SCHOOL OF SCIENCE AND TECHNOLOGY

ASSIGNMENT / PROJECT SUBMISSION FORM

PROGRAMME: Bachelor of Information Systems (Honours) (Data Analytics)

SEMESTER: Jan / [✓ Mar] / Aug 2020

SUBJECT: IST2024 Applied Statistics

DEADLINE: 3rd July 2020

INSTRUCTIONS TO CANDIDATES

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

IMPORTANT

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.

SHAMALAN RAJESVARAN	**18042945**
NEO JUI JIE	**18025536**
NG WEI XIANG	**18033167**
YAP ZI HAN	**18046524**

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

Jui Jie Xiang Zi Han .... (Signature/Date) **[3 JULY 2020]**

Academic Honesty Acknowledgement

SHAMALAN RAJESVARAN
NEO JUI JIE
NG WEI XIANG
YAP ZI HAN

"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."

Jui Jie Xiang Zi Han .... (Student's signature / Date) **[3 JULY 2020]**

Data Protection

The protection of personal data is an important concern to Sunway University and any personal data collected on this form will be treated in accordance with the Personal Data Protection Notice of the institution.

http://sunway.edu.my/pdpa/notice_english (English version)

http://sunway.edu.my/pdpa/notice_bm (Malay version)

Table of Contents

1. Introduction

Our objective for statistical analysis is as follows :

- To determine how the variables MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission and AbsenceDays have an effect individually and collectively on the response variable FinalExamMarks.
- To determine how each gender performs in each subject with regard to passing and failing.
- To better understand the patterns of the individual variables.

As part of our statistical analysis, we have employed **Multiple Linear Regression** as well as **Logistic Regression**.

For the **Multiple Linear Regression**, we have identified FinalExamMarks as our response variable and MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission, and AbsenceDays as the predictor variables.

Variables MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission and AbsenceDays are chosen as the predictor variable as its statistical significance can be identified using the Multiple Linear Regression analysis.

Variables OnTimeSubmission and AbsenceDays are binary categorical variables. Hence, we are able to find out the individual statistical significance by comparing it to the baseline value. For example, the baseline value for variable OnTimeSubmission is 0 = Assigned work not submitted on time and the baseline value for variable AbsenceDays is 0 = Absent for less than 7 days.

As for the **Logistic Regression**, we have identified FinalExamPass as our response variable and Gender and Subject as our predictor variable.

We have decided to include variables Gender and Subject into the Logistic Regression due to several reasons.

Only Gender and Subject are included in the Logistic Regression because as per our 2nd objective, we are only interested in finding out how the genders perform in each subject with regard to passing and failing the final exam.

The variable Subject is included in the Logistic Regression because it is a multi-levelled categorical variable. Hence, we are unable to identify the variable Subject’s statistical significance in the Multiple Linear Regression.

Overall variables MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission and AbsenceDays are included in the Linear Regression. This is because the variables are measurable and it would depend on the performance of the individual students in order to bring value to its data. Hence, it can be more accurately used to determine its significance with variable FinalExamMarks.

Variables Gender and Subject are included in the Logistic Regression. This is because the variables are general and do not depend on any external factors (ie. performance). Hence, it can be more accurately used to determine its significance with variable FinalExamPass.

Finally, when it comes to understanding the patterns of the individual variables, we want to find out the general pattern of the data and it’s normality.

2. Descriptive Analysis

In order to proceed with our Multiple Linear Regression Analysis, we have replaced the values of the dataset while carefully maintaining its accuracy and integrity. The given data set had the values of Subject initially coded as BM, English, IT, Math and Science.

We had carefully provided each Subject with a numbered representation. The representation is as follows:

BM	- 1
English	- 2
IT	- 3
Math	- 4
Science	- 5

As part of our descriptive analysis for each of the variables, we will be examining the measure of central tendency and measure of dispersion for the continuous variables. As for the categorical variables, we will be representing the data in Pie Charts in order to get a brief idea on its representations.

The **measure of central tendency** refers to a summary statistic that is used to represent the centre point of a dataset (Narkhede, 2018). This analysis would include the mean, mode and median.

Both the mean and median indicates the centre of the data. However, the median is less affected by the outlier as compared to the mean. (Minitab, 2019)

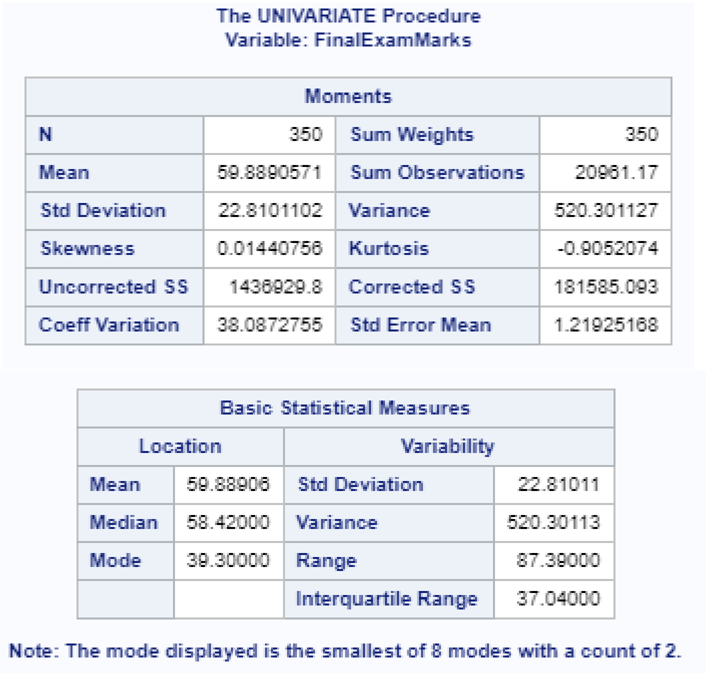
As for the **measure of dispersion**, this measures the variability within the data (Narkhede, 2018). This analysis would include the standard deviation, variance range, interquartile range, skewness, kurtosis as well as the distribution graph.

The range is calculated by taking the largest value and subtracting it by the smallest value. The quartile range, on the other hand, is calculated by subtracting the third quartile value by the first quartile value.

Skewness measures the asymmetry of the probability distribution about its mean. (Narkhede, 2018).

Kurtosis is the measure of whether the data contains an abundance or lack of outliers relative to a normal distribution (Narkhede, 2018).

Response Variable for Linear Regression: FinalExamMarks



The UNIVARIATE Procedure
Variable: FinalExamMarks

Moments			
N	350	Sum Weights	350
Mean	59.8890571	Sum Observations	20961.17
Std Deviation	22.8101102	Variance	520.301127
Skewness	0.01440756	Kurtosis	-0.9052074
Uncorrected SS	1436929.8	Corrected SS	181585.093
Coeff Variation	38.0872755	Std Error Mean	1.21925168

Basic Statistical Measures			
Location		Variability	
Mean	59.88906	Std Deviation	22.81011
Median	58.42000	Variance	520.30113
Mode	39.30000	Range	87.39000
		Interquartile Range	37.04000

Note: The mode displayed is the smallest of 8 modes with a count of 2.

Figure 1. Univariate Analysis of Variable FinalExamMarks

Measure of Central Tendency

- The **mean** is 59.89
  - On average, a student scores 59.8 marks in the final exam.
- The **median** mark is 58.42
- The smallest **mode** is 39.3
  - The value 39.3 has the highest frequency in variable FinalExamMarks.

Measure of Dispersion

- The **standard deviation** is 22.81
- The **variance** is 520.30
- The **range** is 87.39
- The **interquartile range** is 37.04
- The **skewness** value is 0.0144
  - The positive values indicate that it is positively skewed.
  - The skewness value can be said to be relatively low, hence its difference from the normal distribution is relatively low.
- The **Kurtosis** value is -0.9052
  - The negative value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.
  - This is called Platykurtic.

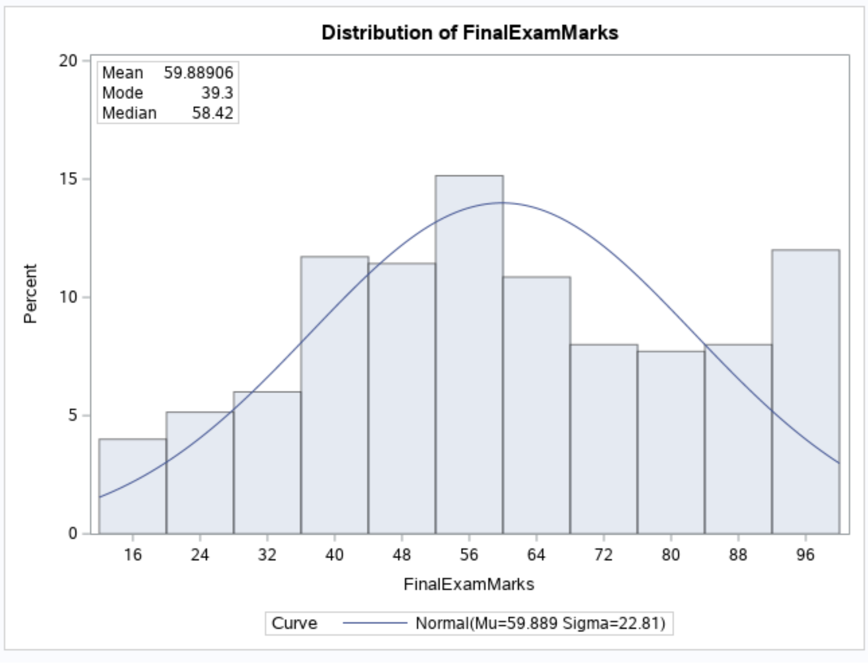
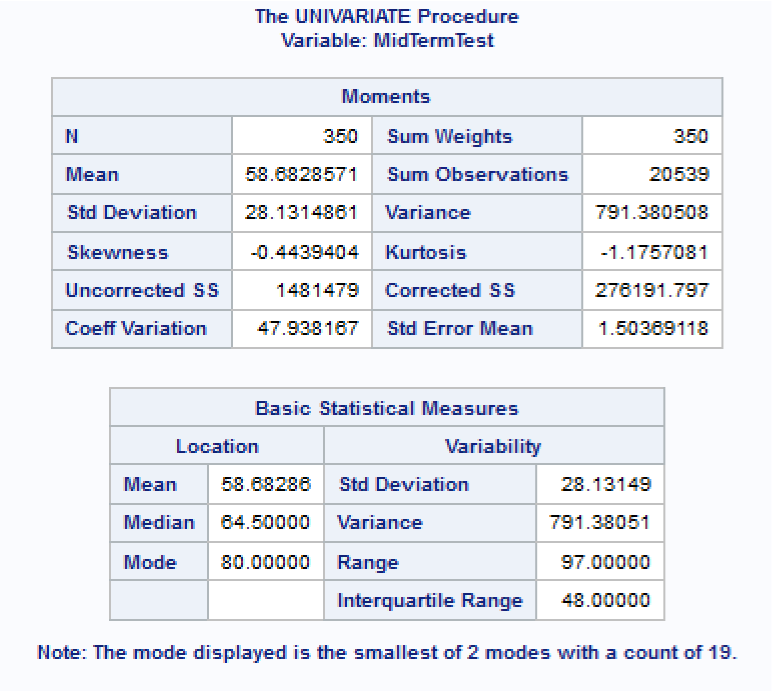


Figure 2. Distribution of FinalExamMarks

Based on the distribution, we can conclude that the low and positive skewness contributes to a slight difference from the normal distribution. Positive skewness is indicated by a distribution that is skewed to the left.
Additionally, the negative kurtosis contributes to a lighter tail and flatter peak than a normal distribution.

Predictor Variable for Linear Regression: MidTermTest



The UNIVARIATE Procedure
Variable: MidTermTest

Moments			
N	350	Sum Weights	350
Mean	58.6828571	Sum Observations	20539
Std Deviation	28.1314861	Variance	791.380508
Skewness	-0.4439404	Kurtosis	-1.1757081
Uncorrected SS	1481479	Corrected SS	276191.797
Coeff Variation	47.938167	Std Error Mean	1.50369118

Basic Statistical Measures			
Location		Variability	
Mean	58.68286	Std Deviation	28.13149
Median	64.50000	Variance	791.38051
Mode	80.00000	Range	97.00000
		Interquartile Range	48.00000

Note: The mode displayed is the smallest of 2 modes with a count of 19.

