SCHOOL OF SCIENCE AND TECHNOLOGY

ASSIGNMENT / PROJECT SUBMISSION FORM

PROGRAMME: BSc (Hons) Information System (Business Analytics)

SEMESTER: Jan / Mar / Aug 2020

SUBJECT: IST2024 Applied Statistics

DEADLINE: 3rd July 2020

INSTRUCTIONS TO CANDIDATE

- This is an ~~individual~~ / group project.

The University requires students to adhere to submission deadlines for any form of assessment. Penalties are applied in relation to unauthorized late submission of work.

- Coursework submitted after the deadline but within 1 week will be accepted for a maximum mark of 40%.
- Work handed in following the extension of 1 week after the original deadline will be regarded as a non-submission and marked zero.

Lecturer's Remark (Use additional sheet if required)

List down the name of the group members and the student IDs here.

Virox Sim : 16066268

Koh Fu Kang : 16078875

Chia Mun Choon : 16074536

Jarrod Tham Kuok Yew : 16034753 (Leader)

I..... (Student's Name) ..... (Student ID) received the assignment and read the comments.

..... (Signature/Date)

Academic Honesty Acknowledgement

"I .....(Student's Name) verify that this paper contains entirely my own work. I have not consulted with any outside person or materials other than what was specified (an interviewee, for example) in the assignment or the syllabus requirements. Further, I have not copied or inadvertently copied ideas, sentences, or paragraphs from another student. I realize the penalties *(refer to page 16, 5.5, Appendix 2, page 44 of the student handbook diploma and undergraduate programme)* for any kind of copying or collaboration on any assignment."

…..... (Student's signature / Date)

Table of content

Introduction

Student academic performance has been the area of interest to an analyst for most of the educational institutions. Investigation on factors related to student academic performance has shown a big growth in the education field(Erum & Ahmad, 2011). This is an educational dataset which is collected by an education institution for recording and analysis purposes of student academic performance. This dataset contains 349 student records and 9 variables. Those variables include the demographic character of a student, academic performance, and attitude towards school and learning. Due to this analysis is to analyze the student performance, The chosen response variable is FinalExamMarks and chosen explanatory variables are Gender, MidTermTest, DiscussionMarks, OnTimeSubmission, and AbsenceDays.

Table 1

Edu Data Variable Description

Variable	**Description**
Gender	0 = Male, 1 = Female
Subject	The subject for which the marks are recorded
MidTermTest	Marks attained for the midterm test
LogIn	Number of times the student logged in to learning management system in the semester
DiscussionMarks	Marks attained for participation in class

OnTimeSubmission	1 = Assigned work submitted on time; 0 = Assigned work not submitted on time
AbsenceDays	1 = Absent for 7 days or more; 0 = Absent for less than 7 days
FinalExamMarks	Marks attained for final exam
FinalExamPass	1 = Passed final exam; 0 = Failed final exam

Descriptive Analysis

Variable	Mean	Std Dev	Minimum	Maximum	Mode	N	N Miss	Median	Coeff of Variation
MidTermTest	59.6103152	26.9439074	0	99.0000000	90.0000000	349	0	64.0000000	45.2000754
LogIn	38.8166189	26.7491682	0	98.0000000	12.0000000	349	0	33.0000000	68.9116387
DiscussionMarks	44.4154728	27.8170726	2.0000000	98.0000000	40.0000000	349	0	40.0000000	62.6292391
OnTimeSubmission	0.5845272	0.4935109	0	1.0000000	1.0000000	349	0	1.0000000	84.4290707
AbsenceDays	0.4011461	0.4908342	0	1.0000000	0	349	0	0	122.3579569
FinalExamMarks	60.6897421	23.0056469	14.2600000	99.8300000	38.3200000	349	0	57.8500000	37.9069776
FinalExamPass	0.6532951	0.4766041	0	1.0000000	1.0000000	349	0	1.0000000	72.9538731
Gender	0.3495702	0.4775188	0	1.0000000	0	349	0	0	136.6017002

Figure 1. Descriptive Statistics for Edu Data.

The figure above shows the descriptive statistics for 8 variables from the Edu data. The descriptive statistics include the mean, standard deviation, minimum, maximum, mode, number of observations, number of missing values, median, and coefficient of variation value. Binary variables are ignored in this table because their central tendency doesn't mean anything because all their values just have '0' and '1' in each of the variables. For categorical variables, the Subject variable is not included in the descriptive analysis because it just contains 5 unique values which are BM, English, IT, Math, and Science.

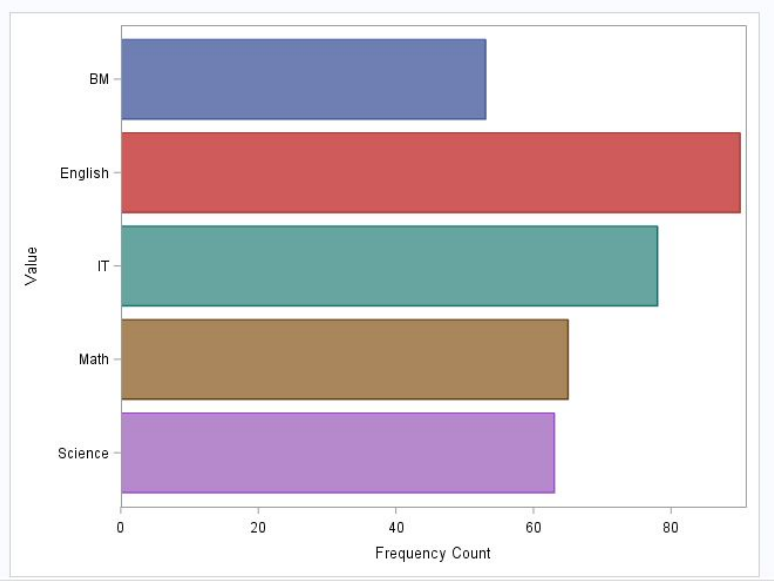

Figure 2. Frequency count for each subject

The horizontal bar chart above shows the frequency count of each subject taken by the student. It can be observed that English subject is being enrolled the most followed by IT.

