

1.Introduction

At the University of Toronto Scarborough campus, students are evaluated on their academic achievements by determining their Cumulative Grade Point Average or commonly abbreviated as CGPA. The CGPA is a number that is used as an assessment tool to evaluate a student's academic performance. This is the most common way for educational institutions to map their students and provide information on how to maximize their skills. In high school, the purpose of a CGPA is to help navigate a student towards their interests and guide them into renowned higher educational institutions.

Although, this is not the case in every institution. Some also follow the highly popular Alphabet Grading Scheme that ranges from A to F which can be further broken down into + and -. The Alphabet Grading Scheme is not the focus of this paper but is a scheme that must be acknowledged due to its popularity.

At the undergraduate level, a student's CGPA defines the students time at the university. With students being given adult like freedom, a student's CGPA allows hiring representatives at firms and recruitment staffs at honorable master's programs to understand what type of student they are recruiting. The higher a student's CGPA, the better their chances are at getting a good job or their desired master's program. However, it might be easy to pinpoint the best student, but the purpose of this study will be to find out which socio-economic and demographic factors have an impact on students' CGPA.

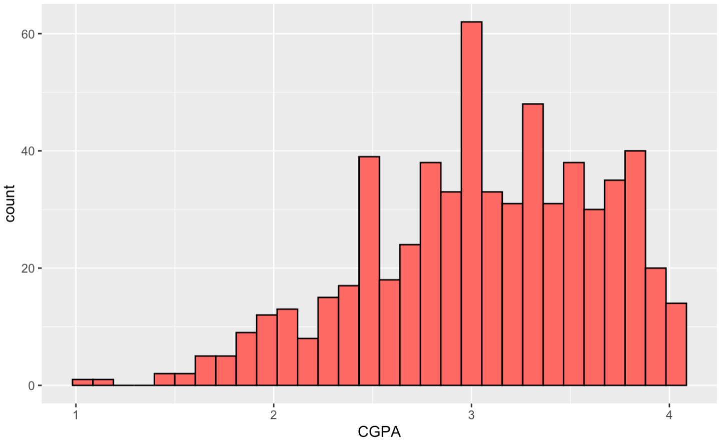
2.Hypothesis

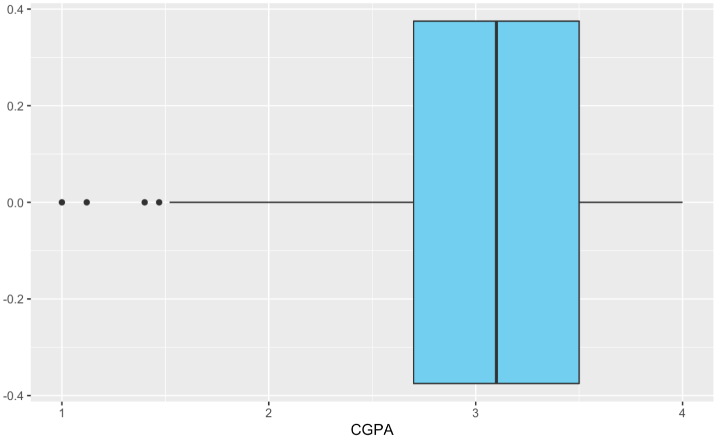
What Socio-economic and demographic factors Affects Cumulative Grade Point Average of Students at the University of Toronto, Scarborough. Students at the University of Toronto come from different walks of life and to understand why their grades are either good or bad can depend on multiple factors. This paper will serve the purpose of identifying these factors by using the data collected of students from the university by various regression learning techniques to create linear models.

3.Description of Data

A CGPA dataset of 635 students was collected from each project team that worked on surveying 30 students each. The students that were represented in this dataset were only limited to third- and fourth-year students. Some of the features that are of important significance out of the 13 columns are status (domestic or international), major type, commute time, living with family/no family, students’ interest in the course, social and co-curricular activities, family’s’ expectation in school and whether the student works part time or not.

