Proceedings of Science and Mathematics
Faculty of Science, Universiti Teknologi Malaysia
Vol. 8, 2022, page 102-108

Analysis COVID-19 Death Cases in Pulau Pinang using Multiple Linear Regression

Nurfarhaniza Ramlee, Norazlina Ismail*
Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 Johor Bahru, Malaysia
*Corresponding author: i-norazlina@utm.my

Abstract

COVID-19 is a serious health hazard for people all over the world including Malaysia. The objective of this study is to identify the significant factors affecting the number of COVID-19 death cases in Pulau Pinang using multiple linear regression. The variables included death cases, new case, total active cases,ICU cases, and cases requiring oxygen, which were all collected from June to August 2021. From the regression analysis by using backward elimination, the new case and ICU case has a statistically significant relationship with COVID-19 death cases. The other independent variables total active case and case need oxygen have no significant impact on death case. Overall, this research found that the new case and ICU case is the factor that affect new COVID-19 death cases in Pulau Pinang.

Keywords: COVID-19; Multiple Linear Regression; Regression Analysis

1. Introduction

COVID-19 is a serious health hazard for people all over the world. Its emergence has drastically altered people's everyday routines, and new societal issues are constantly emerging as a result ofits presence. Since mid-January 2021, there has been a dramatic increase in the number of infections, with around 4000 cases every day. Following that on June 9, 2022, Malaysia recorded 1887 new confirmed cases of COVID-19, reflecting an increase from over 1.6 cases on June 4, 2022 [1]. More than 51.82 million people have been infected, and almost 1280000 people have died as a result of the disease in 215 counties.

Malaysia has taken urgent steps to keep the spread of COVID-19 to a minimal number of cases and to stop the deaths that go along with it from 25 January 2020 until now. To control the spread of the disease, a restricted lockdown called Movement Control Order (MCO) was immediately implemented across the country. On March 18, 2020, the first Movement Control Orderin Malaysia, ordering the shutdown of all companies except those that provide essential services and commodities [2]. On the other hand, the Malaysian government intends to obtain COVID-19 vaccine from a variety of agencies and businesses to vaccinate at least 70% of the population.

Furthermore, the rate of infection had not seemed to slow down in most of the affected countries included Malaysia and it has not been stopped despite various efforts. Therefore, to ultimately combat the emerging COVID- 19 pandemic, the government has planned vaccination treatment, social distancing and ensuing lockdown. However, the effectiveness of the COVID-19 vaccination, social distancing as well as lockdown in maintaining pandemic control was unclear. Hence, it was for this reason that this research was carried out to analyse the risk factors that affecting the number of COVID-19 death cases in Pulau Pinang, Malaysia from June to August 2021.

Information about multiple linear regression in this study can be used as a reference in COVID-19 mitigation strategies or applied in other fields. This research aims to identify the significant factors affecting the number of COVID-19 death cases and perform a multiple linear regression model of COVID-19 cases in Pulau Pinang. The variables are obtained by General of Health Malaysia from June to August 2021.

2. Literature Review

Previous research intends to forecast new COVID-19 cases using multiple linear regression. The data that has been use are total confirmed cases, total active cases, total death cases and total positive cases. To display the trend of the affected cases, regression model techniques are appliedto the data set. According to the findings, 52290 active cases are expected across India by the 15thof August, and 9358 active cases in Odisha by the 15th of August [3].

A study of multiple linear regression model to identify the significant factors affecting the numberof confirmed COVID-19 case and the number of deaths per 100000. This study consists of several independent variable such as proportion of the population, population density, per-capita GDP, population of population with a college degree, population of population that over than 65 years old, monthly flights into the state before travel bans, party in control of the state government's office, proportion of the initial 35 days where distancing restrictions in place and change in mobility index. According to this study, population density is the most influential factor in all the models. Party control is a significant predictor for the number of fatalities and cases at the fifth week, but not at the thirteenth week, and per-capita GDP was significant in all models except the fifth week numberof deaths [4].