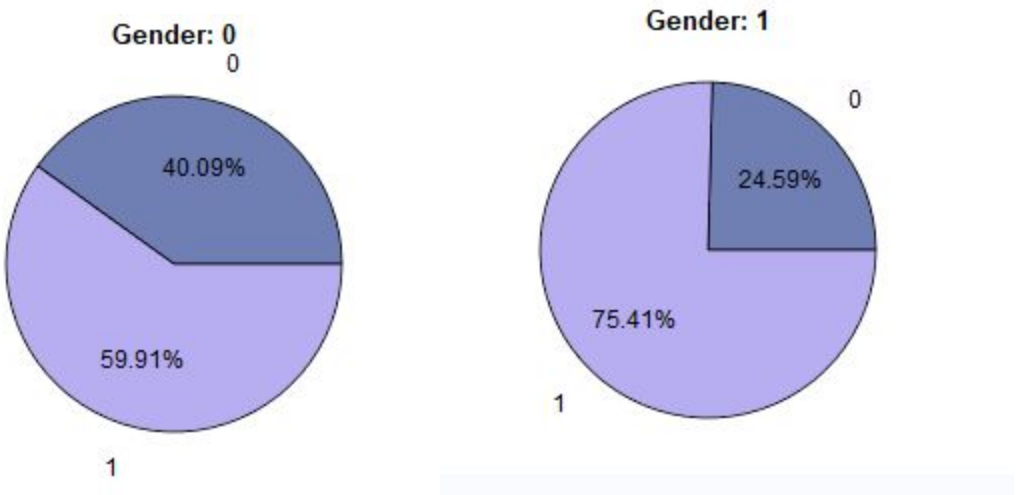

Figure 3. Percentage of students pass or fail grouped by gender.

Figure 3 indicates that students' passing percentage is higher than the failing percentage for both genders. The "0" and "1" in the outer part indicate the student passed or failed their final exam. The passing percentage for male is 59.91 whereas female percentage is 75.41. The chart shows that the difference between passing and failing is pretty huge in females.

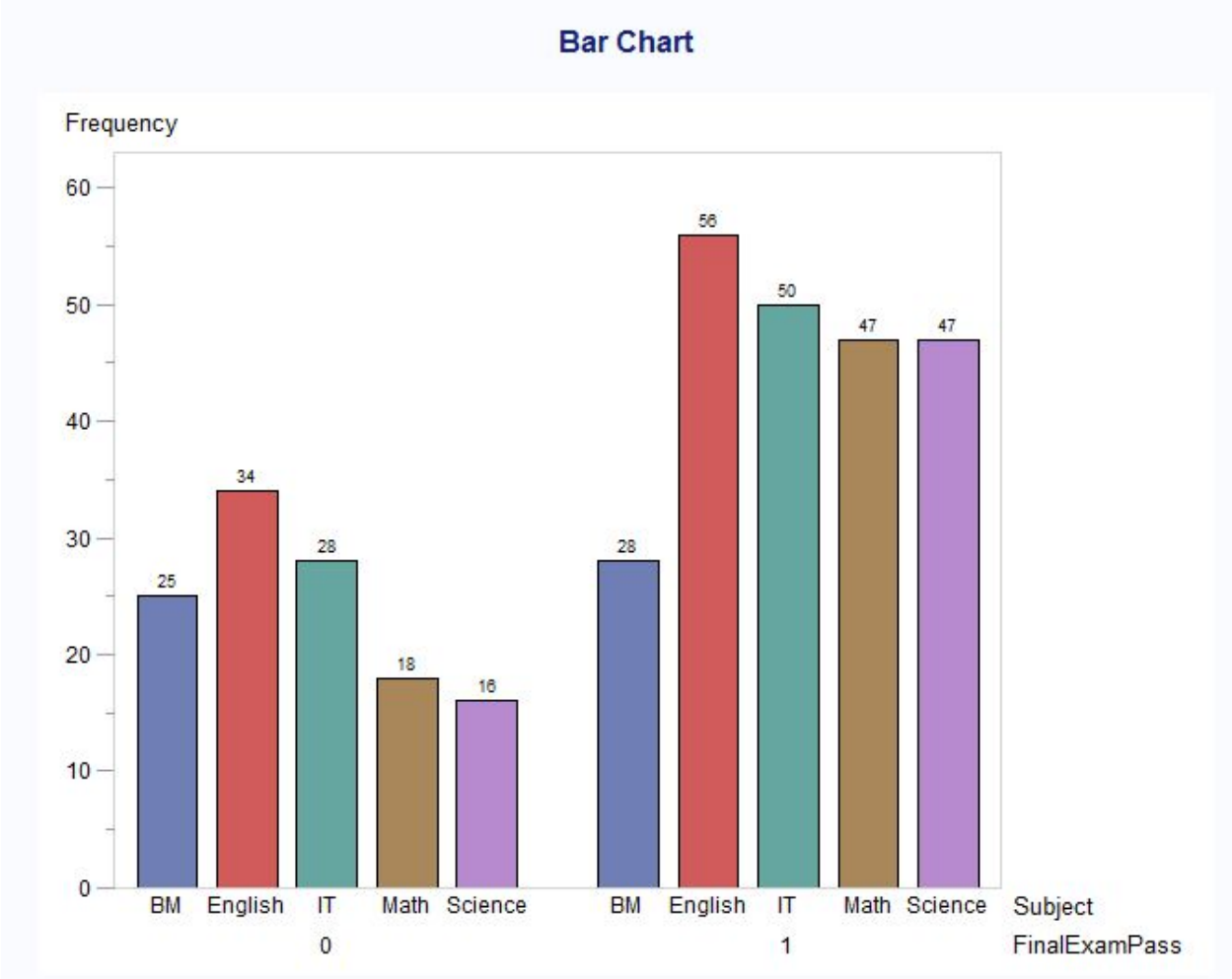

Figure 4. Number of students pass or fail each subject.

