



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SECI 2143/SCSI 2143

PROBABILITY & STATISTICAL DATA ANALYSIS

2019/2020

INDIVIDUAL PROJECT 2

(SURVEY ON THE STATUS OF COVID-19

OUTBREAK IN MALAYSIA)

LECTURER'S NAME	DR. AZURAH ABU SAMAH
STUDENT'S NAME	NG JING ER
MATRIC NO	A19EC0115

TABLE OF CONTENTS

TOPIC	PAGE
INTRODUCTION	1
METHODOLOGY	2
DATA ANALYSIS AND RESULTS: A. 1 SAMPLE TEST B. CORRELATION C. REGRESSION D. GOODNESS OF FIT	3-13
DISCUSSION	14-16
CONCLUSION	17
REFERENCE	18
APPENDIX 1	19-25

INTRODUCTION

Due to the outbreak of the infectious disease: Coronavirus disease (COVID-19), our government has implemented MCO as an initiative to prevent a more serious epidemic. As we all know, Coronavirus disease causes a serious impact to every one of us. Thousands of lives had sacrificed due to this disease. Therefore, I decided to conduct a study about COVID-19.

The virus that causes COVID-19 is mainly transmitted through droplets generated when an infected person coughs, sneezes, or exhales. These droplets are too heavy to hang in the air, and quickly fall on floors or surfaces. That is the reason we may be infected easily just by breathing in the virus if you are within close proximity of someone who has COVID-19, or by touching a contaminated surface and then your eyes, nose or mouth.

Therefore, in this project 2 I have conducted a statistical analysis on COVID-19 to study about the impact of COVID-19 in compiling the national statistics. The purpose of this study mainly about the further analysis of data obtained from secondary data (Department of Statistics, Malaysia (DOSM)) about the status of COVID-19 in Malaysia.

Objectives:

1. To determine whether male has higher risk of death due to COVID-19 than female.
2. To estimate the risk of COVID-19 by determine the relationship between the confirmed cases and deaths.
3. To determine whether the new cases (independent variable) of Covid-19 from different states is able to predict the value of deaths (dependent variable) of Covid-19.
4. To determine whether the confirmed cases of Covid-19 from different states are the same.

METHODOLOGY

All the data used for analysis in this project is collected from secondary source, website of Department of Statistics, Malaysia (DOSM). DOSM has specially developed a landing page as an initiative to inform the public about the status of COVID-19 outbreak in Malaysia.

All the data and statistics about COVID-19 by States in Malaysia that presented on website are based on official data from the Ministry of Health Malaysia and media agencies (primary sources). All the contents are handpicked, filtered, and curated to the best extent to ensure that sources of information for the public to get real-time updates on the virus's reach and impact. The statistics are presented using an interactive visualisation or infographic to help the public to grasp the information.

The population counted in are all the people who stayed in Malaysia. The sample size is 33782300 people from 13 states (Negeri) and 3 federal territories (Wilayah Persekutuan).

There are a few types of variables are selected for analysis such as nominal, ratio and interval. And those variables including deaths, date, confirmed cases, proportion according gender and states. The data for deaths and confirmed cases used is based on the statistic provide by DOSM website on 17th May 2020. The statistical test analysis related to the variables chosen are 1 sample test, correlation, regression and goodness of fit.

R studio is the only medium that used in this project for the purpose to test and analyze the data. And some of the findings are presented with the graph.

DATA ANALYSIS AND RESULTS

A. 1 SAMPLE TEST

1.0 Description Of Case Study And Data

In this example, the dataset of covid19 1 is used. Data are the proportion of deaths of Covid-19 among gender until 17th May 2020.

2.0 Scenario

I found that the number of deaths due to Covid-19 from male is higher than female. So, I decided to further determine the claim whether “male has higher risk of death due to COVID-19” by using 1 sample test.

3.0 Summary Of Analysis Using 1 Sample Test Result

The hypothesis test is conducted at the 5% level of significance to test the claim where “male has higher risk of death due to COVID-19”. 1 sample test is used to test the claim. Here we assumed that male has risk of death, where more than 50% of the deaths are male.

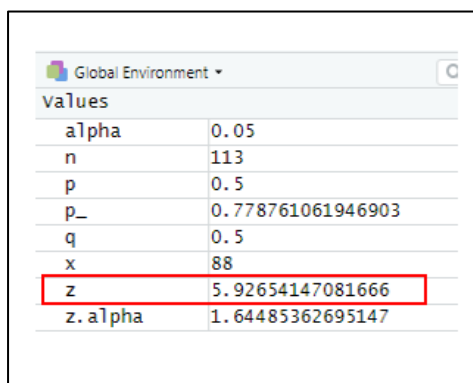
3.1 Solution:

$$H_0: p = 0.5$$

$$H_1: p > 0.5$$

Reject H_0 if the proportion of male deaths is lower or equal to the 0.5.