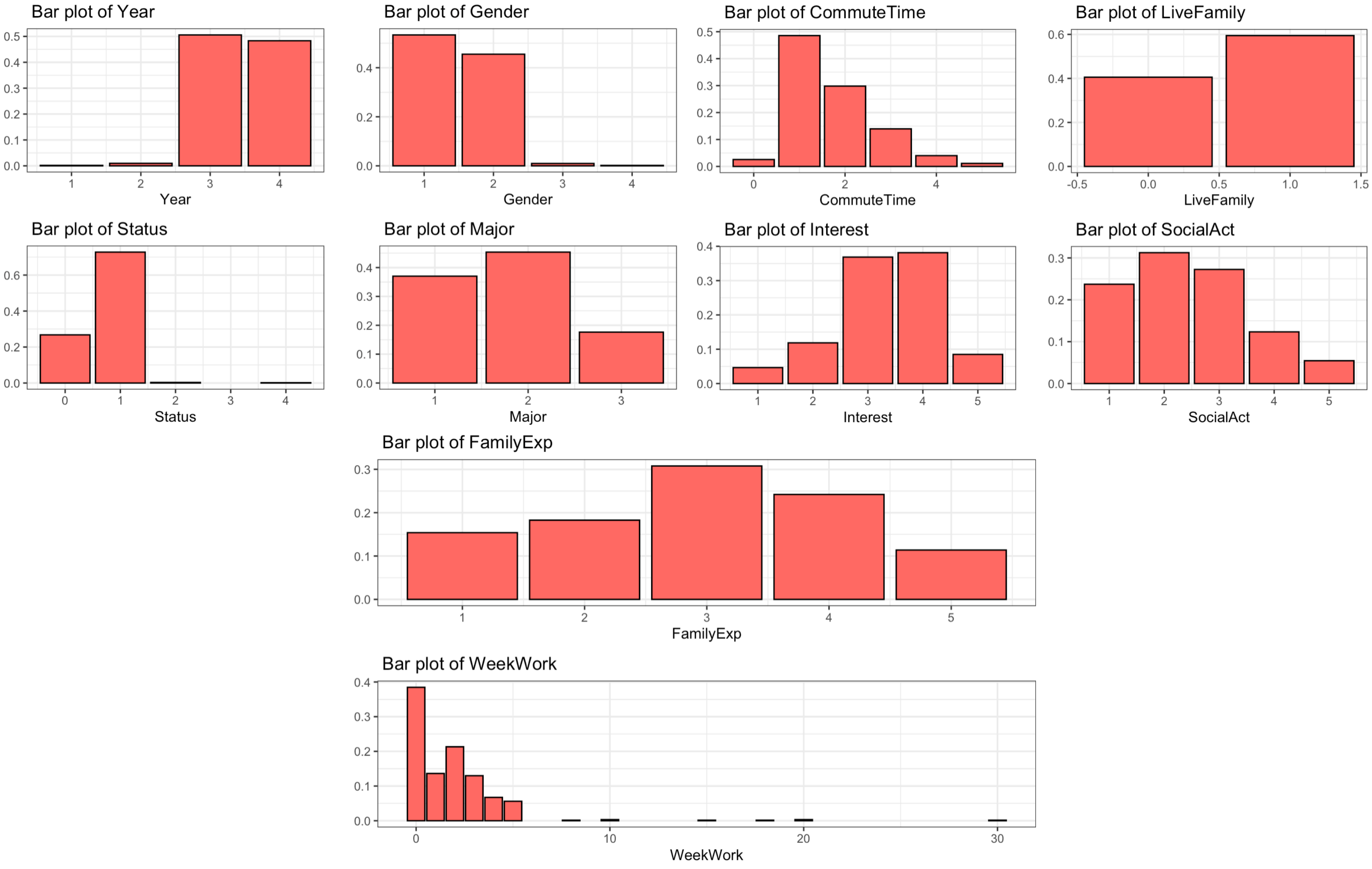
4.Initial Data Analysis



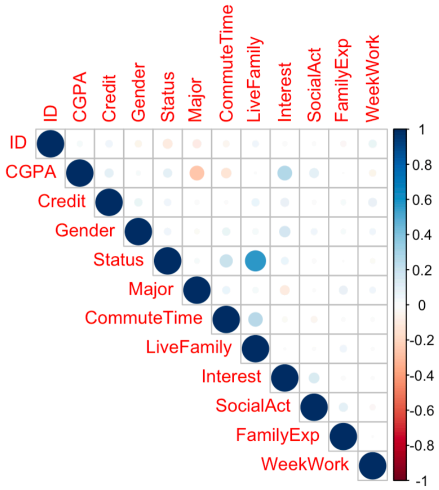


CGPA distribution shows a slightly left skewed distribution with mean 3.1. Observing the the box plot distribution of CGPA lead us to the same findings i.e., left skewed and mean 3.1.

Explanatory Variables



Correlation Among Variables

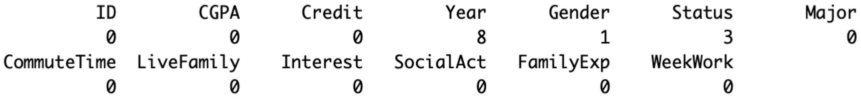


The correlation graph of our dataset indicates the relationship of each variable. The blue circle shows that there is a positive correlation between those variables. As the correlation goes down and becomes negative, its color also changes from blue to orange. An example would be CGPA and Major, based on our data, there is negative correlation between those variables.

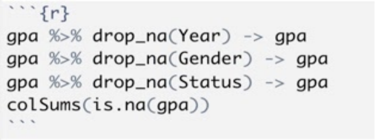
Data Cleaning and Organization

