**Bulletin of Applied Mathematics and Mathematics Education**
Volume 5, Number 2, October 2025 pp. 135-152
https://dx.doi.org/10.12928/bamme.v5i2.14559
e-ISSN 2776-1029, p-ISSN 2776-1002

# Application of Multiple Linear Regression Models for prediction of rice production yields in Central Lampung

**Nadia Fitri Yani[1], Luluk Muthoharoh[1*], Abdy Winardi[2]**

[1]Data Science Study Program, Science Faculty, Sumatera Institute of Technology, Lampung, Indonesia
[2]Department of Food Security, Food Corps, and Holticulture of Lampung Province, Bandar Lampung, Indonesia

*Corresponding E-mail: luluk.muthoharoh@sd.itera.ac.id

ARTICLE INFO

**Article History**
Received 24 September 2025
Revised 15 October 2025
Accepted 31 January 2026

**Keywords**
National food security,
Multiple linear regression,
Rice production.

**How to cite this article:**
Yani, N. F., Muthoharoh, L., & Winardi, A. (2025). Application of Multiple Linear Regression Models for prediction of rice production yields in Central Lampung. *Bulletin of Applied Mathematics and Mathematics Education*, *5*(2), 135-152.

ABSTRACT

Rice production is a crucial component of agricultural sustainability and food security in Indonesia, particularly in Central Lampung. This study aims to analyze the influence of planting area and harvested area on rice production using a multiple linear regression approach. The analysis employs secondary time-series data and applies an ordinary least squares (OLS) method with a logarithmic transformation of the dependent variable to address heteroskedasticity issues. Descriptive statistics and classical assumption tests, including normality, multicollinearity, heteroskedasticity, and autocorrelation tests, were conducted to ensure model validity. The results indicate that harvested area has a statistically significant positive effect on rice production, while planting areas shows a negative but statistically insignificant effect. The regression model demonstrates strong explanatory capability with an R-squared value of 81.27% and is statistically significant based on the F-test. Model evaluation using in-sample error metrics yields a Mean Absolute Error (MAE) of 19,344.89, a Root Mean Squared Error (RMSE) of 46,738.41, and a Mean Absolute Percentage Error (MAPE) of 48.20%, indicating that the model effectively captures general production trends but has limited accuracy for precise quantitative forecasting. These findings suggest that harvested area plays a dominant role in determining rice output, while further improvements in predictive performance may be achieved by incorporating additional explanatory variables and exploring alternative modeling techniques.



## Introduction

Rice production remains a critical component of food security in Indonesia, particularly in major production regions such as Lampung Province. In recent years, increasing demand driven by population growth and urbanization has intensified the need for reliable production planning supported by quantitative modeling approaches (Ambya et al., 2022). The Lampung Province Food Security, Food Crops and Horticulture Office plays an important role in formulating and managing agricultural policies and programs to increase production and farmers' welfare. Food commodities that continue to increase in production to support national food security include rice, corn,

soybeans, cassava, peanuts, and green beans. Among these commodities, rice has an important role because most people still depend on rice as the main food source. Central Lampung Regency is one of the main rice-producing areas in the province, contributing significantly to regional and national rice supply (Abdurrazak et al., 2019).

Data from the Lampung Province Central Bureau of Statistics (BPS) in 2023 showed a significant increase in rice production. Rice production increased by 2.59%, equivalent to an additional 69.74 thousand tons of milled dry grain (MDG), bringing total production to around 2.76 million tons of MDG. This figure reflects positive progress compared to the previous year's production of 2.69 million tons of MDG in 2022. This increase shows the successful implementation of various agricultural programs that encourage land productivity and the use of more efficient agricultural technology (Lampung, 2024). In particular, Central Lampung Regency experienced a remarkable surge in production in 2023 with total production reaching 608.01 thousand tons, much higher than the 2022 achievement of only 101.61 thousand tons. This significant increase emphasizes Central Lampung's role as one of the main rice producing regions in the province. This increase is not only important in supporting regional food security, but also has a direct impact on the welfare of the local community. Considering the increasing demand for rice due to population growth and urbanization, increasing rice production in Central Lampung is very important to ensure adequate and stable food availability in local and national markets (Herliana et al., 2025). This study aims to apply multiple linear regression models to predict rice production in Central Lampung Regency. This region was chosen because of its large contribution to overall rice production in Lampung Province. As one of the national food barns, Lampung, especially Central Lampung, has an important role in meeting the food needs of the Indonesian people.

This model will consider factors that affect rice production, such as harvest area and planting area (Apriyana et al., 2023). Several studies have applied statistical and econometric models to analyze rice production in Indonesia. Multiple linear regression has been widely used to predict production levels and evaluate influencing factors, such as planting area, harvested area, and climatic variables (Swarbawa et al., 2023). For instance, Kharisma S et al. (2025) developed Rice Production Prediction Model Based on Rainfall and Temperature Using Multiple Linear Regression. The results of the study show that the multiple linear regression model is able to predict rice production with a relatively low error rate, so it can be used as an analytical tool to estimate potential harvest yields. Similarly, Nababan and Nugraha (2024) examined the influence of climatic factors and harvested area on rice production in Sumatra and found that harvested area had a significant effect, while climate variables were not statistically significant (Nababan & Nugraha, 2024).

However, despite the growing number of studies, there is still a lack of research that explicitly compares the relative influence of planting area and harvested area within a single, highly productive region using a validated quantitative framework. Most previous studies either focus on broader regional scales or emphasize climatic variables, without sufficiently examining whether planting area or harvested area plays a more dominant role in determining rice production outcomes at the local level. This gap is particularly relevant for Central Lampung Regency, where rapid production growth suggests structural changes in land use and harvesting efficiency.

Therefore, this study aims to analyze the influence of planting area and harvested area on rice production in Central Lampung Regency using a multiple linear regression model. By focusing on a major rice-producing region and evaluating factor dominance through quantitative analysis, this research contributes empirical evidence to support more effective agricultural planning and land management policies at the regional level.