Figure 3. Univariate Analysis of Variable MidTermTest

Measure of Central Tendency

- The **mean** is 58.68
  - On average, a student scores 58.68 marks in their midterm test.
- The **median** mark is 64.5
- The smallest **mode** is 80
  - The value 80 has the highest frequency in the MidTermTest variable.

Measure of Dispersion

- The **standard deviation** is 28.13
- The **variance** is 791.38
- The **range** is 97
- The **interquartile range** is 48
- The **skewness** value is -0.4439
  - The negative values indicate that it is negatively skewed.
  - The skewness value can be said to be moderate, hence its difference from the normal distribution is moderate.
- The **Kurtosis** value is -1.1757
  - The negative value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.
  - This is called Platykurtic.

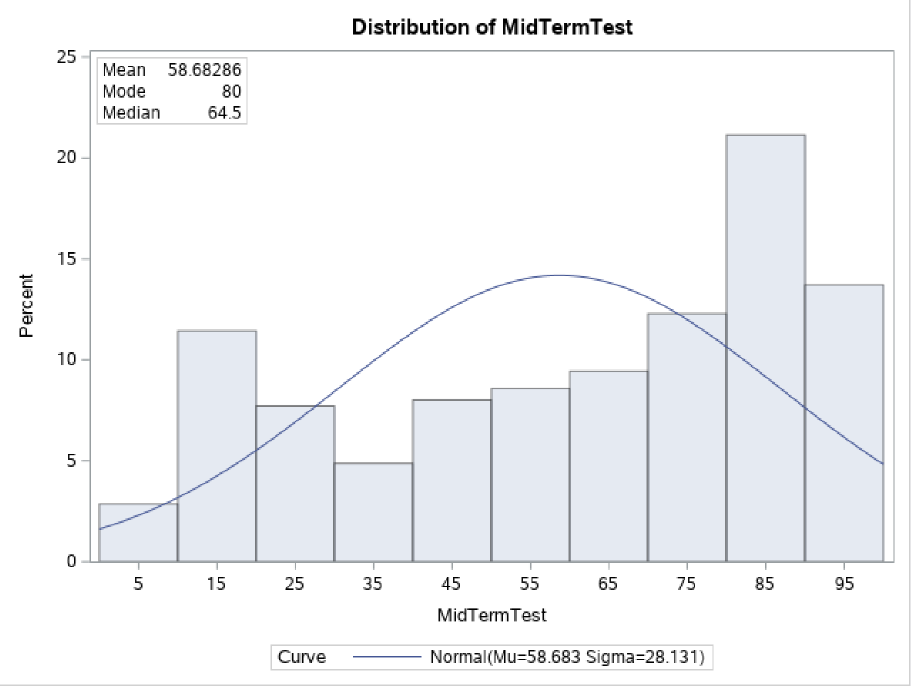
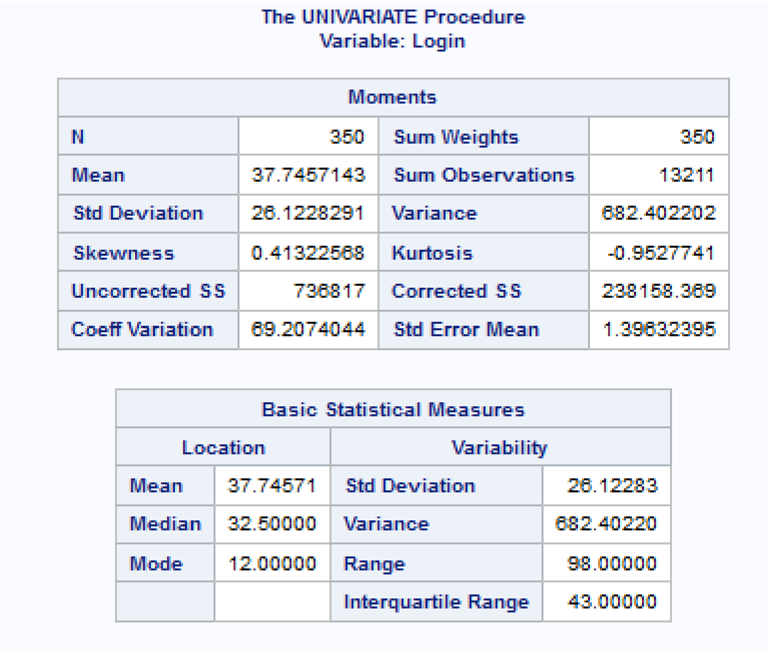


Figure 4. Distribution of MidTermTest

Based on the distribution, we can conclude that the moderate and negative skewness contributes to a slight difference from the normal distribution. Negative skewness is indicated by a distribution that is skewed to the right.
Additionally, the negative kurtosis contributes to a lighter tail and flatter peak than a normal distribution.

Predictor Variable for Linear Regression: LogIn



The UNIVARIATE Procedure
Variable: Login

Moments			
N	350	Sum Weights	350
Mean	37.7457143	Sum Observations	13211
Std Deviation	26.1228291	Variance	682.402202
Skewness	0.41322568	Kurtosis	-0.9527741
Uncorrected SS	736817	Corrected SS	238158.369
Coeff Variation	69.2074044	Std Error Mean	1.39632395

Basic Statistical Measures			
Location		Variability	
Mean	37.74571	Std Deviation	26.12283
Median	32.50000	Variance	682.40220
Mode	12.00000	Range	98.00000
		Interquartile Range	43.00000

Figure 5. Univariate Analysis of Variable LogIn

Measure of Central Tendency

- The **mean** is 37.75
  - On average, a student logs in 38 (rounded up from 37.75) times into the learning management system.
- The **median** mark is 32.5
- The smallest **mode** is 12
  - The value 12 has the highest frequency in the LogIn variable.

Measure of Dispersion

- The **standard deviation** is 26.12
- The **variance** is 682.4
- The **range** is 98
- The **interquartile range** is 43
- The **skewness** value is 0.4132
  - The positive values indicate that it is positively skewed.
  - The skewness value can be said to be moderate, hence its difference from the normal distribution is moderate.
- The **Kurtosis** value is -0.9528
  - The negative value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.
  - This is called Platykurtic.

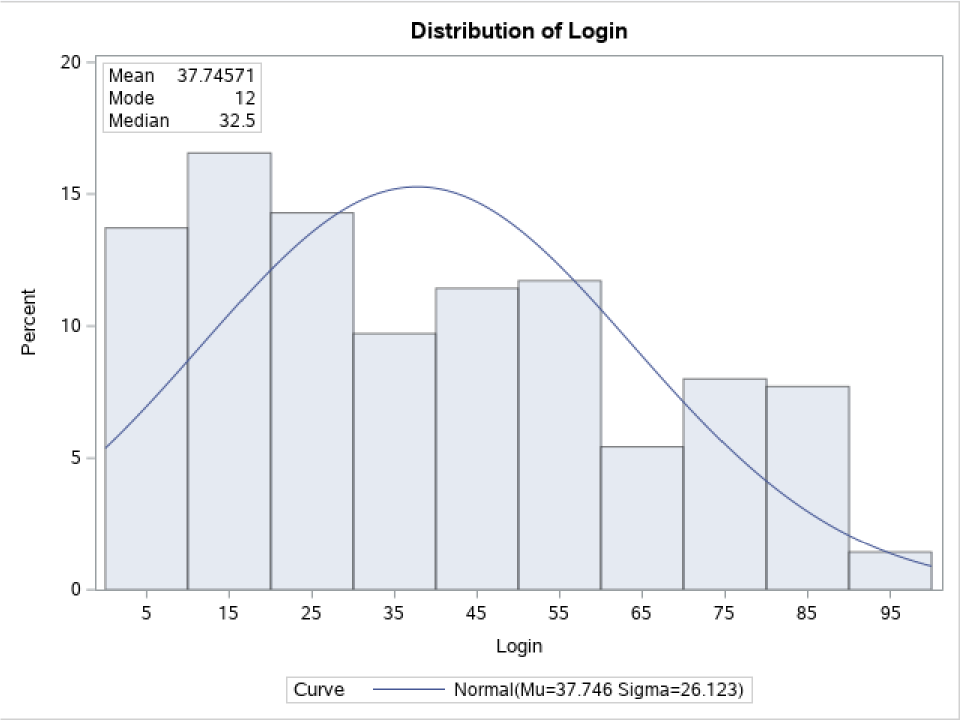
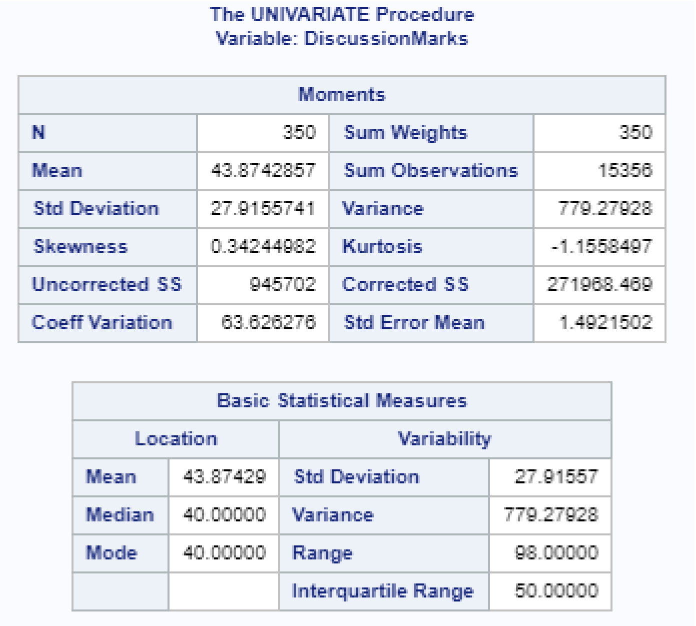


Figure 6. Distribution of LogIn

Based on the distribution, we can conclude that moderate and positive skewness contributes to a slight difference from the normal distribution. Positive skewness is indicated by a distribution that is skewed to the left.

Additionally, the negative kurtosis contributes to a lighter tail and flatter peak than a normal distribution.

Predictor Variable for Linear Regression: DiscussionMarks



The UNIVARIATE Procedure
Variable: DiscussionMarks

Moments			
N	350	Sum Weights	350
Mean	43.8742857	Sum Observations	15356
Std Deviation	27.9155741	Variance	779.27928
Skewness	0.34244982	Kurtosis	-1.1558497
Uncorrected SS	945702	Corrected SS	271968.469
Coeff Variation	63.626276	Std Error Mean	1.4921502

Basic Statistical Measures			
Location		Variability	
Mean	43.87429	Std Deviation	27.91557
Median	40.00000	Variance	779.27928
Mode	40.00000	Range	98.00000
		Interquartile Range	50.00000

Figure 7. Univariate Analysis of Variable DiscussionMarks

<u>Measure of Central Tendency</u>

- The **mean** is 43.87
  - On average, a student scores 43.87 participation marks in class.
- The **median** mark is 40
- The smallest **mode** is 40
  - The value 40 has the highest frequency in the DiscussionMarks variable.

<u>Measure of Dispersion</u>

- The **standard deviation** is 27.92
- The **variance** is 779.28
- The **range** is 98
- The **interquartile range** is 50
- The **skewness** value is 0.3424
  - The positive values indicate that it is positively skewed.
  - The skewness value can be said to be low, hence its difference from the normal distribution is low.
- The **Kurtosis** value is -1.1558
  - The negative value indicates that the distribution has lighter tails and a flatter peak than the normal distribution.
  - This is called Platykurtic.