Given, $\alpha = 0.05$. The formula to find the value of test statistic is:
$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$



Global Environment	
values	
alpha	0.05
n	113
p	0.5
p_	0.778761061946903
q	0.5
x	88
z	5.92654147081666
z.alpha	1.64485362695147

- n denoted as total number of deaths = 113
- x denoted as total number of female deaths = 88
- p denoted as population proportion = 0.5
- q is equal to 1-p
- \hat{p} denoted as sample proportion = 88/11

- z denoted as test statistic, and the result obtained in R studio is 5.9265, such that:

$$z = \frac{\hat{p} - p}{\sqrt{\frac{pq}{n}}}$$

$$= \frac{\frac{88}{113} - 0.5}{\sqrt{\frac{0.5 \times 0.5}{113}}}$$

$$= 5.9265$$

4.0 Justification And Conclusion

4.1 Check with critical value

$$Z_{0.05} = 1.645$$

Reject H_0 if $Z_0 > Z_{0.05} = 1.645$

4.2 Check with p-value

$$P\text{-value} = 1.547e-09$$

Reject H_0 if P-value is less than the significance level=0.05

```
1-sample proportions test without continuity correction
data:  x out of n, null probability p
x-squared = 35.124, df = 1, p-value = 1.547e-09
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.7084368 1.0000000
sample estimates:
      p 
0.7787611
```

Conclusion:

- Since $Z_0 > 1.645$, we reject H_0 at 0.05 of the significance level.
- Since $p\text{-value} < 0.05$, reject H_0 .

Therefore, we reject null hypothesis. There is sufficient evidence to support that male has higher risk of death due to COVID-19.

B. CORRELATION

1.0 Description Of Case Study And Data

In this example, the covid19 2 data set is used. It contains the statistic of Covid-19 on 17th May 2020. The data of Confirmed and Deaths from the dataset covid19 2 are used.

2.0 Scenario

I found that the number of deaths due to Covid-19 is getting rise along with the number of confirmed cases. So, I wish to further determine the relationship between these variables using correlation test. In addition, I want to know whether the statistic can be used to estimate the case fatality rate (aka: CFR) which can determine the risk of death from COVID-19 using the formula (Worldometers.info,2020).

3.0 Summary Of Analysis Using Correlation Test Result

The data used to measure the relationship between variables:

- x denoted as the number of confirmed cases
- y denoted as the number of deaths

Due to the variables (Confirmed cases, deaths) are ratio-type data, therefore we use Pearson's technique to analysis the correlation between variables.

States	Σx	Σy	Σx^2	Σy^2	Σxy
Selangor	1644	21	2,702,736	441	34,524
WP Kuala Lumpur	1566	17	2,452,356	289	26,622
N Sembilan	776	8	602,176	64	6,208
Johor	668	19	446,224	361	12,692
Sarawak	544	17	295,936	289	9,248
Pahang	336	6	112,896	36	2,016
Sabah	331	5	109,561	25	1,655
Perak	255	6	65,025	36	1,530

Melaka	215	5	46,225	25	1075
Kelantan	155	3	24,025	9	465
P Pinang	121	1	14,641	1	121
Terengganu	110	1	12100	1	110
Kedah	95	1	9025	1	95
WP Putrajaya	91	1	8281	1	91
Perlis	18	2	324	4	36
WP Labuan	18	0	324	0	0
Grand total	6943	113	6901855	1583	96488

Where,

Global Environment	
Data	
covid19_2	16 obs. of 5 variables
Values	
r	0.858867623127667
sum_of_X	6943
sum_of_X2	6901855
sum_of_XY	96488
sum_of_Y	113
sum_of_Y2	1583

- sum_of_X = 6943
- sum_of_Y = 113
- sum_of_X2 = 6901855
- sum_of_Y2 = 1583
- sum_of_XY = 96488
- n denoted as total of states = 16

4.0 Justification And Conclusion

- r denoted as the result of correlation coefficient obtained using R studio, with the formula: $\text{cor}(x, y, \text{method} = "pearson")$
- And the result is $r = 0.8588676$, which is same as the formula such that:

$$r = \frac{\sum xy - (\sum x \sum y)/n}{\sqrt{[(\sum x^2) - \frac{(\sum x)^2}{n}][(\sum y^2) - \frac{(\sum y)^2}{n}]}}$$

$$r = \frac{96488 - \frac{6943 \times 113}{16}}{\sqrt{[6901855 - \frac{6943^2}{16}][1583 - \frac{113^2}{16}]}}$$