Before we started fitting any type of linear model our first step was to assess whether the given dataset was in an acceptable form or not. We uncovered through performing basic drop.na functions and categorizing all categorical variables with their respective defined descriptions that there were some discrepancies in the listed values of some of the parameters. Upon further investigation it was deciphered that these inconsistencies took place in two forms.

There were 12 instances in which the "NA" value was detected. As mentioned before, they were removed through the drop function since it was an unmendable discrepancy.

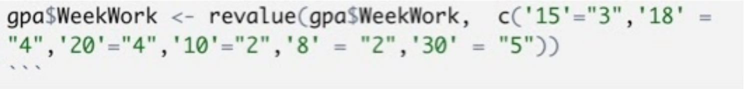


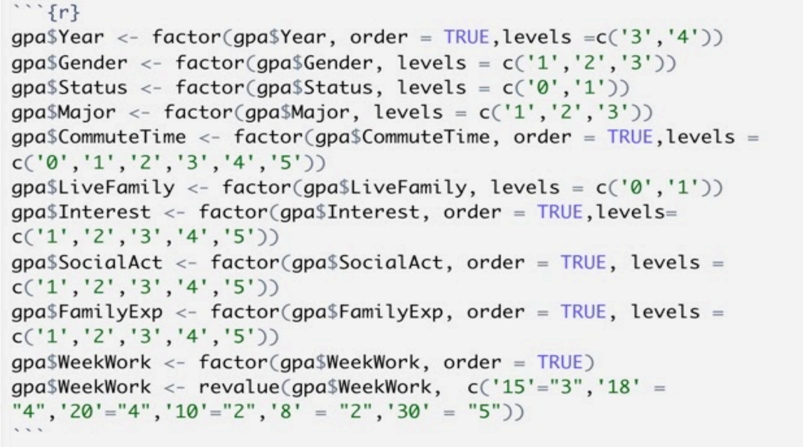
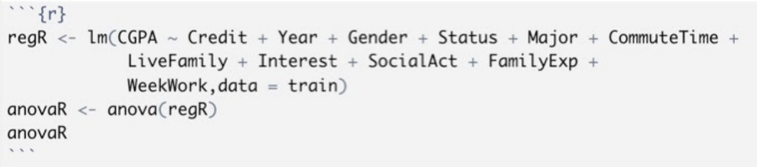
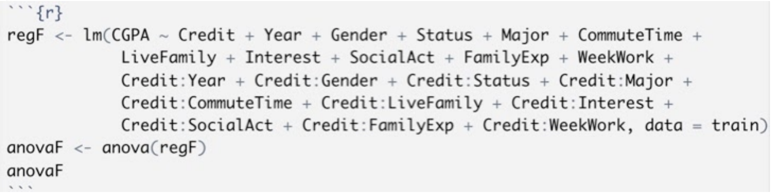
```
         ID        CGPA      Credit        Year      Gender      Status       Major 
          0           0           0           8           1           3           0 
CommuteTime  LiveFamily    Interest   SocialAct   FamilyExp    WeekWork 
          0           0           0           0           0           0 
```



