Available online at http://www.iorajournal.org/index.php/ijgor/index

International Journal of Global Operations Research

e-ISSN: 2722-1016
p-ISSN: 2723-1739

Vol. 4, No. 4, pp. 220-228, 2023

Forecasting Model of the Effect of Vaccination on the Number of Covid-19 Cases in West Java Province

Nazihah[1*], Julita Nahar[2], Sudradjat Supian[3]

[1*]*Mathematics Undergraduate Study Program, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, Indonesia*

[2,3]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran, Jatinangor, Indonesia*

**Corresponding author email: nazihah19001@mail.unpad.ac.id*

Abstract

COVID-19 is a virus caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) and was first discovered in Wuhan, Hubei, China. Apart from wearing masks and practicing social distancing, another way to prevent COVID-19 exposure is by administering the complete COVID-19 vaccine dosage, which includes both the first and second doses of the same vaccine, as well as an additional booster dose for enhanced immunity. Vaccination is a process within the body that makes a person highly resistant or protected against a disease, so that if they get infected with the disease later on, they either won't get sick or only experience mild symptoms. The purpose of this research is to build a forecasting model to determine the impact of vaccination on the number of COVID-19 cases in West Java Province and to predict the number of COVID-19 cases in West Java Province using multiple linear regression method. The research data used consists of the number of COVID-19 cases and the recipients of the first, second, and booster vaccine doses within the period from January 13, 2021, to January 22, 2023. The data was obtained from the official websites of Diskes Jabar and Vaksin Kemkes. The results of the statistical tests with a 5% alpha level indicate that among the first, second, and booster vaccines, only the first and second vaccines have a significant impact, meaning there is an effect of vaccination effectiveness on the number of COVID-19 cases in West Java Province. The forecasted number of COVID-19 cases in West Java Province is estimated to be around 3864 people affected by Covid-19 in week 107.

Keywords: Covid-19, vaccination, multiple linear regression

1. Introduction

Covid-19 is a virus caused by SARS-CoV-2 (Severe Acute Respiratory Syndrome Coronavirus 2) and was first discovered in Wuhan, Hubei, China (Rath et al. 2020). It caused chaos for humans by claiming 523,011 lives worldwide according to the World Health Organization. In general, the symptoms of Covid-19 are fever, cough, and shortness of breath. Other symptoms include difficulty breathing and chest pain.

The West Java Province Covid-19 pandemic occurred in March 2020, it is recorded for now that there are around 1.113 million total people who are confirmed positive for Covid-19 with the number of deaths around 13.769 thousand (Dashboard Jabar, 2023). One of the most effective efforts to reduce this pandemic is the administration of vaccines. Vaccines are one of the most reliable and cost-effective public health interventions ever implemented that save millions of lives each year (Rizky et al. 2021). According to the Indonesian Ministry of Health (2021), Covid-19 vaccination has four benefits: stimulating the immune system, reducing the risk of transmission, reducing the severe impact of the virus, and achieving herd immunity. The number of Indonesians who have received the first vaccine is 203,817,286 people, while those who have received the second vaccine are 174,816,333 people, those who have received the third vaccine are 69,711,156 people and those who have received the fourth vaccine are 1,903,265 people (Ministry of Health RI, 2021). Meanwhile, the total vaccination target until the final stage in Indonesia is 234,666,020 people. West Java is ranked 12th in the first vaccine achievement with a percentage of 86.18%, then for the second vaccine achievement it is ranked 7th with a rate of 76.24%, for the third vaccine it is ranked 6th with a rate of 36.99% and the fourth vaccine achievement is ranked 4th with a rate of 0.10% (Vaccine Dashboard, 2023).

