BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1653–1662
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1653-1662

MODELING CLUSTERWISE LINEAR REGRESSION ON POVERTY RATE IN INDONESIA

Eni Meylisah [1], Dyah Setyo Rini [2*], Herlin Fransiska [3], Winalia Agwil [4], Bagus Sartono [5]

[1,2,3,4] Statistics Study Program, Mathematic Departement, Faculty of Mathematics and Natural Science, Bengkulu University
WR. Supratman Street, Kandang Limun, 38371, Bengkulu, Indonesia

[5] Statistics Study Program, Faculty of Mathematics and Natural Science, IPB University
Dramaga Raya Street,16680, Bogor, Indonesia

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

ABSTRACT

Article History:

Received: 01st May 2023
Revised: 10th August 2023
Accepted: 18th August 2023

Keywords:

Poverty;
Cluster;
Linear regression;
Clusterwise Linear Regression (CLR).

When a person's income is so low that it cannot cover even the most basic living expenses, they are said to be poor. Data on poverty levels and hypothesized causes are used in this study. If the data pattern forms clusters, one of the regression analyses that can be used is Clusterwise Linear Regression (CLR). Therefore, this study aimed to determine the poverty rate modeling in Indonesia with the CLR method. The results showed that the best model is with 3 clusters, that for cluster 1, the factors that significantly affect the percentage of poverty are the percentage of electricity users (X_1), the number of small and micro industries (X_6), and the number of tourist villages (X_7). In cluster 2, the amount of village tours (X_7) $significantly\ affects$ the percentage of poverty. In cluster 3, the percentage of users of electricity (X_1) and the percentage of villages that have mining and quarrying (X_2) $significantly\ affect$ the percentage of poverty.

How to cite this article:

E. Meylisah, D. S. Rini, H. Fransiska, W. Agwil and B. Sartono., " MODELING CLUSTERWISE LINEAR REGRESSION ON POVERTY RATE IN INDONESIA," *BAREKENG: J. Math. & App.*, vol. 17, iss. 3, pp. 1653-1662, September, 2023.

Research Article • Open Access

1. INTRODUCTION

Poverty is a state of economic inability to meet primary needs, education, and health [1]. Poverty is a reality that a person experiences without his will. Poverty can be divided into two types, namely material and non-material. Material poverty includes poor income and non-income education, health, employment, and public services. Meanwhile, non-material poverty is related to personality traits such as emotional, psychological, and spiritual traits [2]. Poverty is a complex and multidimensional problem because it is related to social, economic, cultural and other aspects [3]. The problem of poverty is one of the issues that is the center of attention from the Indonesian government [4]. It can be seen from the *Millennium Development Goals* (MDGs) that tackling poverty and hunger is a top priority. Poverty can hurt the country, such as increased unemployment, early school dropouts, health problems in the community, a decrease in the quality of the next generation, and various criminal acts.

Based on the Central Bureau of Statistics (BPS) publication, the number of poor people amounted to $27\ 55$ million people in September 2020 and increased by $1{,}13$ million compared to March 2020. The percentage of poor people was $10{,}19$ percent in September 2020 and increased by $0{,}41$ percent compared to March 2020. Regionally, the poverty rate in rural areas is still high at 13.20 percent. Meanwhile, poverty in urban areas is much lower at 7.88 percent, which means that the poverty disparity between urban and rural areas is still relatively high, so the data is *cluster* identified [5].

Seeing the adverse effects of the poverty rate on a country, the government is always looking for ways to solve this problem. One way is to find out what factors affect the poverty rate. Many researchers have researched the topic of poverty rates and looked at the factors that influence it with various methods, including [6], which discusses the impact of micro, small and medium enterprises (MSMEs) on alleviating poverty and unemployment, The method used was a quantitative method with a descriptive research type and the result is that the impact of MSMEs has a significant effect on poverty and unemployment. The following research by [7] discusses the effect of tourism on poverty in the tourist areas of Bali; the purposive sampling method was used in this study and found that natural factors, cultural factors, and the lack of poor people working in the tourism industry are contributors to the poverty problem. Other research by [8] examined the causes of poverty in Indonesia over the past five years (through case studies in 33 provinces). This study used the panel data regression method and found that economic growth and TPT do not reduce poverty, while the human development index affects poverty.

Linear regression is one approach to understanding what variables play a role in shaping the poverty rate. Regression analysis is a mathematical model that has the aim of establishing relationships between variables [9]. According to [10] , linear regression analysis has the opportunity to have several regression models. Ordinary linear regression analysis cannot be used to estimate these models because it cannot just use one regression model to estimate them; this is because there are unknown subpopulations. Because the poverty data is identified as having clusters, another method is needed to find these clusters to estimate the subpopulations, namely the clusterwise linear regression (CLR) method.