Lastly, from previous research multiple linear regression was used to find correlations between the spread of the COVID-19 outbreak and climate. Temperature and humidity have been shown to affect the transmission of epidemics in several studies, prompting this study to look into the global impact of environmental conditions on COVID-19. The number of daily confirmed cases, monthly average maximum temperature, monthly average minimum temperature, sea level pressure, wind-speed, elevation, rainfall, dew point temperature, and relative humidity were all used in the analysis. According to the findings, the activity of the COVID-19 has little correlation with elevation, sea level pressure, wind speed and rainfall [5]. The activity of the COVID-19 is correlatedwith temperature maximum, temperature minimum, dew point temperature and relative humidity. The COVID-19's activity was mostly influenced by temperature and humidity.

3. Methodology

3.1. Research Data

The data in this study are collected from June to August 2021 obtained from Ministry of Health (MOH) through the portal that known as Kenyataan Akhbar (kpkesihatan.com). The information taken from health services includes the daily number of COVID-19, the number of deaths, total active cases, ICU cases and cases need oxygen as shown in Table 1.

Table 1: Description of variables

Variables	Data Descriptive
$\boldsymbol{Y}$	**Death Case**
$\boldsymbol{X_1}$	**New Case**
$\boldsymbol{X_2}$	**Total Active Case**
$\boldsymbol{X_3}$	**ICU Case**
$\boldsymbol{X_4}$	**Case Need Oxygen**

In this research, SPSS software was used to analyse data using multiple linear regression. Part of the process involves checking the data to make sure the data can be analysed using multiple linear regression to give a valid result.

3.2. Descriptive Statistic

A summary statistic that summarizes the features of a collection of data, usually in the form of graph of table that giving the overall sample sizes in important subgroups. The mean, standard deviation

and variance were calculated from the daily COVID-19 data.

3.3. Pearson's Correlation Coefficient, r

Pearson correlation also known as Pearson product-moment correlation coefficient was used to measure the linear correlation between two sets of data. It is the ratio between the covariance of two variables and the product of their standard deviations.

$$r = \frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2(y_i-\bar{y})^2}} \quad (1)$$

Where:
$r =$ correlation coefficient
$x_i =$ values of the x-variables in a sample
$\bar{x} =$ mean of the values of the x-variable
$y_i =$ values of the y-variable in a sample
$\bar{y} =$ mean of the value of the y-variable

Correlation coefficients are scaled in range from -1 to +1 where 0 indicates that there is no linear correlation between two sets of data. The strength of association is determined by the correlation coefficient. The value of correlation coefficient r, lies between -1 to +1. The strength of association for different ranges is tabulated below (Table 2).

Table 2: Strength of association correlation coefficient

Range	Strength of association
Close to +1	Strong positive relationship
Close to -1	Strong negative relationship
Close to 0	No relationship

3.4. Model Assumptions

First step before applying linear regression, it is needed to go through few of analysis to make surethe independent variables are suitable for the regression. In order to build the best model of multiplelinear regression analysis, the assumption of multiple linear regression are checking the linearity, multicollinearity, values of residuals are independent, the variance of the residuals is constant and residual are normally distributed.

3.5. Multiple Linear Regression

The regression using a single independent variable is known as univariate regression analysis, whereas the regression using multiple independent variables is known as multivariate regression analysis or multiple regression analysis. Regression analysis is performed to determine the correlation between 2 or more variables having cause-effect relations and to make predictions for the topic by using the relation. Multiple regression is a highly adaptable technique for investigatingthe connection between a number of independent variables (or predictors) and a single dependentvariable (or criterion). The independent variables can be either quantitative or categorical.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon \quad (2)$$

Where:
$Y =$ the predicted value of the dependent variable or response variable
$X_1, X_2, \ldots, X_n =$ the value of independent variable or predictor variables
$\beta_0, \beta_1. \ldots, \beta_n =$ the regression parameter relating the mean of Y to $X_1, X_2, \ldots, X_n$
$\epsilon =$ the error terms

3.6. Backward Elimination

Multiple linear regression s a type of regression where the model depends on several independent variables. In this multiple linear regression model, the variable selection method will be backward elimination or backward deletion. Backward elimination consists of the following stepswhich is select a significant level (p-value less than 0.05) and fit the model with all possible predictors. The variables with the smallest p-value below a specific value (0.05) indicating statisticalsignificance is kept in the model.