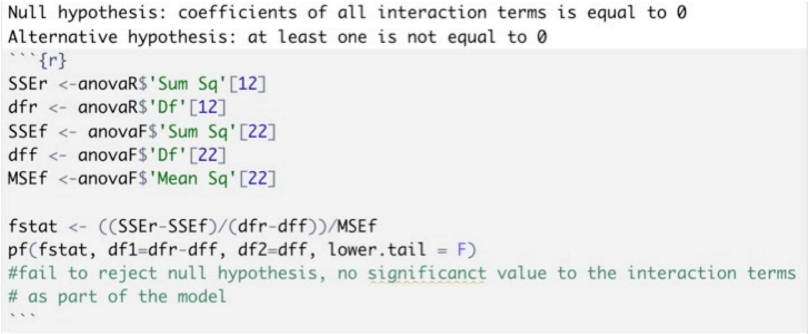
~~~
```{r}
gpa %>% drop_na(Year) -> gpa
gpa %>% drop_na(Gender) -> gpa
gpa %>% drop_na(Status) -> gpa
colSums(is.na(gpa))
```
~~~

There were instances of values in WeekWork not following the 1-5 scale but instead being listed as the actual as the actual number of hours the person works part time in a week. Since this inconsistency was amendable, it was decided that the given values which didn't follow the 1-5 scale would be transformed to fit the 1-5 scale.



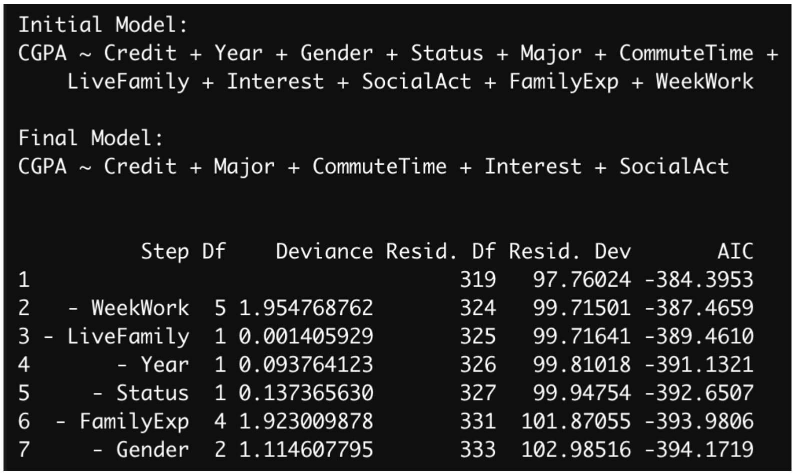
```
gpa$WeekWork <- revalue(gpa$WeekWork,  c('15'="3",'18' =
"4",'20'="4",'10'="2",'8' = "2",'30' = "5"))
```
```

After accounting for these two inconsistencies, the factor function was utilised to transform the parameters Year, Gender, Status, Major, CommuteTime, LiveFamily, Interest, SocialAct, FamilyExp, and WeekWork into categorical variables.



