UTM

UNIVERSITI TEKNOLOGI MALAYSIA

SECI2143 PROBABILITY & STATISTICAL DATA ANALYSIS

SEMESTER II 2020/2021

GROUP PROJECT 2

No.	TEAM OMEGA	
1	AFIF HAZMIE ARSYAD BIN AGUS	A20EC0176
2	AHMAD AIMAN HAFIZI BIN MUHAMMAD	A20EC0177
3	LUQMAN ARIFF BIN NOOR AZHAR	A20EC0202
4	MUHAMMAD IMRAN HAKIMI BIN MOHD SHUKRI	A20EC0213

Lecturer:
DR. CHAN WENG HOWE

Submission Date:
3 JULY 2021

TABLE OF CONTENT

INTRODUCTION

As a student ourselves, we would like to see exactly what would affect one's performance in academics. We were always told that in order to get good results in exams we need to study hard. Personally, we always had a hard time believing that because it almost undermines so many other factors which are often overlooked. As students of these recent years, we now realise that our situation is different from what our parents went through. Though, even after all these years many still believe common lies which we are forced to follow. Effort does not necessarily contribute to a good result, as a student ourselves we would understand that better than anyone else. Therefore, we would like to know exactly what contributes to one’s performance in academics. By doing so, we have set in our mind our plan for this case study if other external factors would affect a student’s performance.

DATA SET

Before we start our analysis, all the group members have decided to choose on one of the dataset given to us. The data utilised in this study was gathered from the dataset "Student Performance" in form of excel file. This dataset provides information on a number of student's mathematics, reading, and writing scores. The data will also contain the number of students as mentioned, gender, race or ethnicity, parental level of education and type of lunch. The 100 data we used in our inference statistical analysis report belong to the 100 students that are selected from the database and are being used for hypothesis testing to see if there is adequate evidence to back up a claim of the null hypothesis. The data sample is distributed uniformly and plotted using the Rstudio application. The datasets contain five qualitative datas which are Gender, Race or ethnicity, Parental level of education , Type of lunch, and Test preparation course. The dataset used also contains a total of three quantitative datas which are Math score, Reading score and Writing score. The usefulness of the data our group used can be seen after this part. However, we have decided to do a testing on reading score against writing score.

Variable	Data Type	Level of Measurement
Number of Student	Quantitative	Nominal
Mathematics Score	Quantitative	Ratio
Reading Score	Quantitative	Ratio
Writing Score	Quantitative	Ratio
Gender	Qualitative	Nominal
Race/Ethnicity	Qualitative	Nominal
Parental level of Education	Qualitative	Nominal
Type of Lunch	Qualitative	Nominal
Test Preparation Course	Qualitative	Ordinal

Table 1

DATA ANALYSIS & DISCUSSION

Hypothesis testing

Based on the dataset "Students Performance" we claim that reading score would influence one's writing score. Therefore, to test that out it would mean that if the student's reading score is higher then so would be their writing score. Hence, the null hypothesis,H_0 and the alternative hypothesis,H_1 would be:

$H_0 : \mu_1 - \mu_2 \neq \Delta_0$

$H_1 : \mu_1 - \mu_2 > \Delta_0$

Where μ is the sample mean for the score achieved by the students. N is the number of data. A random sample size of 100 is used.

```
        Two Sample t-test

data:  datasets$`reading score` and datasets$`writing score`
t = 0.56556, df = 198, p-value = 0.5723
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -3.232906  5.832906
sample estimates:
mean of x mean of y
    64.56     63.26
```

Figure 1

A 95% level of confidence is used to test the claim of this study. The critical value of 95% level of confidence is 1.645. We also assumed the variance for both data are equivalent.

μ_1	μ_2	S_1	S_2	Critical Value	t-value
64.56	63.26	251.1	277.2	1.645	0.56556

Table 2

Since test statistics, t < critical value (0.56556 < 1.645), we fail to reject the null hypothesis. There is not sufficient evidence to support the claim that if the student's reading score is higher then so would be their writing score.

Correlation

Correlation is made up of two words, co- and relation. Co- means together and both combined can be defined as a measure of the statistical connection between two variables or quantities that are comparable. The correlation analysis is as below:

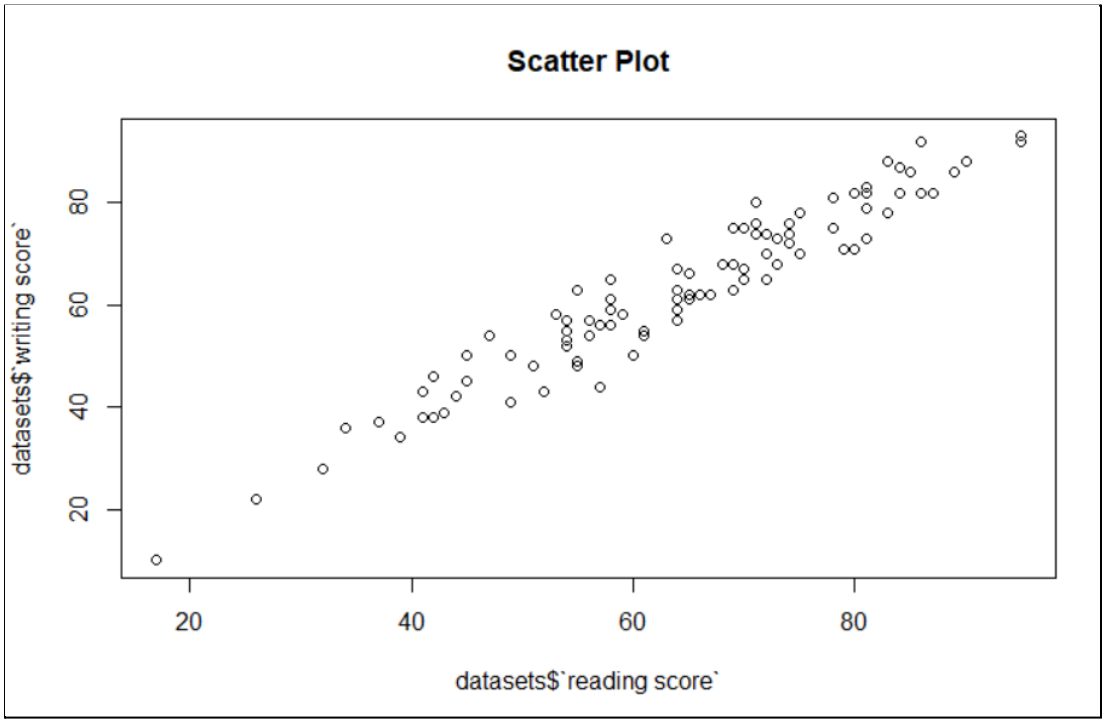

Figure 2: Correlation scatter plot

The scatter plot or scatter diagram shown above is the relationship between two variables, reading score and writing score. The data pattern will determine the relationship type, either positive or negative, with no relationship and curvilinear relationship. In this correlation test, we can measure the strength of the association between reading scores and writing scores with a sample of 100 students. We did the calculation for correlation coefficient of r by using Rstudio. The provided data was in ratio level. Therefore, we calculate the correlation coefficient by using Pearson's technique. Based on the scatter plot that we get from all the datas, we can see that the relation between writing score and reading score has a positive relationship. From the calculation, the value of r is 0.9635399 which is closer to 1 indicates that variables x and y have a strong positive relation. However, we cannot assume that if a student performs well on the writing exam, he or she would perform well on the reading test, but there is enough evidence to suggest that reading and writing scores are related.

```
        Pearson's product-moment correlation

data:  datasets$`reading score` and datasets$`writing score`
t = 35.65, df = 98, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9461957 0.9753639
sample estimates:
      cor
0.9635399
```

Figure 3: Correlation coefficient r analysis

Significance Test for Correlation

H_o : $\rho = 0$; p = population correlation coefficient

H_1 : $\rho \neq 0$

$\alpha = \frac{0.05}{2} = 0.025$

In this test, the test statistics, where sample size n = 100 while r equals 0.9635399 as calculated above. Value of t can be obtained by using the Rstudio which is 35.65.

Test Statistic, t	α	Degree of Freedom	Critical Value
35.65	0.025	98	1.98446745 or -1.98446745

Table 3: Correlation result

Since the test statistic, t > critical value (35.65 > 1.9845), we reject the null hypothesis. There is not enough evidence to claim that there is a linear relationship between the writing score and reading score at 95% level of significance. In other words, there is no linear relationship.

Regression

There are two types of regression models, linear and non-linear models. As we can probably tell from the name, the linear model assumes the relationship between independent variable and independent variable as a straight line. As opposed to that, a non-linear model does the same, assuming both variables, but this time around as a curved line.

In linear regression, we will predict the y-value based on the changes of the independent variable, x. It is worth noting that the early assumption or idea is that there are changes in y-value as x-value changes.

Therefore, the idea is summarised in a mathematical equation:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

where,

y is Dependent Variable,

β_0 is Population Intercept,

β_1 is Population Slope Coefficient,

x is Independent Variable,

ε is Random Error term, or residual.

