Proceeding Konferensi Nasional Matematika (KNM) XXII
Universitas Bengkulu, 15-16 Juli 2024
Dipublikasikan online pada:
PATTIMURA PROCEEDING: Conference of Science and Technology
e-ISSN: 2829-3770, volume 5 issue 1, pp.89-100, Edisi November 2024

Binary Logistic Regression Modeling on Household Poverty Status in Bengkulu Province

Esther Damayanti Sihombing[1], Pepi Novianti[2*], Indah Wahyuliani[3]

[1,2,3]Study Program S2 Statistika, Faculty of Mathematics and Natural Sciences, University of Bengkulu, Jl. Wr Supratman, Bengkulu, 38123, Indonesia

Corresponding author's e-mail: **** pie_novianti@unib.ac.id***

Keywords:
Binary logistic regression;
Poverty;
House hold;
Bengkulu

ABSTRACT

Binary logistic regression is a statistical method used to analyze the relationship between one or more predictor variables and a binary or dichotomous response variable. Poverty is an issue in every province in Indonesia. One of the provinces with a relatively high poverty rate is Bengkulu Province, ranking seventh in Indonesia with a poverty rate of 14.62%. The Central Bureau of Statistics of Bengkulu Province (2023) explains that efforts to reduce poverty must involve all levels of society. Various government programs and policies in various fields such as health, social, and other areas are continuously being implemented to reduce the number of households classified as poor. Identifying the characteristics of households in Bengkulu Province by poverty status is important to study, as it serves as a reference to ensure that government programs are implemented according to the target. One method that can be used to identify household characteristics is binary logistic regression. This study aims to model the poverty status of households in Bengkulu Province using binary logistic regression and to identify the factors that influence it. The data used are social and economic household data from March 2022. The response variable used is household poverty status (poor and not poor), while the predictor variables include the ownership of toilet facilities, the source of lighting, floor area, family size, and per capita calorie consumption. Modeling is done using binary logistic regression with simultaneous and partial parameter significance tests, as well as model fit tests. The analysis results show that the factors significantly influencing household poverty status in Bengkulu Province are the ownership of toilet facilities, the source of household lighting, floor area, family size, and per capita calorie consumption. The formed binary logistic regression model has a classification accuracy of 89.98% with a sensitivity of 18.34% and a specificity of 98.61%.

DOI: https://doi.org/10.30598/ppcst.2024.knmxxii.89-100

1. INTRODUCTION

Binary logistic regression is a statistical method used to analyze the relationship between one or more predictor variables and a binary or dichotomous response variable. Predictor variables in logistic regression can be either categorical or continuous, whereas the response variable is categorical. Binary logistic regression is also considered a mathematical model approach used to analyze the relationship between several factors and a dichotomous (binary) variable. This means that in binary logistic regression, the data for the response variable is binary (0 and 1). These binary numbers represent two opposing data categories, such as 'yes or no', 'success or failure', and so on.

Poverty is a classic problem faced and given attention by every country. This is demonstrated by the international declaration of the Millennium Development Goals (MDGs) by 189 member countries of the United Nations (UN) in September 2000. One of the main goals of this declaration is to eradicate poverty and hunger [1]. Poverty that is effectively addressed is considered to align with efforts to tackle other global issues such as eliminating hunger, achieving health and well-being, providing quality education, ensuring adequate sanitation, and access to clean water [2]. Efforts to alleviate poverty are ongoing, especially in developing countries like Indonesia. Indonesia has a poverty rate of 9.71% of the total population, with 26.50 million people living below the poverty line. (Statistic Indonesia, 2020). This Figure makes Indonesia the 73rd poorest country in the world [3]. Poverty is an issue in every province of Indonesia. One of the provinces with a relatively high poverty rate is Bengkulu Province, which ranks seventh in Indonesia with a poverty rate of 14.62%. This Figure indicates that poverty in Bengkulu Province is still quite high compared to the national poverty rate, even though the economy of Bengkulu Province grew by 3.03% in the second quarter of 2022 (year to year) [4].

The Central Bureau of Statistics of Bengkulu Province (2023) explains that efforts to reduce poverty must involve all levels of society. Various government programs and policies in fields such as health, social welfare, and others are continuously being implemented to reduce the number of households classified as poor. Identifying the characteristics of households in Bengkulu Province based on their poverty status is important to study, as it serves as a reference to ensure that government programs are implemented according to the target. One method that can be used to identify household characteristics is binary logistic regression. This method is very useful when the predicted variable is binary, such as poverty status categorized as "poor" or "not poor." By using binary logistic regression, the influence of various factors on the likelihood of a household being classified as poor can be determined. Therefore, statistical methods are essential for accurately and comprehensively identifying household characteristics in Bengkulu Province, to enable policies or interventions to be more targeted and effective in reducing poverty.