10.12928/bamme.v5i2.14559

## Method

This research uses primary data obtained directly from the Lampung Province Food Security, Food Crops and Horticulture Office. The data collected includes information on planting area (ha), harvest area (ha), and rice production (tons) from the Central Lampung Regency area. The data was collected through observation and data entry from relevant agency documents during the implementation of Practical Work carried out on July 1, 2024 to August 1, 2024 at the Lampung Province Food Security, Food Crops and Horticulture Office located on Jl. ZA. Pagar Alam No.1, Rajabasa, Bandarlampung. The agency's beautiful environment and adequate facilities support the process of collecting and processing data effectively

### Materials

Data processing in this study begins with the data selection process, which selects data that is relevant and in accordance with the needs of the analysis, namely data on planting area, harvest area, and rice production in Central Lampung Regency. The next stage is data cleaning to ensure data quality is maintained, such as removing duplicate data, handling missing values, and harmonizing data formats to be consistent (Rachman et al., 2024; Swarbawa et al., 2023). After the cleaning process, data transformation is performed to convert the raw data into a format that is more ready for analysis. This transformation includes data normalization, value aggregation, and changing the data type if needed (Hendrastuty, 2024). The next stage is data mining to find patterns or relationships between variables.

### Data Analysis

Data analysis in this study was conducted using a multiple linear regression approach to examine the effect of planted area and harvested area on rice production in Central Lampung Regency. All data processing was performed using R statistical software.

*Descriptive Statistics*

The initial stage of the analysis was conducted by presenting descriptive statistics for all research variables, including minimum, maximum, average (mean), and standard deviation values. These descriptive statistics aim to provide a general overview of the data characteristics and detect potential anomalies before conducting inferential analysis (Kaur et al., 2018).

*Multiple Linear Regression Model*

The method used in the analysis is multiple linear regression, which aims to determine the effect of planting area and harvest area on rice production. This method allows simultaneous analysis of two independent variables on one dependent variable (Agusta et al., 2024). The multiple linear regression model used is formulated as follows:

$$Y = a + b_1X_1 + b_2X_2 \quad (1)$$

Description:

$Y$ : rice production

$X_1$ : planting area

$X_2$ : harvest area

$a$ : constant

$b_1, b_2$ : regression coefficient of each independent Variable

The value of the coefficients $a$ and $b$ can be determined using the following formula:

$$b_1 = \frac{\sum(X_1Y)\sum(X_2^2) - \sum(X_2Y)\sum(X_1X_2)}{\sum(X_1^2)\sum(X_2^2) - (\sum X_1X_2)^2} \quad (2)$$

$$b_2 = \frac{\sum(X_2Y)\sum(X_1^2) - \sum(X_1Y)\sum(X_1X_2)}{\sum(X_1^2)\sum(X_2^2) - (\sum X_1X_2)^2} \quad (3)$$

$$a = \bar{Y} - b_1\bar{X}_1 - b_2\bar{X}_2 \quad (4)$$

Parameter estimation is done using the Ordinary Least Squares (OLS) method which minimizes the sum of squared errors.

*Model and Coefficient Significance Test*

To assess the overall feasibility of the regression model, an F-test was conducted to determine whether the independent variables simultaneously have a significant effect on the dependent variable (Rachman et al., 2024). Next, a t-test was used to test the significance of each regression coefficient partially, thus determining the individual contributions of planted area and harvested area to rice production.

*Classical Regression Assumption Test*

To ensure the regression estimation results are BLUE (Best Linear Unbiased Estimator), the following classical regression assumptions were tested (ELSAYIR, 2024; Navianti et al., 2023).

*Residual Normality Test*

Residual normality was tested using the Shapiro–Wilk test to ensure that the model residuals were normally distributed. Residual normality was assessed using the Shapiro–Wilk test. (Midway & White, 2025)

*Heteroscedasticity Test*

To detect non-constant residual variance, the Breusch–Pagan test was used. The model was deemed to meet the homoscedasticity assumption if there were no significant indications of heteroscedasticity. The presence of heteroscedasticity was examined using the Breusch–Pagan test.

*Multicollinearity Test*

Multicollinearity between independent variables was tested using the Variance Inflation Factor (VIF). A VIF value below a certain threshold indicates a lack of high correlation between the independent variables. Multicollinearity among independent variables was evaluated using the Variance Inflation Factor (VIF).

*Autocorrelation Test*

Given the time series nature of the data used, a Durbin–Watson test was performed to detect residual autocorrelation. Given the time-series structure of the data, residual autocorrelation was tested using the Durbin–Watson statistic.


10.12928/bamme.v5i2.14559

*Model Performance Evaluation*

The model's predictive performance was evaluated using several error metrics. Mean Absolute Percentage Error (MAPE) was used to measure the relative error of the predictions in percentage form, making it easily interpretable by policymakers. To strengthen the evaluation, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) were also used as additional metrics that reflect the absolute error of the predictions in the original data units.

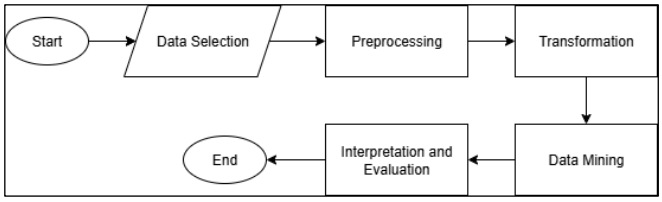



**Figure 1.** Research Flowchart

$$MAPE = \frac{1}{n}\sum_{I=1}^{n}\left|\frac{y_i - \hat{y}_i}{\hat{y}_i}\right| x100\% \quad (5)$$

Description:

$y_i$ : Actual value at index i

$\hat{y}_i$ : Predicted value at index i

$n$ : Amount of the data

$i$ : Data indexing

The smaller the MAPE value, the more accurate the prediction model used (Fitri & Nugraha, 2024; Muharram et al., 2023; Santoso et al., 2022). This evaluation process aims to ensure that the multiple linear regression model is not only theoretically relevant but also reliable in practical application in agriculture, particularly in predicting rice production based on historical data of planting area and harvest area in Central Lampung.