$$r = 0.8589$$

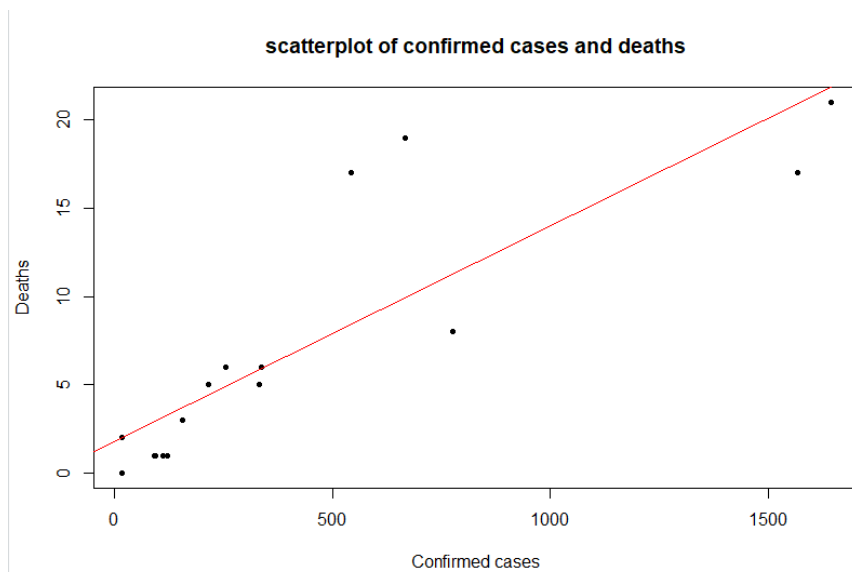


Figure 1 – Scatter graph of deaths versus confirmed cases

Conclusion:

Since $r=0.8589$, it is closer to 1.

Therefore, it is a strong positive linear relationship between the variables.

- The case fatality rate (aka: CFR) which can determine the risk of death from COVID-19 using the formula (Worldometers.info,2020):

$$\begin{aligned}
 CFR &= \frac{\text{total deaths}}{\text{total number of confirmed cases}} \times 100\% \\
 &= \frac{113}{6943} \times 100\% \\
 &= 1.628 \%
 \end{aligned}$$

Unfortunately, due to the situation of fast-spreading of COVID 19, the number confirmed deaths and cases change over time and the CFR obtained cannot be used as it is extremely difficult to make accurate estimates of the true risk of death (Our World in Data, 2020).

C. REGRESSION

1.0 Description Of Case Study And Data

In this example, the covid19 2 data set is used. It contains the statistic of Covid-19 on 17th May 2020. Data are the new cases along with the deaths in case. The number of new cases from 16 different states including 13 states (Negeri) and 3 federal territories (Wilayah Persekutuan) and the observed deaths have been recorded.

2.0 Scenario

I found, based on scatterplot and correlation co-efficient, that *deaths* and *newcase* are related. I wish to further quantify this relationship by developing an equation (using linear regression method) predicting deaths, based on number of new cases from 16 different states. 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 1.

3.0 Summary Of Analysis Using Linear Regression Result

The scatter graph in Figure 1 of Appendix 1 suggests that a linearly increasing relationship between the *deaths* and the *newcase* variables. In this example, the correlation coefficient (0.2168) is not too large, but since the result of correlation coefficient between the two variables using the R function `cor()` is 0.4656686 which greater than 0.2. Therefore, I continue by building a linear model of *y* (*deaths*) as a function of *x* (*newcase*).

3.1 Linear Regression Model

From the output of linear regression, as shown in Figure 2 of Appendix 1,

- the estimated regression line equation can be written as follow:
$$\text{deaths} = 6.0274 + 0.3524 * \text{newcase}$$
- the intercept (b_0) is 6.0274. It can be interpreted as the predicted deaths for a zero new case. This means that, for a new case equal zero, we can expect a death of 6.0274.
- the regression beta coefficient for the variable new cases (b_1), also known as the slope, is 0.3524. This means that, for new cases equal to 100 cases, we can expect an increase of 36 cases ($0.3524 * 100$) in deaths. That is, $\text{deaths} = 6.0274 + 0.3524 * 100 = 41.2674(42)$ cases.

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 are used to check how well the model fits to our data.

The output of the six components are as follows:

```
call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-6.0274 -5.0274 -1.7223  0.4726 12.9726

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0274     1.7380   3.468  0.00377 **
x              0.3524     0.1790   1.969  0.06909 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.626 on 14 degrees of freedom
Multiple R-squared:  0.2168,    Adjusted R-squared:  0.1609
F-statistic: 3.876 on 1 and 14 DF,  p-value: 0.06909
```

3.2.1 Residuals

The median is -1.7223 which is a bit far from zero. While the absolute value of minimum and maximum are far.

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:

- H_0 : the coefficients are equal to zero
(*no relationship between new cases (x) and deaths (y)*)
- H_a : the coefficients are not equal to zero
(*there is some relationship between cases (x) and deaths (y)*)

The line below the table shows the definition of these symbols; ‘.’ means $0.05 < p < 0.1$ and 2 stars means $0.001 < p < 0.01$. The more the stars beside the variable’s p-value, the more significant the variable. The result of a statistically significant coefficient indicates that there is association between the predictor (x) and the outcome (y) variable.

In our example, the p-values for the intercept are highly significant, but the predictor variable is less significant. In conclusion, even though there is a less significant association between the predictor and the outcome variables, we reject the null hypothesis and accept the alternative hypothesis.

3.2.3 Residuals Standard Error

In our example, $RSE = 6.626$, meaning that the observed deaths values deviate from the true regression line by approximately 6.6 cases in average. It is considered as slightly big prediction error.

3.2.4 R squared

In our dataset, the R-squared is 0.2168. This shows that newcase is a weak predictor as it only able to explain 21.68 % of the variation in deaths.

