**Data Analysis Using Correlation and Linear Regression Model**

1.0 **DESCRIPTION OF CASE STUDY AND DATA**

In this case study, the cases data set [datarium package] is used. It contains the impact of total new cases of Covid-19 on total cases. Data are the statistics of the number of cases which reported by the Crisis Preparedness and Response Centre (CPRC). The cases has been recorded since 18 March 2020 until now.

2.0 **SCENARIO**

We found, based on scatterplot and correlation co-efficient, that *total cases* and *total new cases* are positively relatef. We wish to further quantify this relationship by developing an equation (using linear regression method) predicting *total cases,* based on total new cases. Additionally we want to assess the degree to which the equation fits. The detail workings of the data analysis using Linear Regression are described in Appendix A.

3.0 **SUMMARRY OF ANALYSIS USING LINEAR REGRESSION RESULT**

The scatter graph in Appendix A suggests that a linearly increasing relationship between the total cases and the total new cases variables. This is a good thing because one important assumption of the linear regression is that the relationship between the outcome and predictor variables in linear and additive. In this cases study, the correlation coefficient is large enough, so we can continue by building a linear model of y (total cases) ad a function of x (total new cases).

3.1 Linear Regression Model

From the output of linear regression, as shown in Figure 2 on Appendix A

* the estimated regression line equation can be written as follow:

Total cases = 3561.612 + 5.080 \* new cases

* the intercept (b0) is 3561.612. It can be interpreted as the predicted total cases unit for a zero new cases.
* the regression beta coeffiecient for the variable new cases (b1), also known as the slope, is 5.080.

3.2 Summary of Model Assessment

Model Assessment of the generated linear regression model are performed by observing the following six components:

i) Residuals

ii) Co-efficients

iii) Residual Standard Error (RSE)

iv) R squared

v) F-statistic and

vi) P-value

* **Coefficients.** Shows the regression beta coefficients and their statistical significance. Predictor variables, that are significantly associated to the outcome variable, are marked by stars.
* **Residual standard error** (RSE) and the **F-statistic** are metrics that rae used to check how well the model fits to our data.

The output of the six components are as follows

Residuals:

Min 1Q Median 3Q Max

-4268.5 -3544.7 3.6 3083.1 5010.1

Coefficients:

Estimate Std. Error t value

(Intercept) 3561.612 358.920 9.923

`New Cases` 5.080 4.167 1.219

Pr(>|t|)

(Intercept) <2e-16 \*\*\*

`New Cases` 0.225

Residual standard error: 3325 on 151 degrees of freedom

Multiple R-squared: 0.009745,

Adjusted R-squared: 0.003187

F-statistic: 1.486 on 1 and 151 DF,

p-value: 0.2247

3.2.1 Residuals

The median is 3.6 which is far from zero. The absolute value of minimum and maximum are quite close, although there are not equal.

3.2.2 Co-efficients

Co-efficients of the regression model has been explained in 3.1. In this section, their statistical significance are further described.

The statistical hypotheses of our case study are as follow:

* Null hypothesis (H0) : the coefficients are equal to zero (i.e., no relationship between *new cases* (x) and *total cases* (y) )
* Alternative hypothesis (Ha) : the coefficients are not equal to zero (i.e., there is some relationship between *new cases* (x) and *total cases* (y) )

The higher the t-statistic (and the lower the p-value), the more significant the predictor (new cases). The symbols to the right visually specifies the level of significance. The line below the table shows the definition of these symbols; one star means 0.01 < p > 0.05. The more the stars beside the variable’s p-value, the more significant the variable.

A statistically significant coefficient indicates that there is an association between the predictor (x) and the outcome (y) variable.

In this case study, both the p-values for the intercept and the predictor variable are highly significant, so we can reject the null hypothesis and accept the alternative hypothesis, which means that there is a significant association between the predictor and the outcome variables.

3.2.3. Residuals Standard Error

In this case study, RSE = 3325, meaning that the observed sales values deviate from the true regression line by approxiamately 3325 units in average. This is an acceptable prediction error.

3.2.4 R squared

In dataset, the R-suared is 0.009745. This shows that new cases ia a good predictor.

3.2.5 F-statistic

In this case study, the F-statistic equal 1.468 producing a p-value of 0.2247, which is highly significant. As an overall, this shows that the factor we are using i.e. new cases are very relevant.

3.2.6 p-value

**Refer section 3.2.2**

4.0 **SUMMARRY OF ANALYSIS USING CORRELATION RESULT**

The scatter graph in Appendix A suggests that a linearly increasing relationship between the total cases and the new cases variables. This is a good thing because one important assumption of the correaltion is that the relationship between the outcome and predictor variables in linear and additive. In this cases study, the correlation coefficient is large enough, so we can continue by building a linear model of y (total cases) ad a function of x (new cases).

The output are as follows:

cor(`Total Cases`,`New Cases`,method="pearson")

## [1] 0.09871877

cor.test(`Total Cases`,`New Cases`,method="pearson")

## Pearson's product-moment correlation

## data: Total Cases and New Cases

## t = 1.219, df = 151, p-value = 0.2247

## alternative hypothesis: true correlation is not equal to 0

## 95 percent confidence interval:

## -0.06091354 0.25342709

## sample estimates:

## cor

## 0.09871877

Suppose to be the result should be seen the total cases increases as the new cases increases. A scatter plot and correlation analysis of the data indicates that there is positive relationship between the new cases and total cases.

5.0 **SUMMARRY OF ANALYSIS USING ONE SAMPLE TEST RESULT**

The boxplot graph in Figure 4 of Appendix represent the data of Active cases for Covid-19.

The output are as follows:

t.test(Active, mu=0, alternative = "two.sided", level=0.95)

## One Sample t-test

##

## data: Active

## t = 14.69, df = 152, p-value = 1

## alternative hypothesis: true mean is not equal to 0

## 95 percent confidence interval:

## -Inf 1143.083

## sample estimates:

## mean of x

## 1027.346

From the output here, we got the mean of x is 1027.346 with the percent confidence interval is 95%.

5.0 **SUMMARRY OF ANALYSIS USING CHI-SQUARE TEST RESULT**

The barplot graph in Figure 5 of Appendix represent the data of the active cases and recoveries cases of Covid-19 in Malaysia.

The result is follows:

chisq.test(TAB, correct=T)

CHI = chisq.test(TAB, correct=T)

CHI

## Pearson's Chi-squared test

##

## data: TAB

## X-squared = 16673, df = 14274, p-value < 2.2e-16

attributes(CHI)

## $names

## [1] "statistic" "parameter" "p.value" "method" "data.name" "observed"

## [7] "expected" "residuals" "stdres"

##

## $class

## [1] "htest"

The output tell that the value is 16673 with the p-value = 2.2.

6.0 **JUSTIFICATION AND CONCLUSION**

The model adequacy check shows that the regression model fits to our data and that the cases are very significant and relevant. The above analysis also implies that “total new cases” is a good predictor or factor that gives significantly high impact in predicting total cases. The above findings is in parallel with instructions Movement Control Order (MCO). MCO is instructions from government that everyone need to stay at home and isolate themselves unless if it is necessary. This is to prevent the spread of the virus Covid-19. The more you obey the government, the more quick our country free from this virus.