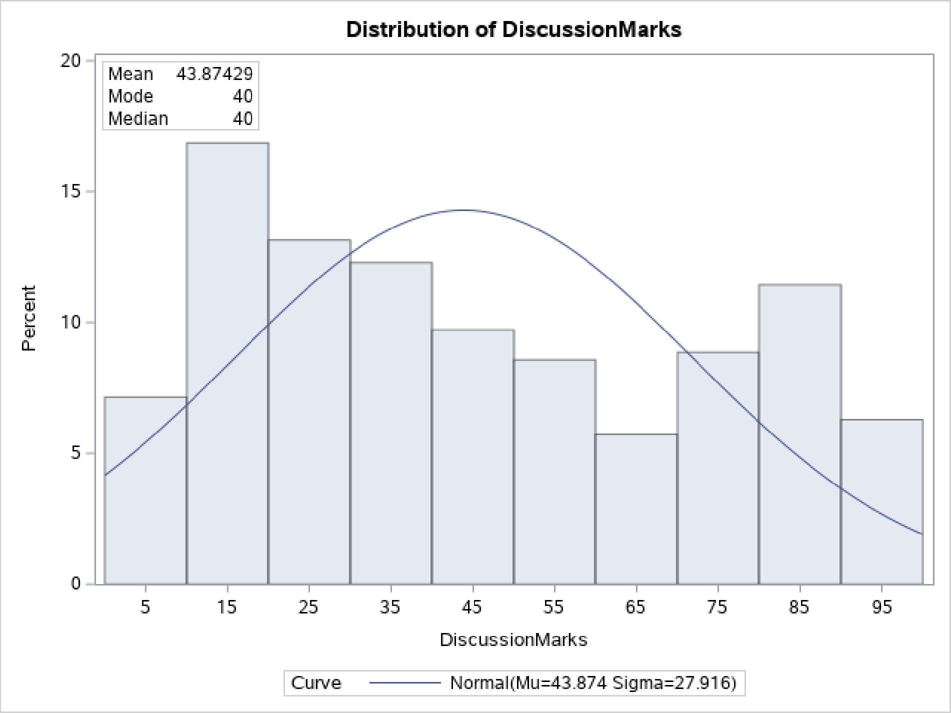
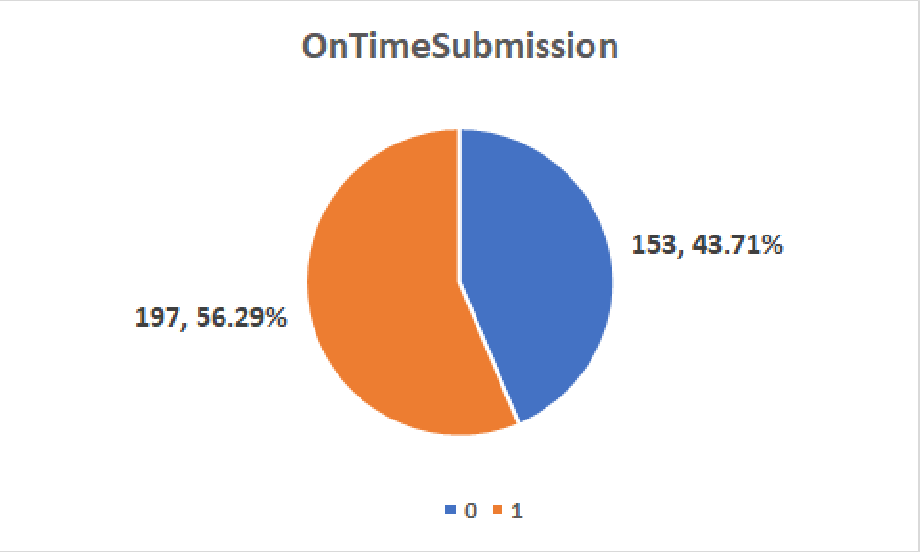


Figure 8. Distribution of DiscussionMarks

Based on the distribution, we can conclude that the low and positive skewness contributes to a slight difference from the normal distribution. Positive skewness is indicated by a distribution that is skewed to the left.

Additionally, the negative kurtosis contributes to a lighter tail and flatter peak than a normal distribution.

Predictor Variable for Linear Regression: OnTimeSubmission

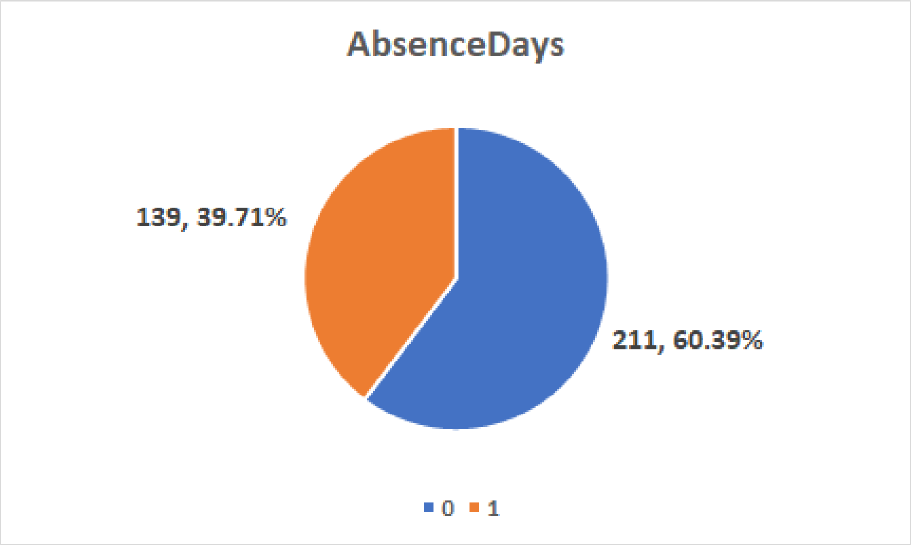


1 = Assigned work submitted on time; 0 = Assigned work not submitted on time

Figure 9. Pie Chart representation of Variable OnTimeSubmission

A pie chart was generated for variable ‘OnTimeSubmission’.
It is seen that a total of 197 (56.29%) of the students submitted their assigned work on time.
It is also found that 153 or 43.71% of the students did not submit their assigned work on time.

Predictor Variable for Linear Regression: AbsenceDays

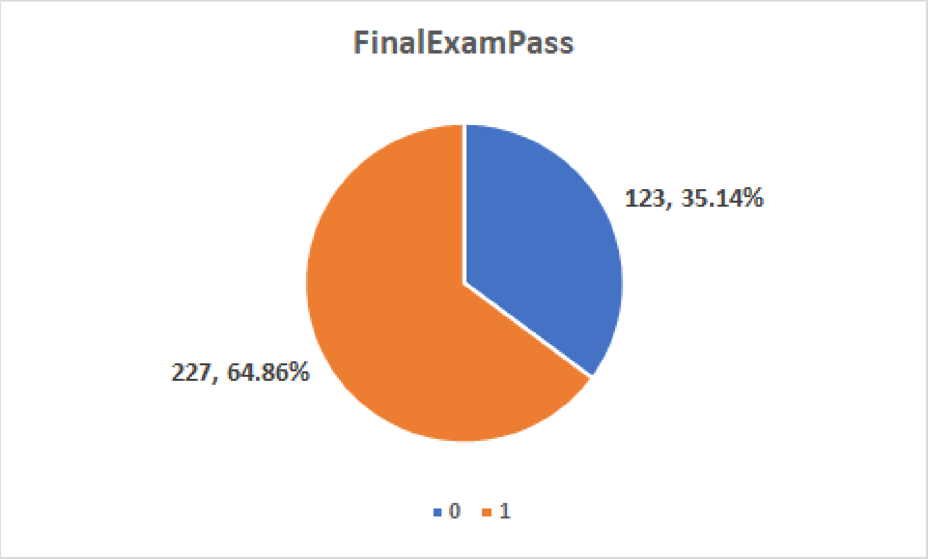


1 = Absent for 7 days or more; 0 = Absent for less than 7 days

Figure 10. Pie Chart representation of Variable AbsenceDays

A pie chart was generated for variable ‘AbsenceDays’.
It is seen that a total of 137 (39.71%) of the students were absent for 7 days or more.
Additionally, it is found that 211 or 60.39% of the students were absent for less than 7 days.

Response Variable for Logistic Regression: FinalExamPass

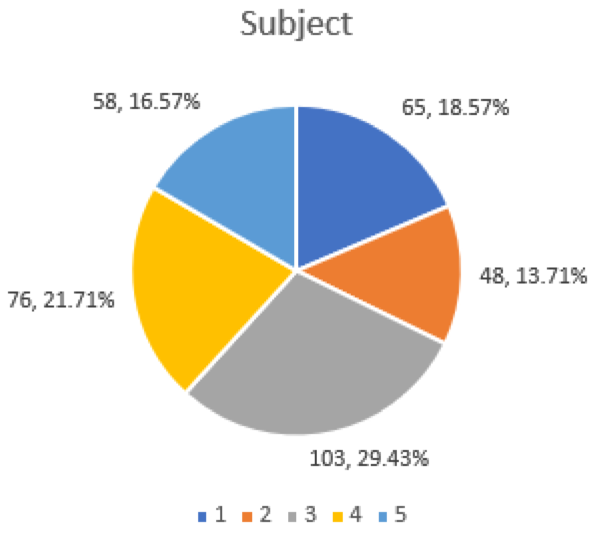


1 = Passed Final Exam; 0 = Failed Final Exam

Figure 11. Pie Chart representation of Variable FinalExamPass

A pie chart was generated for variable ‘OnTimeSubmission’.
It is seen that a total of 123 or 39.71% of the students passed the final exam.
Additionally, it is found that 227or 60.39% of the students failed the final exam.

Predictor Variable for Logistic Regression: Subject

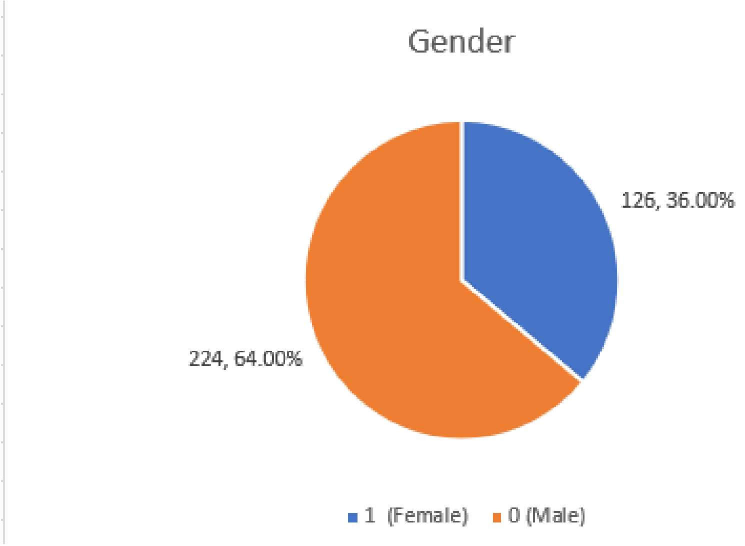


1 = BM, 2 = English, 3 = IT, 4 = Math, 5 = Science

Figure 12. Pie Chart representation of Variable Subject

A pie chart was generated for a descriptive analysis of the variable “Subject”.
The top 2 subjects with the highest number of enrollment are IT and Math. It is seen that 103 or 29.34 % of the students took IT and 76 or 21.65% of the students took Math.
Next, the subjects with a moderate number of enrollment are BM and Science. A total of 65 (18.25%) of the students enrolled in BM while 58 or 16.57% of students enrolled in Science.
Finally, 49 or 14% of the students took the English subject.

Predictor Variable for Logistic Regression: Gender



1 = Female; 0 = Male

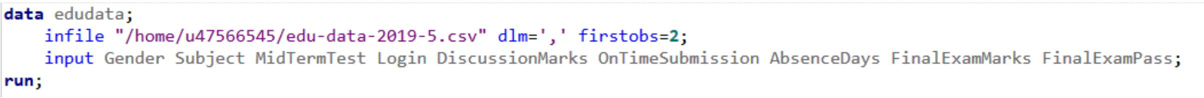
Figure 13. Pie Chart representation of Variable FinalExamPass

A pie chart was generated for the variable ‘Gender’.
It is seen that a total of 126 or 36.00% of the students are female.
Additionally, it is found that 224 or 64.00% of the students are male.