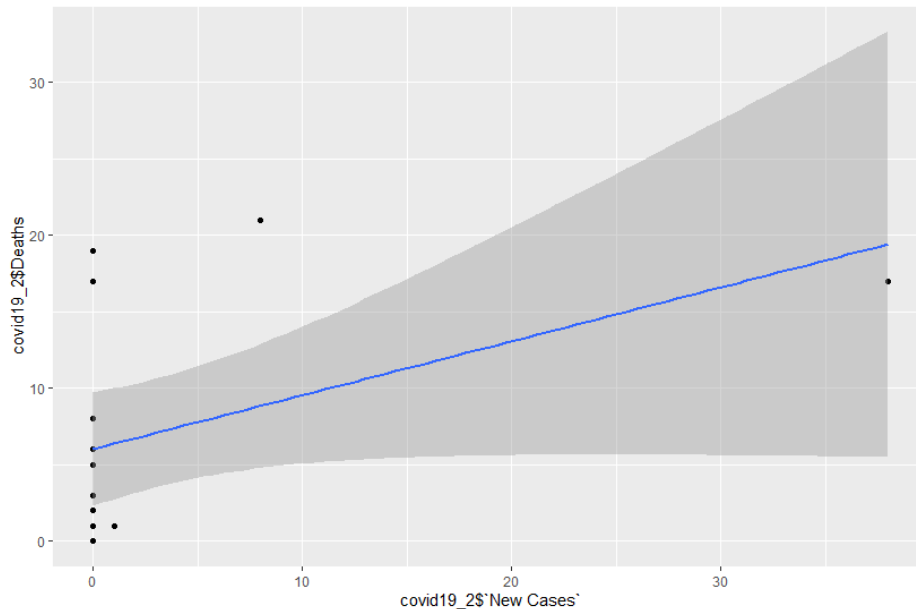
3.2.5 F-statistic

In our example, the F-statistic equal 3.876 producing a p-value of 0.0609, which is less significant. As an overall, this shows that the factor we are using newcase is less relevant.

3.2.6 p-value

Refer section 3.2.2

4.0 Justification And Conclusion



D. GOODNESS OF FIT

1.0 Description Of Case Study And Data

In this example, the covid19 2 data set is used. It contains the statistic of Covid-19 on 17th May 2020. Data are the confirmed cases from 16 states including 13 states (Negeri) and 3 federal territories (Wilayah Persekutuan).

2.0 Scenario

From the statistic obtained from the website DOSM, I found that the number of confirmed cases has significant different among the states in Malaysia. So, goodness of fit test is used to test the hypothesis that the observed proportion claim that the confirmed cases from different states are same at a significant level of 0.05. The expected value will then be calculated by using data of observed proportion.

3.0 Summary Of Analysis Using Linear Regression Result

The null hypothesis and the alternative hypothesis are as follow:

H_0 : The proportions of the confirmed cases from different states are same

H_1 : At least one of the proportions of the confirmed cases is different from other states.

- O represents the observed frequency of confirmed cases.
- E represents the expected frequency of confirmed cases.
- k represents the number of states
- n represents the total frequency of confirmed cases

Since expected value is the same, the value of expected proportion can be calculated by using the formula of mean, $Ei = \frac{n}{k}$.

```
> Ei$expected
[1] 433.9375 433.9375 433.9375 433.9375 433.9375 433.9375
[7] 433.9375 433.9375 433.9375 433.9375 433.9375 433.9375
[13] 433.9375 433.9375 433.9375 433.9375
```

Ei obtained in R studio is 433.9375.

4.0 Justification And Conclusion

- The chi-square result is then calculated by using R Studio with $Ei = 433.9375$, by the test statistic which can be calculated using the formula, $\chi^2 = \frac{(O_i - E_i)^2}{E_i}$

```
chi-squared test for given probabilities
data: result
X-squared = 8962.2, df = 15, p-value < 2.2e-16
```

- χ^2 obtained in R studio is 8962.2 with a degree freedom of 15.