Several studies on poverty have been widely researched, such as by [5] conducted a study aimed at classifying poor households in Pemalang Regency using the C5.0 algorithm. From the study, the sensitivity value could not be defined, indicating that the model failed to classify the data accurately. This was due to an imbalance in the data between the poor and non-poor groups, showing that C5.0 was unsuccessful in predicting the minority class. The study conducted by [6] aimed to classify the status of villages in East Nusa Tenggara Province using MARS with an imbalanced data group proportion, resulting in an accuracy of 99.40%, sensitivity of 99.84%, and specificity of 92.8% on the test data. Considering the importance of this issue, it is therefore necessary to analyze which variables influence poverty so that the policies or interventions implemented by the government can be more targeted and effective. Based on the results of previous studies, the author is interested in conducting research titled "Binary Logistic Regression Modeling on Household Poverty Status in Bengkulu Province.

2. RESEARCH METHODS

The data used in this study are secondary data from the National Socio-Economic Survey (SUSENAS) conducted in March 2022, sourced from the Central Bureau of Statistics of Bengkulu Province. The observation objects in this study are households in Bengkulu Province. This study includes 5 predictor variables and 1 response variable. The response variable in this study is a nominal binary scale, which categorizes households as either poor or not poor. The research steps are as follows:

1. Conduct data exploration to observe the general characteristics of the data.
2. Identify the correlation between the response and predictor variables.
3. Conduct multicollinearity assumption tests.
4. Perform parameter estimation to obtain parameter estimate values β.
5. Conduct parameter significance tests using simultaneous and partial tests.
6. Perform model adequacy test.
7. Calculate the odds ratio for model interpretation.
8. Calculate classification accuracy.

2.1 Regression Analysis

Regression analysis is a statistical method used to investigate and model relationships between variables [7]. Regression analysis provides an explanation of the relationship pattern (model) between a response variable and predictor variables. The main goal of regression analysis is to estimate the form of the regression curve. There are three approaches to regression analysis for estimating the form of the regression curve: parametric regression, nonparametric regression, and semiparametric regression. Parametric regression is used for modeling when the form of the regression curve between the response and predictor variables follows or is assumed to follow a specific function such as linear, quadratic, cubic, exponential, and others. Nonparametric regression is used when the form of the regression curve is unknown or tends not to form a specific pattern, and the regression curve is assumed to be smooth [8]. Nonparametric regression is an approach that offers high flexibility, as it allows data to determine the shape of the regression curve estimation independently of subjective researcher influence. Another understanding of nonparametric regression is that this method refers to distribution-free methods. Semiparametric regression combines parametric and nonparametric regression approaches. The general equation for regression is typically expressed as [8] :

$$y_i = f(x_i) + \varepsilon_i, i = 1, 2, \dots, n \tag{1}$$

2.2 Regression Logistic Analysis

Logistic regression analysis is used to examine the functional relationship between a categorical response variable (binary and multinomial) and predictor variables that can be qualitative (nominal or ordinal) or quantitative (interval or ratio). A response variable with two categories in logistic regression is called binary logistic regression, while a response variable with more than two categories is termed multinomial or ordinal logistic regression. The difference lies in the nature of the response variable. If the response variable has more than two nominal categories and there is a reference category designated as a baseline against which other categories are weighted, multinomial logistic regression is used [9]. However, if the response variable consists of two ordinal categories, meaning there is a ranking involved, ordinal logistic regression is used [10]. The logistic regression model with predictor variables and a response variable with two categories, also known as the binary logistic regression model, is as follows:

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)} \tag{2}$$

Where $\beta_1, \beta_2, \dots, \beta_k$ are unknown parameters. Thus, the linear function form of the logistic regression function is as follows:

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \tag{3}$$

Parameter estimation in logistic regression using the maximum likelihood method [11]. Estimators with a dichotomous response variable distributed Bernoulli can be written as follows:

$$P(Y_i = y_i) = \pi(x)^{y_i}[1 - \pi(x)]^{1-y_i} \tag{4}$$

Then the logarithm likelihood function for is:

$$\text{In } L(\beta) = l(\beta) = \prod_{i=1}^{n} \pi(x)^{y_i}[1-\pi(x)]^{1-y_i} \tag{5}$$

The value of is obtained by maximizing the logarithm likelihood function, finding the stationary value of the logarithm likelihood function, where the first derivative of the logarithm likelihood function with respect to equals zero. Because the first derivative of this function is nonlinear, parameter estimation of is achieved using the Newton-Raphson iteration method [12].