Clusterwise linear regression (CLR) is a new machine-learning algorithm that improves the accuracy of classical linear regression by partitioning the training space into subspaces [11]. *Clusterwise linear regression* (CLR) is a clustering method that follows the properties of regression parameters to obtain the structure contained in random samples from a population with unknown subpopulations [12]. CLR is a combination of two different techniques, namely cluster technique and regression analysis. In the CLR method, the parameters that need to be estimated are the number of *clusters* and the regression coefficient for each cluster [13]. The purpose of this method is to simultaneously obtain the optimal cluster and regression model that can maximize the data [14].

Therefore, this research discusses how to model the poverty rate in Indonesia using the *clusterwise linear regression* method. The purpose of this research is to determine the poverty rate modeling in Indonesia using clusterwise linear regression.

2. RESEARCH METHODS

2.1 Data Source

Secondary data from SUSENAS and PODES from Central Bureau of Statistics (BPS) for each district and city in Indonesia for 2020 was used for this analysis. There are 514 districts/municipalities in Indonesia used as observations in this dataset based on 2020 data.

2.2 Research Variables

In this analysis, Y is the dependent variable, while X is the independent variable. Here is the list of variables included in this analysis:

Table 1. Research Variables

No.	Variables	Description
1	Y	Poverty Percentage (%)
2	X_1	Percentage of Electricity Users (%)
3	X_2	Percentage of Villages with Mining and Quarrying (%)
4	X_3	Percentage of Villages with Livestock and Fishery Commodities (%)
5	X_4	Percentage of Villages with Plantation Commodities (%)
6	X_5	Percentage of Villages with Food and Horticultural Crops (%)
7	X_6	Number of Small and Micro Industries
8	X_7	Number of Tourism Villages

2.3 Clusterwise Linear Regression

Clusterwise linear regression (CLR) is a clustering method that follows the properties of regression parameters to obtain the structure contained in random samples from a population with unknown subpopulations. The general model of *clusterwise linear regression* is

$$y_i = \sum_{k=1}^{K}\sum_{p=1}^{P} a_{ik}\, x_{ip} b_{pk} + e_i \tag{1}$$

Where y_i is the response variable at the observation $i-th$, k is the number of clusters $(1,2,\dots,K)$, p is the independent variable $(1,2,\dots,P)$, a_{ik} is a dummy variable which has the value 1 on the observation $i-th$ pada *cluster* k-th and has value 0 for others. x_{ip} is the observation of the i-th for independent variable p, b_{pk} is the regression coefficient p-th on the *cluster* k-*th*, and e_i is the error of the observation $i-th$, $e_i \sim N(0,\sigma_i^2)$ [15].

2.4 Maximum Likelihood Estimation

The estimated value of the parameter in the maximum likelihood method is sought whose expected value is the same as the parameter value (unbiased). Besides that, the estimator using the maximum likelihood method is considered consistent and efficient [16]. The following are observations in a sample of size n assumed to have a finite sum distribution or a mixture of conditional univariate normal densities.

$$y_i \sim \sum_{k=1}^{K} \lambda_k\, f_{ik}\left(y_i \middle| x_{ip,}\sigma_k^2, b_{pk}\right) \tag{2}$$

$$= \sum_{k=1}^{K} \lambda_k \left(2\pi\sigma_k^2\right)^{-1/2} exp\left[\frac{-(y_i - x_i' b_k)^2}{2\sigma_k^2}\right]$$

Where $x_i' = \left(\left(x_p\right)\right)_i$ and $b_k = \left(\left(b_p\right)\right)_k$, the meaning is that it is assumed to be an independent sample of observations of the dependent variable $y_1, y_2, \dots, y_n$ taken randomly from a mixture of conditional normal densities. The following is the likelihood function for a sample size of n.

$$L = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \lambda_k \left(2\pi\sigma_k^2\right)^{-1/2} exp \left[\frac{-(y_i - \boldsymbol{x}_i' \boldsymbol{b}_k)^2}{2\sigma_k^2} \right] \right] \tag{3}$$

The ln likelihood function is used to linearize the likelihood function, the following is a log of the likelihood function

$$\ln L = \sum_{i=1}^{n} ln \left[\sum_{k=1}^{K} \lambda_k \left(2\pi\sigma_k^2\right)^{-1/2} exp \left[\frac{-(y_i - \boldsymbol{x}_i' \boldsymbol{b}_k)^2}{2\sigma_k^2} \right] \right] \tag{4}$$