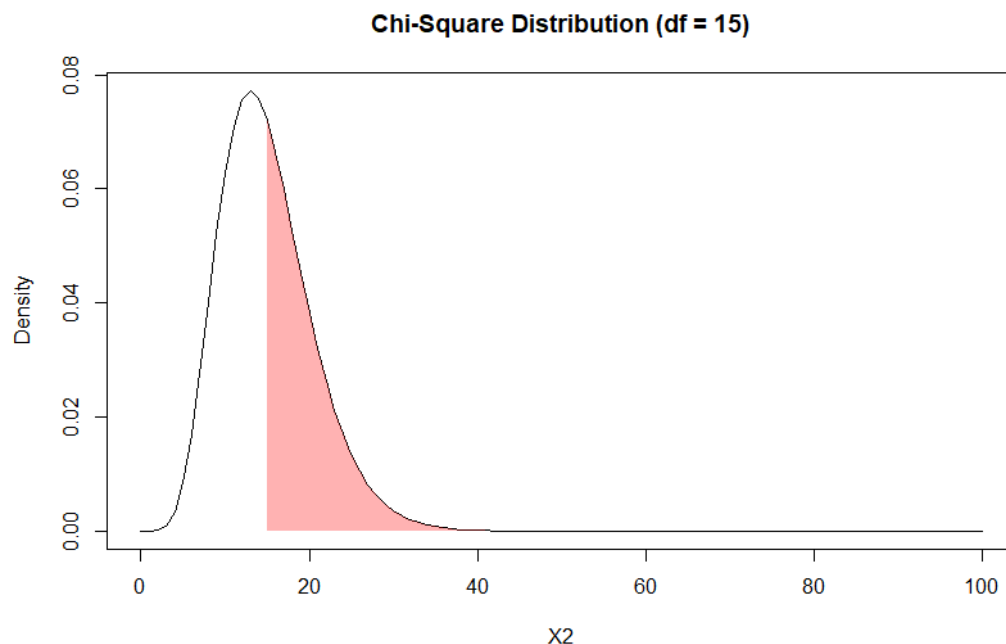


Figure 3 – Chi-squared distribution

Conclusion:

Test statistic value ($\chi^2 = 8962.2$) much greater than the critical value ($\chi^2 = 24.996$, $df=15$, $\alpha = 0.05$), that is it falls within critical region. Thus, we reject H_0 .

There is sufficient evidence to reject the claim that the proportions of the confirmed cases from different states are same.

DISCUSSION

1. 1 SAMPLE TEST

According to the data released showing that more than three-quarter of those who died of Covid-19 in Malaysia are men, and the proportion is similar observed across the world (Free Malaysia Today, 2020). And from the analysis of data in this project, it is found out that there is sufficient evidence to support that male has higher risk of death due to COVID-19.

Death rates still on climbing for both men and women, but the numbers for men tend to outstrip women. And this may due to other health conditions and also their lifestyle habits which could put them to higher risk (Michelle Roberts). According to FMT News, a study about British medical journal published in Lancet stated that emerging evidence suggested that the reasons for more male deaths lay in sex-based immunological or gendered differences such as lifestyle habits and prevalence of smoking.

From the data collected from DOSM website, there are 113 deaths on 17th May 2020 due to COVID-19. Among the 113 deaths, 88 of the cases are from male while only 25 of them are female. Based on the result, the proportion of male deaths are relatively higher than female. The test statistic, Z_0 calculated using R-studio is 5.9265 which is greater than 1.645. Thus, we reject H_0 . Therefore, from the results generated in test analysis of 1 sample test, it is said to have sufficient evidence to support that male has higher risk of death due to COVID-19 male has higher risk of death due to COVID-19 at significance level of 0.05.

2. CORRELATION

Second test analysis is correlation to test about the relationship between variables of confirmed cases and deaths on 17th May 2020. Based on the test analysis of correlation, we manage the determined that there is a strong relationship between the variables, where x denoted as confirmed cases while y denoted as deaths.

And since both the variables are ratio-type data, therefore Pearson's technique is used to analysis the correlation between variables. From the result obtained in R-studio, r is 0.8589 which is closer to 1. And a scatter plot also generated which showing that it is a strong positive linear relationship. In this part, I tried to calculate the case fatality rate (aka: CFR) which can determine the risk of death from COVID-19 using the formula (Worldometers.info,2020):

$$\begin{aligned}
 CFR &= \frac{\text{total deaths}}{\text{total number of confirmed cases}} \times 100\% \\
 &= \frac{113}{6943} \times 100\% \\
 &= 1.628 \%
 \end{aligned}$$

Unfortunately, due to the situation of fast-spreading of COVID 19, the number confirmed deaths and cases change over time and the CFR obtained cannot be used as it is extremely difficult to make accurate estimates of the true risk of death (Our World in Data, 2020).

3. REGRESSION

The relationship between new cases and deaths is determined by using the $\text{lm}(y \sim x)$ functions in R. Where residual standard deviation is also referred to as the standard deviation of points around a fitted line or the standard error of estimate (Will Kenton, 2019). The residual standard obtained is greater than 0, which means it is not that closer of the estimate data to fit the actual data. Besides, the coefficient of determination, also called R^2 is the portion of the total variation in the dependent variable that is explained by variation in the independent variable. The value of R^2 obtained is 0.2168 which means some but not all of the variation in y (deaths) is explained by variation in x (new cases). It shows only 21.68% of the variation in deaths is explained by variation in new cases. Therefore, it may be not a good example of model as it is less predictive.

4. GOODNESS OF FIT

The last test analysis is Goodness of fit test, which is used to test whether the confirmed cases from different states are same at a significant level of 0.05. The expected value, E_i obtained is 433.9375 and the chi-square result calculated by using R-studio is 8962.2. Since the chi-square result is greater than the test statistic 24.996, we reject null hypothesis. There is insufficient evidence to support that the confirmed cases from different states including 13 states and 3 federal territories are same, there is at least one of the proportions of the confirmed

cases from a particular state is different from others. So, since there is a large disagreement between observed and expected values, it will lead to a large value of χ^2 (8962.2) and a small p- value ($2.2e-16$). Thus, significantly large value of χ^2 cause to a rejection of the null hypothesis of no difference between the observed and the expected confirmed cases. And this may due to the Coronavirus disease has resulted in an ongoing pandemic, there are different clusters identified in different states, which may result some of the areas have relatively higher cases compare to others.

