UTM

UNIVERSITI TEKNOLOGI MALAYSIA

University of Technology, Malaysia

Faculty of Engineering

School of Computing

Semester 2019-2020 2

SECI2143-02

Probability Statistical Data Analysis

Project 2 Report

“Student Grade Data Prediction”

Lecturer:

Dr. Chan Weng Howe

Name:

Mohammad Safwan Bin Azhar

(A19EC0191)

Submission Date:

27/6/2020

Abstract

Every year, teachers and instructors in schools always do their best to predict and find solutions on how they can improve their students’ grades. In this study,an evaluation of different types of data analysis for factors affecting secondary school students’ final grade using several functions in R Studio.However, the selection of types of data analysis is dependent on the quality of data that made it almost impossible to easily choose the models. In addition, most current models are hard for educators or teachers to predict their students’ grade. The result from this evaluation can be used by teachers to make further planning about syllabus in schools, identify students potential and boost evidence in between the instructors. The analysis of this data finally brings to a conclusion that the final grades of students can be related to certain aspects like their previous grades, attendance and their study time.

Table Of Contents

1.0 Introduction

1.1 Introduction to R

R is a free, open-source software for statistical computing with RStudio as an integrated development environment (IDE) for R. It was founded in 1995 at the University of Auckland as an environment for graphics and computing statistics. Since then, R has become the most used software environment in terms of statistical data analysis where a lot of statistical functions can be found in R.

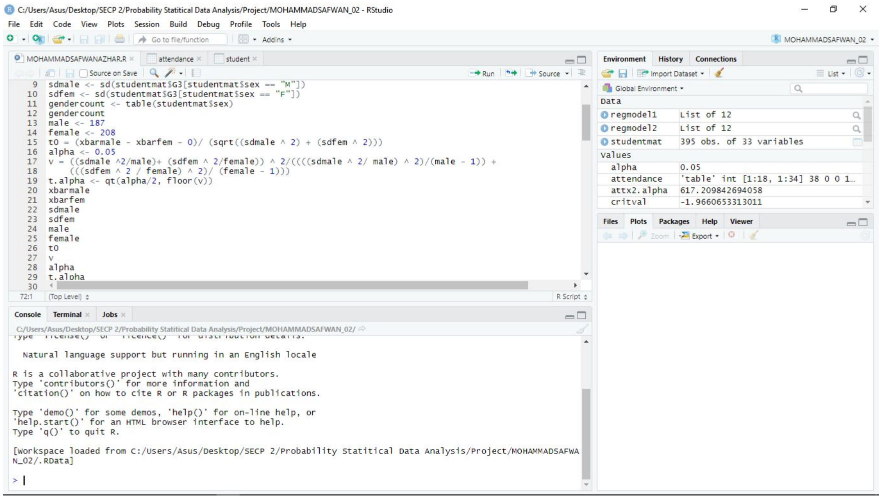
1.1.1 RStudio

RStudio is an integrated development environment (IDE) that allows users to do all statistical analysis and graphics. RStudio is similar to RGui, but is considered as a more user friendly software where users can easily find functions and help information in RStudio. It is partly written in C++ language but a bigger percentage of the code is written in Java.

1.1.1.1 Interface/Screens

Upon launching RStudio, users will see some windows and panels on top of their screen. They are:

1. Source
2. Console
3. Environment/History
4. Files/Plots/Packages/Help


C:/Users/Asus/Desktop/SECP 2/Probability Statitical Data Analysis/Project/MOHAMMADSAFWAN_02 - RStudio
File Edit Code View Plots Session Build Debug Profile Tools Help
Go to file/function
Addins
MOHAMMADSAFWAN_02
MOHAMMADSAFWANAZHAR.R
attendance
student
Source on Save
Run
Source
72:1
(Top Level)
R Script
Console
Terminal
Jobs
C:/Users/Asus/Desktop/SECP 2/Probability Statitical Data Analysis/Project/MOHAMMADSAFWAN_02/
Environment
History
Connections
Import Dataset
List
Global Environment
Files
Plots
Packages
Help
Viewer
Zoom
Export

```
sdmale <- sd(studentmat$G3[studentmat$sex == "M"])
sdfem <- sd(studentmat$G3[studentmat$sex == "F"])
gendercount <- table(studentmat$sex)
gendercount
male <- 187
female <- 208
t0 = (xbarmale - xbarfem - 0)/ (sqrt((sdmale ^ 2) + (sdfem ^ 2)))
alpha <- 0.05
v = ((sdmale ^2/male)+ (sdfem ^ 2/female)) ^ 2/((((sdmale ^ 2/ male) ^ 2)/(male - 1)) +
    (((sdfem ^ 2 / female) ^ 2)/ (female - 1)))
t.alpha <- qt(alpha/2, floor(v))
xbarmale
xbarfem
sdmale
sdfem
male
female
t0
v
alpha
t.alpha
```

```
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

[Workspace loaded from C:/Users/Asus/Desktop/SECP 2/Probability Statitical Data Analysis/Project/MOHAMMADSAFWA
N_02/.RData]

> 
```