Previous studies that have been conducted regarding forecasting Covid-19 cases include Ogundokun et al. (2020) predicts Covid-19 cases in Nigeria with the Y variable used as the number of active Covid-19 cases in Nigeria while the X variables used are travel history before and after lockdown, as well as the number of contacts made by Covid-19

patients. Rath et al. (2020) predicted active Covid-19 cases in Odisha and India with the Y variable used as the number of active Covid-19 cases in Odisha and India while the X variables used were the number of positive Covid-19 cases, Covid-19 recovery cases, and Covid-19 death cases. Argawu et al. (2021) predicted new Covid-19 cases using multiple linear regression in Ethiopia with the Y variable used as the number of new Covid-19 cases in Ethiopia and the X variables used were the number of days, daily laboratory tests, and the number of new patients from Addis Ababa city.

Based on the above background, this thesis is entitled Forecasting Model of the Effect of Vaccination on the Number of Covid-19 Cases in West Java Province. One way to model the data is to use multiple linear regression models with the help of the R script in data processing. Therefore, in this study, the authors intend to use multiple linear regression in modeling the forecasting effect of vaccination on the number of covid-19 cases in West Java Province.

2. The Theory of Forecasting Concepts

2.1. Forecasting

Forecasting is predicting the future as accurately as possible with all available information, such as historical data and knowledge of future events that may affect the forecast (Hyndman and Athanasopoulos, 2018). Forecasting is necessary to determine when an event will occur or a need arises, so that appropriate action can be taken (Makridakis et al. 1997). Forecasting is an important problem that covers many fields including business and industry, government, economics, environmental science, medicine, social science, politics, and finance (Montgomery et al. 2015). Forecasting methods are divided into two types based on data availability, namely qualitative and quantitative forecasting methods (Hyndman and Athanasopoulos, 2018). Based on the period of the problem, forecasting is classified into three types, namely short, medium, and long-term (Montgomery et al. 2015).

2.2. Correlation Coefficient

The correlation coefficient can be used to show the linear relationship between two continuous variables in a multiple linear regression model (Argawu et al. 2021). The estimated correlation coefficient formula is as follows:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2}\ \sqrt{\sum(Y_i - \bar{Y})^2}} \tag{1}$$

If the relationship between X and Y is perfect, the correlation coefficient value is 1, a correlation coefficient of -1 indicates a perfect inverse relationship, and a correlation coefficient of 0 means there is no relationship. A positive correlation coefficient is said to be a positive correlation, while a negative correlation coefficient is said to be a negative correlation.

2.3. Multiple Linear Regression

Multiple linear regression is a statistical method used to model the relationship between one dependent variable and two or more independent variables (Gujarati, 2003). The purpose of multiple linear regression is to understand the extent to which the independent variable affects the dependent variable and to make predictions based on that relationship. The multiple linear regression model is (Rath et al, 2020)

$$Y_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik} + e_i \tag{2}$$

where,

Y_i	: dependent variable
$b_0, b_1, b_2, \dots, b_k$	: estimated regression parameters
$X_{i1}, X_{i2}, \dots, X_{ik}$	: independent variable
e_i	: *error*

2.4. Ordinary Least Squares (OLS)

The parameter estimator aims to obtain a multiple linear regression model that will be used in the analysis. In this study, the method used to estimate the parameters of multiple linear regression models Ordinary Least Squares (OLS). According to Ogundokun et al. (2020), the multiple linear regression estimation model with k variables:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik} \tag{3}$$

because,

$$e_i = Y_i - \hat{Y}_i \tag{4}$$

so that,

$$e_i = Y_i - (b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_k X_{ik}) \tag{5}$$

The objective function of the Ordinary Least Squares (OLS) method, minimizing the squared error, can be expressed as follows:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2} - \cdots - b_k X_{ik})^2 \tag{6}$$

The estimation for regression parameters in matrix form can be formulated as follows:

$$\hat{b} = (X^T X)^{-1} X^T Y \tag{7}$$

2.5. Normality Test

The normality test is used to determine whether the model and its variables have a normal distribution or not (Theofani and Sediyono, 2022). Detecting whether the data is normal or not can be seen visually by looking at the distribution of data on the diagonal sources of the Normal Q-Q Plot graph and can also use the Kolmogorov-Smirnov test.