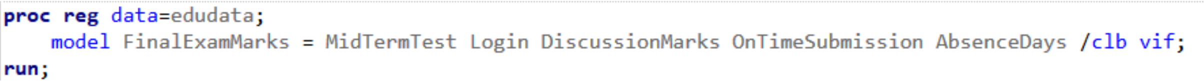
3. Regression Analysis

Multiple Linear Regression

CODE :

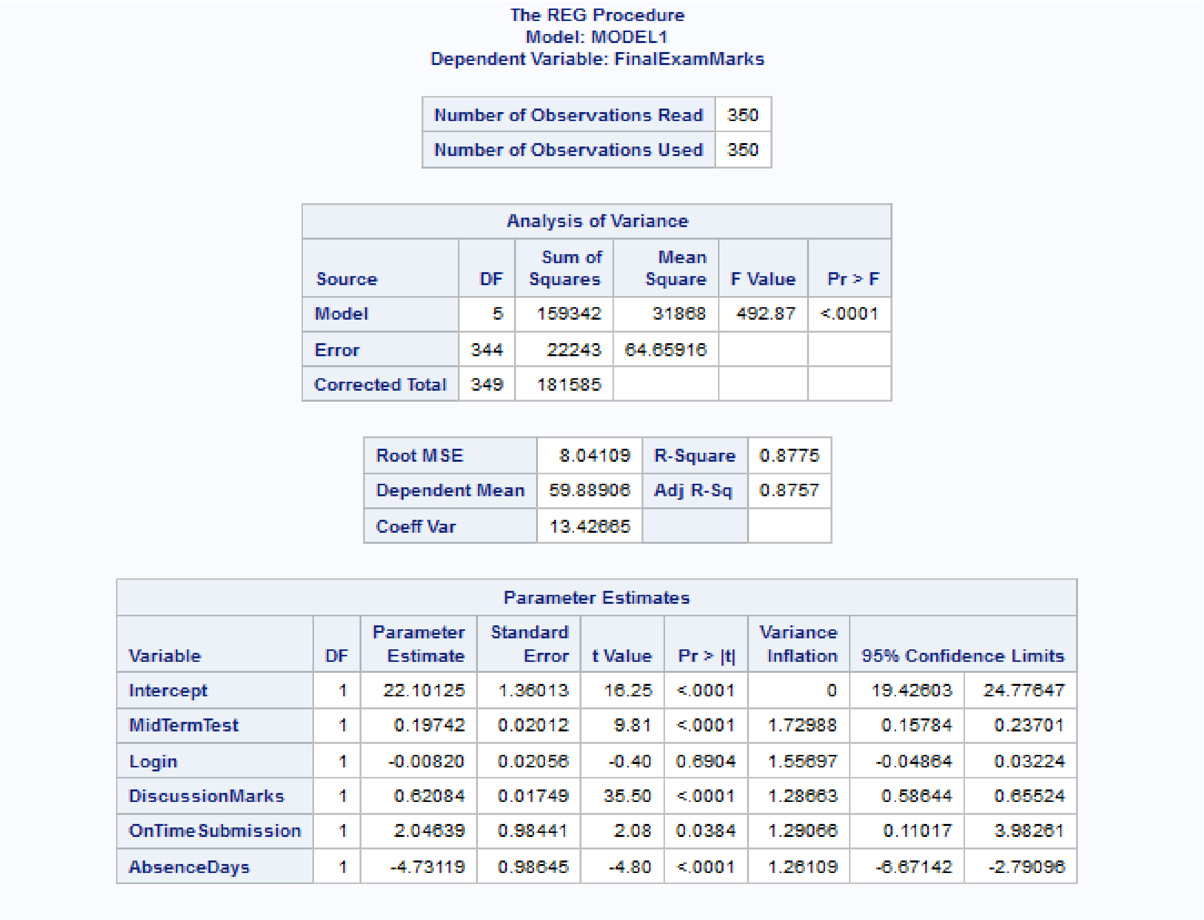


```
data edudata;
    infile "/home/u47566545/edu-data-2019-5.csv" dlm=',' firstobs=2;
    input Gender Subject MidTermTest Login DiscussionMarks OnTimeSubmission AbsenceDays FinalExamMarks FinalExamPass;
run;
```



```
proc reg data=edudata;
    model FinalExamMarks = MidTermTest Login DiscussionMarks OnTimeSubmission AbsenceDays /clb vif;
run;
```

OUTPUT:



The REG Procedure
Model: MODEL1
Dependent Variable: FinalExamMarks

Number of Observations Read	350
Number of Observations Used	350

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	159342	31868	492.87	<.0001
Error	344	22243	64.65916		
Corrected Total	349	181585			

Root MSE	8.04109	R-Square	0.8775
Dependent Mean	59.88906	Adj R-Sq	0.8757
Coeff Var	13.42665		

Parameter Estimates								
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation	95% Confidence Limits	
Intercept	1	22.10125	1.36013	16.25	<.0001	0	19.42603	24.77647
MidTermTest	1	0.19742	0.02012	9.81	<.0001	1.72988	0.15784	0.23701
Login	1	-0.00820	0.02056	-0.40	0.6904	1.55697	-0.04864	0.03224
DiscussionMarks	1	0.62084	0.01749	35.50	<.0001	1.28663	0.58644	0.65524
OnTimeSubmission	1	2.04639	0.98441	2.08	0.0384	1.29066	0.11017	3.98261
AbsenceDays	1	-4.73119	0.98645	-4.80	<.0001	1.26109	-6.67142	-2.79096

Figure 14. SAS Studio Output

Based on the output, the linear regression equation is

$y = 22.1013 + 0.1974x_1 - 0.0082x_2 + 0.6208x_3 + 2.0464x_4 - 4.7312x_5$.

Based on this plot, we can determine that the **sample mean** is 59.89.

Next, the **Root MSE** is the standard deviation of the error term, which is 8.04.

After that, we observed the **coefficient of variation** is 13.43 where the residuals are defined to be the root mean square error divided by the mean of the dependent variable. The ratio of the standard deviation to the mean indicates that the dispersion from the distribution is 13.43%.

The **R**^{2} is recorded as 0.8775. In other words, 87.75% of the variation in FinalExamMarks is explained by the variation in the independent variables.

The **adjusted value of R**^{2} is 0.8757 which means that 87.57% of the variation in final exam marks is explained by the multiple regression model, adjusted for the number of independent variables and sample size.

To test if the independent variables collectively have a statistically significant effect on the response variable, **FinalExamMarks**.

From the output, ***F*** = 492.87 with corresponding **p-value** < 0.0001.

Since the **p-value** is < 0.0001, thus H_0 is rejected at the level of significance ($\alpha = 0.05$). We can say that at least one of the explanatory variables have a significant effect on the response variable.

Since the **p-value** is < 0.05, H_0 of β_1, β_3, β_4, and β_5 is rejected at significance level $\alpha = 0.05$. There is strong evidence that **Gender**, **MidTermTest**, **DiscussionMarks**, **OnTimeSubmission**, and **AbsenceDays** is related to the **FinalExamMarks**.

Since the **p-value** is > 0.05, H_0 of β_2 is accepted at significance level $\alpha = 0.05$. There is no evidence that **LogIn** is related to the **FinalExamMarks**.

Residual Plot

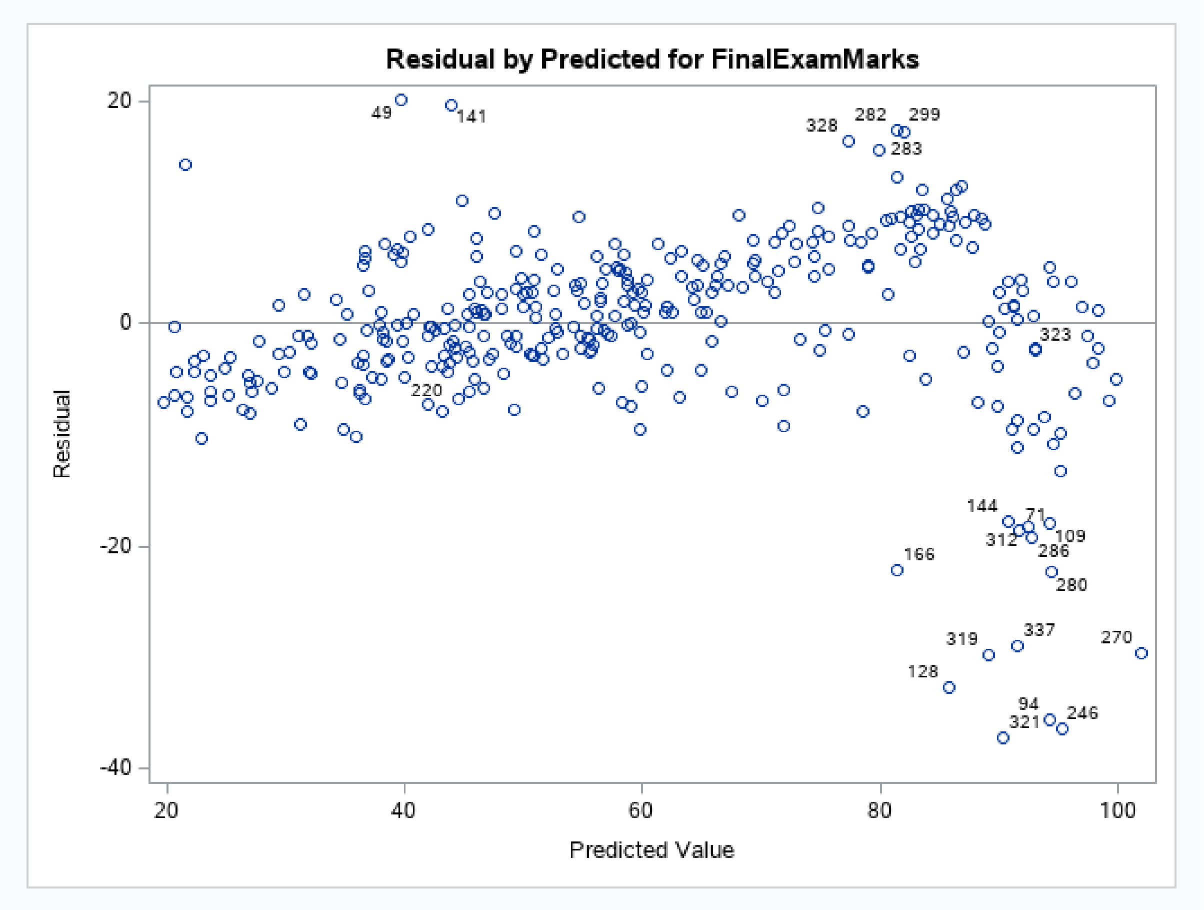


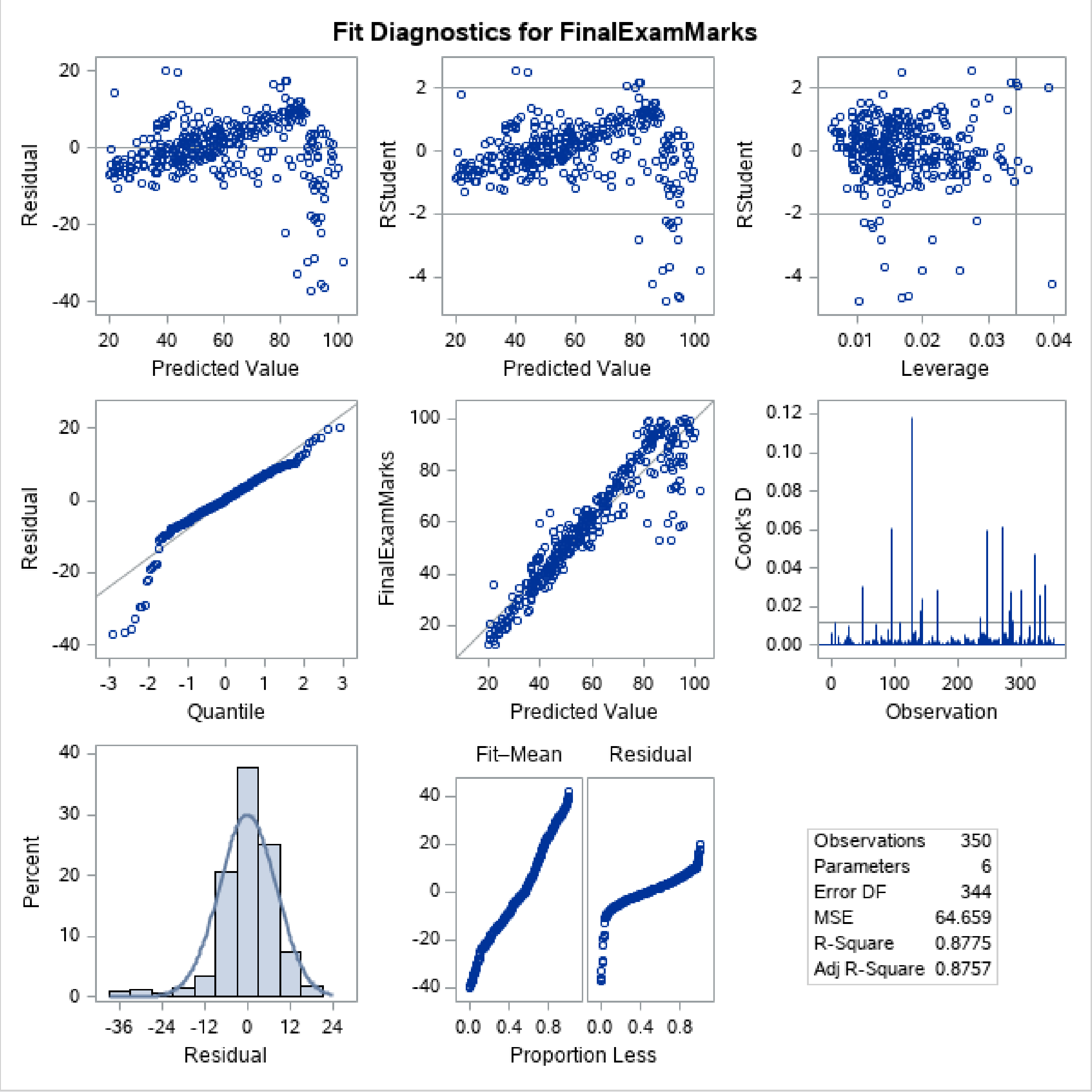
Figure 15. The plot of residuals versus predicted values

The plot of residuals shows that residuals seem to be grouped together and increase in the lower predicted values.