Assuming the sample error is taken randomly from the normal density function of each cluster whose proportions are not yet known $\lambda_1, \lambda_2, \dots, \lambda_k$. This function will be maximized to get the estimated value λ_k, σ_k dan b_{pk}, with constraints $0 \leq \lambda_k \leq 1, \sum_k^K \lambda_k = 1$ and $\sigma_k > 0$, for all values $k = 1,2, \dots, K$. Then to place the i-th observations into the k-th clusters is done by selecting the posterior probability value in each of the k-th clusters that is the largest:

$$\hat{p}_{ik} = \frac{\hat{\lambda}_k f_{ik}\left(y_i \middle| X_{ip}, \hat{\sigma}_k^2, \hat{b}_{pk}\right)}{\sum_{k=1}^{K} \hat{\lambda}_k f_{ik}\left(y_i \middle| X_{ip}, \hat{\sigma}_k^2, \hat{b}_{pk}\right)} \tag{5}$$

With

$$f_{ik}\left(y_i \middle| X_{ip}, \hat{\sigma}_k^2, \hat{b}_{pk}\right) = \left(2\pi\sigma_k^2\right)^{-1/2} exp \left[\frac{-\left(y_i - \boldsymbol{x}_i' \hat{\boldsymbol{b}}_k\right)^2}{2\hat{\sigma}_k^2} \right] \tag{6}$$

The function to be maximized will be solved using the Lagrange multiplier method. Then we get an estimator for λ_k, σ_k and b_{pk}.

$$\hat{\lambda}_k = \frac{\sum_{i=1}^{n} \hat{p}_{ik}}{n} \text{ dan } \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \hat{p}_{ik} \left(y_i - \boldsymbol{x}_i' \boldsymbol{b}_k\right)^2}{\sum_{i=1}^{n} \hat{p}_{ik}} \tag{7}$$

$$b_k = (\boldsymbol{X'W_kX})^{-\boldsymbol{1}}(\boldsymbol{X'W_kY}) \; dengan \; \boldsymbol{W_k} \tag{8}$$

$$= \begin{pmatrix} p_{1k} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & p_{ik} \end{pmatrix}$$

[15]

2.5 Expectation Maximization (EM) Algorithm

The iteration process of the EM algorithm consists of two stages, namely: the Expectation stage and the Maximization stage [17].

1. The Expectation Stage

   The expectation stage aims to find the log likelihood function of the estimated values of parameters λ_k, σ_k, and b_{jk} in the first iteration obtained from the initialization values

2. The Maximization Stage

   The Maximization stage aims to find estimated new parameter values λ_k^*, σ_k^*, and b_{jk}^*. Before calculating the new estimated parameter values, the Bayesian posterior probability value p_{ik} is calculated using the parameter values in the E-stage.

2.6 Research Steps

Clusterwise linear regression (CLR) modeling in this study used the *flex mix package* available in R *software*. This study used the maximum likelihood method implemented with the *Expectation Maximization* (EM) algorithm to estimate parameters. The EM algorithm has the advantage that it can solve several problems in the field of statistics, such as estimating parameters for combinations of functions and parameters from incomplete data [18]. In CLR modeling, iteration was carried out to get the model. Iteration continued until a convergent model was obtained. Then, in determining the optimal number of *clusters*, the number of *clusters* was selected when the *Bayesian Information Criterion* (BIC) was the smallest. The following are the research steps :

1. Perform descriptive statistics of research variables.
2. Determining the number of clusters (k) that are optimal $(k = 2,3,4,5)$. The number of *clusters is* chosen when the *Bayesian Information Criteria* (BIC) is minimum. Bayesian Information Criterion (BIC) is a criterion for determining the best model from several existing models [19]. BIC is consistent as the best model selection. When there are several models, including the original model, the probability that BIC will select the correct model will approach one as the number of samples approaches infinity [20]. Mathematically, the BIC value can be calculated using the formula

$$BIC = -2lnL + p\ln n \tag{9}$$

3. Perform data analysis using *Clusterwise Linear Regression* with the following steps:
   a. Determining *cluster* members and parameter estimation values using the maximum likelihood method with the EM algorithm used in the *flexmix package* in R *software* which produces *cluster* members and models that match the number of clusters k
   b. Conducting hypothesis testing using the *Wald test* on regression parameters in each model. The Wald test was carried out on each predictor variable to determine the effect each predictor variable had on the value of the response variable [21]. With test statistics

$$W = \frac{\hat{b}_p}{SE(\hat{b}_p)} \tag{10}$$

   c. Interpreting the model obtained

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistical Analysis of Research Variables