The bar graph above indicates the number of students who pass and fail for each subject. Students have a higher passing rate for each subject compared to the failure rate in the final exams.

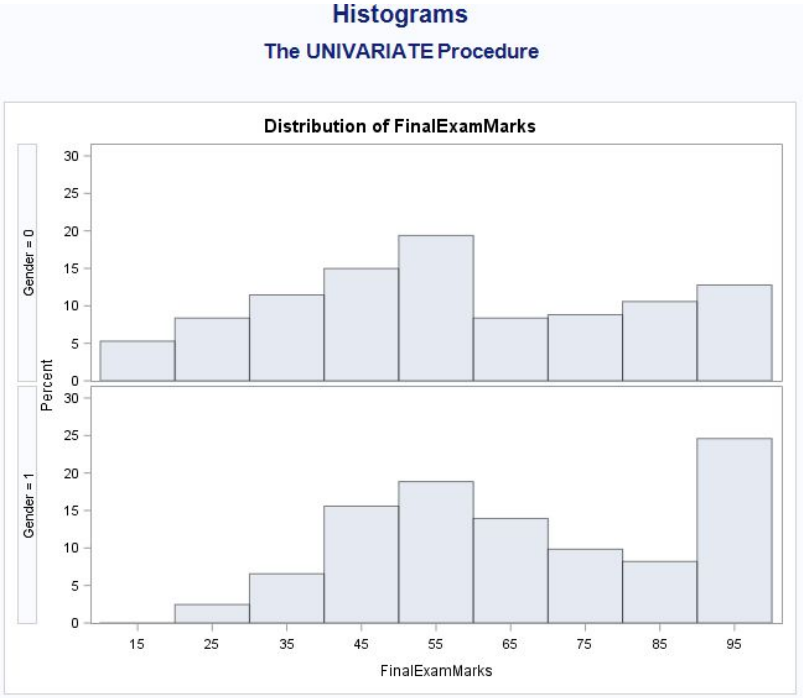

Figure 5. Distribution of FinalExamMarks by gender.

Figure 5 shows that 25% of the female and 12% of the male had scored more than 90 marks and this indicates that most of the females had scored higher marks than males. None of the females scored below 20 marks meanwhile there are 5% of the male who scored below 20 marks.

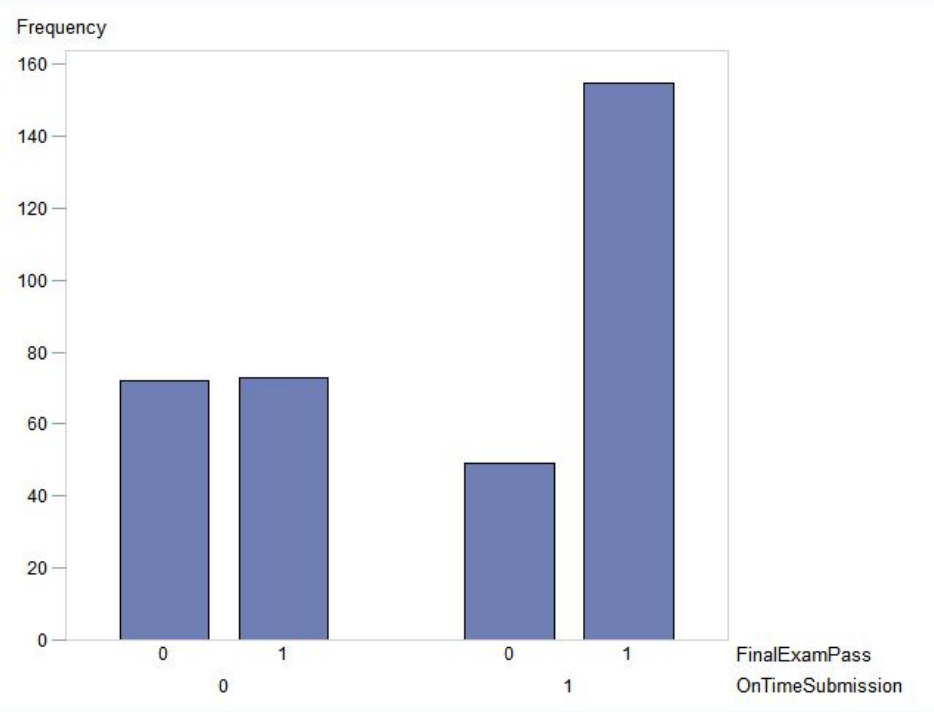

Figure 6. Frequency of student pass or fail their final exam based on OnTimeSubmission

The figure above indicates students that submit their assignment on time will most likely pass in their final exam. The bar chart shows that around 150 students pass their exams when they submit their assigned work on time.

Regression Analysis

Multiple Linear Regression

Multiple Linear Regression was chosen as the regression model for this analysis because the response variable is quantitative data and has multiple explanatory variables. The stepwise selection was used as model selection to compute each variable's significance in the model.