````
```{r}
gpa$Year <- factor(gpa$Year, order = TRUE,levels =c('3','4'))
gpa$Gender <- factor(gpa$Gender, levels = c('1','2','3'))
gpa$Status <- factor(gpa$Status, levels = c('0','1'))
gpa$Major <- factor(gpa$Major, levels = c('1','2','3'))
gpa$CommuteTime <- factor(gpa$CommuteTime, order = TRUE,levels =
c('0','1','2','3','4','5'))
gpa$LiveFamily <- factor(gpa$LiveFamily, levels = c('0','1'))
gpa$Interest <- factor(gpa$Interest, order = TRUE,levels=
c('1','2','3','4','5'))
gpa$SocialAct <- factor(gpa$SocialAct, order = TRUE, levels =
c('1','2','3','4','5'))
gpa$FamilyExp <- factor(gpa$FamilyExp, order = TRUE, levels =
c('1','2','3','4','5'))
gpa$WeekWork <- factor(gpa$WeekWork, order = TRUE)
gpa$WeekWork <- revalue(gpa$WeekWork,  c('15'="3",'18' =
"4",'20'="4",'10'="2",'8' = "2",'30' = "5"))
```
````

Lastly, the dataset was split into training and testing dataset for the purpose of validation.

````
```{r}
set.seed(1234)
gpa.sample <- sample(1:length(gpa$ID), 350, replace = FALSE)
train <- gpa[gpa.sample,]
test <- gpa[-gpa.sample,]
```
````

# 5.Process of Finding for Suitable Model

Using the training data, we began the journey of finding the most suitable model by creating two regression models: RegR & RegF.



````
```{r}
regR <- lm(CGPA ~ Credit + Year + Gender + Status + Major + CommuteTime +
           LiveFamily + Interest + SocialAct + FamilyExp +
           WeekWork,data = train)
anovaR <- anova(regR)
anovaR
```
````



````
```{r}
regF <- lm(CGPA ~ Credit + Year + Gender + Status + Major + CommuteTime +
           LiveFamily + Interest + SocialAct + FamilyExp + WeekWork +
           Credit:Year + Credit:Gender + Credit:Status + Credit:Major +
           Credit:CommuteTime + Credit:LiveFamily + Credit:Interest +
           Credit:SocialAct + Credit:FamilyExp + Credit:WeekWork, data = train)
anovaF <- anova(regF)
anovaF
```
````
````

RegF accounted for all possible interactions between our parameters while RegR ignored all possibilities of interactions between variables. Next, we formulated a Null and Alternative hypothesis to reveal what model fits our dataset better. After computing using the full model – reduced model approach, we get that the *F-stat = 0.8989691* and we are failing to reject the null hypothesis and conclude that coefficients of all interaction terms is equal to 0.



~~~
Null hypothesis: coefficients of all interaction terms is equal to 0
Alternative hypothesis: at least one is not equal to 0
```{r}
SSEr <-anovaR$'Sum Sq'[12]
dfr <- anovaR$'Df'[12]
SSEf <- anovaF$'Sum Sq'[22]
dff <- anovaF$'Df'[22]
MSEf <-anovaF$'Mean Sq'[22]

fstat <- ((SSEr-SSEf)/(dfr-dff))/MSEf
pf(fstat, df1=dfr-dff, df2=dff, lower.tail = F)
#fail to reject null hypothesis, no significanct value to the interaction terms
# as part of the model
```
~~~

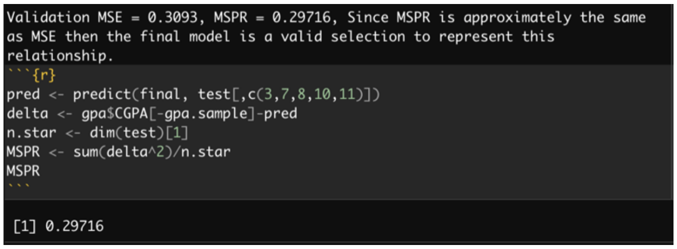
After failing to reject the null hypothesis, we concluded that there is no significant value to the interaction terms in our model. We then proceed to find the final model using the stepAIC() function. The final model reveals that Credit, Major, CommuteTime, Interest, and SocialAct are the important variables in our dataset.



```
Initial Model:
CGPA ~ Credit + Year + Gender + Status + Major + CommuteTime +
    LiveFamily + Interest + SocialAct + FamilyExp + WeekWork

Final Model:
CGPA ~ Credit + Major + CommuteTime + Interest + SocialAct


          Step Df    Deviance Resid. Df Resid. Dev       AIC
1                                   319   97.76024 -384.3953
2   - WeekWork  5 1.954768762       324   99.71501 -387.4659
3 - LiveFamily  1 0.001405929       325   99.71641 -389.4610
4       - Year  1 0.093764123       326   99.81018 -391.1321
5     - Status  1 0.137365630       327   99.94754 -392.6507
6  - FamilyExp  4 1.923009878       331  101.87055 -393.9806
7     - Gender  2 1.114607795       333  102.98516 -394.1719
```