Hypothesis:

H_0 : error is normally distributed.

H_1 : error is not normally distributed.

If the p-value $> \alpha$, then H_0 is not rejected, meaning that the error is normally distributed.

2.6. Heteroscedasticity Test

Heteroscedasticity is the presence of inequality of variance of the residuals for all observations in the regression model. A good regression model fulfils the assumption of homoscedasticity where the variance value of the residuals (e) is the same for all observations of the regression model. The easiest way to detect the presence or absence of heteroscedasticity is to look at the scatterplot graph and can also be done with the Breusch-Pagan test.

Hypothesis:

H_0 : there are no symptoms of heteroscedasticity.

H_1 : there are symptoms of heteroscedasticity.

If the p-value $> \alpha$, then H_0 is not rejected, meaning that there are no symptoms of heteroscedasticity.

2.7. Multicollinearity Test

A multicollinearity test is needed to determine whether there is a correlation between independent variables. The best model is a model that does not have multicollinearity, detecting multicollinearity can be done by looking at the Variance Inflation Factor (VIF).

Hypothesis:

H_0 : the model does not have multicollinearity.

H_1 : the model occurs multicollinearity.

The formula for calculating the VIF value is as follows:

$$VIF = \frac{1}{1 - R_j^2}; j = 1,2, \dots, k \tag{8}$$

If the VIF $\leq$ 10 then H_0 is not rejected, meaning that the model does not occur multicollinearity. After the multicollinearity test is fulfilled then proceed to the autocorrelation test.

2.8. Autocorrelation Test

The autocorrelation test is conducted to determine whether there is a correlation between each independent and independent and dependent variable separately or in combination (Theofani and Sediyono, 2022). The test in this study uses the Durbin-Watson (DW) test.

Hypothesis:

H_0 : there is no autocorrelation problem.

H_1 : there is an autocorrelation problem.

If the p-value $> \alpha$, then H_0 is not rejected, meaning that there is no autocorrelation problem.

2.9. F-Test

The F test is needed to determine the significance of the influence between all independent variables on the dependent variable (Theofani and Sediyono, 2022).

H_0 : there is no effect of the independent variables simultaneously on the dependent variable.

H_1 : there is at least one independent variable that simultaneously affects the dependent variable.

The results of the F calculation can be compared with the F_{table} using a certain significance level. If $F_{count} > F_{table}$ at the α confidence level, then reject H_0, which means that the independent variables simultaneously affect the independent variables.

2.10. T-Test

The t-test is used to see the effect of each independent variable individually on the dependent variable (Bhattacharyya et al., 1979).

H_0 : there is no effect of the independent variable partially on the dependent variable.

H_1 : there is an effect of the independent variable partially on the dependent variable.

If the p-value $< \alpha$ then H_0 is rejected, which means that there is a partially significant effect of the independent variable on the independent variable.

2.11. Coefficient of Determination

The Coefficient of Determination (R^2) is carried out to determine the strength of the model in explaining the variables. The higher the R^2 value, the better the prediction model made in the study. According to Gujarati (2003), the formula for calculating the coefficient of determination is as follows:

$$R^2 = \frac{\sum\left(\hat{Y}_i - \bar{Y}\right)^2}{\sum(Y_i - \bar{Y})^2} \tag{9}$$

3. Materials and Methods

3.1. Materials

The objects in this study are data on the number of Covid-19 cases in West Java Province from 13 January 2021 to 22 January 2023 obtained through the official website https://diskes.jabarprov.go.id/covid19/data and vaccination data in West Java Province for everyday from 13 January 2021 to 22 January 2023 through the official website http://vaksin.kemkes.go.id.