The REG Procedure
Model: MODEL1
Dependent Variable: FinalExamMarks

Number of Observations Read	349
Number of Observations Used	349

Stepwise Selection: Step 1

Variable DiscussionMarks Entered: R-Square = 0.7860 and C(p) = 231.6997

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	144770	144770	1274.59	<.0001
Error	347	39413	113.58141		
Corrected Total	348	184182			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	28.12317	1.07589	77606	683.27	<.0001
DiscussionMarks	0.73323	0.02054	144770	1274.59	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable MidTermTest Entered: R-Square = 0.8574 and C(p) = 41.2574

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	157922	78961	1040.35	<.0001
Error	346	26261	75.89848		
Corrected Total	348	184182			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	17.78930	1.17889	17282	227.70	<.0001
MidTermTest	0.25414	0.01931	13152	173.28	<.0001
DiscussionMarks	0.62481	0.01870	84731	1116.37	<.0001

Bounds on condition number: 1.2407, 4.9627

Figure 7. First and second iteration for the regression model.

The first iteration shows the DiscussionMarks variable entered the model with p-value <.0001 and MidTermTest entered the model in the second iteration with p-value < 0.0001. Both p-values were less than the default significance level 0.15 of the stepwise selection model.

Stepwise Selection: Step 3

Variable AbsenceDays Entered: R-Square = 0.8668 and C(p) = 18.0275

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	159646	53215	748.24	<.0001
Error	345	24537	71.12062		
Corrected Total	348	184182			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	22.21333	1.45244	16635	233.90	<.0001
MidTermTest	0.21489	0.02032	7956.10346	111.87	<.0001
DiscussionMarks	0.62337	0.01810	84318	1185.57	<.0001
AbsenceDays	-5.03681	1.02294	1724.26011	24.24	<.0001

Bounds on condition number: 1.4664, 11.823

Stepwise Selection: Step 4

Variable OnTimeSubmission Entered: R-Square = 0.8698 and C(p) = 11.9806

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	160196	40049	574.35	<.0001
Error	344	23987	69.72871		
Corrected Total	348	184182			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.63266	1.45294	15457	221.68	<.0001
MidTermTest	0.19717	0.02108	6097.92209	87.45	<.0001
DiscussionMarks	0.62024	0.01796	83154	1192.53	<.0001
OnTimeSubmission	2.82541	1.00607	549.94013	7.89	0.0053
AbsenceDays	-4.72662	1.01889	1500.58479	21.52	<.0001

Bounds on condition number: 1.6106, 21.34

Figure 8. Third and fourth iteration for the regression model.

In the third iteration, AbsenceDays entered the model with p-value <.0001 and OnTimeSubmission entered with p-value 0.0053 which both are lower than 0.15 significance level. No removal of variables on both iterations.

Stepwise Selection: Step 5

Variable Gender Entered: R-Square = 0.8727 and C(p) = 6.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	160741	32148	470.40	<.0001
Error	343	23441	68.34189		
Corrected Total	348	184182			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	20.77379	1.47020	13645	199.65	<.0001
Gender	2.70987	0.95925	545.40622	7.98	0.0050
MidTermTest	0.19145	0.02097	5695.53404	83.34	<.0001
DiscussionMarks	0.61851	0.01779	82592	1208.51	<.0001
OnTimeSubmission	3.05574	0.99935	638.97512	9.35	0.0024
AbsenceDays	-4.24071	1.02327	1173.78005	17.18	<.0001

Bounds on condition number: 1.6258, 32.323

All variables left in the model are significant at the 0.1500 level.

All variables have been entered into the model.

Summary of Stepwise Selection								
Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	DiscussionMarks		1	0.7860	0.7860	231.700	1274.59	<.0001
2	MidTermTest		2	0.0714	0.8574	41.2574	173.28	<.0001
3	AbsenceDays		3	0.0094	0.8668	18.0275	24.24	<.0001
4	OnTimeSubmission		4	0.0030	0.8698	11.9806	7.89	0.0053
5	Gender		5	0.0030	0.8727	6.0000	7.98	0.0050

Figure 9. Stepwise selection output.

The stepwise selection shows up to Step 5 with the default significant level for entering and staying resulting in the best model with all variables DiscussionMarks, MidTermTest, AbsenceDays, OnTimeSubmission, and Gender. In all 5 iterations, no variables were removed.

The equation of the fitted linear regression is $\hat{y} = 20.78 + 2.71x1 + 0.19x2 + 0.62x3 + 3.06x4 - 4.24x5$, where x1 = Gender, x2 = MidTermTest, x3 = DiscussionMark, x4 = OnTimeSubmission, and x5 = AbsensceDays. This shows that, for example, 0.19 is interpreted as the increase change in the final exam marks ($\hat{y}$) corresponding to every unit change in midterm test (x2), when the gender (x1), discussion marks (x3), on time submission (x4), and absence days (x5) is held constant.

The coefficient of determination, R-Square is 0.8727. Approximately 87.27% of the variation of FinalExamMarks can be explained by the variation in Gender, MidTermTest, DiscussionMark, OnTimeSubmission, and AbsenceDays.

The typical difference between actual final exam marks of a student and the predicted marks using the regression model is 8.27 and the average FinalExamMarks of the regression is 60.69.

The collective regression coefficient states that with an F-value = 470.40 with a corresponding p-value < 0.0001. Thus, $H0$ is rejected at the 0.05 level of significance. We can conclude that at least one of the predictor variables has a significant relationship with the response variable.

Based on the *Figure 9*, all explanatory variables have a p-value of less than the 0.05 level of significance. Thus, $H0$ is rejected. We can conclude there is strong evidence that there is a significant relationship between FinalExamMarks and all explanatory variables, which include Gender, MidTermTest, DiscussionMarks, OnTimeSubmission, and AbsenceDays.