In the greater predicted values, the residuals seem to disperse and then move towards a negative slope.

Hence, there is no independence of residual errors. This means that we can assume for the linear regression, there is no constant variance across all levels of all independent variables.

Fit Diagnostic



Observations	350
Parameters	6
Error DF	344
MSE	64.659
R-Square	0.8775
Adj R-Square	0.8757

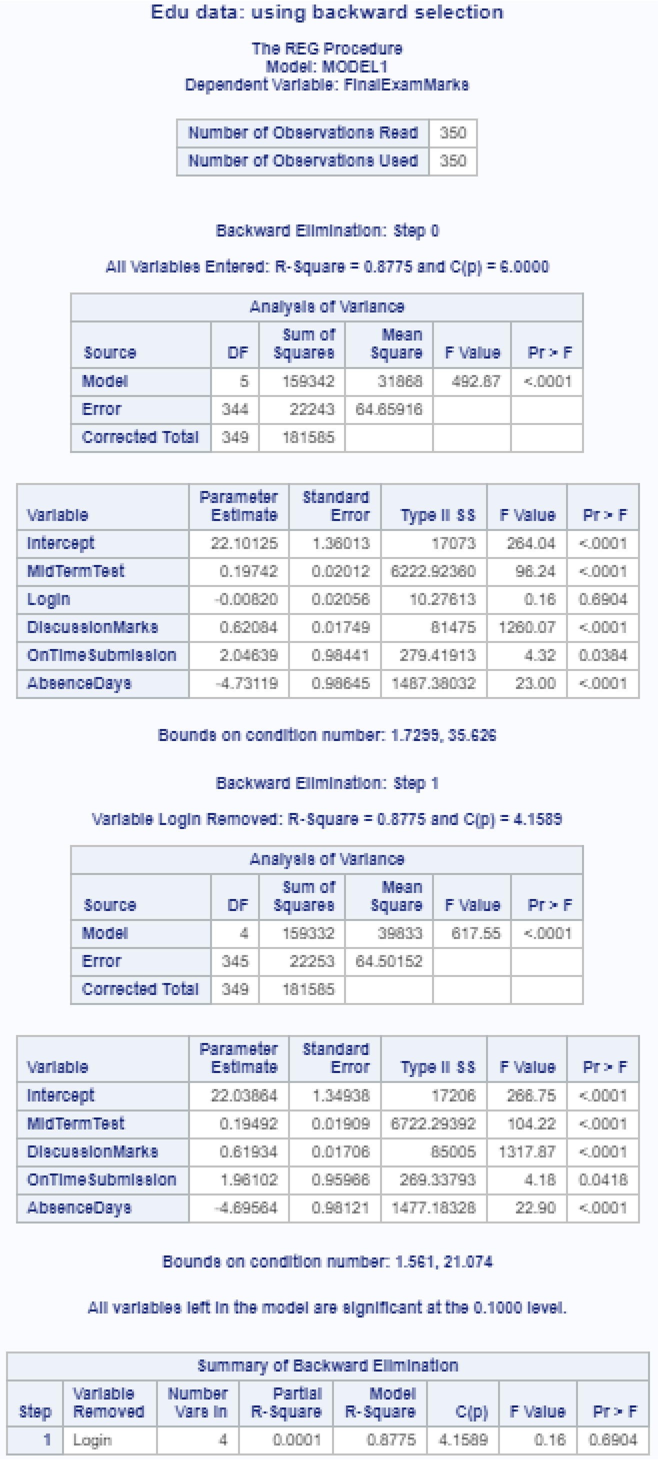
Figure 16. Fit Diagnostics

When we look at the Quantile – Quantile (Q-Q) plot, it does not appear to follow a straight diagonal line which signals that assumption of normality is not met.

To further prove it, the histogram also did not show a symmetrical bell shape.

Hence, it is proven that the **assumption of normality is not met**.

Selection Method: Backward Selection Method



Edu data: using backward selection

The REG Procedure
Model: MODEL1
Dependent Variable: FinalExamMarks

Number of Observations Read	350
Number of Observations Used	350

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8775 and C(p) = 6.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	159342	31868	492.87	<.0001
Error	344	22243	64.65916		
Corrected Total	349	181585			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	22.10125	1.36013	17073	264.04	<.0001
MidTermTest	0.19742	0.02012	6222.92360	96.24	<.0001
Login	-0.00820	0.02056	10.27613	0.16	0.6904
DiscussionMarks	0.62084	0.01749	81475	1260.07	<.0001
OnTimeSubmission	2.04639	0.98441	279.41913	4.32	0.0384
AbsenceDays	-4.73119	0.98645	1487.38032	23.00	<.0001

Bounds on condition number: 1.7299, 35.626

Backward Elimination: Step 1

Variable Login Removed: R-Square = 0.8775 and C(p) = 4.1589

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	159332	39833	617.55	<.0001
Error	345	22253	64.50152		
Corrected Total	349	181585			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	22.03864	1.34938	17206	266.75	<.0001
MidTermTest	0.19492	0.01909	6722.29392	104.22	<.0001
DiscussionMarks	0.61934	0.01706	85005	1317.87	<.0001
OnTimeSubmission	1.96102	0.95966	269.33793	4.18	0.0418
AbsenceDays	-4.69564	0.98121	1477.18328	22.90	<.0001

Bounds on condition number: 1.561, 21.074

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	Login	4	0.0001	0.8775	4.1589	0.16	0.6904

Figure 17. Backward Selection

We loaded in a total of **5 predictor variables** into the model for selection at the default significance level of 0.1.

At step 0, we can see that the predictor variable **LogIn** has a p-value of **0.6904** which is much greater than the default significance level at 0.1.

Thus, at step 1, we can see that the variable "**LogIn**" is removed from the model. After that, we do not see any other variables that have a p-value of more than 0.1.

Hence, we can conclude the remaining 4 variables (**MidTermTest**, **DiscussionMarks**, **OnTimeSubmission**, and **AbsenceDays**) are the significant variables for the model.

Studentized Analysis

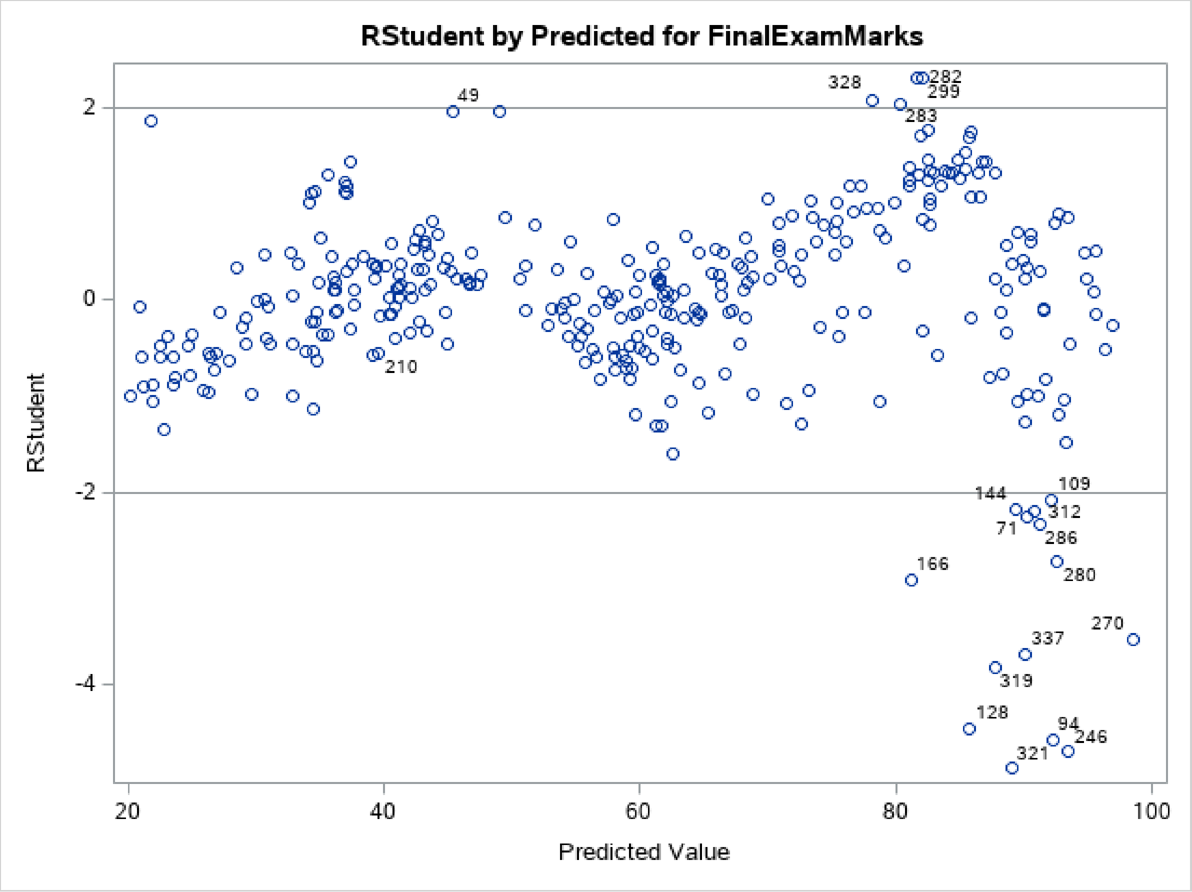


Figure 18. Studentized Analysis

There is a **pattern** on the Residual plot. The points from the left to the middle are closer whereas the points on the right are further apart moving downwards.

Based on this, the variance is smaller when the Predicted values are small whereas the variance is larger as the Predicted values get bigger.

Constant Variance assumption may not be met. There are multiple outliers. The plot of the residuals versus the values of the independent variables, **MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission, AbsenceDays** is shown above.

Yes, there are outliers as indicated by the evidence in the Residual by Predicted. They are indicated by the plots outside of the boundary set.