Data	
regmodel1	List of 12
regmodel2	List of 12
studentmat	395 obs. of 33 variables
Values	
alpha	0.05
attendance	'table' int [1:18, 1:34] 38 0 0 1…
attx2.alpha	617.209842694058
critval	-1.9660653313011

Figure 1.0 Windows and Panels in RStudio

1.2 Introduction to Case Study and Dataset

Planning a smooth increment of performance of students in secondary school has become very significant for the teachers. They will have a lot of ways to do it and one of them is by analysing how students do in their examination. This includes taking into account their grades and what affects it. In order to carry out this inferential statistics about student grades prediction, the secondary data is obtained from this website:

- https://www.kaggle.com/dipam7/student-grade-prediction

This data from Paulo Cortez from University of Minho, Portugal is about how various factors can affect students' grades. This data approaches student achievement from two secondary Portugese schools. The data attributes consist of student demographic, social, school and family related features and student grades. The dataset was modeled under five-level measurement scales.

1.2.1 Attributes Information

Here are some of the attributes information and their measurement levels.

Attribute	Measurement level
Sex	Nominal
Study time	Interval
Internet	Nominal
Absences	Ratio
G1	Ratio
G2	Ratio
G3	Ratio

The target population for this dataset is secondary school students. The types of inferential statistics that will be done are two sample hypothesis tests, correlation, regression and chi-square test of independence.

1.2.2 Purpose/Objective of this Study

The purposes of why this study is being conducted are:

- To find out whether if gender is the contribution for student grade
- To find out whether if different study times can affect student grades
- To find out whether if student attendance can affect their grades
- To find out whether if students previous grades can affect their final grades

2.0 Data Analysis

2.1 Two-way Hypothesis Sampling Test

Let,
μ_1 = Population mean of final grade for male
μ_2 = Population mean of final grade for female,

Assume unknown variances are unequal

Hypothesis:

$H_0 : \mu_1 - \mu_2 = 0$
$H_1 : \mu_1 \neq \mu_2$

Significance level, $\alpha = 0.05$

Test Statistics:

General Formula :

Test statistics :

$$t = \frac{(\bar{x} - \bar{y}) - (\mu_x - \mu_y)}{\sqrt{\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}}}$$

Degree of freedom :

$$v = \frac{\left(\frac{s_x^2}{n_x} + \frac{s_y^2}{n_y}\right)^2}{\frac{\left(\frac{s_x^2}{n_x}\right)^2}{n_x - 1} + \frac{\left(\frac{s_y^2}{n_y}\right)^2}{n_y - 1}}$$

Result:

Test statistics = 0.1470
Critical value = (-1.9661, 1.9661)

Conclusion:

Based on the calculation done in R Studio, the test statistic(t_0) obtained is 0.1470 while the critical value($t_{0.025,\ 390}$,$t_{0.025,\ 390}$) obtained is (-1.9661, 1.9661). Since -1.9661 < 0.1470 < 1.9661, we fail to reject the null hypothesis. There is sufficient evidence to conclude that the mean for final grade for both male and female are the same.

