BAREKENG: Journal of Mathematics and Its Applications
September 2026 Volume 20 Issue 3 Page 2099-2116
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol20iss3pp2099-2116

BETA REGRESSION MODELING ON POVERTY DATA IN INDONESIA 2019 - 2022

Muhammad Arib Alwansyah[1*], **Sigit Nugroho**[2], **Ramya Rachmawati**[3]

[1]*Department of Mathematics Education, Faculty of Mathematics and Natural Sciences, Universitas Negeri Jakarta*
Jln. Rawamangun Muka, Jatinegara Kaum, Pulo Gadung, ADM. Jakarta Timur, 13220, Indonesia

[2,3]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Bengkulu*
Jln. WR. Supratman Kandang Limun, Muara Bangkahulu, Bengkulu, 38122, Indonesia

*Corresponding author's e-mail: * **muhammadarib@unj.ac.id***

Article Info

Article History:

Received: 28th July 2025
Revised: 7th December 2025
Accepted: 16th March 2026
Available online: 8th April 2026

Keywords:

Beta Regression Model;
Newton-Raphson Method;
Percentage of poor population.

ABSTRACT

The Central Statistics Agency (BPS) reported that the percentage of poor people in Indonesia increased from 2019 to 2021, reaching 10.14 percent. This condition highlights the need for an analytical approach capable of accurately modeling percentage data that are naturally bounded between 0 and 1. This study introduces a new approach by applying the Beta regression model to analyze the factors influencing poverty levels across Indonesian provinces. The novelty of this research lies in the application of the Beta regression model to panel data on poverty, which remains rarely explored in empirical studies on Indonesia's socio-economic indicators. The model was chosen because it provides more efficient and unbiased parameter estimates than the ordinary least squares (OLS) method, especially when the dependent variable exhibits asymmetry and heteroskedasticity. Parameter estimation was conducted using the Maximum Likelihood Estimation (MLE) method with the Newton–Raphson iterative algorithm to ensure convergence and estimation efficiency. The data used in this study are provincial-level poverty data sourced from official publications by the BPS. The analysis results indicate that the model meets the model suitability criteria for 2019 and 2020. Factors that significantly influenced the percentage of poor people in both years included the percentage of the population with health insurance and the literacy rate. Meanwhile, in 2021 and 2022, factors that significantly influenced the percentage of the poor population included the average years of schooling, the percentage of the population with health insurance, and the literacy rate. This study contributes to the field of applied statistics by demonstrating that the Beta regression model offers a novel and robust alternative for analyzing bounded and asymmetric socio-economic data. Furthermore, it provides new empirical insights into the statistical modeling of poverty in Indonesia, offering a methodological advancement over traditional regression approaches.

How to cite this article:

M. A. Alwansyah, S. Nugroho, and R. Rachmawati, "BETA REGRESSION MODELING ON POVERTY DATA IN INDONESIA 2019 – 2022", *BAREKENG: J. Math. & App.*, vol. 20, no. 3, pp. 2099-2116, Sep, 2026.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article · Open Access

1. INTRODUCTION

Indonesia is an archipelagic country that is rich in natural resources. Natural wealth that is utilized well will certainly make this country even better. However, as a developing country, Indonesia still has several social, cultural, and economic challenges that require serious attention. One of the main problems that continues to be faced is poverty [1]. Statistics Indonesia shows that the percentage of poor people in Indonesia has increased from 2019 to 2021, reaching 10.14 percent in 2021. The problem of poverty remains a focus that needs to be addressed further [2]. The emergence of poverty problems is caused by some residents who have not yet achieved an adequate level of welfare to improve their quality of life [3].

The high poverty level cannot be separated from the many factors that influence it. According to [4], the causes of poverty can come from individual characteristics and household characteristics, which are grouped into three categories, namely demographic categories, economic categories, and social categories. The poverty level of a region can be seen from the proportion of poor people, with a high percentage of poor people reflecting a high level of poverty in that region. The percentage of poor people is those living below the poverty line.

This study uses data from 2019–2022 because that period coincided with the COVID-19 pandemic, which significantly impacted the social and economic conditions of Indonesians. The focus of this study is not on long-term trends, but rather on the impact of the pandemic on poverty and its contributing factors in the short term. Furthermore, it should be emphasized that although the poverty rate decreased after the pandemic, the presence of people still living below the poverty line remains a critical issue that requires ongoing attention. Research on poverty conducted by [5] shows that the percentage of poor people is influenced by factors such as literacy rate, number of workers in the agricultural sector, economic growth, unemployment rate, and School Enrollment Rate. Further research conducted by [6] found that the factors that influence poverty are the human development index and the unemployment rate.

One of the most widely used methods to determine the factors that influence the percentage of poor people is the classical regression method. However, in some studies, the use of classical regression models cannot always be applied perfectly, because classical regression is not suitable for estimating data in the form of proportions. Proportion data, such as poverty rates, are naturally bounded between 0 and 1, and often exhibit asymmetric distributions and non-constant variance (heteroskedasticity). These characteristics violate the basic assumptions of the ordinary least squares (OLS) method, which assumes normally distributed and homoscedastic residuals. To overcome these limitations, one appropriate approach is to use the Beta regression model [7]. The Beta regression model is specifically designed to handle dependent variables that follow a Beta distribution, which is inherently flexible and capable of modeling data with various shapes-symmetric, right-skewed, or left-skewed-within the $(0,1)$ interval [8]. This makes the Beta regression model a powerful alternative that provides more accurate and efficient parameter estimates compared to classical regression, especially for analyzing proportion data such as the percentage of poor people.