4. Results and discussion

4.1. Descriptive Statistics

Table 3 shows the descriptive statistic for daily data COVID-19 cases in Pulau Pinang from June 2021 to August 2021.

Table 3: Descriptive statistics of COVID-19 data in Pulau Pinang

Variables	Mean	Std. Deviation	Variance
Death Case	5..76	7.851	61.634
New Case	816.08	511.565	261699.258
Total Active Case	157068.89	138671.510	42356.89
ICU Case	966.23	74.807	5596.024
Case Need Oxygen	478.80	42.886	1839.214

4.2. Correlation Coefficient

The degree of linear association between all variables is computed by Pearson's Correlation Coefficient. In this study, it can be observed that there is a positive correlation between death case and new case, total active case, ICU case and case need oxygen.

4.3. Model Assumption

A multiple linear regression model is said to be unsuitable for determining the relationship between the response variable and predictor variables if one of the assumptions is violated. Several types of statistical tests will be considered in this study, which are Durbin-Watson test, Breusch-Pagan-Godfrey test, Kolmogorov-Smirnov test and Variance Inflation Factor (VIF) test. The assumption underlying the multiple linear regression model are stated below:

1. The relationship between the response variable and the predictor variables is linear.
2. There is no autocorrelation in the residuals.
3. The variance of errors is constant.
4. The errors follow a normal distribution.
5. There is no multicollinearity in the data.

4.4. Multiple Linear Regression Model

Model 1:

According to the result shown in the Table 4, new case has value 0.001, total active case has value 0.002, ICU case has value 0.002 and case need oxygen has value 0.001 then for each 1-unit in the predictor variable, the outcome variable will increase by 0.001 for new case, 0.002 for total active case, 0.002 for ICU case and 0.001 for case need oxygen. there is a statistically significant association betweennew cases since there exist p-value less than 0.05, whereas other variables such as total active case, ICU case and case need oxygen have no significant impact on COVID-19 death cases in Pulau Pinang.

Table 4: Coefficient backward selection model 1

Variables	Coefficient	Std. Error	t-Statistic	p-value
(Constant)	-0.693	0.577	-1.200	0.234

New Case	0.001	0.000	5.661	0.000
Total Active Case	0.002	0.000	0.812	0.419
ICU Case	0.002	0.002	1.280	0.205
Case Need Oxygen	0.001	0.002	-0.177	0.860

$$Y = -0.693 + 0.001X_1 + 0.002X_2 + 0.002X_3 + 0.002X_4 + 0.001X_5$$

Where:
$Y =$ Death Case
$X_1 =$ New Case
$X_2 =$ Total Active Case
$X_3 =$ ICU Case
$X_4 =$ Case Need Oxygen

Model 2:
According to the result shown in the Table 4.6, new case has value 0.001, total active case has value 0.002 and ICU case has value 0.002, then for each 1-unit in the predictor variable, the outcome variable will increase by 0.001 for new case, 0002 for total active case and 0.002 for ICU case.