Regression Diagnostics

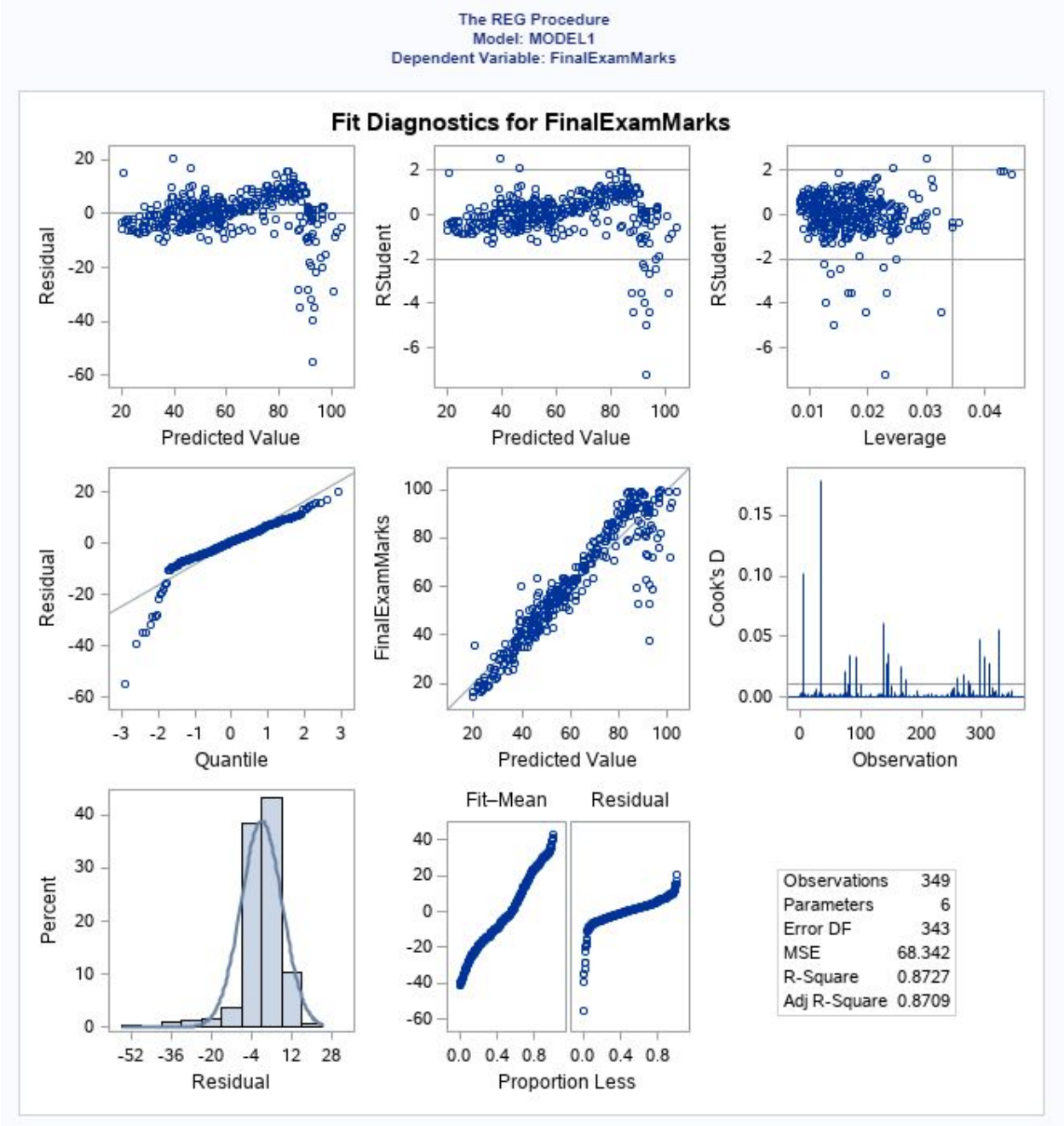

Figure 10. Regressions diagnostics plot.