Indonesia consists of 514 regencies/cities divided into 416 districts and 98 cities. The data is standardized because each variable in this study has a different unit value. The following Table 2 displays the descriptive statistics of each variable after the data is standardized

Table 2. Descriptive Statistics

Variables	*Mean*	Variance	Minimum	Maximum
Percentage of Poverty (Y)	0	1	-1.3595	3.9727
Percentage of Electricity users (X_1)	0	1	-3.7954	0.6154
Percentage of Villages with Mining and Quarrying (X_2)	0	1	-0.2401	12.9450
Percentage of Villages with Livestock and Fishery Commodities (X_3)	0	1	-0.5080	5.0216
Percentage of Villages with Plantation Commodities (X_4)	0	1	-0.8688	2.2934
Percentage of Villages with Food and Horticultural Crops (X_5)	0	1	-1.6155	1.2026
Number of Small and Micro Industries (X_6)	0	1	-0.7168	6.0631
Number of Tourism Villages (X_7)	0	1	-0.6579	7.8172

A scatter plot between predictor variable (X) and the response variable (Y) can be seen in Figure 1.

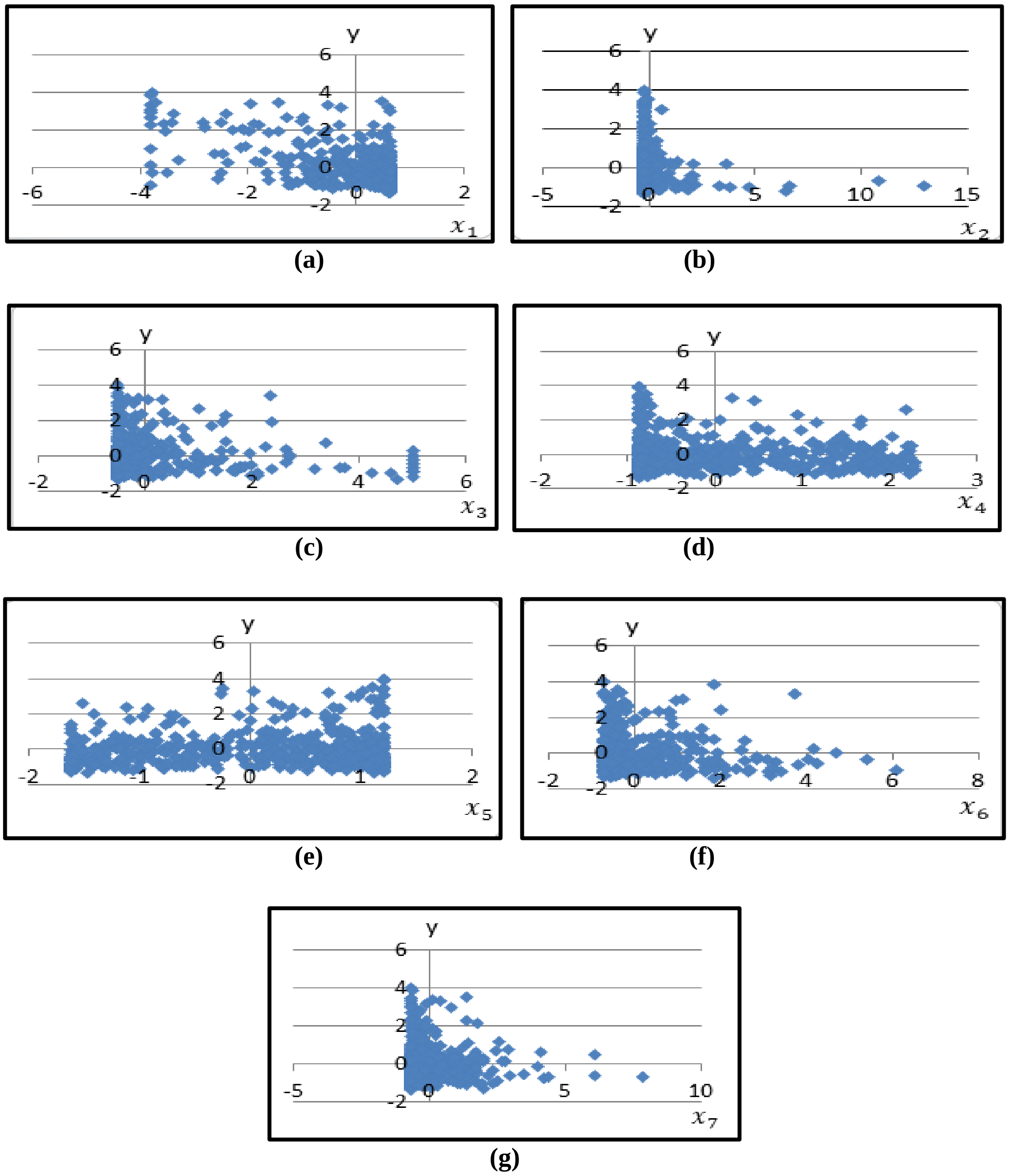

Figure 1. ***Scatter Plot*** **for Each Variable, (a) variable (X_1) with (Y), (b) variable (X_2) with (Y), (c) variable (X_3) with (Y), (d) variable (X_4) with (Y), (e) variable (X_5) with (Y), (f) variable (X_6) with (Y), (g) variable (X_7) with (Y).**

Based on Figure 1, the variable percentage of villages that have plantation commodities (X_4) and the variable percentage of villages with food crops and horticulture (X_5) are not used in the analysis because the data pattern is spread and indicates that there is no relationship between the variable (Y) with variables (X_4) and (X_5). Therefore, this study only uses five predictor variables, namely the percentage of electricity users (X_1), percentage of villages that have mining and quarrying (X_2), the