6.Model Validation and Remedial Measures

After getting to the final model for our dataset, we then proceeded to do the model validation. In order for our model to be validated, we got the Mean Square Prediction Error or MSPR. It is the average squared difference between independent observations and predictions from the fitted model.

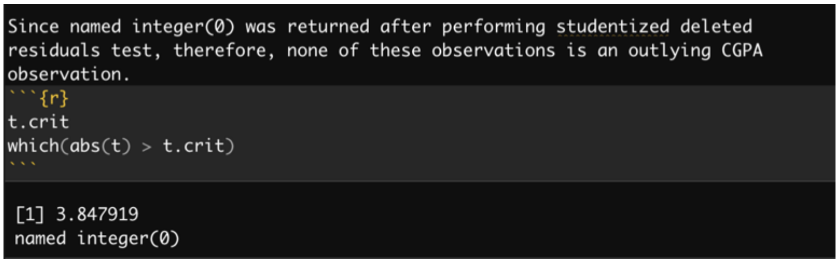


Validation MSE = 0.3093, MSPR = 0.29716, Since MSPR is approximately the same as MSE then the final model is a valid selection to represent this relationship.

~~~
```{r}
pred <- predict(final, test[,c(3,7,8,10,11)])
delta <- gpa$CGPA[-gpa.sample]-pred
n.star <- dim(test)[1]
MSPR <- sum(delta^2)/n.star
MSPR
```

[1] 0.29716
~~~

The MSPR that we calculated is roughly the same as the MSE of the final model. Thus, we concluded that the final model is a valid selection to represent this relationship.

A problem may arise when one influential or outlying point influences the regression model in a way that it pulls the regression model towards the potential outlier so that it would not be tagged as an outlier (*Deleted Residuals, PennState*). To identify outlying points in our data, using the studentized deleted residuals formula will help us identify which points are outlying if there are any. The studentized deleted residuals will delete the observations one by one and refit the model each time.