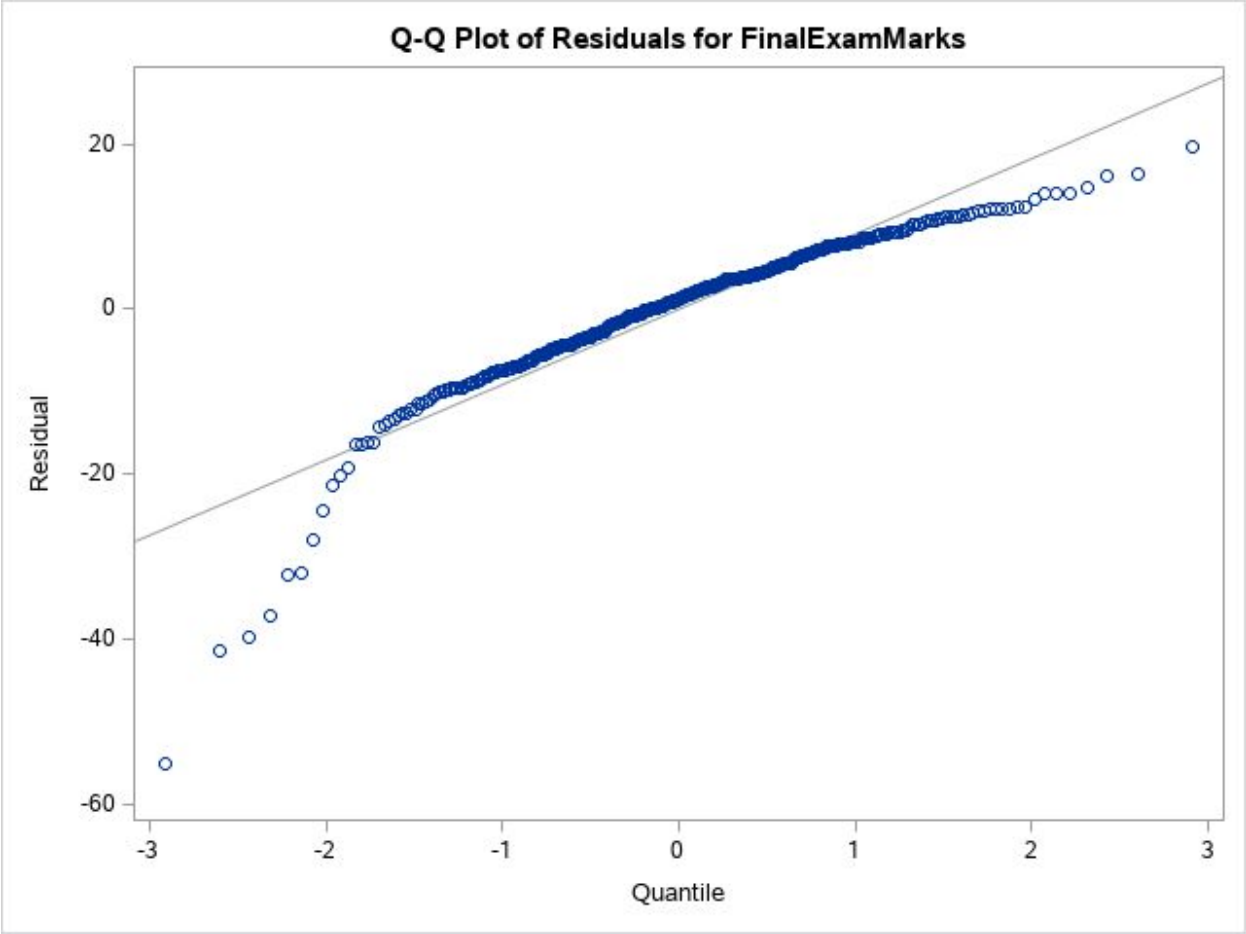

Figure 11. Q-Q plot of residuals for FInalExamMarks.

The plot of the residuals against the normal quantiles (Q-Q plot) is shown above. If the residuals are normally distributed, the plot should appear to follow closely a straight, diagonal line. However, the Q-Q plot shows there is a slight violation of normality assumption at the beginning and the end of the tail but it is much stable in the middle quantile of the regression model.

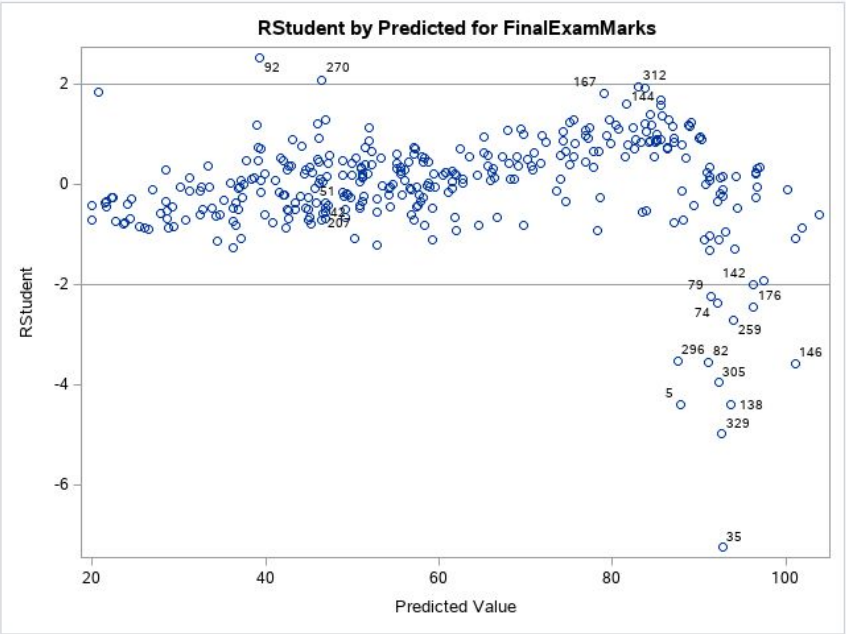

Figure 12. RStudent plot predicted for FinalExamMarks.

The residual plot shows there are many outliers beyond the 2 and -2 standard errors from the mean of 0. It seems that number 35, 329, and 5 are labeled in this plot because they have the most extreme predictor variable values.

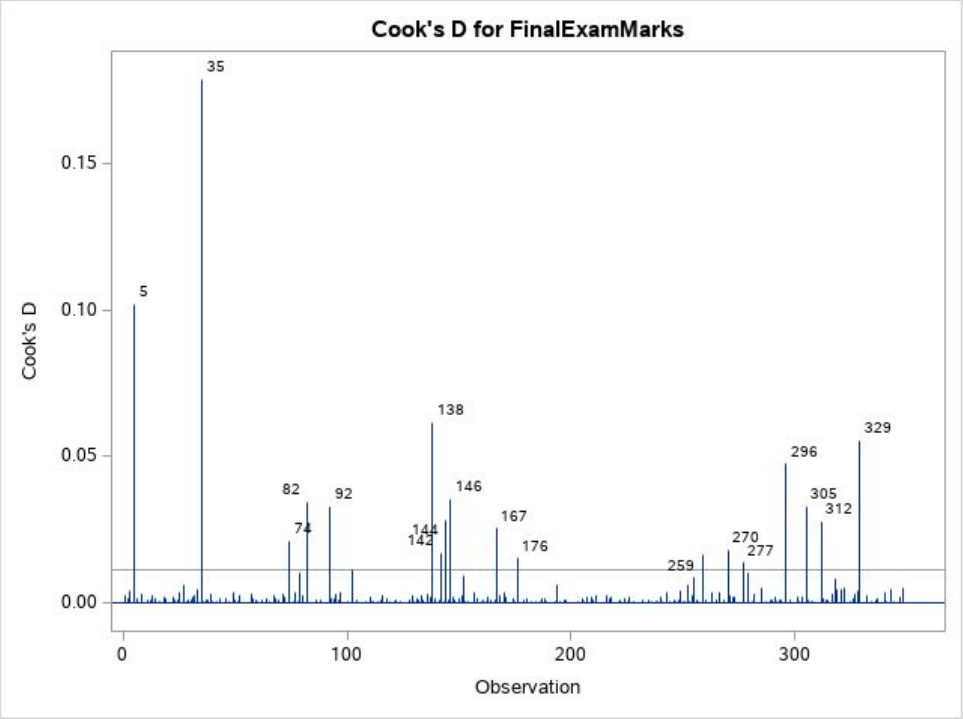

Figure 13. Cook's D plot for FinalExamMarks.