percentage of villages that have livestock and fishery commodities (X_3), the number of small and micro industries (X_6), and the number of tourism villages (X_7) because it illustrates that the data collect in an area and looks to form a *cluster* then indicates that there is a relationship between the variable and the variable (Y) with variables (X_1), (X_2), (X_3), (X_6) and (X_7).

3.2 Modeling with Clusterwise Linear Regression

Modeling is usually done without observing *clusters* in the data, but the model obtained could be better when observations form a *cluster* [22]. This study's use of *clusterwise linear regression* (CLR) is expected to provide good modeling results. By using the EM algorithm, the results of *clusterwise linear regression* are obtained with $(k = 2,3,4,5)$ as in Table 3 below:

Table 3. Results of *Clusterwise Linear Regression* with $(k = 2, 3, 4, 5)$

k	AIC	BIC
2	1187.516	1251.149
3	1112.66	1210.231
4	1109.903	1241.412
5	1080.89	1246.337

Furthermore, changes in the BIC value must be considered to determine the optimal number of *clusters* to choose from. The best model is the model with the lowest BIC value. Table 3 shows that the BIC value continues to increase when $k > 3$, so the iteration process k is cut at $k = 4$. Then, the optimal *cluster* is determined by the smallest BIC value, namely 1210.231. Therefore, the most recommended model is the model with three *clusters*.

By using $k = 3$, clusters with sizes as shown in Table 4 below are obtained:

Table 4. *Cluster* Size

Cluster 1	*Cluster 2*	*Cluster 3*
225	23	266

In Cluster 1, there are 225 regencies/cities*; in Cluster* 2, there are 23 regencies/cities*; in Cluster* 3, there are 266 regencies/cities. Visually, the poverty rate *cluster* can be seen in Figure 2 below.