## Results and Discussion

The research data is a time series data covering the period from January 2022 to December 2023, thus depicting production conditions sequentially over a two-year observation period (see Table 1).

**Table 1.** Rice Production Dataset Used

| Period | Planting Area | Harvest area | Production Results |
|---|---|---|---|
| January-2022 | 32311,89 | 3779 | 18724 |
| February-2022 | 17327,79 | 4225 | 22183 |
| March-2022 | 4581,85 | 14330 | 81146 |
| April-2022 | 1400,52 | 27393 | 158344 |
| May-2022 | 8757,06 | 17881 | 100059 |
| June-2022 | 18377,81 | 1853 | 10413 |
| July-2022 | 6829,2 | 1150 | 6464 |

| Period | Planting Area | Harvest area | Production Results |
|---|---|---|---|
| August-2022 | 1562,6 | 7628 | 42545 |
| September-2022 | 1503,62 | 15162 | 82339 |
| October-2022 | 3054,13 | 6232 | 33653 |
| November-2022 | 13220,04 | 721 | 3834 |
| December-2022 | 25695,48 | 1257 | 6897 |
| January-2023 | 20516,62 | 1871 | 11083 |
| February-2023 | 11818,14 | 7173 | 38045 |
| March-2023 | 6664,56 | 16631 | 97476 |
| April-2023 | 4587 | 23030 | 136199 |
| May-2023 | 14265 | 13477 | 73888 |
| June-2023 | 15761 | 5022 | 27775 |
| July-2023 | 9181 | 4573 | 24589 |
| August-2023 | 4303 | 14044 | 77403 |
| September-2023 | 680 | 12353 | 66864 |
| October-2023 | 1048 | 7344 | 40341 |
| November-2023 | 2682 | 2253 | 12373 |
| December-2023 | 22868 | 359 | 1972 |

Descriptive statistics are shown in Figure 2. The planting area variable has a minimum value of 680 and a maximum of 23,312, with an average value of 10,375 and a median of 7,793. The difference between the average and median values indicates that the distribution of planting area tends to be asymmetrical and skewed to the right, indicating that some periods have significantly larger planting areas than others.

```
     Period                    Planting Area    Harvest area    Production Results
Min.   :2022-01-01 00:00:00   Min.   :  680   Min.   :  359   Min.   :  1972
1st Qu.:2022-06-23 12:00:00   1st Qu.: 2961   1st Qu.: 2158   1st Qu.: 12050
Median :2022-12-16 12:00:00   Median : 7793   Median : 6702   Median : 35849
Mean   :2022-12-16 00:00:00   Mean   :10375   Mean   : 8739   Mean   : 48942
3rd Qu.:2023-06-08 12:00:00   3rd Qu.:16153   3rd Qu.:14116   3rd Qu.: 78339
Max.   :2023-12-01 00:00:00   Max.   :32312   Max.   :27393   Max.   :158344
```

**Figure 2**. Descriptive Statistic

The harvested area variable shows a relatively similar pattern, with a minimum value of 359 and a maximum of 27,393. The average value of 8,739, which is higher than the median of 6,702, indicates significant variation in harvested area between periods. This reflects that not all planting periods produce the same harvest, which can be influenced by seasonal factors, weather, and agricultural policies. Meanwhile, the production results variable has a very wide range of values, with a minimum value of 1,972 and a maximum of 158,344. The average value of 48,942, which is significantly greater than the median of 35,849, indicates that the distribution of production results is also asymmetrical and tends to be skewed to the right. This condition indicates a surge in production during certain periods, which substantially contributes to the increase in average production values. Overall, descriptive statistics indicate significant fluctuation and variability


10.12928/bamme.v5i2.14559

across all study variables, therefore testing the assumptions of classical regression is necessary to ensure the validity and reliability of the regression model estimation results used.

The multiple linear regression estimation results indicate that the developed model performs very well in explaining variations in production yields, as demonstrated by the OLS results in Figure 3. The coefficient of determination ($R^2$) of 0.998 and the adjusted $R^2$ of 0.9978 indicate that approximately 99.8% of the variation in the production yield variable can be explained by the planted area and harvested area variables. This indicates that the model has a very high level of goodness of fit. The simultaneous significance test using the F-test yielded an F-statistic of 5.195 with a significance value of $p < 2.2\times10^{-16}$, confirming that the independent variables collectively have a significant effect on production yield.

```
Residuals:
    Min      1Q  Median      3Q     Max
-3535.8 -1818.7    96.8  1260.4  4635.7

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.027e+03  1.125e+03  -1.803   0.0858 .
PlantingArea   3.193e-02  5.751e-02   0.555   0.5846
HarvestedArea  5.794e+00  6.761e-02  85.707   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2048 on 21 degrees of freedom
Multiple R-squared:  0.998,     Adjusted R-squared:  0.9978
F-statistic:  5195 on 2 and 21 DF,  p-value: < 2.2e-16
```

**Figure 3**. OLS Estimationd

Partially, the t-test results indicate that the harvested area variable has a positive and highly significant effect on production yield, with a coefficient value of 5.794 and a p-value < 0.001. This means that each one-unit increase in harvested area will increase the average production yield by 5.794 units, assuming other variables remain constant. In contrast, the planting area variable has a positive coefficient of 0.0319, but it is not statistically significant at the 5% level (p = 0.5846). This indicates that planting area does not significantly contribute to yield after harvested area is accounted for in the model.

The constant (intercept) is negative at -2.027 and significant at the 10% level (p = 0.0858), indicating that when both planting area and harvested area are zero, yield is estimated to be negative. Although this condition has no direct practical significance, the presence of a constant is necessary to maintain the mathematical equilibrium of the model. Furthermore, the residual standard error of 2.048 indicates that the average model estimation error based on the observed data is relatively small, confirming the model's accuracy in representing the relationship between the independent variables and yield.

Overall, these results indicate that harvested area is the primary factor determining yield, while planting area does not have a direct significant effect once harvested area is accounted for. These findings indicate that the effectiveness of the harvesting process and actual harvest realization play a greater role in determining production output than simply the size of the planned planting area.