The Cook's D and DFFits plot shows there are many influential points. Especially, number 35 and 5 are shown to be the highest influential points in the regression model.

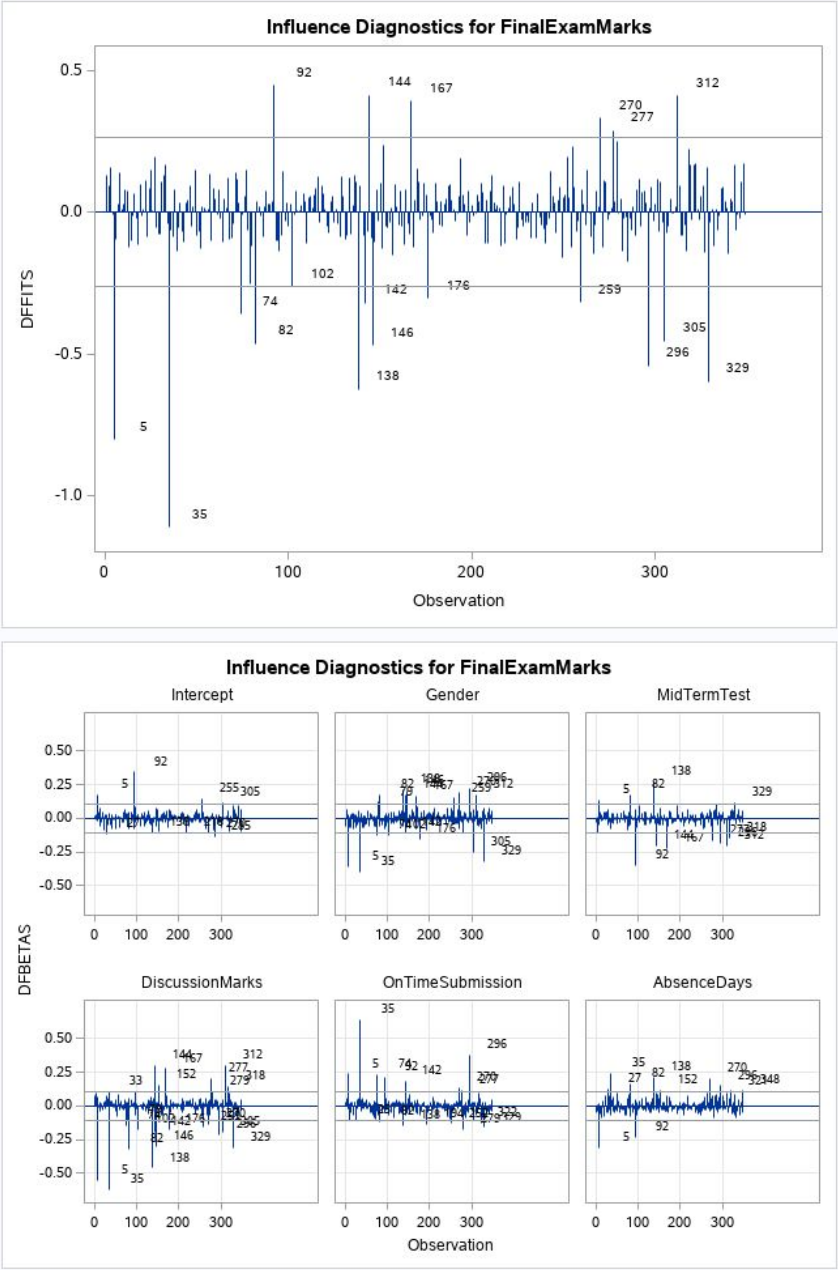

Figure 14. Influence diagnostics plot.

Figure 14 shows that number 35, 5, and others appear once again as influential points based on the values on DFFITS. By looking at the influence diagnostics, there are a total of 18 influential

points for the regression model. Apparently, Number 35 and 5 are influential because of the effects on all the predictors.

Variance Inflation

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	20.77379	1.47020	14.13	<.0001	0
Gender	1	2.70987	0.95925	2.82	0.0050	1.06841
MidTermTest	1	0.19145	0.02097	9.13	<.0001	1.62580
DiscussionMarks	1	0.61851	0.01779	34.76	<.0001	1.24726
OnTimeSubmission	1	3.05574	0.99935	3.06	0.0024	1.23857
AbsenceDays	1	-4.24071	1.02327	-4.14	<.0001	1.28451

Figure 15. Variance Inflation Factor

Based on *Figure 15*, there is no collinearity in the model because the variance inflation for all variables is less than 10. Therefore, we can conclude that no additional changes or removal of variables are made for the regression model.

Conclusion

In this assignment, we can determine different factors that are related to students' academic performance. They should make an effort to take initiatives in academic activities such as have their assignment submitted on time, attend the class with few absence days, do well in the midterm exam, and attain participation in class in order to enhance their academic performance.