2.3 Identification of Variable Correlation

The correlation between the response variable and the predictor needs to be identified in order to conduct further analysis to see if there is a correlation between the two variables. The correlation between two categorical variables can be tested using a test χ^2. Using the χ^2 test statistic [13]:

$$\chi^2 = \sum_{i=1}^{a}\sum_{j-1}^{b} \frac{\left[p_{ij} - m_{ij}\right]^2}{m_{ij}} \tag{6}$$

The test statistic value χ^2 compared with the chi-square Table with degrees of freedom $(a-1)(b-1)$ where a is the number of rows and is the number of columns. H_0 rejected if the test statistic value χ^2 is greater than the Table value $\chi^2_{a;db}$ or $p - value < 5\%$ which means there is a correlation between the two variables.

Meanwhile, to identify variables and observe correlation between two variables where one is binary and the other is numeric, you can conduct a point-biserial correlation test. Using the point-biserial correlation test statistic [14]:

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{\bar{s}_y}\sqrt{p_0 p_1}\sqrt{\frac{n}{n-1}} \tag{7}$$

$$\bar{s}_y = \frac{\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y)^2}{n}}{n-1} \tag{8}$$

The test statistic value r_{pb} compared with the Table value with freedom degree $(n-2)$. The rejection criteria is at the significance level $\alpha = 5\%$, H_0 rejected if the test statistic r_{pb} is greater than the Table value r or $p - value < \alpha$ which means there is a relationship between the two variables.

2.4 Multicollinearity Test

Multicollinearity test according to [15], the purpose is to test whether there is correlation between predictor variables. A good regression model should not exhibit correlation among predictor variables. To detect the presence of multicollinearity, we use the Variance Inflation Factor (VIF). If the VIF value is greater than 10, multicollinearity is present.

2.5 Parameter Significance Test

Testing the model parameters is conducted to examine whether predictor variables have a significant role in the model. Simultaneous testing is performed using the Likelihood Ratio Test, and partial testing is done using the Wald Test.

2.5.1 Simultaneous Test

Simultaneous testing is conducted to determine the significance of the parameters β on the response variable as a whole. Based on [11], Likelihood Ratio Test with statistics test G^2 has the following formula:

$$G^2 = -2\text{In}\left[\frac{L_0}{L_1}\right] \tag{9}$$

2.5.2 Partial Test

Partial testing is conducted to identify the significance of the parameters on each response variable individually. Additionally, this test is conducted to show the appropriateness of including a predictor variable in the model. Based on [11], the Wald Test with the test statistic has the following formula:

$$W = \left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2 \tag{10}$$

2.6 Model Fit Test

Evaluating the adequacy of a logistic regression model is done using goodness-of-fit testing. This test aims to assess how well the estimated model fits the observed data. The Hosmer-Lemeshow test is one of the most commonly used tests to evaluate the fit of a logistic regression model. This test groups the data into several categories based on the probabilities predicted by the model, then compares the observed and expected frequencies in each group. Based on [11], goodness of fit test has the following formula:

$$\hat{C} = \sum_{p=1}^{p} \frac{(o_p - n_p\bar{\pi}_p)^2}{n_p\bar{\pi}_k(1-\bar{\pi}_k)} \tag{11}$$

2.7 Odds Ratio

The odds ratio is a measure of the likelihood of experiencing a particular event between one category and another. [11]. The odds ratio is denoted by which is the ratio of two odds values as follows:

$$\theta = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}} = \frac{\left[\frac{\exp(\beta_0+\beta_m)}{1+\exp(\beta_0+\beta_m)}\right]\left[\frac{1}{1+\exp(\beta_0)}\right]}{\left[\frac{1}{1+\exp(\beta_0+\beta_m)}\right]\left[\frac{\exp(\beta_0)}{1+\exp(\beta_0)}\right]} = \frac{\exp(\beta_0+\beta_m)}{\exp(\beta_0)} = \exp(\beta_m) \tag{12}$$

A decision is made that there is no relationship between the predictor variable and the response variable if the odds ratio is equal to 1. If the odds ratio is less than 1, there is a negative relationship, indicating that each change in the predictor variable (x) decreases the odds of the response variable. Conversely, if the odds ratio is greater than 1, a positive relationship exists, meaning that each change in the predictor variable (x) increases the odds of the response variable [16].