The results of the residual normality test using the Kolmogorov–Smirnov method show a D statistic value of 0.12141 with a significance value (p-value) of 0.8299 as shown in Figure 4. A p-value that is much greater than the 5% significance level (α = 0.05) indicates that there is insufficient statistical evidence to reject the null hypothesis. Thus, it can be concluded that the

residuals of the regression model are normally distributed. Fulfillment of the residual normality assumption indicates that the regression parameter estimates obtained using the Ordinary Least Squares (OLS) method are valid, so that the coefficient significance test and statistical inference carried out can be trusted.

```
        Exact one-sample Kolmogorov-Smirnov test

data:  scale(residuals_model)
D = 0.12141, p-value = 0.8299
alternative hypothesis: two-sided
```

**Figure 4.** Residual Normality Test Using the Kolmogorov–Smirnov

Based on the results of the Shapiro–Wilk normality test on the residual model, the W statistic value was obtained at 0.96821 with a p-value of 0.6231 as shown in Figure 45. A p-value greater than the general significance level (α = 0.05) indicates that there is insufficient evidence to reject the null hypothesis, namely that the residuals are normally distributed. Thus, it can be concluded that the residual model meets the normality assumption. This indicates that the regression model used is feasible in terms of the residual normality assumption, so that the results of the estimation and further statistical testing can be considered valid and can be interpreted with more confidence.

```
        Shapiro-Wilk normality test

data:  residuals_model
W = 0.96821, p-value = 0.6231
```

**Figure 5**. Shapiro–Wilk Normality Test on The Residual Model

The results of the multicollinearity test using the Variance Inflation Factor (VIF) in Figure 6 show that the VIF values for the PlantingArea and HarvestedArea variables are 1.424, respectively. These VIF values are well below the common limit used in regression analysis, which is 10 (and even below the conservative limit of 5). This indicates that there are no significant multicollinearity problems among the independent variables in the regression model. Thus, each independent variable provides relatively independent information in explaining the variation in the dependent variable ProductionResults, so that the estimated regression coefficients can be considered stable and can be interpreted individually without any distortion due to the strong linear relationship between the independent variables.

```
PlantingArea HarvestedArea
    1.424471      1.424471
```

**Figure 6.** Variance Inflation Factor (VIF) value

Based on the results of the studentized Breusch–Pagan test, the BP statistic value was 8.4075 with degrees of freedom (df) = 2 and a p-value of 0.01494, as shown in Figure 7. Because the p-value is smaller than the 0.05 significance level, the null hypothesis stating that the residual variance is constant (homoscedasticity) is rejected. Thus, it can be concluded that there is an indication of heteroscedasticity in the regression model. This means that the error variance is not constant across all observations, so one of the classical assumptions of linear regression is not met. This condition can cause the estimation of standard errors to be inefficient and the statistical tests (t-test and F-test) to be less reliable, so further treatment needs to be considered, such as data transformation or the use of regression methods that are robust to heteroscedasticity.


10.12928/bamme.v5i2.14559

```
        studentized Breusch-Pagan test

data:  model
BP = 8.4075, df = 2, p-value = 0.01494
```

**Figure 7.** Breusch–Pagan test

After performing a logarithmic transformation on the dependent variable, the regression model is re-estimated, and all classical assumptions are re-tested. The estimation results of the multiple linear regression model with logarithmic transformation on the dependent variable in Figure 8 indicate that the model has good explanatory power for variations in Production Results. The Adjusted R-squared value of 0.7949 indicates that approximately 79.49% of the variation in production results can be explained jointly by the Planting Area and Harvested Area variables, while the remainder is influenced by other factors outside the model. The simultaneous significance test (F-test) yielded an F-value of 45.57 with a p-value of $2.297\times10^{-8}$, which is smaller than the 5% significance level. Therefore, it can be concluded that the overall regression model is significant and suitable for further analysis.

The estimation results of the multiple linear regression model with logarithmic transformation on the dependent variable in Figure 8 indicate that the model has good explanatory power for variations in Production Results. The Adjusted R-squared value of 0.7949 indicates that approximately 79.49% of the variation in production results can be explained jointly by the Planting Area and Harvested Area variables, while the remainder is influenced by other factors outside the model. The simultaneous significance test (F-test) yielded an F-value of 45.57 with a p-value of $2.297\times10^{-8}$, which is smaller than the 5% significance level. Therefore, it can be concluded that the overall regression model is significant and suitable for further analysis.

Partially, the t-test results indicate that the Harvested Area variable has a positive and significant effect on production results at the 5% significance level, with a coefficient value of $1.267\times10^{-4}$ (p-value = $4.34\times10^{-7}$). This indicates that increasing harvested area tends to increase production yields, assuming other variables remain constant. Conversely, the PlantingArea variable has a negative coefficient of $-1.999\times10^{-5}$, but it is not statistically significant (p-value = 0.197), so its effect on production yield is not strong enough to be empirically proven in this model.

The Residual Standard Error value of 0.5337 indicates that the deviation of the model prediction from the actual value is relatively small on a logarithmic scale, indicating fairly good model estimation accuracy. Furthermore, the significant intercept coefficient indicates that when the independent variable value approaches zero, the model still produces a statistically significant baseline production value. With the classical assumptions met after the logarithmic transformation, the coefficient estimation results and significance testing in this model can be considered valid and reliable as a basis for drawing research conclusions.