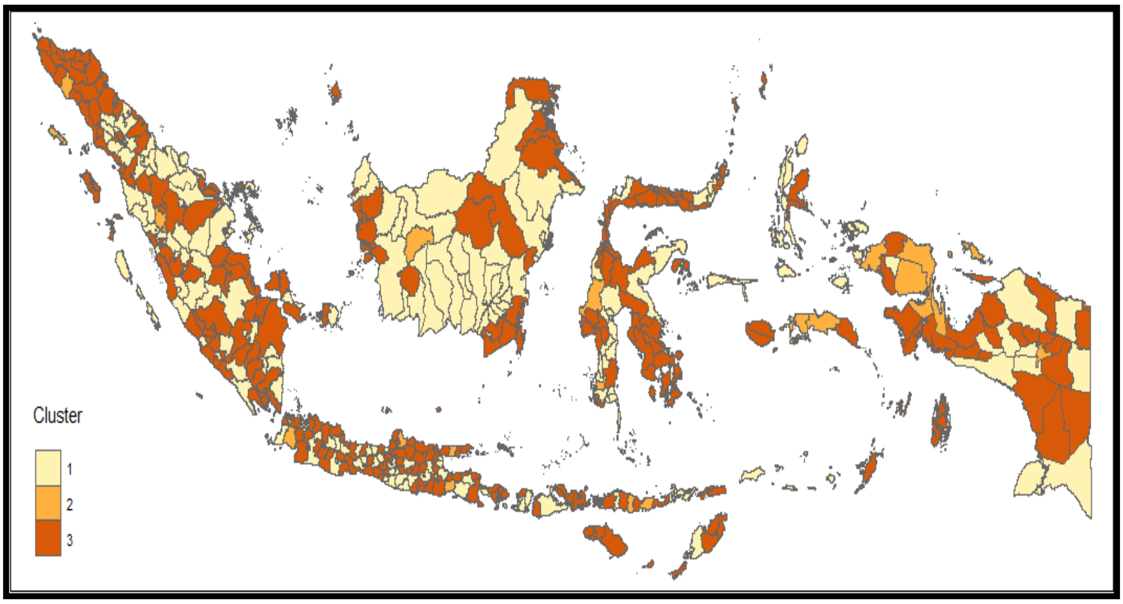

Figure 2. Poverty Rate *Cluster*

Based on **Figure 2,** most of the area of Cluster 1 is on Sumatra Island, most of the area of Cluster 2 is on Papua Island, and most of the area of Cluster 3 is on Sumatra Island.

The following are the parameter estimates of each *cluster* presented in Table 5.

Table 5. Parameter Estimation

Variables	Parameter Estimation of Each *Cluster*		
	1	2	3
(*Intercept*)	−0.5012	1.6414	0.1238
X_1	−0.1553	−0.3887	−0.7131
X_2	−0.0341	−0.3030	−0.1887
X_3	−0.0020	0.4593	−0.0229
X_6	−0.0651	−0.0715	−0.0608
X_7	0.3541	0.8364	−0.0351

From Table 5, the estimated model for the regency/city poverty case in Indonesia is as follows.

$$\hat{y}_i = -0.5012a_{i1} + 1.6414a_{i2} + 0.1238a_{i3} - 0.1553a_{i1}X_{i1} - 0.3887a_{i2}X_{i1} - 0.7131a_{i3}X_{i1} - 0.0341a_{i1}X_{i2} - 0.3030a_{i2}X_{i2} - 0.1887a_{i3}X_{i2} - 0.0020a_{i1}X_{i3} + 0.4593a_{i2}X_{i3} - 0.0229a_{i3}X_{i3} - 0.0651a_{i1}X_{i6} - 0.0715a_{i2}X_{i6} - 0.0608a_{i3}X_{i6} + 0.3541a_{i1}X_{i7} + 0.8364a_{i2}X_{i7} - 0.0351a_{i3}X_{i7} \quad (11)$$

The variable a in **Equation 11** is a *dummy* variable that describes the *cluster* in question. For example, for *cluster* 1, the value of $a_{i1} = 1$, while a_{i2} and a_{i3} are each worth 0.

3.3 Parameter Testing

The importance of each parameter was tested to see if the predictor variable has a meaningful effect on the model, known as "parameter testing." Wald test was used to test the hypothesis of parameters in CLR; this test was performed on each **predictor** variable.

1. Testing hypothesis

$H_0: b_1 = 0$ vs $H_1: b_1 \neq 0;$

$H_0: b_2 = 0$ vs $H_1: b_2 \neq 0;$

$H_0: b_3 = 0$ vs $H_1: b_3 \neq 0;$

$H_0: b_6 = 0$ vs $H_1: b_6 \neq 0;$

$H_0: b_7 = 0$ vs $H_1: b_7 \neq 0;$

2. The real test level (α) used is equal to 5%

3. Test Statistics:

$$W = \frac{\hat{b}_p}{SE(\hat{b}_p)}$$

4. Rejection criteria:

Reject H_0 if $p_{value} \leq \alpha$

Accept H_0 if $p_{value} > \alpha$

5. Draw Conclusions

By using R *software, the Wald* Test value for *Cluster* 1 is obtained as follows:

Table 6 The Results of *the Wald* Test on *Cluster* 1

Variables	*Std.Error*	*z-value*	*p-value*
X_1	0.0314	-4.9502	$7.414e-07$ ***
X_2	0.0187	-1.7881	0.0738.
X_3	0.0281	-0.0982	0.9217
X_6	0.0279	-2.3309	0.0198 *
X_7	0.0453	7.8859	$3.123e-15$ ***
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 1			

Based on **Table 6** for *cluster* 1, it is known that a significance level of (α) by 5% indicates that the influential variables in the model are the percentage of electricity users (X_1), the number of small and micro industries (X_6), and the number of tourist villages (X_7).