2.8 Classification Accuracy

Classification accuracy is assessed using the accuracy value generated by the confusion matrix. An effective classification method will yield minimal classification errors or a low probability of errors. Classification results can be evaluated using metrics such as the Apparent Error Rate (APER), Total Accuracy Rate (TAR), sensitivity, and specificity. The classification accuracy is calculated based on the confusion matrix, which provides information by comparing the classification results from the system (model) with the actual classification outcomes [16].

Table 1. Confusion matrix

Predict	Actual	
	Group 1	**Group 2**
Group 1	n_{11}	n_{12}
Group 2	n_{21}	n_{22}

n_{11} : Total observations predicted as group 1 correctly classified as group 1
n_{12} : Total observations predicted as group 1 but classified incorrectly as group 2
n_{21} : Total observations predicted as group 2 but classified incorrectly as group 1
n_{22} : Total observations predicted as group 2 correctly classified as group 2

Apparent Error Rate (APER) is an evaluation procedure used to assess classification errors made by a classification function. Total Accuracy Rate (TAR) is used to calculate the accuracy of classification in clustering observations. The Total Accuracy Rate (TAR) value can represent the proportion of observations correctly classified. The values of Apparent Error Rate (APER) and Total Accuracy Rate (TAR) are calculated as follows:

$$APER(\%) = \frac{n_{12} + n_{21}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\% \tag{13}$$

$$TAR(\%) = \frac{n_{11} + n_{22}}{n_{11} + n_{12} + n_{21} + n_{22}} \times 100\% \tag{14}$$

In a classification Table like Table 1, it's also important to consider the sensitivity, which describes the accuracy of observations in the positive group, and specificity, which describes the accuracy of observations in the negative group [7]. The ability to measure sensitivity and specificity effectively indicates that the classification method is good at predicting observations for each group.

$$\text{sensitivity}(\%) = \frac{n_{22}}{n_{12} + n_{22}} \times 100\% \tag{15}$$

$$\text{specificity}(\%) = \frac{n_{11}}{n_{11} + n_{21}} \times 100\% \tag{16}$$

2.9 Poverty

2.9.1 Definition of Poverty

Badan Pusat Statistik (2023) Viewing poverty from an economic perspective involves meeting basic needs, both food and non-food, measured by expenditure. Poverty can be understood as the inability of an individual or group to meet economic needs, living below the poverty line set by the government. The definition of the poverty line in the BPS (2023) description is the expenditure in rupiah per person per month needed to meet basic needs other than food (non-food poverty line), plus the expenditure in rupiah per person to meet a daily intake of 2,100 kcal per capita (food poverty line). This calorie measure, agreed upon globally by the FAO (Food and Agriculture Organization) and WHO (World Health Organization) based on extensive research by experts, recommends that the minimum threshold for human survival and ability to work is around 2,100 kilocalories [17].

Poverty is a condition where the standard of living is not met due to the inability to fulfill basic or essential needs of life. These basic needs include food, healthcare, clothing, shelter or housing, and education. The inability to meet these basic needs or achieve a decent standard of living is referred to as poverty according to the basic needs approach.

2.9.2 Characteristics of Poor Households

Poor households per household can be observed from various characteristics as follows:

1. According to BPS (2023), factors influencing poverty include internal factors. Characteristics of poor households can often be identified through the use of toilet facilities. Poor households usually have limited access to proper sanitation facilities. Many of them still use pit latrines (simple hole in the ground) or even lack dedicated facilities, resorting to open defecation in places like rivers, fields, or gardens. This is due to economic constraints that prevent them from building or installing adequate toilets and proper sanitation. This situation not only poses high health risks due to disease spread but also reflects low levels of well-being and education regarding the importance of good sanitation.
2. According to [17] Identifying poor households can also be assessed based on the source of lighting they use, whether it's electric or non-electric. Poor households often have limited or no access to electricity. Many of them still rely on traditional lighting sources such as oil lamps, candles, or torches.
3. The number of household members influences poverty, as an increasing number of dependents can be caused by several factors such as having many children, having nonproductive family

members, and difficulties for family members in obtaining employment during productive ages, thereby causing poverty.

4. Adequate calorie consumption can serve as an indicator to assess the nutritional status of the population and the government's success in integrated food, agricultural, health, and socio-economic development. High levels of calorie sufficiency indicate a high level of public health, thereby enhancing community productivity. This in turn contributes to poverty reduction. The recommended calorie intake is 2,150 kilocalories per capita per day.