We performed stepwise selection up to step 5 with the default significant level. After performing the stepwise selection, none of the variables was removed from the model. Through Regression analysis, there are several influential points that exceed the suggested cut-off values and Cook's D plot and DFFIT have shown two highest influential points in the regression model (the higher the Cook's D is the more influential the point is). Based on the variance inflation test, all of the six variables show no collinearity in the model.

References

1. Shahzadi, Erum & Ahmad, Z.. (2011). A STUDY ON ACADEMIC PERFORMANCE OF UNIVERSITY STUDENTS. 10.13140/2.1.3949.3126.

Appendix A

Meeting Record Template

Date	Attended by	Items Discussed
25/05/2020	All Members	The first meeting was conducted and all the members started discussing and looked through the given dataset.
27/05/2020	All Members	SAS Studio code was shared by Jarrod and everyone reviewed the code together.
14/06/2020	All Members	Our team had discussed several topics such as choosing which graph to use in descriptive analysis.
20/06/2020	All Members	Our team discussed the variable in the dataset and decided which variables are going to be used for our response variable and explanatory variables.
20/06/2020	All Members	Our team had reviewed the Introduction of the assignment. Decision-making to choose between stepwise selection, backward elimination or forward elimination
20/06/2020	All Members	Our team had discussed the regression analysis part and either choosing to do Multiple Linear Regression or Logistic Regression.
21/06/2020	All Members	Our team plotted the diagnostic plots and variance inflation plot for the regression model and interpreted them.
25/06/2020	All Members	Our team concluded the report with unfinished points due to confusion and seeking the lecturer's feedback before proceeding.
01/07/2020	All Members	Our team has attended a consultation with the lecture regarding assignment progression.
03/07/2020	All Members	Our team edited the report and added important information needed based on the feedback given by the lecturer. We also double-checked the entire report for any typos, grammatical errors, etc and fixed them.

Appendix B

Activity Log Record Template

Date	Progress / Task / Activity	Recorded by
27/5/20	Created a Google Docs template that includes necessary information for group members to start the assignment and share our codes and ideas.	Jarrod Tham Kuok Yew
27/5/20	Using SAS Studio: ● Coded the import statements for the CSV file. ● Created a permanent library to store the SAS and CSV files. ● Performed basic data exploration and data cleaning to check if there are any missing values in the dataset. ● Created macros for easier data manipulation.	Jarrod Tham Kuok Yew
14/6/20	Using SAS Enterprise Guide Perform ● Characterize Data ● Summary Statistics ● One-way Analysis	Koh Fu Kang
20/6/20 - 24/6/20	Using SAS Studio: ● Perform Multiple Linear Regression ● Stepwise Selection Model ● Generate Fit Diagnostics Plots ● Check Variance Inflation	Chia Mun Choon
20/6/20 - 24/6/20	Written Interpretations ● Introduction ● Descriptive Analysis ● Plots and Graphs for Descriptive Analysis	Koh Fu Kang and Virox Sim
20/6/20 - 24/6/20	Written Interpretations ● Multiple Linear Regression ● Regression Diagnostics ● Variance Inflation	Chia Mun Choon and Jarrod Tham Kuok Yew
25/06/2020	Written the conclusion of the report	Virox Sim
03/07/2020	Final updates and interpretations based on the lecturer's feedback	All Members

Appendix C

SAS Studio Code

```
/* Generated Code (IMPORT) */

/* Source File: edu-data-2019-3.csv */

/* Source Path: /home/u42888972/BSBA Sem 6 Applied Statistics/AS Assignment */

/* Code generated on: 27/5/2020 */


/*

I start by uploading the edu-data-2019-3.csv file to my folder (AS Assignment, stated at the
Source Path).

Then I created a permanent library called ASLIB20 and path it to the folder .

*/


%web_drop_table(ASLIB20.ASCW20);


FILENAME REFFILE '/home/u42888972/BSBA Sem 6 Applied Statistics/AS
Assignment/edu-data-2019-3.csv';


PROC IMPORT DATAFILE=REFFILE

        DBMS=CSV

        OUT=ASLIB20.ASCW20;
```

```
        GETNAMES=YES;
RUN;

%web_open_table(ASLIB20.ASCW20);

*Created Macro for easier data manipulation;
%let library = ASLIB20;
%let filename = ASCW20;

/** 1. BASIC DATA EXPLORATION & DATA CLEANING/VALIDATION. **/

/*
Using PROC FORMAT AND FREQ to find out which variables have missing data.
(I did not use PROC MEANS because it can only be used to check num variables.)
*/

PROC CONTENTS DATA = ASLIB20.ASCW20;
RUN;

PROC FORMAT;
```

```
        VALUE $missing_char
                ' ' = 'Missing'
                other = 'Present';

        VALUE missing_num
                . = 'Missing'
                other = 'Present';
RUN;



TITLE 'Listing of Present and Missing Data for Each Variable';
PROC FREQ DATA = &library..&filename;
        TABLES _all_ / missing;
        FORMAT _character_ $missing_char. _numeric_ missing_num.;
RUN;
TITLE;


options nolabel;
PROC MEANS DATA = &library..&filename N NMISS MIN MAX MEAN;
RUN;


*Multiple Linear Regression;
```

```
*Model Selection with Stepwise Selection;

proc reg data = &library..&filename;

        model FinalExamMarks = Gender MidtermTest DiscussionMarks OnTimeSubmission
AbsenceDays / selection = stepwise;

run;



proc reg data = &library..&filename;

        model FinalExamMarks = Gender MidtermTest DiscussionMarks OnTimeSubmission
AbsenceDays / VIF;

run;


ods graphics on;

proc reg data = &library..&filename

        plots(only label)=rstudentbypredicted

        plots(only label)=cooksd

        plots(only label)=dffits

        plots(only label)=dfbetas

;

model FinalExamMarks = Gender MidtermTest DiscussionMarks OnTimeSubmission
AbsenceDays;

run;
```