By using R *software, the Wald* test value for *Cluster* 2 is obtained as follows:

Table 7 The Results of *the Wald* Test on *Cluster* 2

Variables	*Std.Error*	*z-value*	*p-value*
X_1	0.2897	-1.3183	0.1874
X_2	0.3711	-0.8264	0.4086
X_3	0.3686	1.2242	0.2209
X_6	0.2070	-0.3565	0.7215
X_7	0.3326	2.4819	0.0131 *
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 1			

Table 7 shows that a significance level (α) by 5% indicates that the influential variable in the model is the number of tourist villages. (X_7).

By using R *software, the Wald* test value for *Cluster* 3 is obtained as follows:

Table 8 The Results of *the Wald* Test on *Cluster* 3

Variables	*Std.Error*	*z-value*	*p-value*	
X_1	0.0517	-13.7961	$2.2e-16$	***
X_2	0.0549	-3.4261	0.0006	***
X_3	0.0474	-0.4637	0.6428	
X_6	0.0433	-1.3925	0.1638	
X_7	0.0387	-0.8782	0.3799	

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' 1

Based on **Table 8,** it is known that a significance level (α) by 5% indicates that the influential variables in the model are the percentage of electricity users (X_1) and the percentage of villages that have mining and quarrying (X_2).

Based on the modeling, it can be explained that in *cluster* 1, most areas are on Sumatra Island. In *cluster 1, the* decline in poverty is influenced by electricity users and total small and micro industries. It means that the more electricity users and total small and micro industries located in regencies/cities in Indonesia in *cluster* 1 will reduce the poverty rate. Then, it is also obtained from the model that the number of tourist villages does not guarantee that the poverty rate can decrease in Indonesian districts/cities in *cluster* 1. Moreover, *cluster* 2, with the most territory on the Papua island, explains that the number of tourist villages does not guarantee the poverty rate in Indonesian districts/cities. Then, *cluster* 3, with the most territory on Sumatra island, explains that the poverty reduction is influenced by electricity users and villages that have mining and quarrying, where the more electricity users and the more villages that have mining and quarrying located in regencies/cities in Indonesia will reduce the poverty rate of districts/cities in Indonesia.

4. CONCLUSION

1. Referring to the results of analyzing and discussing, the best *clusters linear regression* model with 3 *clusters* for the poverty rate in Indonesia in 2020 is as follows:

$$\hat{y}_i = -0.5012a_{i1} + 1.6414a_{i2} + 0.1238a_{i3} - 0.1553a_{i1}X_{i1} - 0.3887a_{i2}X_{i1} - 0.7131a_{i3}X_{i1} - 0.0341a_{i1}X_{i2} - 0.3030a_{i2}X_{i2} - 0.1887a_{i3}X_{i2} - 0.0020a_{i1}X_{i3} + 0.4593a_{i2}X_{i3} - 0.0229a_{i3}X_{i3} - 0.0651a_{i1}X_{i6} - 0.0715a_{i2}X_{i6} - 0.0608a_{i3}X_{i6} + 0.3541a_{i1}X_{i7} + 0.8364a_{i2}X_{i7} - 0.0351a_{i3}X_{i7}$$

2. From the *clusters linear regression* model, using (α) by 5%, It is found that for *cluster* 1, the influential variables in the model are the percentage of electricity users (X_1), the number of small and micro industries (X_6), and the number of tourist villages (X_7). Meanwhile, *cluster* 2 shows that the influential variables in the model are the number of tourist villages. (X_7). Meanwhile, *cluster* 3 shows that the influential variables in the model are the percentage of electricity users (X_1) and the percentage of villages with mining and quarrying (X_2).