3. RESULTS AND DISCUSSION

3.1 Characteristics of Data on Poor Households

The following discussion provides a general overview of household data comprising poor and non-poor households in Bengkulu Province, reviewed across several factors. In March 2022, a survey of households in Bengkulu Province identified a total of 5,731 households, with 616 households classified as poor and 5,115 households classified as non-poor. Figure 1 indicates a significant difference in numbers between poor and non-poor households.

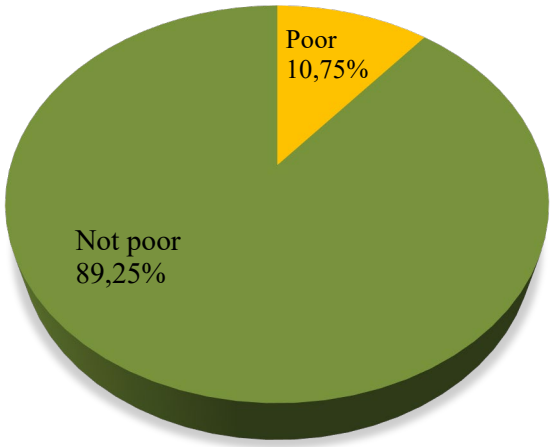


Figure 1. Percentage of poor and non-poor households

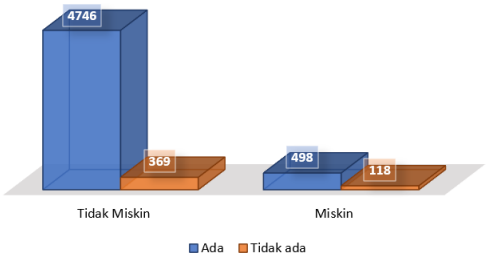
An overview of poor and non-poor households categorized by predictor variables is presented in Table 2, which shows predictor variables according to household poverty status based on the count and percentage of each predictor variable category. A bar chart is also provided to compare categories of nominal predictor variables based on household poverty status.

Table 2. Categories of categorical predictor variables according to poverty status

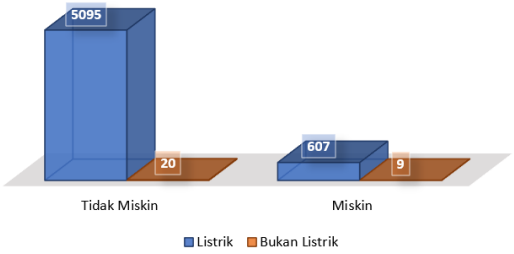
Variable predictor	Categoric	Not Poor		Poor	
		Total	%	Total	%
X_1	Defecation facilities are available	4746	82.81	498	8.69
	No defecation facilities	369	6.44	118	2.06
X_2	Electricity	5095	88.90	607	10.59
	No electricity	20	0.35	9	0.16

Table 2 shows an overview of predictor variables based on count and percentage according to household poverty status. According to Table 2, in the non-poor household group, the majority of households have access to toilet facilities (82.81%), while only a small percentage do not have such facilities (6.44%). On the other hand, in the poor household group, households with toilet facilities comprise

8.69%, with 2.06% lacking these facilities. This indicates that access to toilet facilities is more common among non-poor households. Regarding the lighting source variable, it shows that the majority of non-poor households use electricity for lighting (88.90%), while only a small percentage use non-electric sources (0.35%). Conversely, in the poor household group, 10.59% use electricity for lighting and 0.16% use non-electric sources. This suggests that electricity for lighting is more prevalent among non-poor households.



(a)



(b)

Figure 2. (a) Proportion of toilet facilities ownership (b) Proportion of lighting sources.

Next, descriptive statistics of predictor variables are presented in Table 3 as follows:

Table 3. Descriptive statistics of predictors based on household poverty status

Variable	Mean	Minimum	Maximum	Median
Floor Area				
Not poor	75	4	800	60
Poor	54,75	9	450	48
Number of Families				
Not poor	1,09	0	6	1
Poor	1,21	0	3	1
Calorie Consumption Per Capita				
Not poor	2272	1003	4483	2162
Poor	1599	1004	2855	1549

The characteristics of the floor area variable (X_3) in Table 3 indicate that households classified as poor have living spaces ranging from 9 m^2 to 450 m^2 ranging from 4 m^2 to 800 m^2 , while non-poor households have living spaces . The characteristics of the family size variable (X_4) in Table 4.2 show that both poor and non-poor households have similar ranges in family size. The most common family size for poor households is 3, and for non-poor households, it is 6. The characteristics of the per capita calorie consumption variable in Table 3 show that daily calorie consumption per capita for poor households ranges from 1004 kcal to 2855 kcal, while for non-poor households, it ranges from 1003 kcal to 4483 kcal.