Cook’s D

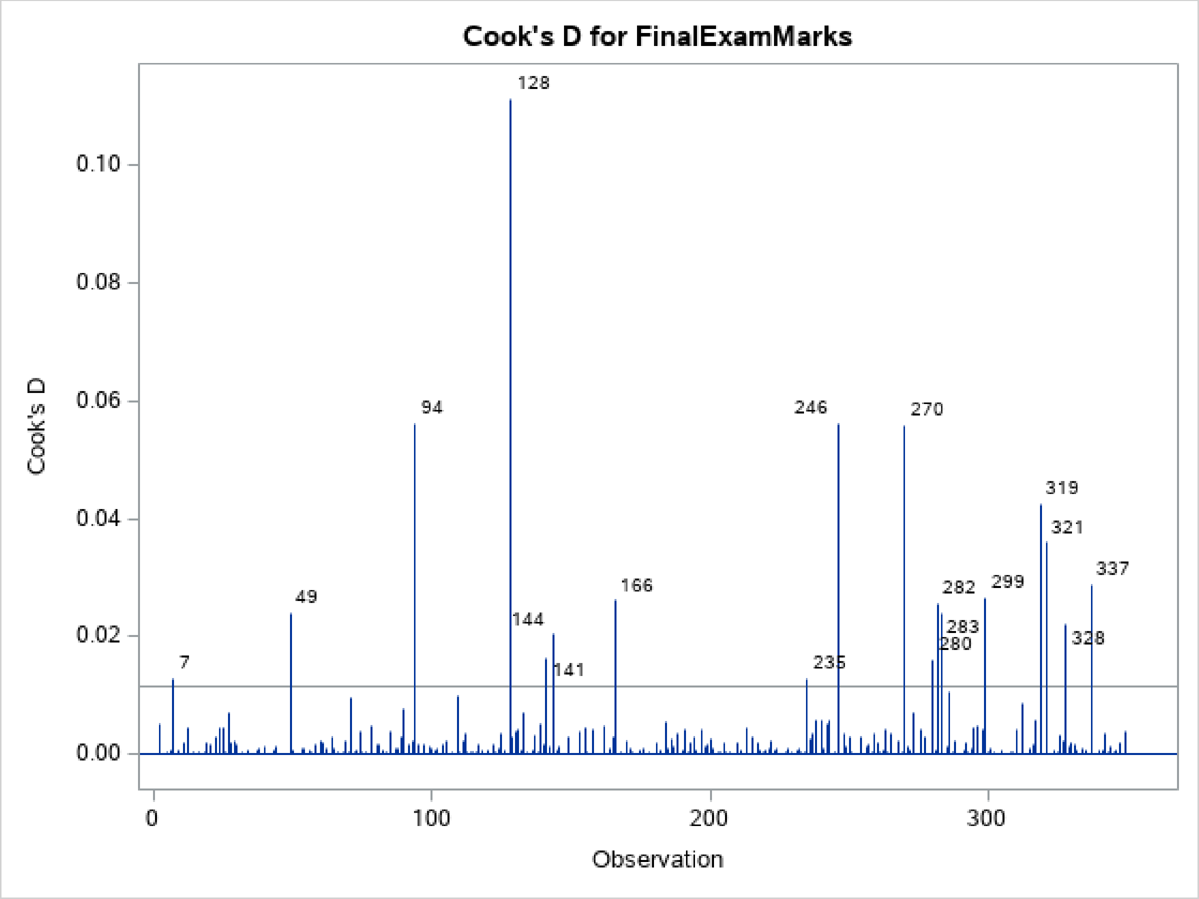


Figure 19. Cook’s D Plot

Cook’s D Plot is used to estimate the influence of a data point when performing a least-squares regression analysis. The Cook’s D Plot shows **Observation 128** to be the main influential point. There are numerous other influential points which have crossed the boundary.

Outlier and Leverage Diagnostic

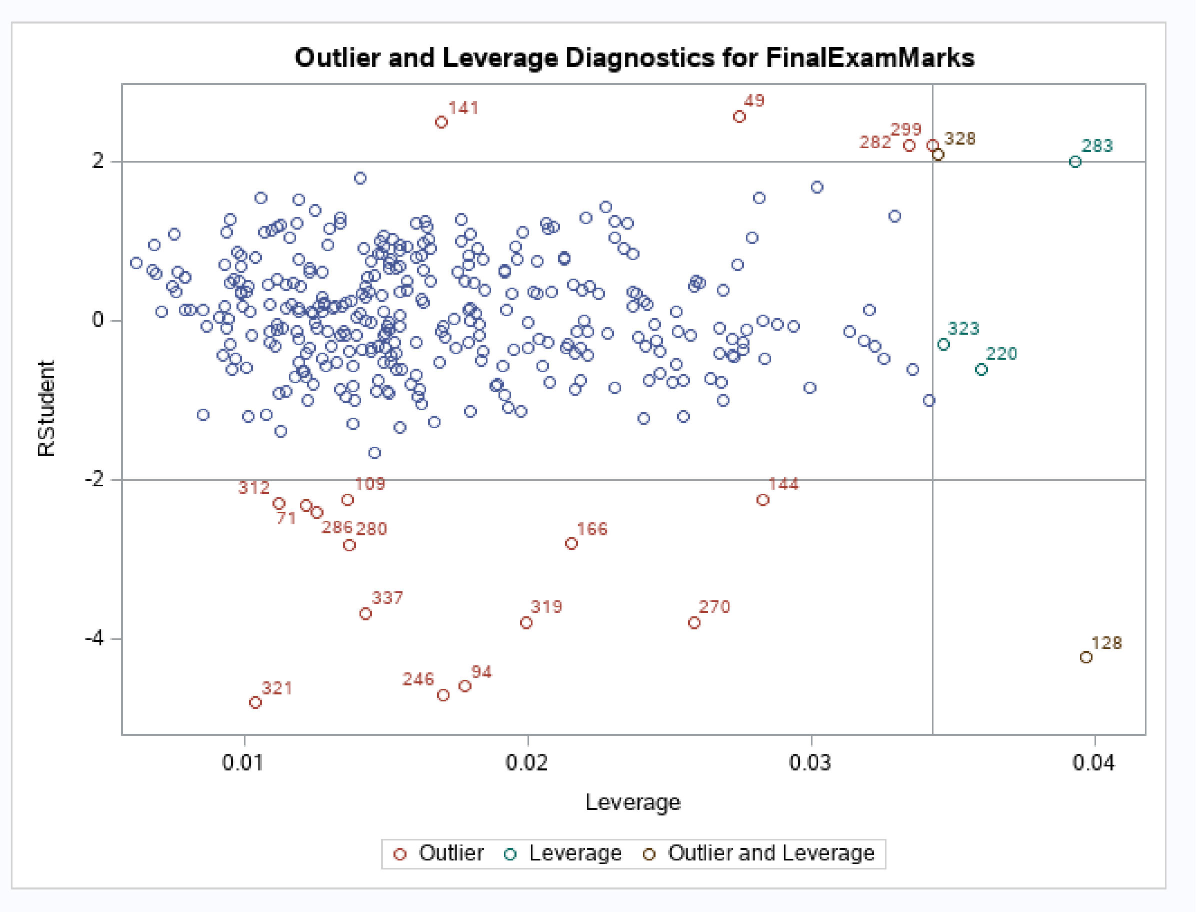


Figure 20. Outlier and Leverage

There are outliers and leverage points as indicated by the outlier and leverage diagnostics. They are indicated by the plots outside of the boundary set.

Observations 128, 328 are shown to be both an outlier and leverage.

DFFITS

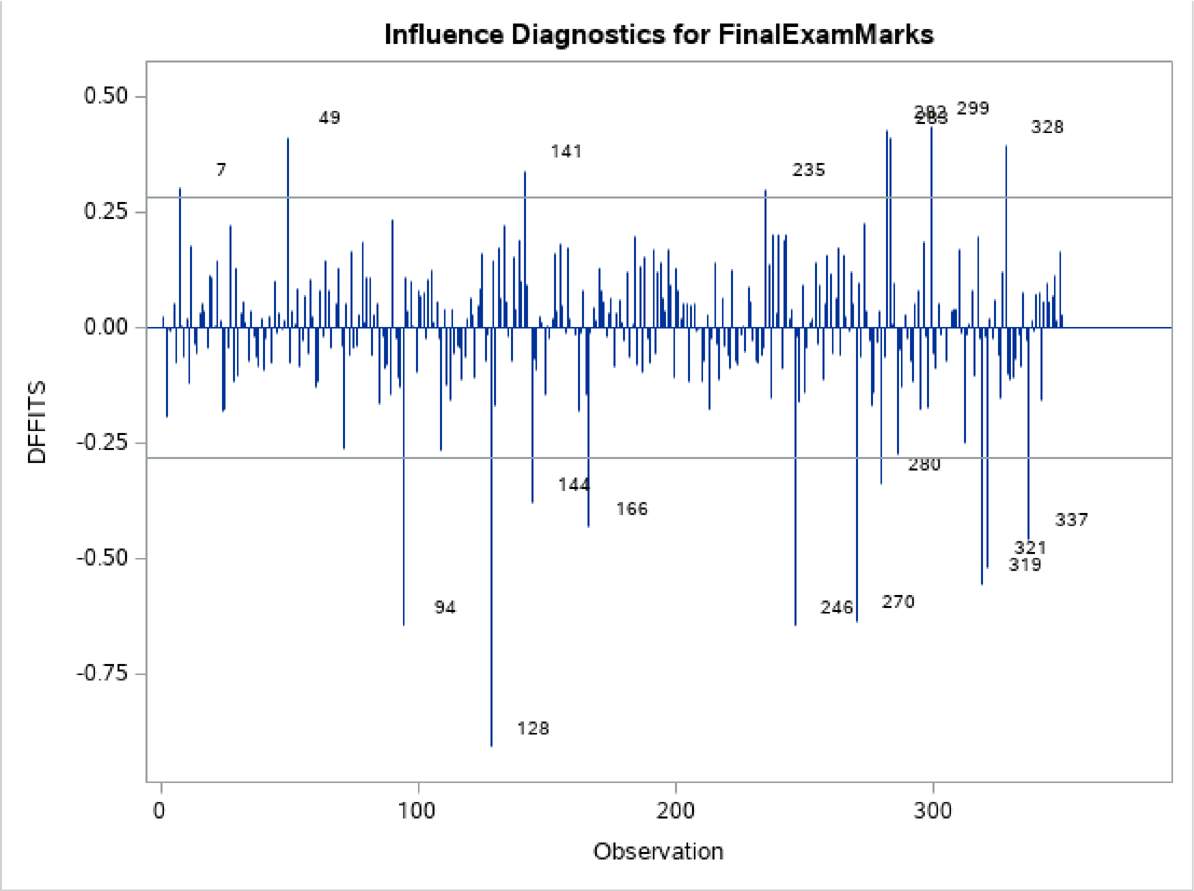


Figure 21. Influence Diagnostics (DFFITS)

DFFITS test indicates how influential a point is in a statistical regression analysis.

According to DFFITS, **Observation 128** is shown to be the most influential point in the test.