Boxplot:

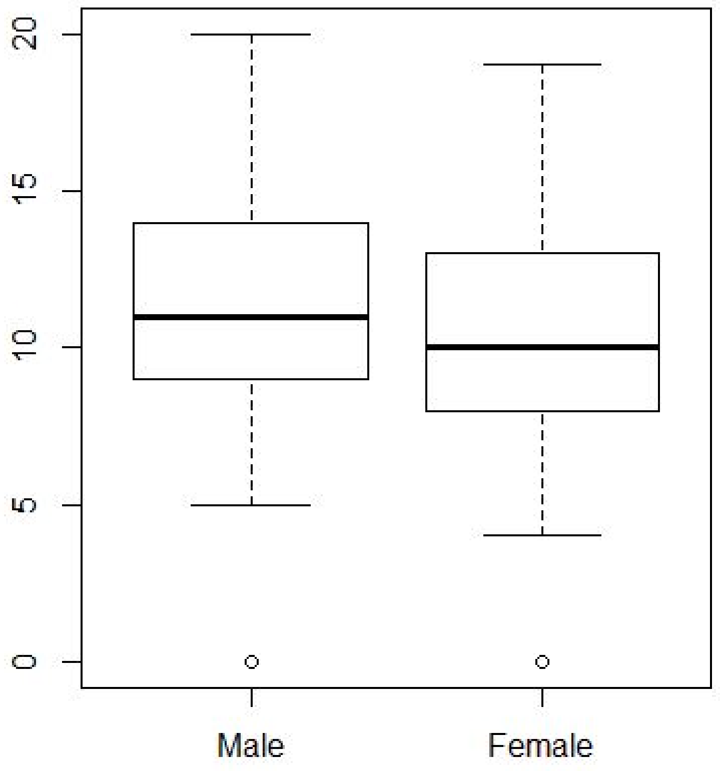


Figure 2.0 Boxplot for Male and Female Student Grades

2.2 Correlation

1.1 Correlation between First Period Grade (G1) and Final Period Grade (G3)

Scatter Plot:

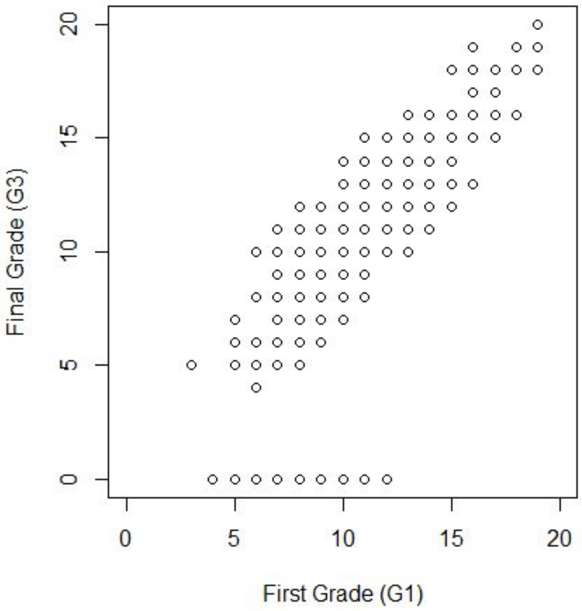


Figure 3.0 Scatter plot between G1 and G3

Result:

It can be seen from the scatter plot that when the G1 increases, the G3 increases. The correlation value for Final Grade with the First Period Grade is 0.8015. From this value, we can conclude that there is a strong relationship between the Final Period Grade and the First Period Grade. The scatter plot and the correlation analysis of the data indicate that there is a positive relationship between G1 and G3. Correlation technique used is Pearson's product-moment correlation coefficient.

1.2 Correlation between Second Period Grade (G2) and Final Period Grade (G3)

Scatter Plot:

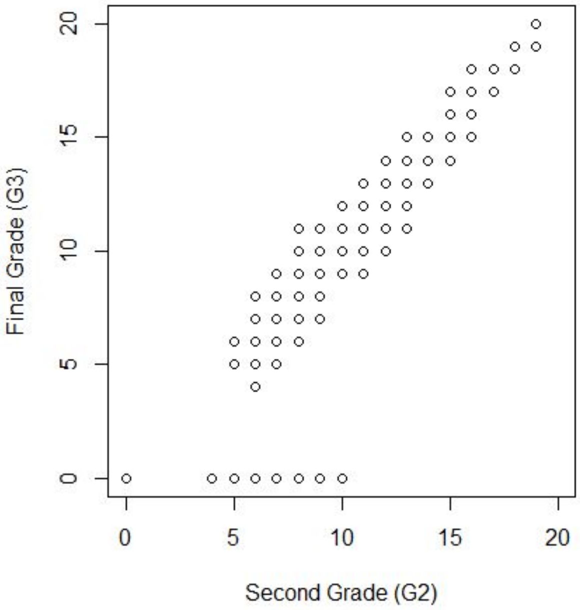


Figure 4.0 Scatter plot between G2 and G3

Result:

It can be seen from the scatter plot that when the G2 increases, the G3 increases. The correlation value for Final Period Grade with the Second Period Grade is 0.9049. From this value, we can conclude that there is a strong relationship between the Final Period Grade and the Second Period Grade. The scatter plot and the correlation analysis of the data indicate that there is a positive relationship between G2 and G3. Correlation technique used is Pearson's product-moment correlation coefficient.

Correlation Conclusion

We can say that between these two correlations, a stronger relationship of Final Period Grade can be found with First Period Grade than that of Second Period Grade due to its lower value of correlation coefficient. Although, the relationships between the variable are all strong

2.3 Regression

2.3.1 Between First Grade (G1) and Final Grade (G3)

Estimated Regression Model:

$$\bar{y} = -1.653 + 1.106x$$

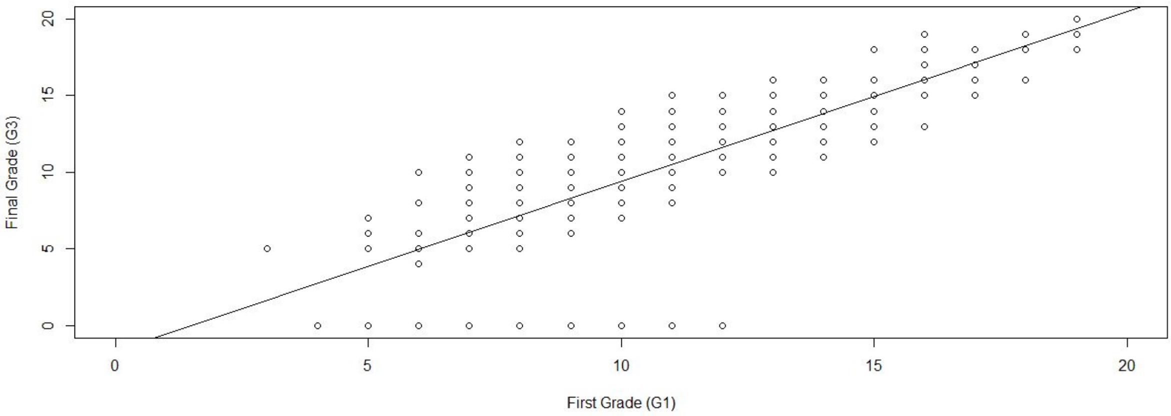


Figure 5.0 Scatter Plot

Result:

From the analysis, we can conclude that the value of intersection coefficient is -1.653 indicating that when the value of first grade is zero. The value for the estimated change in final grade can increase by 1.106 when the first grade increases by one.

Coefficient of Determination, $R^2 = 0.6424$

64.24% of the Final Period Grade variation is explained by the First Period Grade from the dataset.

2.3.2 Between Second Grade (G2) and Final Grade (G3)

Estimated Regression Model:

$$\bar{y} = -1.393 + 1.102x$$

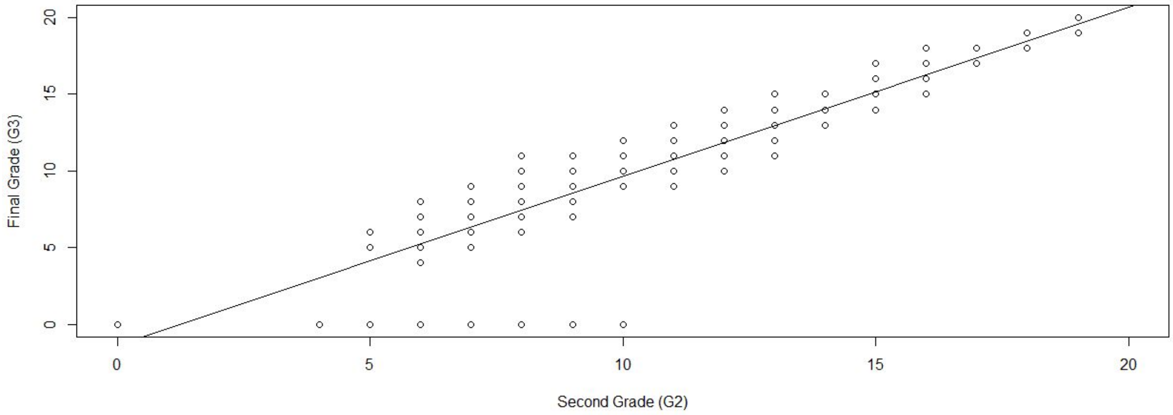


Figure 6.0 Scatter Plot

Result:

From the analysis, we can conclude that the value of intersection coefficient is -1.393 indicating that when the value of second grade is zero. The value for the estimated change in final grade can increase by 1.102 when the second grade increases by one.

Coefficient of Determination, $R^2 = 0.8183$

81.83% of the Final Period Grade variation is explained by the Second Period Grade from the dataset.

2.3.2 Between Study Time and Final Grade (G3)

Estimated Regression Model:

$$\bar{y} = 9.328 + 0.534x$$

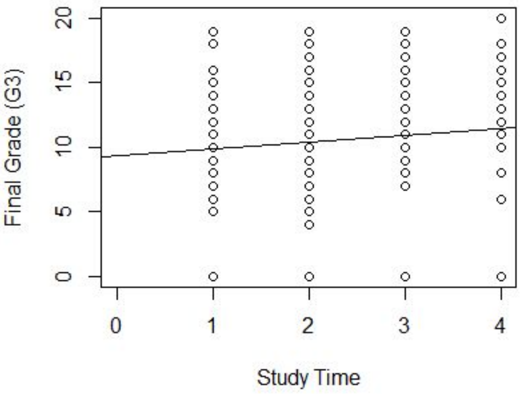


Figure 7.0 Scatter Plot

Result:

From the analysis, we can conclude that the value of intersection coefficient is 9.328 indicating that when the value of study time is zero. The value for the estimated change in final grade can increase by 0.534 when study time increases by one.

Coefficient of Determination, $R^2 = 0.007049$

0.70% of the Final Period Grade variation is explained by the student Study time from the dataset.

Regression Conclusion

Based on the plots generated, we can see that both First Period Grade and Second Period Grade have strong positive linear relationship with Third Period Grade. Study time on the other hand has a really weak positive linear relationship with Final Grade students. This is due to the straight line produced based on the scatter plots. The regression model involved also can be called as simple regression because involve only one independent variable in all regression

2.4 Chi Square Test of Independence

2.4.1 Independency between Final Grade and Study Time

Hypothesis:

H_0 = The final grade of students is independent of the students' study time
H_1 = The final grade of students is dependent on the students' study time

Significance level, $\alpha = 0.05$

Test Statistics:

$$\chi^2 = 59.432$$
$$\chi^2_{0.05,\ 57} = 75.6237$$

Since the test statistic ($\chi^2 = 59.432$) < critical value ($\chi^2_{0.05,\ 57} = 75.6237$), we fail to reject the null hypothesis (H_0). There is not enough evidence to support the claim that the final grade of students is dependent on the students' study time.

2.4.2 Independency between Final Grade and Students' Attendance

Hypothesis:

H_0 = The final grade of students is independent of the students' attendance
H_1 = The final grade of students is dependent on the students' attendance

Significance level, $\alpha = 0.05$

Test Statistics:

$$\chi^2 = 810.71$$

$$\chi^2_{0.05,\ 561} = 617.2098$$

Since the test statistic ($\chi^2 = 810.71$) > critical value ($\chi^2_{0.05,\ 561} = 617.2098$), we reject the null hypothesis (H_0). There is enough evidence to support the claim that the final grade of students is dependent on the students' attendance.

3.0 Conclusion

Based on the inferential analysis done above, we finally yield to conclusion that:

1. The population mean for both female and male are the same.
2. There is a strong positive relationship between First Period Grade (G1) and FInal Period Grade (G3)
3. There is a strong positive relationship between Second Period Grade (G2) and FInal Period Grade (G3)
4. G3 increases when both G1 and G2 increase.
5. The Final Grades of students are independent of their study time
6. The Final Grades of students are dependent on their attendance