3.2. Methods

The following are the steps taken to complete this research:

1. Inputting Covid-19 data (Y) as a dependent variable and then the first vaccine (X_1), second vaccine (X_2), and booster vaccine (X_3) as independent variables.
2. Calculation of the correlation coefficient using equation (1).
3. Building multiple linear regression models from these variables using equation (2).
4. Testing multiple linear regression models with classical assumption tests consisting of normality, heteroscedasticity, multicollinearity, and autocorrelation.
5. Testing the feasibility of the model using the F test, t-test, and coefficient of determination.
6. Forecasting the Covid-19 cases.
7. Interpretation of the forecasting model of the effect of vaccination on Covid-19 cases in West Java Province.

4. Results and Discussion

4.1 Correlation coefficient

The calculation of the correlation coefficient using equation (1) obtained the following results as Table 1.

Table 1: Correlation of Research Variables

	Y
X_1	-0.03269382
X_2	0.24757163
X_3	-0.01017952

Based on Table 1, it is known that the correlation coefficient value shows that there is a significant positive correlation between Covid-19 cases and the second vaccination and there is a negative correlation between Covid-19 cases and the first vaccination and booster vaccination.

4.2 Multiple Linear Regression Model

Multiple linear regression models using equation (2) obtained the following results as Table 2.

Table 2: Multiple linear regression results

	Estimation	Standard Error	t-value	Pr(> |t|)
Intercept	4289	2088.2	2.054	0.04378
X_1	-0.015082	0.0024834	$-6{,}073$	$6{,}072 \times 10^{-8}$
X_2	0.023029	0.00345	6,675	$5{,}134 \times 10^{-9}$
X_3	-0.027267	0.0017078	$-1{,}597$	0.11491
R^2		0.4166		
F_{count}		16.4		

Based on Table 2, the multiple linear regression model is obtained as follows:

$$\hat{Y} = 4{,}289 - 0.015082X_1 + 0.023029X_2 - 0.027267\,X_3 \tag{10}$$

4.3 Normality Test

The normality test uses the help of the Kolmogorov-Smirnov test and the Q-Q Plot. The results of the data normality test using the Q-Q Plot can be seen in Figure 1.

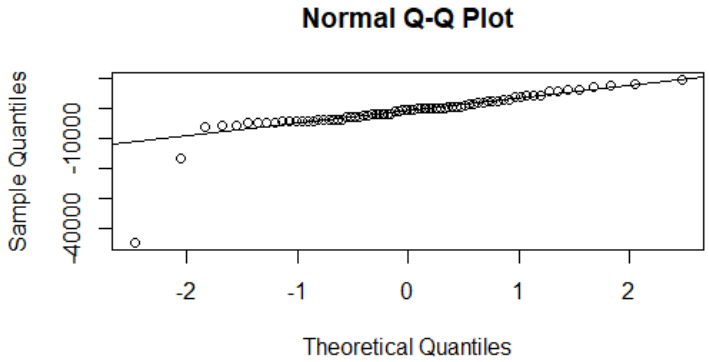

Figure 1: Q-Q Plot

Figure 1 is the result of the Q-Q plot which shows that the points spread around the diagonal line, so the error is normally distributed. The Q-Q plot obtained can also be validated using the Kolmogorov-Smirnov test with a significance level of 5% or 0.05. The results obtained from the Kolmogorov-Smirnov test with a value of $D = 0.083914$ and $p - value = 0.6436$ so that $p - value = 0.6436 > 0.05$ which means H_0 is not rejected. Based on the results of the Q-Q plot and Kolmogorov-Smirnov test, it can be concluded that the error is normally distributed so that the first assumption has been fulfilled.

4.4 Heteroscedasticity Test

The heteroscedasticity test uses the help of the Breusch-Pagan test and Scatter Plot. The results of the heteroscedasticity test using the Scatter Plot can be seen in Figure 2.