REFERENCES

[1] M. J. Pandu, "Pemodelan Persentase Kemiskinan di Provinsi Nusa Tenggara Timur dengan Menggunakan Model Ekonometrika Spasial Data Panel Dinamis," Thesis, Institut Teknologi Sepuluh Nopember, Surabaya, 2016.
[2] P. S Gopal, M. A. Abdul Rahman, N. M. Malek, P. S. Jamir Singh, and L. Chee Hong, "Kemiskinan Adalah Satu Fenomena Multidimensi: Suatu Pemerhatian Awal," *Malaysian J. Social Sciences Humanities (MJSSH).*, vol. 6, no. 1, pp. 40–51, 2021, doi: 10.47405/mjssh.v6i1.608.
[3] E. H. Jacobus, P. Kindangen, and E. N. Walewangko, "Analisis Faktor-faktor yang Mempengaruhi Kemiskinan Rumah Tangga di Sulawesi Utara," *J. Pembangan Ekonomi dan Keuangan Daerah.*, vol. 19, no. 3, pp. 86–103, 2018.
[4] M. Ramdani, "Determinan Kemiskinan di Indonesia Tahun 1982-2012," *Economics Development Analysis Journal.*, vol. 4, no. 1, pp. 58–64, 2015.
[5] BPS, "Profil Kemiskinan di Indonesia September 2020,". Jakarta: Badan Pusat Statistika, 2021.
[6] A. Husniyah, N. A. Basyah, and A. Mustofa, "Dampak Usaha Mikro Kecil dan Menengah (UMKM) Terhadap Pengentasan Kemiskinan dan Pengangguran di Kota Sibolga,"*Journal of Economica Didactica.*, vol. 2, no. 1, 2022.
[7] I. N. Sudiarta and I. W. Suardana, "Dampak Pariwisata terhadap Kemiskinan di Kawasan Pariwisata di Bali," *J. Kajian Bali*, vol. 6, no. 2, pp. 209–228, 2016.
[8] N. Zuhdiyaty and D. Kaluge, "Analisis Faktor-Faktor yang Mempengaruhi Kemiskinan di Indonesia Selama Lima Tahun Terakhir (Studi Kasus Pada 33 Provinsi)," *J. Ilmiah Bisnis dan Ekonomi Asia*, vol. 11, no. 2, pp. 27–31, 2017.
[9] Yaziz, D. Kusnandar, and S. W. Rizki, "Analisis Regresi Robust Estimasi-M dengan Menggunakan Pembobotan Bisquare dan Welsch dalam Mengatasi Data Outlier," *Buletin Ilmiah Mat, Stat, dan Terapannya (Bimaster)*, vol. 08, no. 4, pp. 799–804, 2019.

[10] M. suharsono Putri, “Analisis regresi linear gerombol dengan algoritma pertukaran (exchange algorithm),” Thesis, Institut Pertanian Bogor, Bogor, 2015.
[11] B. Ari and H. A. Guvenir, “Clustered linear regression,” *Elsevier*, vol. 15, no. 3, pp. 169–175, 2002, doi: 10.1016/S0950-7051(01)00154-X.
[12] G. Qian and Y. Wu, “Estimation and Selection in Regression Clustering,” *European Journal of Pure Applied Math.*, vol. 4, no. 4, pp. 455–466, 2011.
[13] A. M. Bagirov, A. Mahmood, and A. Barton, “Prediction of monthly rainfall in victoria, Australia: Clusterwise linier regression approach,” *Atmospheric Research.*, vol. 188, pp. 20–29, 2017, doi: 10.1016/j.atmosres.2017.01.003.
[14] De Sarbo W.S., Oliver R.L., and Rangaswamy“A simulated Annealing Methodology for Clusterwise Linear Regression,” *J. Psychometrika.*, vol. 54, no. 4, pp. 707–736, 1989.
[15] De Sarbo W.S., and W. L. Cron, “A maximum likelihood methodology for clusterwise linear regression,” *Journal of Classification.*, vol. 5, no. 2, pp. 249–282, 1988, doi: 10.1007/BF01897167.
[16] K. M. Ramachandran and C. P. Tsokos, “Mathematical Statistics With Applications”, New York (US): *Elsevier Academic Press*, 2009.
[17] K. T and G. Mclachlan, “The EM Algorithm,” *ResearchGate*, 2014.
[18] T. A. Kusuma and Suparman, “Algoritma Expectation-Maximization (EM) Untuk Estimasi Distribusi Mixture,” *J. Konvergensi*, vol. 4, no. 2, pp. 65–83, 2014.
[19] Y. D. Nugroho and I. Sandyawan, “Analisis Kasus Covid-19 Berbasis Google Trends di Indonesia pada Masa New Normal,” in *Statistics in the New Normal : a Challenge of Big Data and Official Statistics*, 2020, pp. 349–358.
[20] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning Data Mining, Inference, and Prediction” *Springer*. 2009.
[21] M. M. Majore, D. T. Salaki, and J. D. Prang, “Penerapan Regresi Binomial Negatif dalam Mengatasi Overdispersi Regresi Poisson Pada Kasus Jumlah Kematian Ibu,” *J. Mat. dan Apl.*, vol. 9, no. 2, pp. 133–139, 2020.
[22] R. Ardinsyah, “Implementasi Algoritma EM pada Metode Kemungkinan Maksimum untuk Pemodelan Regresi Linear Gerombol,” Skripsi, Institut Pertanian Bogor, Bogor, 2014.