The results of the residual normality test using the Kolmogorov–Smirnov method on the standardized residuals from the logarithmic regression model show a D statistic value of 0.20658 with a significance value (p-value) of 0.2241 in Figure 9. A p-value greater than the 5% significance level ($\alpha = 0.05$) indicates that there is insufficient statistical evidence to reject the null hypothesis that the residuals are normally distributed. Thus, it can be concluded that the residuals of the regression model have met the normality assumption.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.386e+00  2.931e-01  32.025  < 2e-16 ***
PlantingArea  -1.999e-05  1.499e-05  -1.334    0.197
HarvestedArea  1.267e-04  1.762e-05   7.192 4.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5337 on 21 degrees of freedom
Multiple R-squared:  0.8127,    Adjusted R-squared:  0.7949
F-statistic: 45.57 on 2 and 21 DF,  p-value: 2.297e-08
```

**Figure 8.** The Estimation Results of The Multiple Linear Regression Model with Logarithmic Transformation

Fulfillment of this assumption indicates that the parameter estimation and significance testing of the regression coefficients carried out on the logarithmic model are valid, so that the resulting statistical inferences are reliable.

```
        Exact one-sample Kolmogorov-Smirnov test

data:  scale(res_log)
D = 0.20658, p-value = 0.2241
alternative hypothesis: two-sided
```

**Figure 9.** Kolmogorov–Smirnov from the Logarithmic Regression Model

The results of the residual normality test using the Shapiro–Wilk method showed a W statistic of 0.87129 with a significance value (p-value) of 0.005577, as shown in Figure 10. A p-value smaller than the 5% significance level (α = 0.05) indicates that the null hypothesis stating that the residuals are normally distributed is rejected. Therefore, based on the Shapiro–Wilk test, the residuals of the logarithmic regression model do not fully meet the assumption of normality.

However, the difference in results between the Shapiro–Wilk and Kolmogorov–Smirnov tests require careful consideration. The Shapiro–Wilk test is known to be very sensitive to relatively small sample sizes and the presence of small deviations or outliers in the residual distribution. Meanwhile, the Kolmogorov–Smirnov test on standardized residuals showed insignificant results, so in general, the residual distribution can still be considered approximately normal. Furthermore, based on the Central Limit Theorem, with sufficient sample size and a model that has undergone a logarithmic transformation and meets the assumptions of homoscedasticity and non-multicollinearity, the regression coefficient estimates and significance tests (t-test and F-test) remain consistent and reliable.

Therefore, although the Shapiro–Wilk test indicates a deviation from normality, this result does not substantially weaken the suitability of the regression model used. The logarithmic regression model can still be used for further analysis, provided that the results are interpreted with caution and supported by the fulfillment of other classical assumptions.

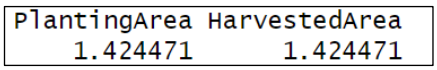
```
        Shapiro-Wilk normality test

data:  res_log
W = 0.87129, p-value = 0.005577
```

**Figure 10.** Shapiro–Wilk from the Logarithmic Regression Model


10.12928/bamme.v5i2.14559

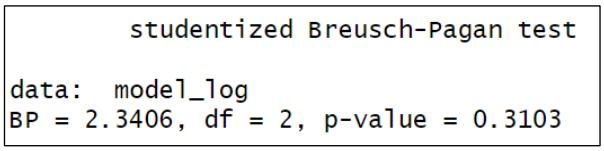
The results of the multicollinearity test using the Variance Inflation Factor (VIF) in the regression model with logarithmic transformation show that the VIF values for the PlantingArea and HarvestedArea variables are 1.424, respectively. These VIF values are shown in Figure 11. These values are well below the commonly used threshold of 10 (even below the conservative limit of 5), so it can be concluded that there is no multicollinearity problem between the independent variables in the model. This indicates that the linear relationship between the independent variables is relatively weak, so that each variable can explain the variation in the dependent variable ProductionResults independently. Thus, the estimated regression coefficients in the logarithmic model are stable and can be interpreted individually without any distortion due to strong correlations between the independent variables.



```
PlantingArea HarvestedArea
    1.424471      1.424471
```

**Figure 11**. Variance Inflation Factor (VIF) in The Regression Model with Logarithmic Transformation

The results of the heteroscedasticity test in Figure 12 using the studentized Breusch–Pagan method on the regression model with logarithmic transformation show a BP statistic value of 2.3406 with degrees of freedom (df) = 2 and a significance value (p-value) of 0.3103. A p-value greater than the 5% significance level (α = 0.05) indicates that there is insufficient statistical evidence to reject the null hypothesis that the residual variance is constant. Thus, it can be concluded that the logarithmic regression model has met the homoscedasticity assumption. These results indicate that the logarithmic transformation on the dependent variable is effective in overcoming the heteroscedasticity problem detected in the initial model, so that the regression coefficient estimation and significance testing in this model can be considered valid and reliable.

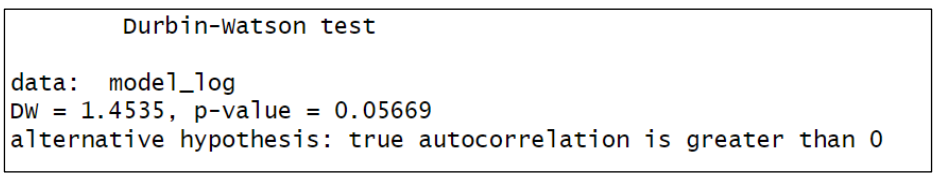


```
        studentized Breusch-Pagan test

data:  model_log
BP = 2.3406, df = 2, p-value = 0.3103
```

**Figure 12.** Breusch–Pagan method on the Regression Model with Logarithmic Transformation

The results of the autocorrelation test in Figure 13 using the Durbin–Watson method on a regression model with logarithmic transformation show a DW statistical value of 1.4535 with a significance value (p-value) of 0.05669. The p-value is slightly greater than the 5% significance level (α = 0.05) indicating that there is insufficient statistical evidence to reject the null hypothesis stating that there is no positive autocorrelation in the residuals. Thus, statistically no indication of significant autocorrelation was found in the logarithmic regression model. Although the DW value is slightly below the ideal value approaching 2, this result is still acceptable within the tolerance limit, so the assumption of residual independence can be considered fulfilled.