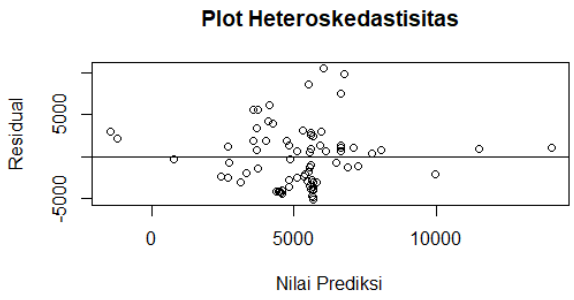

Figure 2: Heteroscedasticity Scatter Plot

Figure 2 is the result of the heteroscedasticity scatter plot which shows that there is no clear pattern and the residuals spread above and below the line, so all residuals from the regression equation model fulfil the assumption of homoscedasticity or no heteroscedasticity. The scatter plot obtained can be validated using the Breusch-Pagan test with a significance level of 5% or 0.05. The Breusch-Pagan test results obtained BP value = 4.5762, $df = 3, p - value = 0.2056$ so that $p - value = 0.2056 > 0.05$ which means H_0 is not rejected. Based on the results of the Scatter Plot and Breusch-Pagan test, it can be concluded that the second assumption has been fulfilled.

4.5 Multicollinearity Test

The VIF calculation uses equation (8). The results of the VIF value test can be seen in Table 3.

Table 3: VIF Test

X_1	X_2	X_3
5.288352	5.476242	1.635107

Based on Table 3, all VIF values of independent variables (X_1, X_2, and X_3) ≤ 10, which means that H_0 is not rejected, meaning that there is no multicollinearity.

4.6 Autocorrelation Test

The Durbin-Watson test results can be seen in Table 4.

Table 4: Durbin-Watson Test

Durbin Watson Value	*p-value*
1,979	0,4151

Based on Table 4 using the Durbin Watson test results in a p-value of 0.4151 so that the $p - value = 0.609 > 0.05$ which means H_0 is not rejected. It can be concluded that the model does not have autocorrelation so that the fourth assumption has been fulfilled.

4.7 F test

Based on Table 2, the value of F_{count} is 16.4. The F_{table} value with a significance level of 5% or 0.05 for a lot of data 74 is 2.74. This means that the $F_{count} > F_{table}$ value, so reject H_0, which means that there is a significant influence between the independent variables (first vaccine, second vaccine, and booster vaccine) simultaneously on positive cases of Covid-19 in West Java Province.

4.8 T-test

Based on Table 2, the p-value for variable X_1 is 6.072×10^{-8} with a significance level of 5% or 0.05. This means that the p-value for X_1 is smaller than the significance level, which means that H_0 is rejected so that the first vaccine has a partially significant effect on positive cases of Covid-19 in West Java Province. The p-value for variable X_2 is 5.134×10^{-9}, which means the p-value for X_2 is smaller than the significance level, which means H_0 is rejected, so the second vaccine has a partially significant effect on positive cases of Covid-19 in West Java Province. The p-value for variable X_3 is 0.11491, this means that the p-value for X_3 is greater than the significance level, which means that H_0 is accepted so that the booster vaccine does not have a partially significant effect on positive cases of Covid-19 in West Java Province.

4.9 Coefficient of Determination

Based on Table 2, the R-squared value is 0.4166. This value indicates that the selected independent variables have an influence of 41.66%, while the rest is influenced by other variables.

4.10 Linear Regression Model Based on t-Test

The following are the regression results based on a significant t-test as Table 5.

Table 5: Regression results based on t-test

	Estimation	Standard Error	t-value	$\Pr(>\|t\|)$
Intercept	3712.5	2434.4	1.525	0.1318
X_1	-0.01453	0.0024722	-5.877	1.288×10^{-7}
X_2	0.0218	0.0033629	6.482	1.082×10^{-8}
R^2		0.4009		

The linear regression model based on the t-test is as follows:

$$\widehat{Y_t} = 3712.5 - 0.01453X_1 + 0.0218X_2 \tag{11}$$

4.11 Forecasting Covid-19 Cases

The multiple linear regression equation models in equation (10) can be made to forecast the number of positive cases of COVID-19 in West Java Province each week. The following results of forecasting the number of COVID-19 cases in West Java Province in weeks 107 to 111 can be seen in Table 6.