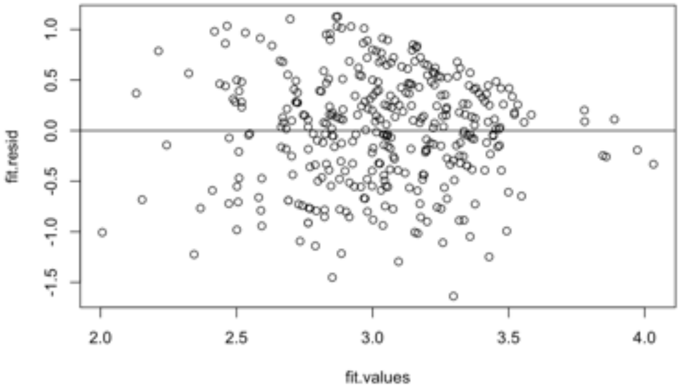


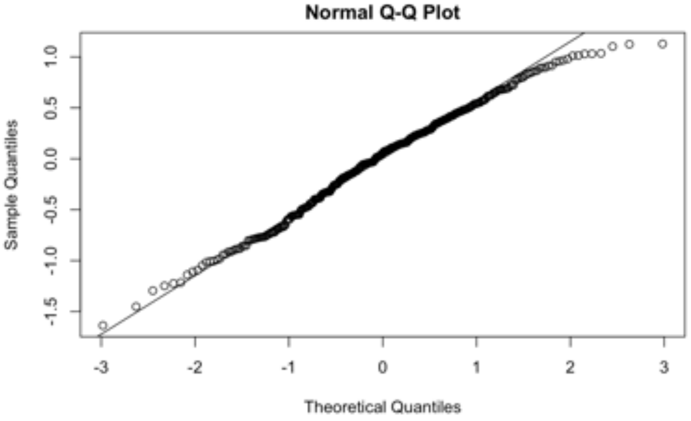
Since named integer(0) was returned after performing studentized deleted residuals test, therefore, none of these observations is an outlying CGPA observation.

~~~
```{r}
t.crit
which(abs(t) > t.crit)
```

 [1] 3.847919
 named integer(0)
~~~

The outcome of the studentized deleted residuals that we computed concluded that there are no outlying GPA observations in our dataset.