CONCLUSION

People of all ages can be infected by the new coronavirus (2019-nCoV), especially older people, and people with pre-existing medical conditions. However, according to the statistic today shows that most of the patients are men. One of the reasons causes the number of deaths for men tend to outstrip women may due to other health conditions and also their lifestyle habits which could put them to higher risk (Michelle Roberts). From the result based on analysis of data in the hypothesis testing (1 sample test), we can conclude that it is statistically showing men have higher risk of death due to Covid-19 compare to women. However, the factors on men comprise higher percentage of those dying from Covid-19 in Malaysia still need to be confirmed.

Although the result of CFR calculated previously seems to be not accurate, however I found a lot of discussion about Mortality Risk of Covid-19 from a more proper and professional source of statistic and research. From the findings displayed in “Our World in Data”, I understand that it is a challenging task for the measurement and the definitions to interpret estimates of the CFR for COVID-19, particularly those relating to an ongoing outbreak.

Since the result of regression shows that only 21.68% of the variation in deaths is explained by variation in new cases. And the graph generated and the findings are less suitable and less predictive to estimate a variable.

The result of the test (goodness of fit) shows that the proportions of confirmed cases are different among the states. There are many factors cause the number of confirmed cases is different among the states since the Coronavirus disease so easy to spread among people. And this may result some of the areas have relatively higher cases compare to others. The number of red zones also depends on the active cases according to that particular district or a locality (New Straits Times, 2020).

REFERENCES

1. Secondary data source: Department of Statistics, Malaysia (DOSM)
Source of dataset (URL): <https://ukkdosm.github.io/covid-19>
Data obtained on 17/5/2020.
2. Thomas, J., 2020. Seeking Answers For Why More Men Die Of Covid-19. [online]
Free Malaysia Today. Retrieved from
<https://www.freemalaysiatoday.com/category/nation/2020/04/13/seeking-answers-for-why-more-men-die-of-covid-19/>.
3. Our World in Data. 2020. Mortality Risk Of COVID-19 - Statistics And Research.
[online] Retrieved from: <https://ourworldindata.org/mortality-risk-covid#>
4. Max Roser, Hannah Ritchie, Esteban Ortiz-Ospina and Joe Hasell, 2020 - "Coronavirus Pandemic (COVID-19)". [online] OurWorldInData.org.
Retrieved from: <https://ourworldindata.org/coronavirus>
5. Worldometers.info. 2020. Coronavirus Death Rate (COVID-19) - Worldometer.
[online]
Retrieved from: <https://www.worldometers.info/coronavirus/coronavirus-death-rate/>
6. Will Kenton, 2019. *How The Residual Standard Deviation Works*. [online]
Investopedia. Retrieved from
<https://www.investopedia.com/terms/r/residual-standard-deviation.asp#:~:text=The%20residual%20standard%20deviation%20is%20simply%20the%20standard%20deviation%20of,spread%20around%20the%20regression%20line.>
7. Michelle Roberts, 2020. Coronavirus: What is the risk to men over 50? [online] BBC News online. Retrieved from <https://www.bbc.com/news/health-52197594>
8. Veena Babulal, Dawn Chan, 2020. 8 red zones for 2 straight days. [online] New Straits Times. Retrieved from <https://www.nst.com.my/news/nation/2020/05/592509/8-red-zones-2-straight-days>

APPENDIX 1

a. VISUALIZATION

- Correlation Coefficient

I compute the correlation coefficient between the two variables using the R function cor():

```
> cor(covid19_2$Deaths,covid19_2$`New Cases` )  
[1] 0.4656686
```

The correlation coefficient measures the level of the association between two variables x and y. Its value ranges between -1 (perfect negative correlation: when x increases, y decreases) and +1 (perfect positive correlation: when x increases, y increases). A value closer to 0 suggests a weak relationship between the variables. A low correlation ($-0.2 < x < 0.2$) probably suggests that much of variation of the outcome variable (y) is not explained by the predictor (x). In such case, we should probably look for better predictor variables.

In our example, although the correlation coefficient is only 0.4656686 but it is greater than 0.2, so I continue by building a linear model of y as a function of x.

- Data Frame

A data frame is made using the formula data.frame(), then display the data as follow:

```
> df <- data.frame(states=c("Selangor","WP Kuala Lumpur","N Sembilan","Johor","Sarawak",  
+ "Pahang","Sabah","Perak","Melaka","Kelantan","P Pinang",  
+ "Terengganu","Kedah","WP Putrajaya","Perlis","WP Labuan"),  
+ newcase=c(8,38,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0),  
+ deaths=c(21,17,8,19,17,6,5,6,5,3,1,1,1,1,2,0) )  
> df
```

	states	newcase	deaths
1	Selangor	8	21
2	WP Kuala Lumpur	38	17
3	N Sembilan	0	8
4	Johor	0	19
5	Sarawak	0	17
6	Pahang	0	6
7	Sabah	0	5
8	Perak	0	6
9	Melaka	0	5
10	Kelantan	0	3
11	P Pinang	0	1
12	Terengganu	0	1
13	Kedah	0	1
14	WP Putrajaya	1	1
15	Perlis	0	2
16	WP Labuan	0	0