Table 5: Coefficient backward selection model 2

Variables	**Coefficient**	**Std. Error**	**t-Statistic**	***p*-value**
(Constant)	-0.653	0.528	-1.236	0.220
New Case	0.001	0.000	6.100	0.000
Total Active Case	0.002	0.000	0.825	0.412
ICU Case	0.002	0.001	2.942	0.004

$$Y = -0.653 + 0.001X_1 + 0.002X_2 + 0.002X_3$$

Where:
$Y =$ Death Case
$X_1 =$ New Case
$X_2 =$ Total Active Case
$X_3 =$ ICU Case

Model 3:
Consequently, it is necessary to proceed with the best subset regression modelformulation with only considers the most important predictor variables from the previous regression model, namely new case, and ICU case. The coefficients of the parameters and a summary of the best subset regression model as listed below (Table 6)

According to the result shown in the Table 6, new case has value 0.001 and ICU case has value 0.002, then for each 1-unit in the predictor variable, the outcome variable will increase by 0.001 for new case and 0.002 for ICU case.

Table 6: Coefficient backward selection model 3

Variables	**Coefficient**	**Std. Error**	**t-Statistic**	***p*-value**
(Constant)	-0.709	0.523	-1.356	0.179
New Case	0.001	0.000	6.477	0.000
ICU Case	0.002	0.001	3.0125	0.003

$$Y = -0.709 + 0.001X_1 + 0.002X_3$$

Where:

$Y =$ Death Case

$X_1 =$ New Case

$X_3 =$ ICU Case

According to Table 7, for Model 1 the R^2 value of 0.630 indicates that the predictor variables in the model such as new case, total active case, ICU case and case need oxygen account form 63% of the variance in the response variable (death case), while the remaining is explained by other factors that are not included in this analysis. In addition, Model 2 discovered that three predictor variables such as new case, total active case and ICU case accounted for 63% of the variance in the response variable (death case). While Model 3 indicates that new case and ICU case accounted for 62.6% of the variance in the response variable (death case). Lastly, this table can conclude that Model 3 is the best model rather than Model 1 and Model 2 since the value for adjusted R^2 in Model 3 much higher than Model 1 and Model 2.

Table 7: R, R-Squared and Adjusted R-Squared

Model	R	R-Squared	Adjusted R-Squared
1	0.794	0.630	0.609
2	0.794	0.630	0.614
3	0.791	0.626	0.616

Conclusion

Based on the main objective of this research, which is to identify the significant factors affecting the number of death case, these findings reveal that each of the predictor variables has a unique relationship with the new COVID-19 cases in Pulau Pinang. Overall, it can be concluded that new case and ICU case has a significant relationship towards death cases, while the other predictor variables such as total active case and case need oxygen have a negligible relationship with the number of COVID-19 death case in Pulau Pinang. Overall COVID- 19 death cases in Pulau Pinang is significantly affected by new case and ICU case. In future study, researchers should do this research with a larger sample size of data (monthly or yearly), collect more characteristic that could be the factor affecting the new COVID-19 death case and collect COVID-19 data every state in Malaysia that would provide a clearer picture of the analyze patterns

Acknowledgement

We thank the Ministry of Health Malaysia (MOH) for providing COVID-19 data in Pulau Pinang through the portal that known as Kenyataan Akhbar (kpkesihatan.com).

References

[1] Hirschmann, R. (2022, June 10). Malaysia: Covid-19 daily cases 2022. Statista. Retrieved June 14, 2022, from https://www.statista.com/statistics/1110785/malaysia- covid-19-daily-cases/

[2] Tang, K. H. D. (2020). Movement control as an effective measure against COVID-19 spread in Malaysia: an overview. *Journal of Public Health,* 1-4. https://dx.doi.org/10.1007%2Fs10389-020-01316-w

[3] Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model.

[4] Tenenholtz, J., George, F., & Gulati, S. (2021). Some multiple regression models for the number of COVID-19 cases and deaths in the United States. *International Journal of Statistics and Probability, 10(1*). https://doi.org/10.5539/ijsp.v10n1p28

[5] Lin, S., Fu, Y., Jia, X., Ding, S., Wu, Y., & Huang, Z. (2020). Discovering correlations between the COVID-19 epidemic spread and climate. *International Journal of Environmental Research and Public Health 2020, 17*(21), 7958. https://doi.org/10.3390/ijerph17217958