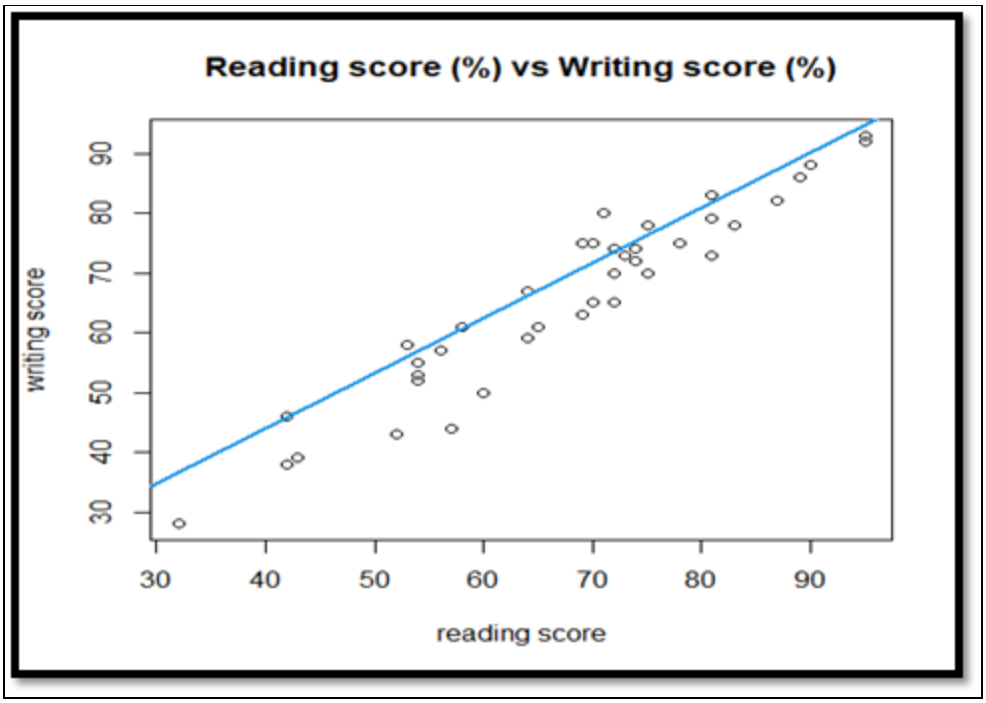

Figure 4: A scatter plot of writing score against reading score.

```
Call:
lm(formula = `reading score` ~ `writing score`)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4910  -2.8174   0.0123   3.2141  10.1004

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.55407    1.68199   3.897 0.000178 ***
`writing score`  0.91694    0.02572  35.650  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.261 on 98 degrees of freedom
Multiple R-squared:  0.9284,    Adjusted R-squared:  0.9277
F-statistic:  1271 on 1 and 98 DF,  p-value: < 2.2e-16
```

Figure 5: Regression test result

The regression of the two variables are analysed to predict the reading score and writing score of 40 students. The independent variable, x is reading score (%) while the dependent variable, y is the writing score (%).

By using RStudio, the estimated regression model is calculated, obtaining $\hat{y}$ = 7.12399 + 0.92075x. The intersection coefficient, $b_0$is 7.12399 which indicates that, for students with reading score within the range observed, 7.12399 is the portion of the writing score when no students had 0 reading score. Meanwhile, the slope coefficient, b_1= 0.92075 informs us that the average writing score increases by 0.92075 on average for each increment of reading score.

Next, we calculate the coefficient of determination, in RStudio. As a result, 0.9162 or 91.62% of the variation in writing score is affected by the variation of reading score.

Optional Test

ANOVA

ANOVA is a method of testing the equality of three or more population means by analyzing sample variances. The purpose of ANOVA is to test for significant differences between means.

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

	Reading score	Writing score
N	100	100
ΣX	6456	6326
Mean	64.56	63.26
ΣX^2	441658	427632
S	15.8461	16.6513
F	0.31986	

Table 4: ANOVA calculation result

By using the formula $F = \left((nS^{2}_{\bar{x}}) / (S^{2})\right)$ we are able to find the wanted value which is F. the value for F obtained is 0.31986. Since F statistic < F critical value (0.31986< 3.8415) therefore we fail to reject the null hypothesis.

CONCLUSION

Based on the hypothesis, we fail to reject the null hypothesis. There is not sufficient evidence to support the claim that if the student's reading score is higher then so would be their writing score.

Also, we obtained 0.9635399 as the value of r and we know that the relationship between writing and reading score has a strong positive relation.

The estimated regression model was produced using RStudio and we obtain $\hat{y} = 7.12399 + 0.92075x$. This regression model has helped us to predict the students' writing score based on their reading score.

REFERENCE

Statistics How To. (2021). *Correlation Coefficient: Simple Definition, Formula, Easy Steps*. [online] Available at: https://www.statisticshowto.com/probability-and-statistics/correlation-coefficient-formula/ [Accessed 21 June 2021].