```
        Durbin-Watson test

data:  model_log
DW = 1.4535, p-value = 0.05669
alternative hypothesis: true autocorrelation is greater than 0
```

**Figure 13.** Durbin–Watson method on a Regression Model with Logarithmic Transformation

The significance test results for the regression coefficients in Figure 14 using robust standard errors show that the intercept coefficient has an estimated value of 9.3856 and is statistically significant at the 1% level (p-value < $2.2\times10^{-16}$). This indicates the existence of a logarithmic baseline for production yields when all independent variables are assumed constant. Partially, the HarvestedArea variable has a positive coefficient of $1.2672\times10^{-4}$ and is statistically significant at the 1% level (p-value = $2.93\times10^{-6}$). This finding indicates that increasing harvested area significantly contributes to increased production yields, even after standard error correction to address potential heteroscedasticity.

Conversely, the PlantingArea variable has a negative coefficient of $-1.9987\times10^{-5}$, but it is not statistically significant (p-value = 0.2043). This indicates that planted area does not have a strong enough influence on production yield in this logarithmic regression model. The consistency of the t-test results between the OLS estimate and the estimate with robust standard errors indicates that the conclusions regarding the influence of each independent variable are stable and insensitive to potential violations of the heteroscedasticity assumption. Thus, the HarvestedArea variable is the main factor that significantly influences production yield, while PlantingArea does not make a significant contribution partially.

```
t test of coefficients:

                 Estimate  Std. Error t value  Pr(>|t|)
(Intercept)    9.3856e+00  2.7624e-01 33.9759 < 2.2e-16 ***
PlantingArea  -1.9987e-05  1.5255e-05 -1.3102    0.2043
HarvestedArea  1.2672e-04  2.0074e-05  6.3125  2.93e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Figure 14.** The significance test results for the regression coefficients

Based on the ANOVA test results on the regression model with the log(ProductionResults) response in Figure 15, it was found that both independent variables, PlantingArea and HarvestedArea, simultaneously contributed significantly to variation in production yields. The PlantingArea variable had an F statistic of 39.407 with a p-value of $3.168\times10^{-6}$, while the HarvestedArea variable showed an F statistic of 51.725 with a p-value of $4.344\times10^{-7}$. These p-values, which were significantly lower than the 0.05 significance level, indicated that each variable significantly explained variation in log production yields.

Furthermore, the Sum of Squares value for HarvestedArea (14.7333) was greater than that for PlantingArea (11.2245), indicating that harvested area had a relatively greater explanatory contribution to variation in production yields compared to planted area. Meanwhile, the Residual Sum of Squares of 5.9816 with 21 degrees of freedom indicates that some of the data variation is still explained by other factors outside the model. Overall, the ANOVA results confirm that the constructed logarithmic regression model has good ability in explaining the relationship between planted area, harvested area, and production yield, and supports the previous F-test results which stated that the model is simultaneously significant.

```
Analysis of Variance Table

Response: log(ProductionResults)
              Df  Sum Sq Mean Sq F value    Pr(>F)
PlantingArea   1 11.2245 11.2245  39.407 3.168e-06 ***
HarvestedArea  1 14.7333 14.7333  51.725 4.344e-07 ***
Residuals     21  5.9816  0.2848
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

**Figure 15.** ANOVA Test Results on The Regression Model with The Log(Productionresults)


10.12928/bamme.v5i2.14559

Based on the estimation results of the linear regression model with logarithmic transformation on the dependent variable (log(ProductionResults)) in Figure 16, the overall model is significant in explaining variations in production yields. This is indicated by the F-statistic of 45.57 with a p-value of $2.297\times10^{-8}$, which is less than the 0.05 significance level. Thus, it can be concluded that the PlantingArea and HarvestedArea variables simultaneously have a significant effect on production yields.

Partially, the HarvestedArea variable has a significant effect on log(ProductionResults) with a coefficient of $1.267\times10^{-4}$ and a p-value of $4.34\times10^{-7}$. This positive coefficient indicates that increasing harvested area will increase production yields, with each unit increase in harvested area estimated to increase production yields by approximately 0.0127%, assuming other variables remain constant. In contrast, the PlantingArea variable has a negative coefficient of $-1.999\times10^{-5}$, but it is not statistically significant (p-value = 0.197 > 0.05). This indicates that planted area has not significantly influenced yield after being controlled by the harvested area variable, which could be caused by harvest efficiency, crop failure, or differences in land productivity.

The Multiple R-squared value of 0.8127 indicates that approximately 81.27% of the variation in yield can be explained by the combination of planted area and harvested area variables in the model, while the remaining 18.73% is influenced by other factors outside the model. Meanwhile, the Adjusted R-squared value of 0.7949 indicates that the model still has strong explanatory power after considering the number of independent variables used. The residual standard error value of 0.5337 indicates that the deviation of the model predictions from the actual values is relatively small on a logarithmic scale, thus the model can be said to have good estimation performance.

Overall, these results indicate that the constructed logarithmic regression model has met the analysis objectives, with harvested area as the main factor that consistently and significantly influences production results, while planting area does not provide a significant partial influence in the final model.

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    9.386e+00  2.931e-01  32.025  < 2e-16 ***
PlantingArea  -1.999e-05  1.499e-05  -1.334    0.197
HarvestedArea  1.267e-04  1.762e-05   7.192 4.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5337 on 21 degrees of freedom
Multiple R-squared:  0.8127,    Adjusted R-squared:  0.7949
F-statistic: 45.57 on 2 and 21 DF,  p-value: 2.297e-08
```

**Figure 16**. Estimation Results of The Linear Regression Model with Logarithmic Transformation

The regression model used in this study is a semi-log model, where the dependent variable is transformed using the natural logarithm. The model equation is expressed in the form of ln(ProductionResults) as a function of planted area and harvested area.