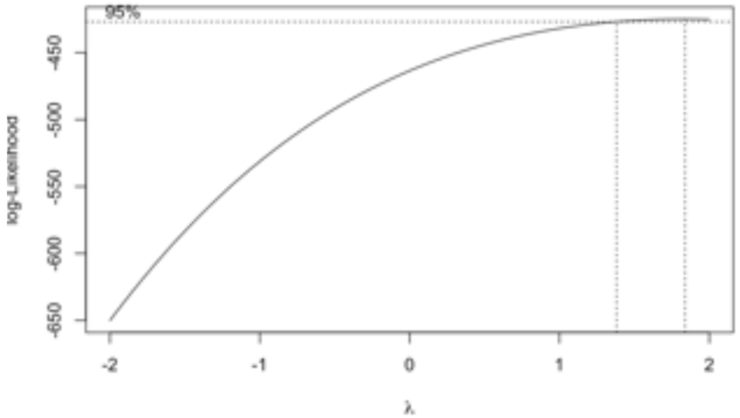
7.Final Model & Residuals

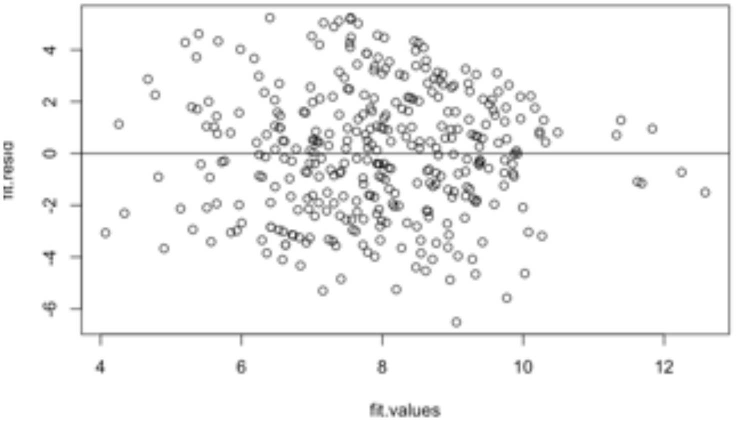


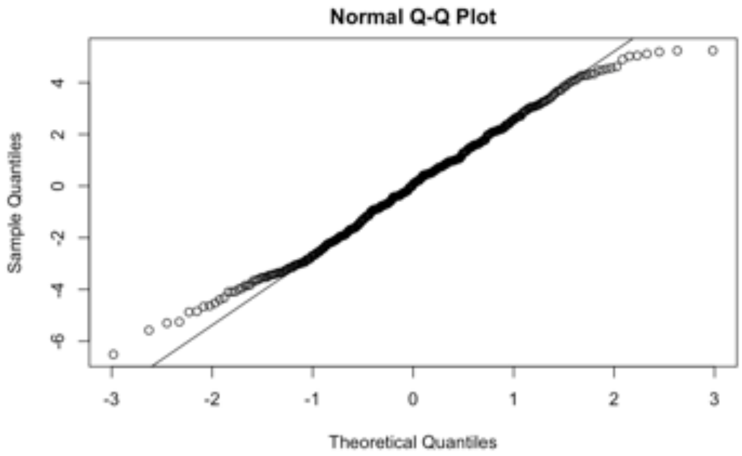


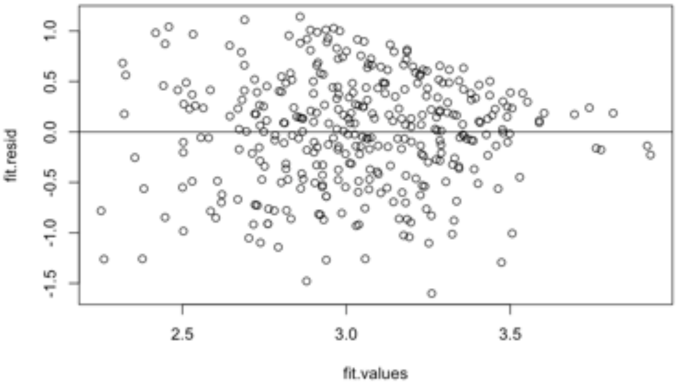
For the final model we selected, we can see from the residual plot that the residuals do not look like a random scatter, they seem to be more concentrated near the center of the graph, indicating potential non-constant variance. The Normal Q-Q plot is not close to the line near the top right most of the graph. So, after conducting the Shapiro-Wilk test, where the null hypothesis is that the residuals are normally distributed in some population, the *p-value < 0.05*. Thus, we reject the null-hypothesis and conclude that the residuals are not normally distributed in some population.

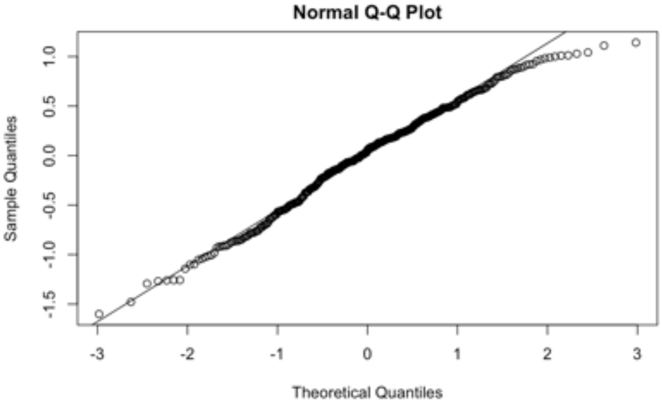
As a result, we went on to conduct a box cox transformation, which unfortunately did not resolve our non-normality of residuals since that Shapiro-Wilk test also had a *p-value < 0.05*.











We went on to conduct a weighted least squares model, which also did not solve for the non-normality of residuals we were experiencing since that Shapiro-wilk test also had a *p-value* < *0.05*. We recognize that when the sample size is large, the Shapiro-Wilk's test can be very sensitive to detect normality, hence we also used a QQplot and found that it also reveals that there still exists non-normality of residuals.

8.Conclusion/Discussion

Our final model indicated that:

- The number of Credits, the type of major you're enrolled in
- Your one-way commute time from home to campus
- How closely your major aligns with your interests
- Your degree of participation at social activities in school

Are the Socio-economic and demographic factors that affects Cumulative Grade Point Average of students at the University of Toronto, Scarborough.

Students who are searching for concrete advice which can be used to bring about informed changes in their lifestyle rather than their studying habits with the ultimate goal of increasing their CGPA at UTSC may find this case study to be the advice they were searching for all along.

Coming to the limitations of our model answering our hypothesis, we've exhausted all the remedial measures needed to make our residuals look random, but neither one of them worked. Hence, We have decided to leave this for future work, as the knowledge needed for solving this is beyond the topics taught to us for this course.

References:

Notes:

Baguley, Thomas. Serious stats: A guide to advanced statistics for the behavioral sciences. Palgrave Macmillan, 2012. (page 402)

11.4 - Deleted Residuals | STAT 501. (n.d.). PennState: Statistics Online Courses. Retrieved April 12, 2021, from https://online.stat.psu.edu/stat501/lesson/11/11.4