- Graph

We want to predict future deaths on the basis of the number of new cases of Covid-19 in different states.

```
ggplot(covid19_2, aes(x = covid19_2$`New Cases`, y = covid19_2$Deaths)) +  
  geom_point() +  
  stat_smooth(method = lm)
```

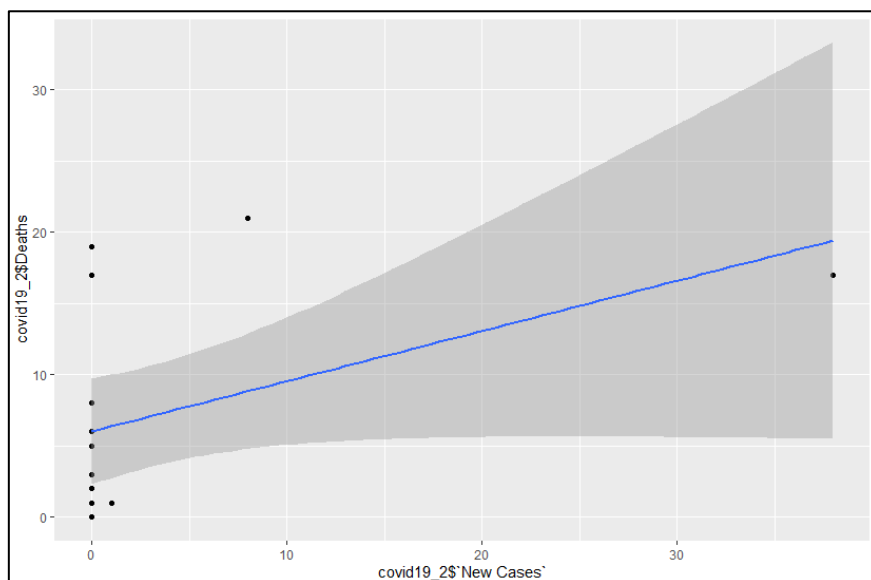


Figure 2 – Scatter graph of new cases versus deaths

The graph above suggests only a slightly linear increasing relationship between the deaths and the new cases variables.

b. COMPUTATION

The simple linear regression tries to find the best line to predict deaths on the basis of newcase.

The linear model equation can be written as follow: $\text{deaths} = b_0 + b_1 * \text{newcase}$.

- The R function `lm()` can be used to determine the beta coefficients of the linear model:

```
> relation <- lm(y~x)  
> print(relation)  
  
Call:  
lm(formula = y ~ x)  
  
Coefficients:  
(Intercept)          x  
    6.0274         0.3524
```

The results show the intercept and the beta coefficient for the newcase variable.

- The function predict() also been used to predict the value of deaths.

```
> #predict deaths with new cases=100
> a <- data.frame(x =100)
> result <- predict(relation,a)
> print(result)
      1
41.26337
```

c. INTERPRETATION

From the output above:

- the estimated regression line equation can be written as follow:
deaths = 6.0274 + 0.3524*new cases
- the intercept (b0) is 6.0274. It can be interpreted as the predicted deaths for a zero new case. This means that, for a new case equal zero, we can expect the deaths of 6.0274.
- the regression beta coefficient for the variable new cases (b1), also known as the slope, is 0.3524. This means that, for new cases equal to 100 cases, we can expect an increase of 36 cases (0.3524*100) in deaths. That is, deaths = 6.0274 + 0.3524*100 = 41.2674(42) cases.

d. MODEL ASSESSMENT

- In the previous section, a linear model of deaths is built as a function of newcase of Covid-19 from different states: deaths = 6.0274 + 0.3524*100 cases.
- Relation summary: Display the statistical summary of the model using the R function:
print(summary(relation))

```
Call:
lm(formula = y ~ x)

Residuals:
    Min       1Q   Median       3Q      Max
-6.0274 -5.0274 -1.7223  0.4726 12.9726

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   6.0274     1.7380   3.468  0.00377 **
x              0.3524     0.1790   1.969  0.06909 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.626 on 14 degrees of freedom
Multiple R-squared:  0.2168,    Adjusted R-squared:  0.1609
F-statistic: 3.876 on 1 and 14 DF,  p-value: 0.06909
```

The summary outputs show 6 components, including:

- **Call.** Shows the function call used to compute the regression model.
- **Residuals.** Provide a quick view of the distribution of the residuals, which by definition have a mean zero. Therefore, the median should not be far from zero, and the minimum and maximum should be roughly equal in absolute 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), R-squared (R2)** and the **F-statistic** are metrics that are used to check how well the model fits to our data.

5. COEFFICIENTS SIGNIFICANCE

The coefficients table, in the model statistical summary, shows:

- a. the estimates of the **beta coefficients**
- b. the **standard errors** (SE), which defines the accuracy of beta coefficients. For a given beta coefficient, the SE reflects how the coefficient varies under repeated sampling. It can be used to compute the confidence intervals and the t-statistic.
- c. the **t-statistic** and the associated **p-value**, which defines the statistical significance of the beta coefficients.