Based on summary(model_log) output, the model shows in Equation 6

$$\ln(\text{ProductionResults}) = 9.386 - 1.999\times10^{-5}\,\text{PlantingArea} + 1.267\times10^{-4}\,\text{HarvestedArea} + \varepsilon \quad (6)$$

By returning to the original scale:

$$\text{ProductionResults} = \exp(9.386 - 1.999\times10^{-5}\,\text{PlantingArea} + 1.267\times10^{-4}\,\text{HarvestedArea}) \times \exp(\varepsilon) \quad (7)$$

Or it can be written as:

$$\text{ProductionResults} = e^{9.386} \cdot e^{-1.999\times10^{-5}\,\text{PlantingArea}} \cdot e^{1.267\times10^{-4}\,\text{HarvestedArea}} \cdot u \quad (8)$$

with $u = e^{\varepsilon}$

Based on the regression model performance evaluation results in Figure 17, the Mean Absolute Error (MAE) was 19,344.89, the Root Mean Squared Error (RMSE) was 46,738.41, and the Mean Absolute Percentage Error (MAPE) was 48.20%. The MAE value indicates that, on average, the absolute difference between actual production results and the model's predictions is approximately 19,000 units of production, providing a direct indication of the magnitude of the prediction error in the original data units. Meanwhile, an RMSE value greater than the MAE indicates several observations with relatively large prediction errors, as RMSE imposes a higher penalty on large errors. The MAPE value of 48.20% indicates that, on average, the model's prediction error is nearly half the actual value, which, based on general accuracy evaluation criteria, falls into the low to moderate accuracy category.

However, it is important to note that the MAPE, MAE, and RMSE values are calculated using training data (in-sample). Therefore, these error metrics are used to evaluate the model's fit to the available data, not to measure the model's generalizability in predicting new data. Therefore, the obtained error values better reflect how well the model explains the relationship patterns between the independent and dependent variables in the research sample.

Overall, although the logarithmic regression model is statistically significant and meets the classical regression assumptions, the error evaluation results indicate that this model is more appropriate for relationship analysis and inference purposes, rather than as a precise prediction tool. The relatively large error values indicate that to improve prediction accuracy, future research could consider adding additional explanatory variables, separating training and test data, or using alternative modeling approaches.

```
> MAE_manual
[1] 19344.89
> RMSE_manual
[1] 46738.41
> MAPE_manual
[1] 48.1971
```

**Figure 17.** The Regression Model Performance Evaluation Calculated Using Training Data (In-Sample).

Table 2 are the results of comparing the actual and predicted values of rice production. From the comparison between actual and predicted data for the period January 2022 to December 2023, it can be seen that the prediction model is quite accurate in some months, but there are also some significant deviations in certain months. The comparison between actual and predicted values shows that the log-linear regression model is able to capture the general pattern of rice production well, especially in the medium data range. However, there is still a tendency to overestimate production data for both very large and very small data sets, which results in increased RMSE and MAPE values.


10.12928/bamme.v5i2.14559

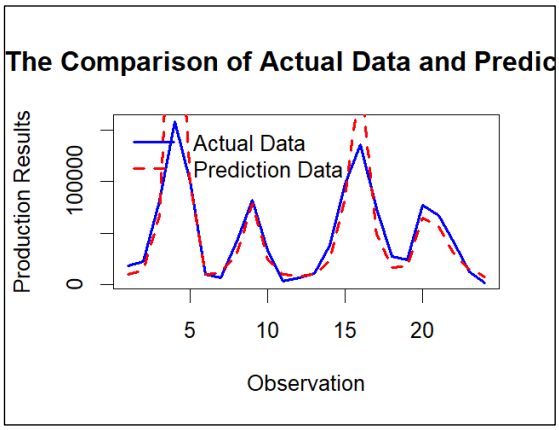



**Figure 18**. The Comparison of the Actual and Predicted Data

**Table 2.** The Comparison of the Actual and Predicted Data

| ProductionResults_Actual | ProductionResults_Prediction |
|---|---|
| 18724 | 10082,9174 |
| 22183 | 14394,54426 |
| 81146 | 66824,717 |
| 158344 | 372766,7643 |
| 100059 | 96408,19861 |
| 10413 | 10436,31335 |
| 6464 | 12025,56785 |
| 42545 | 30361,26324 |
| 82339 | 78966,70537 |
| 33653 | 24691,52652 |
| 3834 | 10023,56044 |
| 6897 | 8360,417093 |
| 11083 | 10022,40734 |
| 38045 | 23348,47057 |
| 97476 | 85800,02869 |
| 136199 | 201220,4541 |
| 73888 | 49424,40889 |
| 27775 | 16430,80505 |
| 24589 | 17703,78683 |
| 77403 | 64806,48276 |
| 66864 | 56235,1721 |
| 40341 | 29590,86221 |
| 12373 | 15024,6803 |
| 1972 | 7895,008765 |

**Relationship between Variables**

In Figure 19, the scatter plot illustrates the relationship between planted area and rice production. While logically, a larger planted area should result in higher production, the pattern observed in the graph appears inconsistent. Several data points indicate low production outcomes despite uncertainties until harvest. This finding is consistent with previous research, which found that although predicted planted area experienced a decline, harvested area remained stable and served as a more reliable indicator in forecasting rice production. The regression model used in the study

demonstrated a low prediction error rate, making it accurate and dependable for future production projections. The study concluded that multiple linear regression can be effectively used to improve production efficiency and assist strategic decision-making by understanding the contribution of each variable to rice production outcomes (Muharromah et al., 2023). This suggests that planted area alone is not a sufficient predictor for accurately explaining rice production. External factors such as weather conditions, pest attacks, cultivation techniques, and access to agricultural inputs may prevent the planted area from being fully realized as harvested area, thus leading to varying production results.

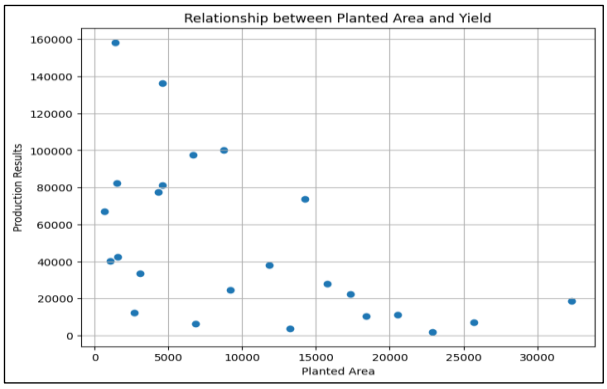



**Figure 19.** Relationship between Planted Area and Yield

Conversely, Figure 20 shows a much more consistent pattern: there is an almost linear positive relationship between harvested areas and rice production. In other words, the greater the harvested area, the greater the resulting production, with very minimal data dispersion. This strong relationship is understandable, as harvested areas represent the portion of crops that successfully grew to maturity. Harvested areas reflect the actual outcome of agricultural activities, unlike planted areas which still carry.