DFBETAS

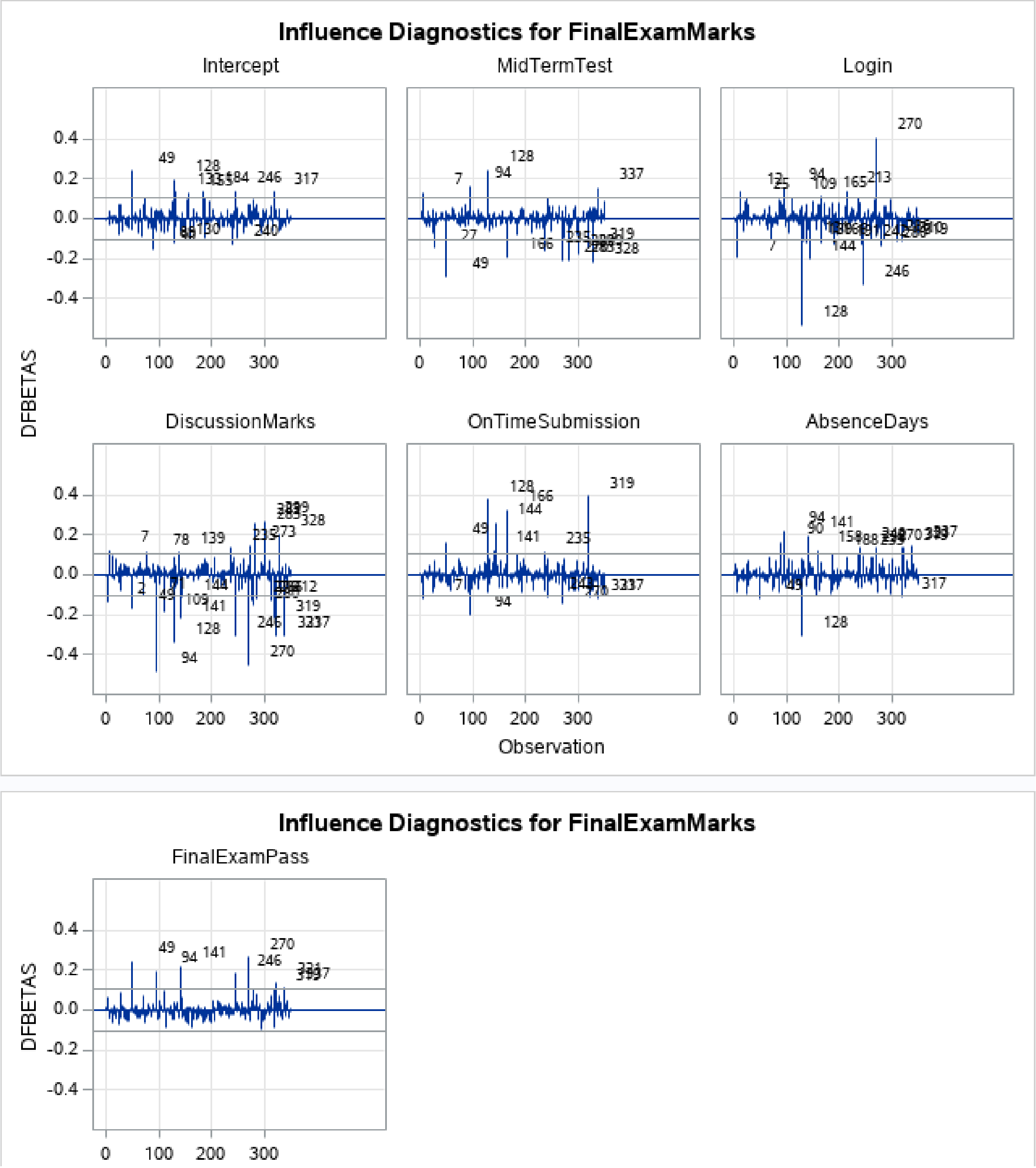


Figure 22. Influence Diagnostics (DFBETAS)

Observation 128 is shown to be an influential point because it affects most of the response variables and the independent variables, **MidTermTest, LogIn, DiscussionMarks, OnTimeSubmission, AbsenceDays**. It also appears in all the variables that influence diagnostic plots.

Test for Homoscedasticity

With reference to Figure 15 Residual plot, it is shown to exhibit heteroscedasticity.

Heteroscedasticity refers to residuals for a regression model that does not have a constant variance.

The residuals are close and increasing in the lower predicted values but as the predicted values slowly increase the residuals take a downward turn.

Residuals at the lower predicted values have a lower variance.

Residuals at the higher predicted values have a higher variance.

Hence, there is no constant variance of the residuals.

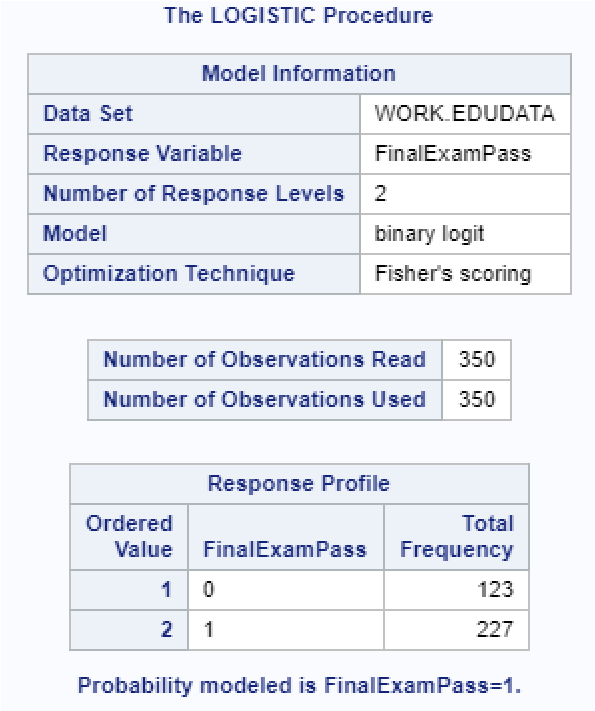
Collinearity Diagnostics

Collinearity refers to the strong linear correlation between two or more predictors in the model.

With reference to the SAS output in Figure 14, none of the variance inflation values is larger than 10.

There is no collinearity problem with any variables.

Logistic Regression



The LOGISTIC Procedure

Model Information	
Data Set	WORK.EDUDATA
Response Variable	FinalExamPass
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	350
Number of Observations Used	350

Response Profile		
Ordered Value	FinalExamPass	Total Frequency
1	0	123
2	1	227

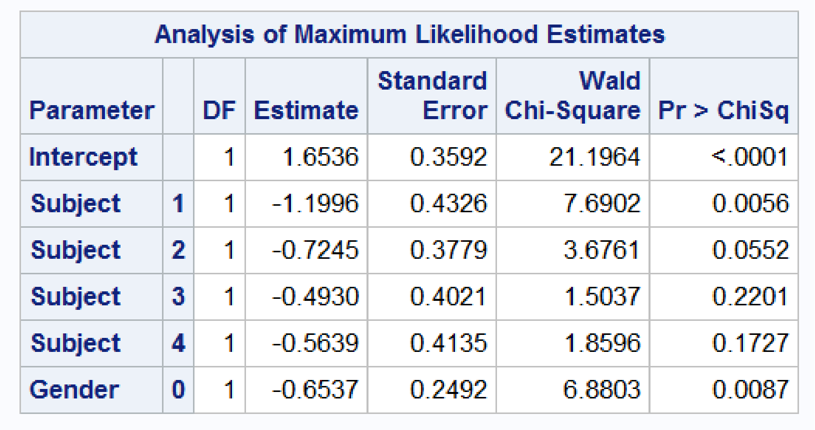
Probability modeled is FinalExamPass=1.

Figure 23. Model Information and Response Profile Table

The Model Information Table describes the Logistic Regression process. Its description also includes the number of response level. In this case, variable FinalExamPass is set as the Response Variable. The variable has 2 levels of values that are 1 = Passed final exam and 0 = Failed final exam.

The Response Profile Table indicates the total frequency of the 2 values of variable FinalExamPass.

It also indicates that the probability is being modelled to the **students passing their final exam**.



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.6536	0.3592	21.1964	<.0001
Subject	1	1	-1.1996	0.4326	7.6902	0.0056
Subject	2	1	-0.7245	0.3779	3.6761	0.0552
Subject	3	1	-0.4930	0.4021	1.5037	0.2201
Subject	4	1	-0.5639	0.4135	1.8596	0.1727
Gender	0	1	-0.6537	0.2492	6.8803	0.0087

Figure 24. Analysis of Maximum Likelihood Estimates

Logistic Regression Equation:

Logit ($\hat{\pi}$)=$\beta_0+\beta_1*X_{Subject1}+\beta_2*X_{Subject2}+\beta_3*X_{Subject3}+\beta_4*X_{Math}+\beta_5*X_{Gender0}$

Based on the information provided in the Maximum Likelihood Estimates table, we are able to produce a sample logistic regression.

Logit($\hat{\pi}$) = $1.6536 - 1.1996*X_{Subject1} - 0.7245*X_{Subject2} - 0.4930*X_{Subject3} - 0.5639*X_{Subject4} - 0.6537*X_{Gender0}$

We have employed reference cell coding. Hence, each variable would be measured against a reference level.

In this case, SAS Studio takes the last category with the highest value. For example, Gender is represented as 0 and 1. SAS would automatically take 1 (Female) as the reference level.

Gender | 0 (Male) shows the difference in logits between the 0 (Male) and the 1 (Female).

Subject | 1 (BM) shows the difference in probability of logistic regression models between 1 (BM) and 5 (Science)

Subject | 2 (English) shows the difference in probability of logistic regression models between 2 (BM) and 5 (Science)

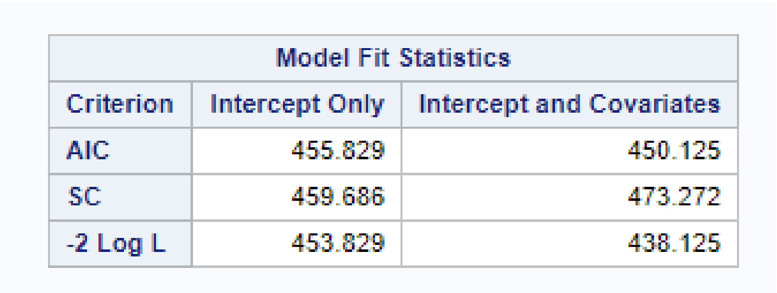
Subject | 3 (IT) shows the difference in probability of logistic regression models between 3 (IT) and 5 (Science)

Subject | 4 (Math) shows the difference in probability of logistic regression models between 4 (Math) and 5 (Science)

Based on the output, we can observe that the p-value of Subject 1 (BM) v Subject 5 (Science) and Gender 0 (Male) and Gender 1 (Female) is lesser than 0.05.

Hence, at a 0.05 significance level, we can conclude that **Subject 1 v Subject 5 and Gender 0 v Gender 1 are statistically significant in the model**.

Model Fit Statistics



Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	455.829	450.125
SC	459.686	473.272
-2 Log L	453.829	438.125

Figure 25. Model Fit Statistics

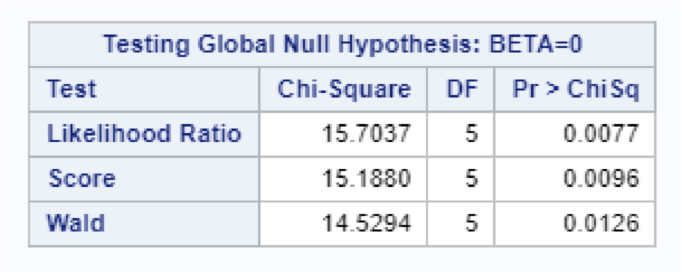
The goodness of fit measures is measured by comparing the difference between Intercept Only and Intercept and Covariates.

Based on the output result, criterion -2 Log L has an Intercept only value of 453.828 whereas the intercept and covariates have a value of 438.125.

The model has a good fit as there is a difference greater than 5 between the intercept only and intercept and covariates.

Thus, the model fit statistics indicate that **Gender** and **Subjects** as predictor variables gives a better fit than an empty model.

Test For Collective Significance



Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.7037	5	0.0077
Score	15.1880	5	0.0096
Wald	14.5294	5	0.0126

Figure 26. Testing Global Null Hypothesis Table

This table is used to identify the collective significance of the predictor variables in the model.

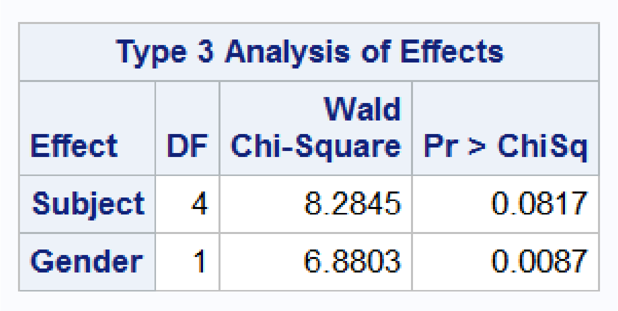
H_0 : All the regression coefficients are 0.

H_1 : At least one of the regression coefficients is not 0.

Based on the output, all 3 tests - Likelihood Ratio, Score and Wald has a p-value of <0.05.

At the 0.05 significance level, H_0 is rejected. Hence, we can conclude that the predictor variables in this logistic regression model are collectively significant.

Type 3 Analysis of Effect



Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Subject	4	8.2845	0.0817
Gender	1	6.8803	0.0087

Figure 27. Type 3 Analysis of Effects

The type 3 analysis of effects table is generated when a predictor variable is defined as a classification variable (**Gender**, **Subject**).

H_0 : $\beta_j = 0$

H_1 : $\beta_j \neq 0$

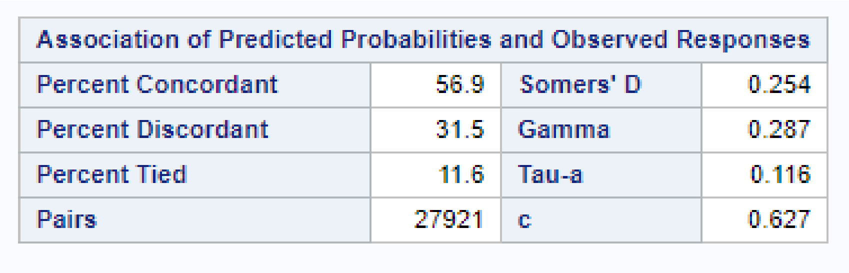
Variable **Subject** has a p-value greater than 0.05. Hence, H_0 is not rejected.