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.0274	1.7380	3.468	0.00377 **
x	0.3524	0.1790	1.969	0.06909 .

- The statistical hypotheses are as follow:
 - Null hypothesis (H0): the coefficients are equal to zero
(no relationship between x and y)
 - Alternative Hypothesis (Ha): the coefficients are not equal to zero
(there is some relationship between x and y)
- a. Mathematically, for a given beta coefficient (b), the t-test is computed as $t = (b - 0) / SE(b)$, where SE(b) is the standard error of the coefficient b. The t-statistic measures the number of standard deviations that b is away from 0. Thus, a large t-statistic will

produce a small p-value. The higher the t-statistic (and the lower the p-value), the more significant the predictor. The symbols to the right visually specify the level of significance.

In our example, the p-values for the intercept are highly significant and the predictor variable is a bit less significant. So, we reject the null hypothesis and accept the alternative hypothesis, which means that there is a small significant association between the predictor and the outcome variables.

b. t-statistic and p-values:

For a given predictor, the t-statistic (and its associated p-value) tests whether or not there is a statistically significant relationship between a given predictor and the outcome variable, that is whether or not the beta coefficient of the predictor is significantly different from zero.

c. Standard errors and confidence intervals:

The standard error measures the variability/accuracy of the beta coefficients. It can be used to compute the confidence intervals of the coefficients.

For example, the 95% confidence interval for the coefficient b_1 is defined as $b_1 \pm 2 \cdot SE(b_1)$, where:

- the lower limits of $b_1 = b_1 - 2 \cdot SE(b_1) = 0.3524 - 2 \cdot 0.1790 = -0.0056$
- the upper limits of $b_1 = b_1 + 2 \cdot SE(b_1) = 0.3524 + 2 \cdot 0.1790 = 0.7104$

That is, there is approximately a 95% chance that the interval $[0.042, 0.052]$ will contain the true value of b_1 . Similarly, the 95% confidence interval for b_0 can be computed as $b_0 \pm 2 \cdot SE(b_0)$.

To get this information, type:

```
> confint(relation)
              2.5 %      97.5 %
(Intercept) 2.29978152 9.7551079
x            -0.03148235 0.7362008
```

6. MODEL ACCURACY

- Goodness-of-fit

The overall quality of the linear regression fit can be assessed using the following three quantities, displayed in the model summary:

- a. The Residual Standard Error (RSE).
- b. The R-squared (R²)
- c. F-statistic

```
##      rse r.squared f.statistic  p.value
## 1  6.626   0.2168      3.876   0.06909
```

- a. Residual standard error (RSE).

The RSE (also known as the model sigma) is the residual variation, representing the average variation of the observation points around the fitted regression line. This is the standard deviation of residual errors. RSE provides an absolute measure of patterns in the data that can't be explained by the model. When comparing two models, the model with the small RSE is a good indication that this model fits the best the data. Dividing the RSE by the average value of the outcome variable will give you the prediction error rate, which should be as small as possible.

In our example, RSE = 6.626, meaning that the observed deaths values deviate from the true regression line by approximately 6.6 units in average. Whether or not an RSE of 6.6 units is an acceptable prediction error is subjective and depends on the problem context. However, we can calculate the percentage error.

In our data set, the mean value of deaths is 7.065, and so the percentage error is $6.6 \times 100\% / 7.065 = 93.42\%$.

```
> sigma(relation)*100/mean(covid19_2$Deaths)
[1] 93.82494
```

- b. R-squared and Adjusted R-squared:

The R-squared (R²) ranges from 0 to 1 and represents the proportion of information (variation) in the data that can be explained by the model. The adjusted R-squared adjusts for the degrees of freedom. The R² measures, how well the model fits the data. For a simple linear regression, R² is the square of the Pearson correlation coefficient.

A high value of R^2 is a good indication. However, as the value of R^2 tends to increase when more predictors are added in the model, such as in multiple linear regression model, you should mainly consider the adjusted R-squared, which is a penalized R^2 for a higher number of predictors.

- An (adjusted) R^2 that is close to 1 indicates that a large proportion of the variability in the outcome has been explained by the regression model.
- A number near 0 indicates that the regression model did not explain much of the variability in the outcome.
- In our dataset, the R-squared is 0.2168. This shows that newcase is a weak predictor as it only able to explain 21.68 % of the variation in deaths

c. F-Statistic:

The F-statistic gives the overall significance of the model. It assesses whether at least one predictor variable has a non-zero coefficient.

In a simple linear regression, this test is not really interesting since it just duplicates the information in given by the t-test, available in the coefficient table. In fact, the F test is identical to the square of the t test: $3.876 = (1.969)^2$. This is true in any model with 1 degree of freedom. A large F-statistic will correspond to a statistically significant p-value ($p < 0.01$).

In our example, the F-statistic equal 3.876 producing a p-value of 0.06909, which is less significant. As an overall, this shows that the factor we are using newcase is less relevant.

7. SUMMARY

These metrics give the overall quality of the model.

- RSE: Closer to zero the better
- R-Squared: Higher the better
- F-statistic: Higher the better