3.2 Identification of Variable Correlation

Exploratory data analysis previously described the condition of the data to be analyzed, including both categorical and numerical data. Before further analysis, it is necessary to determine the correlation between the response variable and predictor variables. The correlation between a categorical response variable and categorical predictor variables is tested using the chi-square test (χ^2). Meanwhile, the

correlation between a categorical response variable and a numeric predictor variable is tested using the point-biserial correlation test. The results of the testing between the response variable and predictor variable are presented in Table 4 below:

Table 4. Results of the correlation test between the response variable and predictor variable

No	Variable	*Chi-square* (χ^2)	r_{pb}	p_{value}	Conclusions
Numeric Predictor Variable					
1	X_1	99.303	-	2.2×10^{-16}	Reject H_0
2	X_2	10.468	-	0.001	Reject H_0
Numeric Predictor Variable					
3	X_3	-	0.114	2.2×10^{-16}	Reject H_0
4	X_4	-	-0.092	3.9×10^{-12}	Reject H_0
5	X_5	-	0.315	2.2×10^{-16}	Reject H_0

Chi-square test (χ^2) Aims to identify whether there is a correlation between the categorical response variable and the categorical predictor variable. The test statistic used is as found in the equation (2.6), with the rejection criteria being reject H_0 if the value $\chi^2_{hitung} > \chi^2_{tabel}$ or $p_{value} < \alpha$ at the significance level $\alpha = 5\%$. Results from testing in Table 4 indicate that the variable X_1 dan X_2 produces p_{value} For the statistical test, it is less than $\alpha = 5\%$. Therefore, it can be concluded that H_0 It is rejected, which means that variables X_1 and X_2 are correlated with the response variable.

Correlation test point biserial (r_{pb}) aims to identify whether there is a correlation between the categorical response variable and the numerical predictor variable. The statistical test used is given in Equation (2.7), with the rejection criteria being to reject H_0 if the value $r_{pb} > r_{0,05;5729}$ where $r_{0,05;5729} = 0.025$ or $p_{value} < \alpha$ at the sifnificance level $\alpha = 5\%$. The test results in Table 4 show that the variable $X_3, X_4,$ and X_5 yield p_{value} for the statistical test is less than $\alpha = 5\%$. for the statistical test is less than H_0 is rejected, which means the variable $X_3, X_4,$ and X_5 is correlated with the response variable.

3.3 Multicollinearity Test

The purpose of the multicollinearity test is to detect the presence of a high linear relationship between predictor variables in a model. In this case, the VIF value is used to detect multicollinearity. The test results are as follows:

Table 5. Multicollinearity Test

Variable	VIF
X_1	1.084
X_2	1.029
X_3	1.079
X_4	1.041
X_5	1.028

Based on the results of the multicollinearity test using the Variance Inflation Factor (VIF), it can be seen that all VIF values for the variables dan are less than 10. Generally, a VIF value smaller than 10 indicates that there is no multicollinearity issue in the model. Therefore, these variables can be considered to not have a high linear relationship with each other, making the regression coefficient estimates reliable.

3.4 Parameter Significance Test

The parameter estimates of the binary logistic regression model are presented in Table 6 as follows:

Table 6. Estimation of the binary logistic regression model parameters

Variable	Q	W	*p-value*	exp(Q)
Intercept	3.436	12.477	2.2×10^{-16}	31.069
X_1	1.159	8.237	2.2×10^{-16}	3.188

Variable	Q	W	*p-value*	exp(Q)
X_2	0.954	1.961	0.049	2.595
X_3	-0.012	-7.651	1.9×10^{-14}	0.988
X_4	0.619	5.686	1.3×10^{-8}	1.858
X_5	-0.003	-22.196	2.2×10^{-16}	0.997

After obtaining the parameter estimation results of the model, a simultaneous and partial parameter significance test is conducted as follows:

3.4.1 Simultaneous Test

A simultaneous test is conducted to determine the significance of the parameters with respect to the response variable as a whole. The test statistic values obtained are as follows:

$$G^2 = -2\text{In}\left[\frac{L_0}{L_1}\right] = 1041.95$$

Based on the value $G^2 > \chi^2_{0,05;k}$ which $1041.95 > 11.071$ and $p - value < 0.05$, therefore H_0 rejected. This means that at the 5% significance level, that at least one predictor variable affects the response variable.