At a 0.05 significance level, it can be concluded that Subject is not statistically significant in this model.

Variable **Gender** has a p-value lesser than 0.05. Hence, H_0 is rejected.

At a 0.05 significance level, it can be concluded that Gender is statistically significant in this model.

Concordance Statistic Value

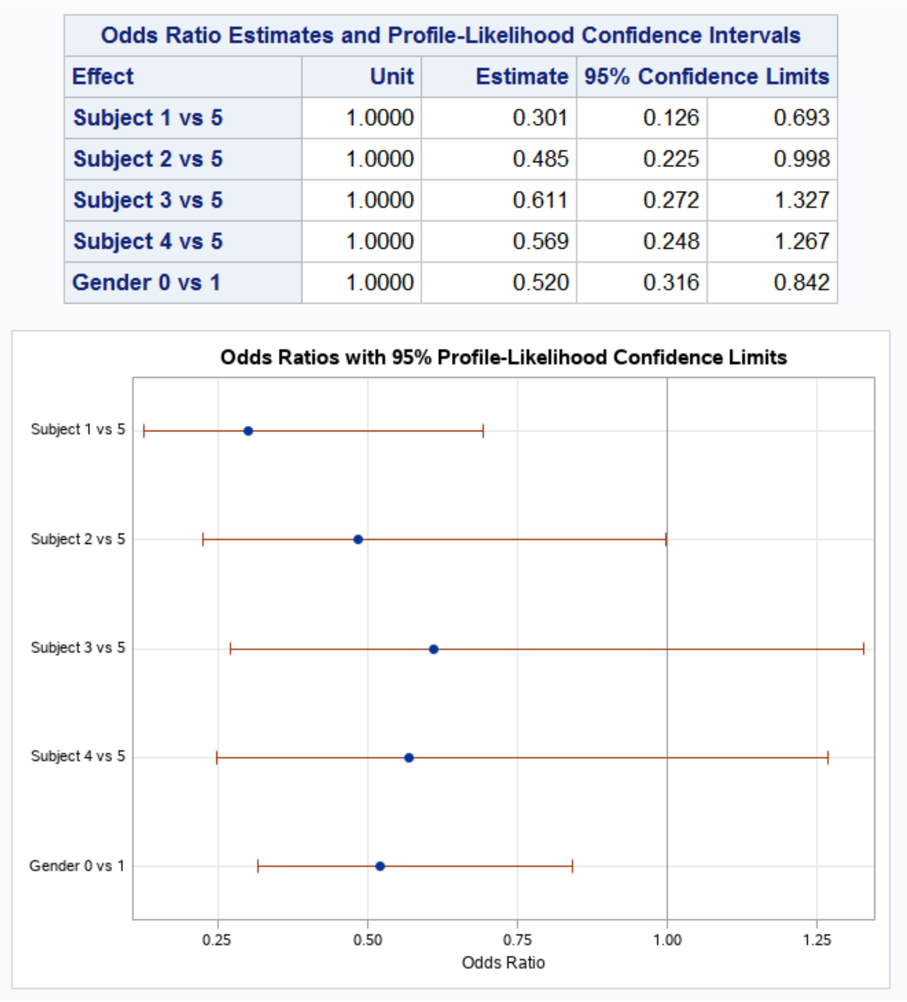


Association of Predicted Probabilities and Observed Responses			
Percent Concordant	56.9	Somers' D	0.254
Percent Discordant	31.5	Gamma	0.287
Percent Tied	11.6	Tau-a	0.116
Pairs	27921	c	0.627

Figure 28. Association of Predicted Probabilities

The c (concordance) statistic value is 0.627 for this model, indicating that the model can correctly classify the outcome at a percentage of 62.70%.

Odds Ratio Plot



Odds Ratio Estimates and Profile-Likelihood Confidence Intervals				
Effect	Unit	Estimate	95% Confidence Limits	
Subject 1 vs 5	1.0000	0.301	0.126	0.693
Subject 2 vs 5	1.0000	0.485	0.225	0.998
Subject 3 vs 5	1.0000	0.611	0.272	1.327
Subject 4 vs 5	1.0000	0.569	0.248	1.267
Gender 0 vs 1	1.0000	0.520	0.316	0.842

Figure 29. Odds Ratio Estimates and Odds Ratio Plot

At a 95% confidence level, Subject 1(BM) v Subject 5(Science) and Gender 0(Male) v Gender 1(Female) is significant as it does not cross the reference line.

We are 95% confident that the effect of the odds ratio of Subject 1 v Subject 5 is between 0.126 and 0.693. Additionally we are also 95% confident that the effect of the odds ratio of Gender 0 v Gender 1 is between 0.316 and 0.842.

Effects Plot

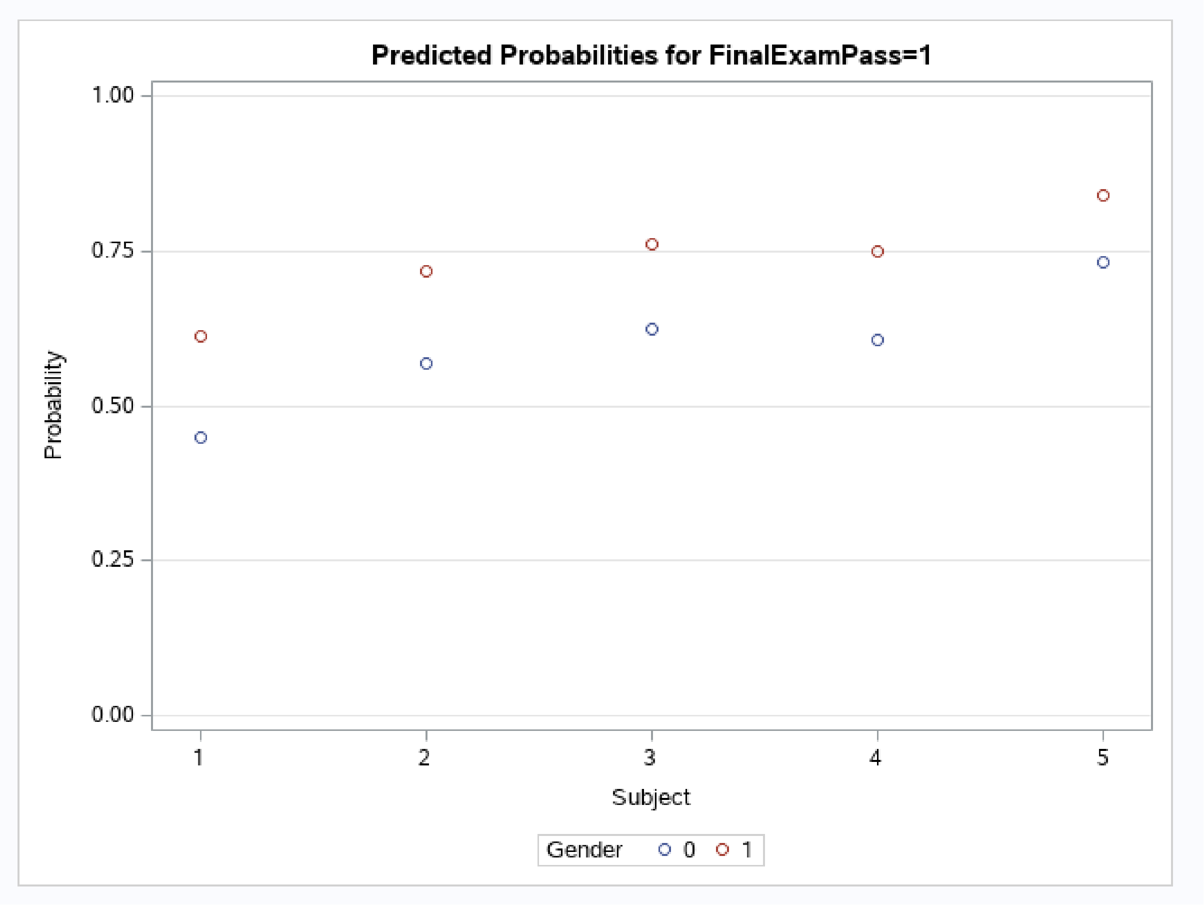


Figure 30. Effects Plot

The effects plot above shows the probability of **FinalExamPass** = 1 (passing Final Exam) across all the different combinations of categories and levels of predictor variables.

The plot shows that the order of the subjects from the lowest probability of passing to the highest probability of passing is the same across male and female. For example, the subjects from the lowest probability of passing to the highest probability of passing for Gender 0 (Male) is Subject 1, Subject 2, Subject 4, Subject 3 and Subject 5. The same order can be observed when we look at the data for Gender 1 (Female).

It is shown that Subject 5 has the highest probability of passing while Subject 1 has the lowest probability of passing across both genders.

Conclusion

In conclusion, we carried out both Multiple Linear Regression and Logistic Regression.

We have identified whether the specific predictor variable has a significant effect on the response variable.

In order to improve the statistical quality of our analysis, we have included two different types of regression analysis methods.

In the Multiple Linear Regression, we have examined which predictor variable has a significant effect on the response variable - FinalExamMarks. The analysis included full selection fitted model, backward selection, regression diagnostic to identify influential points, test for homoscedasticity and collinearity test.

We have concluded that variables MidTermTets, DiscussionMarks, OnTimeSubmission and AbsenceDays are statistically significant in predicting the students' final exam marks (FinalExamMarks) while variable LogIn is not.

In the Logistic Regression, we have identified how gender performs in each subject with regard to passing and failing. This allows us to understand better the passing and failings of each specific subject. This analysis can best be achieved using Logistic Regression. As part of our statistical analysis, we included model fit statistics, type 3 analysis of effect, odds ratio plot, concordance statistic value and effects plot. With the output from this analysis, we are able to identify and compare the probability of female and male in passing and failing the specific subjects.

We have concluded that variable Gender is statistically significant in predicting whether the student passes or fails the exam (FinalExamPass) while variable Subject is not.

References

(2019, June 15). Retrieved from Minitab: https://support.minitab.com/en-us/minitab-express/1/help-and-how-to/basic-statistics/summary-statistics/descriptive-statistics/interpret-the-results/all-statistics-and-graphs/

Naird, W. (2019, May 5). Retrieved from Statalogy: https://www.statology.org/how-to-interpret-the-c-statistic-of-a-logistic-regression-model/#:~:text=The%20c%2Dstatistic%2C%20also%20known,0.5%20indicates%20a%20poor%20model.&text=The%20closer%20the%20value%20is,is%20at%20correctly%20classifying%20outcomes.

Narkhede, S. (2018, June 6). *medium.com*. Retrieved from TowardsDataScience: https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291

Appendix

The meeting record template is as follows:

Date	Attended By	Items Discussion
8/6/2020	• Shamalan Rajesvaran • Ng Wei Xiang • Neo Jui Jie • Yap Zi Han	• Discussion on the response variable and predictor variable for a linear regression analysis on the dataset. • We interpreted the output results from the regression models fitted. • We delegated the work equally amongst all team members.
14/6/2020	• Shamalan Rajesvaran • Ng Wei Xiang • Neo Jui Jie • Yap Zi Han	• Clarification of the analysis of each of our work. • Discussion on the output results of the Linear Regression.
17/6/2020	• Shamalan Rajesvaran • Ng Wei Xiang • Neo Jui Jie • Yap Zi Han	• Discussion on the items to include in Descriptive analysis. • Discussion on the explanation of categorical variables. • Further discussion on the possibility of including Logistic Regression. We decided to make a few tweaks to our initial plan.
1/7/2020	• Shamalan Rajesvaran • Ng Wei Xiang • Neo Jui Jie • Yap Zi Han	• Discussion about the overall output. We established a clear objective line as well as an introduction. • Compilation and final check on the interpretation of the output result.
2/7/2020	• Shamalan Rajesvaran • Ng Wei Xiang • Neo Jui Jie • Yap Zi Han	• Finalization of our report so that it is ready for submission