Table 6: Covid-19 Forecasting Results

Weeks	Forecasting
107	3.864
108	3.894
109	3.893
110	3.950
111	3.978

4.12 Model Interpretation

Based on the discussion above, the regression model for forecasting weekly COVID-19 cases. The regression model can be seen in equation (11) from the results of estimating the parameters of the multiple linear regression model to determine the effect of vaccination using three variables, namely the first vaccine (X_1), second vaccine (X_2), and booster vaccine (X_3) which simultaneously and partially affect Covid-19 cases. COVID-19 cases are influenced by the first and second vaccines, The first vaccine is -0.01453 units which indicates a negative relationship between the number of recipients of the first vaccine and the number of COVID-19 cases so that if the number of recipients of the first vaccine increases, it is estimated that the number of positive Covid-19 cases will decrease. The second vaccine is 0.0218 units which indicates a positive relationship between the number of recipients of the second vaccine and the number of COVID-19 cases if the number of recipients of the second vaccine increases, it is estimated that the number of positive COVID-19 cases will increase.

5. Conclusion

Based on the results and discussion in this study, it can be concluded that:

a) In the estimated results of the multiple regression model by considering the t-test COVID-19 cases in West Java Province were influenced by the first and second vaccines by 40.09%.
b) Based on the results of forecasting COVID-19 cases in West Java Province using multiple linear regression, it is estimated that around 3864 people will be affected by COVID-19 cases in week 107.

References

Argawu, A. S., Gobebo, G., Bedane, K., Senbeto, T., Lemessa, R., & Galdassa, A. (2021). Prediction of COVID-19 New Cases Using Multiple Linear Regression Model Based on May to June 2020 Data in Ethiopia. *Journal of Pharmaceutical Research International*, *33*, 54–63. https://doi.org/10.9734/jpri/2021/v33i51a33468

Baykal, T. M., Colak, H. E., & Kılınc, C. (2022). Forecasting future climate boundary maps (2021–2060) using exponential smoothing method and GIS. *Science of the Total Environment*, *848*, 157633. https://doi.org/10.1016/j.scitotenv.2022.157633

Bhattacharyya, H. T., Kleinbaum, D. G., & Kupper, L. L. (1979). Applied Regression Analysis and Other Multivariable Methods. In *Journal of the American Statistical Association*, *74*(367). https://doi.org/10.2307/2287012

Gujarati, D. N. (2003). Basic Economitrics. In *McGraw-Hill Companies*.

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting : Principles and Practice* (2nd Editio). OTexts.

Makridakis, S., Wheelwright, S. C., & Hyndman, R. J. (1997). *Forecasting Methods and Applications*. New York: John Wiley & Sons.

Montgomery, D. C., Jennings, C. L., & Kulahci, M. (2015). *Introduction Time Series Analysis and Forecasting* (Second Edi). United States of America: John Wiley & Sons.

Ogundokun, R. O., Lukman, A. F., Kibria, G. B. M., Awotunde, J. B., & Aladeitan, B. B. (2020). Predictive modelling of COVID-19 confirmed cases in Nigeria. *Infectious Disease Modelling*, *5*, 543–548. https://doi.org/10.1016/j.idm.2020.08.003

Rath, S., Tripathy, A., & Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes and Metabolic Syndrome: Clinical Research and Reviews*, *14*(5), 1467–1474. https://doi.org/10.1016/j.dsx.2020.07.045

Rizky, A. A., Rahmawati, N., El-Faruqi, A., & ... (2021). Trials and Progress Prediction of Covid-19 Vaccine Using Linear Regression and SIR Parameters. *Indonesia Journal of Computing*, *6*(December), 35–46. https://doi.org/10.34818/indojc.2021.6.3.594

Theofani, G., & Sediyono, E. (2022). Multiple Linear Regression Analysis on Factors that Influence Employees Work Motivation. *SinkrOn*, *7*(3), 791–798. https://doi.org/10.33395/sinkron.v7i3.11453