3.4.2 Partial Test

The partial test is conducted to identify the significance of the parameter β with respect to the response variable individually. According to [11], the Wald Test, with test statistic W, has the following formula.

$$W = \left[\frac{\hat{\beta}_j}{se(\hat{\beta}_j)}\right]^2$$

The Wald Test statistic value for each parameter is as follows:

Table 7. Statistic Wald Test

Variable	W	*p-value*
Intercept	12.477	2.2×10^{-16}
X_1	8.237	2.2×10^{-16}
X_2	1.961	0.049
X_3	-7.651	1.9×10^{-14}
X_4	5.686	1.3×10^{-8}
X_5	-22.196	2.2×10^{-16}

The Wald Test statistic values for each parameter presented in Table 7 indicate that all five predictor variables used significantly affect the response variable.

3.5 Goodness of Fit Test

This test aims to evaluate how well the estimated model fits the observed data. The obtained statistic for the model fit test is as follows:

$$\hat{C} = \sum_{p=1}^{P} \frac{\left(o_p - n_p\bar{\pi}_k\right)^2}{n_p\bar{\pi}_k(1-\bar{\pi}_k)} = 11.998$$

The value $\hat{C} < \chi^2_{0.05;p-2}$ which $11.998 < 15.507$ and $p - value > 0.05$ which is 0.151, therefore H_0 is accepted. This means that at the 5% significance level, that the model fits well, indicating that there is no significant difference between the observed results and the model's predicted outcomes.

3.6 Interpretation of Binary Logistic Regression Model

Based on the analysis conducted, it is found that the model fits well and passes the significance tests for both simultaneous and partial parameters. Therefore, the binary logistic regression model for household poverty status in Bengkulu Province is as follows:

$$g(X) = 3.436 + 1.159X_1 + 0.954X_2 - 0.12X_3 + 0.619X_4 - 0.003X_5$$

To interpret the parameters using odds ratio values, the following are used

1. 3.188 means that households without toilet facilities have a tendency that is 3.188 times greater to be classified as poor households compared to households with toilet facilities, assuming all other predictor variables are held constant.
2. 2.595 means that households whose lighting source is not electricity have a tendency that is 2.595 times greater to be classified as poorer households compared to households whose lighting source is electricity, assuming all other predictor variables are held constant.
3. 0.988 means that for every 1 unit increase in square meters of floor area in a household's residence, the chance of being classified as a poor household decrease by 0.988, assuming all other predictor variables are held constant.
4. 1.858 means that for every 1 unit increase in the number of family members in a household's residence, the likelihood of being classified as a poor household increase by 1.858, assuming all other predictor variables are held constant.
5. 0.997 means that for every 1 unit increase in daily per capita calorie consumption of a household, the chance of being classified as a poor household decrease by 0.997, assuming all other predictor variables are held constant.

3.7 Evaluation of Classification Accuracy

A good classification method will result in few classification errors or a small chance of classification errors. Accuracy of classification is determined by calculations in the confusion matrix table. Here is the confusion matrix Table formed based on the binary logistic regression model:

Table 8. Confusion matrix model regression logistic biner

Predict	Actual	
	Not poor	Poor
Not poor	5044	503
Poor	71	113

Based on Table 8, the model evaluation can be calculated as follows:

$$APER(\%) = \frac{503 + 71}{5044 + 503 + 71 + 113} \times 100\% = 10.01\%$$

$$TAR(\%) = \frac{5044 + 113}{5044 + 503 + 71 + 113} \times 100\% = 89.98\%$$

$$\text{Specificity}(\%) = \frac{5044}{5044 + 71} \times 100\% = 98.61\%$$

$$\text{Sensitivity}(\%) = \frac{113}{503 + 113} \times 100\% = 18.34\%$$

APER value in the model results in a classification error rate of 10.01% and achieves a classification accuracy of 89.98%. The 10.01% error rate indicates that the model correctly predicts whether households are poor or non-poor for most of the tested data. In other words, out of every 100 predictions made by the model, approximately 90 are correct and align with the actual status of the households. This shows that the

model has a reasonably high reliability in classification. However, it's important to note that accuracy alone does not provide information about how well the model detects classes that are proportionally imbalanced.

Therefore, in evaluating the model, it's crucial to also consider other metrics such as sensitivity and specificity. The specificity value describes the accuracy in predicting the negative group, which is households classified as non-poor, with a specificity of 98.61%. The sensitivity value describes the accuracy in predicting the positive group, which is households classified as poor, with a sensitivity of 18.34%. The sensitivity value of 18.34% indicates that only about 18.34% of actual positive observations were correctly identified by the model. Thus, while the model has a high specificity rate, its sensitivity is relatively low, suggesting that the model may be better at identifying negative-value observations than positive ones. These insights are useful for further development and improvement of the model for future researchers, focusing on enhancing sensitivity to ensure better identification of positive-value observations.