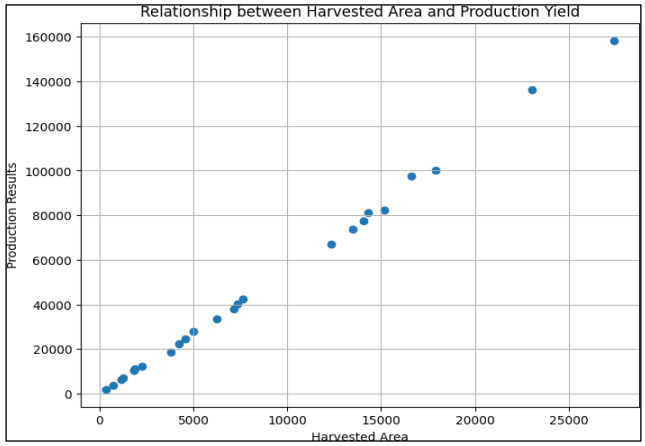



**Figure 20.** Relationship between Harvested Area and Production Yield

## Conclusion

This study concludes that harvested area is the dominant factor influencing rice production in Central Lampung, showing a strong positive relationship with yield, while planting area exhibits a weaker and negative effect. The multiple linear regression model demonstrates adequate explanatory power, with an R-squared value of 81.27%, indicating that the selected variables explain a substantial proportion of production variability. However, the relatively high MAE (19,344.89), RMSE (46,738.41), and MAPE (48.20%) suggest that the model is more effective in


10.12928/bamme.v5i2.14559

capturing general production trends than in providing precise forecasts. Therefore, future research should incorporate additional variables and alternative modeling approaches to improve prediction accuracy.

## References


Abdurrazak, M. A.-M., Zakaria, J., & Mapparenta. (2019). *Keunggulan Komparatif Tanaman Pangan di Kabupaten Manggarai Timur*.

Agusta, G. E., Dewanto, A., & Astriawati, N. (2024). *Multiple Linear Regression Model for Analyzing the Determinants of Rice Production in Sumatra*. *03*(02), 44–57.

Ambya, Fitriani, & Bellapama, I. A. (2022). Sektor Pertanian untuk Pertumbuhan Ekonomi Regional Lampung Agriculture Sector to Support Lampung Regional Economic Growth. *Journal of Food System and Agribusiness*, *6*(1), 102–111.

Apriyana, Y., Rejekiningrum, P., Alifia, A. D., & Ramadhani, F. (2023). *The Transformation of Rice Crop Technology in Indonesia : Innovation and Sustainable Food Security*. 1–14.

Elsayir, H. A. (2024). Overview on comparative methodology of classical ols and two-stage techniques in regression analysis model. *Journal of Jilin University (Engineering and Technology Edition)*, *43*(10), 245–251. Https://doi.org/10.5281/zenodo.14196326

Fitri, E., & Nugraha, S. N. (2024). Optimasi kinerja linear regression, random forest regression dan multilayer perceptron pada prediksi hasil panen. *Inti Nusa Mandiri*, *18*(2), 210–217.

Hendrastuty, N. (2024). *Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa*. *3*, 46–56.

Herliana, S., Ratnaningtyas, S., Aina, Q., Zuraida, U., & Sutardi, A. (2025). *Supply and Demand of Rice in Indonesia : A Critical Review*. *8*(1), 80–96.

Kaur, P., Stoltzfus, J., & Yellapu, V. (2018). *Descriptive statistics*.Lampung, B. P. S. (2024). *No Title*.

Midway, S., & White, J. W. (2025). *Testing for normality in regression models : mistakes abound ( but may not matter )*.

Muharram, A., Purnamasari, A. I., & Ali, I. (2023). Prediksi jumlah produksi daging unggas tahun 2023-2027 menggunakan regresi linier. *JATI (Jurnal Mahasiswa Teknik Informatika)*, *7*(6), 3093–3099.

Muharromah, O., Suarna, N., & Prihartono, W. (2023). Implementasi Algoritma Regresi Linear Berganda Untuk Prediksi Produksi Padi Di Kabupaten Cirebon. JATI (Jurnal Mahasiswa Teknik Informatika), 7(6), 3815-3820.

Nababan, Y., & Nugraha, I. (2024). Penerapan Data Mining Produksi Padi di Pulau Sumatera Menggunakan Analisis Regresi Linear. *JUTIN : Jurnal Teknik Industri Terintegrasi*, *7*(1), 262–272.

Navianti, D. R., Ayu, P., Krisna, G., Ryanto, S. S., Transportasi, P., Bali, D., & Kangin, B. (2023). *Identification of loading and unloading process time at denpasar goods terminal*. *4*(1), 57–66.

Rachman, R., Kusdinar, A. B., & Indrayana, D. (2024). Penerapan Regresi Linear Berganda Dalam Prediksi Dan Optimalisasi Persediaan Barang Toko Mungil. JATI (Jurnal Mahasiswa Teknik Informatika), 8(5), 10499-10506.

Santoso, A. B., Supriana, T., & Girsang, M. A. (2022). *Pengaruh Curah Hujan pada Produksi Padi Gogo di Indonesia ( Precipitation Impact on Upland Rice Yield in Indonesia )*. *27*(4), 606–613. Https://doi.org/10.18343/jipi.27.4.606

Swarbawa, I. B. M., Wibawa, I. G. A., & Suhartana, I. K. G. (2023). *Prediksi Hasil Panen Padi Di Kabupaten Jembrana Dengan Metode Linear Regression*. *11*(3), 671–678.

10.12928/bamme.v5i2.14559