4. CONCLUSIONS

Based on the analysis results obtained, several conclusions were drawn, including:

1. The classification model for household poverty status in Bengkulu Province using binary logistic regression method is as follows:

$$g(X) = 3.436 + 1.159X_1 + 0.954X_2 - 0.12X_3 + 0.619X_4 - 0.003X_5$$

2. The factors influencing household poverty status in Bengkulu Province are ownership of toilet facilities, household lighting source, floor area, number of family members, and daily per capita calorie consumption.

REFERENCES

[1] M. P. Todaro, and S. C. Smith, Economic Development, 11th ed. USA: Pearson Education, 2015.

[2] W. Agwil, D. Agustina, H. Fransiska, dan N. Hidayati, "Klasifikasi karakteristik kemiskinan di Provinsi Bengkulu tahun 2020 menggunakan metode pohon klasifikasi gabungan," Jurnal Aplikasi Statistika & Komputasi Statistik, vol. 14, pp. 23–32, 2022.

[3] W. Bank, Era Baru dalam Pengentasan Kemiskinan di Indonesia. Jakarta: The World Bank Office Jakarta. 2007.

[4] Indonesia. Badan Pusat Statistik: Penghitungan dan Analisis Kemiskinan Makro Indonesia. Jakarta: Badan Pusat Statistik; 2019. [Online]. Available: https://www.bps.go.id/publication/2019/12/20/60138aa2d7b9b78802991240/penghitungan-dan-analisis-kemiskinan-makro di-indonesia-tahun-2019.

[5] F. N. Umma, B. Warsito, and D. A. I. Maruddani, "Klasifikasi Status Kemiskinan Rumah Tangga dengan Algoritma C5.0 di Kabupaten Pemalang," Jurnal Gaussian, vol. 10, pp. 221-229, May 2021.

[6] O. Tamonob, "Analisis Multivariate Adaptive Regression Splines (MARS) untuk Mengklasifikasikan Status Desa di Provinsi Nusa Tenggara Timur," undergraduate thesis, Institur Pertanian Bogor, Bogor, 2020.

[7] P. Sprent, "An Introduction to Categorical Data Analysis," Journal of the Royal Statistical Society Series A: Statistics in Society, vol. 170, pp. 1178, October 2007.

[8] R. L. Eubank, Spline Smoothing and Nonparametric Regression. New York: Marcel Dekker, 1999.

[9] M. Rezapour, and K. Ksaibati, "Application of Multinomial and Ordinal Logistic Regression to Model Injury Severity of Truck Crashes, using Violation and Crash Data," Journal of Modern Transportation, vol. 26, pp. 268–277, June 2018.

[10] Renate H. M. de Groot, Rebecca Emmett, B. J. M. (2019). This peer-reviewed article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI. 10.1017/S0007114519000138. Cambridge University Press, May 2020, 1–24. https://doi.org/10.1017/S0950268820001533.

[11] D. W. Hosmer, and S. Lemeshow, Applied Logistic Regression. USA: John Willey and Sons, 2000.

[12] S. Burhan, and A. K. Jaya, " Penaksiran Parameter Regresi Linier Logistik dengan Metode Maksimum Likelihood Lokal pada Resiko Kanker Payudara di Makassar," Jurnal Matematika Statistika dan Komputasi, vol. 14, pp. 159-165, January 2018.

[13] S. Nugroho, Statistika Multivariat Terapan. Bengkulu: UNIB Press Bengkulu. 2008.

[14] D. J. Sheskin, Handbook of Parametric and Nonparametric Statistical Procedures, fifth ed. London: Chapman & Hall, 2011.

[15] S. Santoso, Statistik Multivariat Konsep dan Aplikasi dengan SPSS, Jakarta: Elek Media Komputindo, 2014.

[16] T. M. T. Nisva, and V. Ratnasari, "Analisis Regresi Logistik Biner pada Faktor-Faktor yang Mempengaruhi Jenis Perceraian di Kabupaten Lumajang," Inferensi, vol. 3, pp. 2721-3862, March 2020.

[17] Y. H. Sa'diyah, and F. Arianti, "Analisis Kemiskinan Rumah Tangga Melalui Faktor-Faktor yang Mempengaruhinya," Diponegoro Journal of Economics, vol. 1, pp. 99-109, October 2012.