Previous research on the application of the Beta regression model has demonstrated its effectiveness in analyzing proportion data. For example, research conducted by [8] found that the cumulative grade point average (GPA) influences student graduation rates, while research by [9], using district/city-level neonatal mortality rate data in West Java Province in 2020, showed that low birth weight was a significant factor influencing the proportion of neonatal deaths. Both studies demonstrate that the Beta regression model is particularly suitable for use when the dependent variable is a proportion or percentage with a value between 0 and 1, as in this situation, the normality assumption of the classical linear regression model is not met.

Theoretically, the Beta regression model is developed based on the Beta distribution, which has a probability density function: $f(x;\alpha,\beta)=\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\quad 0<x<1$. With $\alpha>0$ and $\beta>0$ as shape parameters that control the skewness and kurtosis of the distribution [10]. This distribution is very flexible because it can represent various data distribution shapes, whether symmetrical, right-skewed, or left-skewed, making it suitable for modeling proportion data that is not always normally distributed. The Beta regression model is generally expressed as: $g(\mu_i)=\boldsymbol{x}_i'\boldsymbol{\beta}$. Where $\boldsymbol{\mu_i}=\boldsymbol{E}(\boldsymbol{Y_i})$ is the expected value of the dependent variable $\boldsymbol{Y_i}$, $\boldsymbol{X_i}$ is the vector of covariates, and β is the regression parameter estimated using the Maximum Likelihood Estimation (MLE) method. The link function $\boldsymbol{g}(\cdot)$ that is often used is logit, probit, or cloglog, depending on the characteristics of the data being analyzed [11]. Thus, Beta regression allows researchers to explain the relationship between explanatory variables and the dependent variable in the form of proportions more accurately than classical linear models, especially when there is heteroscedasticity or asymmetric data distribution.

Parameter estimation in Beta regression in this study used the maximum likelihood method with an iterative approach, namely Newton-Raphson iteration. Because the data analyzed were the percentage of the poor population, Beta regression was chosen due to its superior and more precise parameter estimation capabilities compared to classical regression methods. Beta regression is designed to handle situations where the dependent variable is proportionate and follows a Beta distribution. This model provides more accurate and efficient parameter estimation, especially when the data exhibits symptoms of an asymmetric distribution or heteroscedasticity. Therefore, the application of Beta regression in this study is very appropriate, because it aims to determine the factors that influence the poverty rate in Indonesia, which is measured in percentage form.

The novelty of this research lies in the application of the Beta regression model to analyze provincial-level poverty data, a practice rarely employed in previous studies. Most previous studies tend to use linear regression or other classical regression models, even though percentage data that do not meet the assumptions of normal distribution or homogeneity of variance require a more appropriate approach. Furthermore, this research also provides an empirical contribution by identifying specific factors that significantly influence poverty across different periods (2019–2022), such as the human development index, average years of schooling, literacy rate, and health insurance coverage.

2. RESEARCH METHODS

2.1. Regression Analysis

Regression analysis is a statistical method used to describe, measure, and analyze the functional relationship between one or more independent variables and a dependent variable. According to [12], regression analysis is used not only to determine the direction and strength of the relationship between variables but also to build predictive models that can be used in data-driven decision-making. Therefore, regression analysis plays a crucial role in understanding the patterns of relationships between quantitative phenomena. Meanwhile, [13] explains that regression analysis is one of the most widely used statistical analysis techniques in empirical research due to its ability to explain the causal relationship between several independent variables and a dependent variable. Through parameter estimation and significance testing, regression analysis can be used to assess the relative influence of each independent variable, thus helping researchers identify the most dominant factors in explaining variation in the dependent variable. Therefore, regression analysis is a key tool in various fields such as economics, social sciences, and other applied sciences, as it can provide a quantitative picture of complex inter-variable relationships.

2.2 Beta Distribution

Beta distribution is one of the continuous probability distributions limited to the range $(0,1)$. This distribution is determined by two positive parameters, symbolized as α and β, and functions as an exponent of the random variable that forms the characteristics of the Beta distribution. The following Eq. (1) is the equation that describes the probability density function of the Beta distribution:

$$f(x;\alpha,\beta)=\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1},\quad 0<x<1 \tag{1}$$

In the Beta distribution, determining the expected values of X and X^2 will be easier if we first determine the k-th moment around the center point. Because the Beta distribution is continuous, the expected value uses an integral. Here are the k-th moments around the center point of the Beta distribution with $k=1,2,3,\ldots$:

$$E\left(X^k\right)=\int_{-\infty}^{\infty}x^k f(x)\,dx=\int_0^1 x^k\frac{1}{B(\alpha,\beta)}x^{\alpha-1}(1-x)^{\beta-1}\,dx$$

$$=\frac{1}{B(\alpha,\beta)}\int_0^1 x^{k+\alpha-1}(1-x)^{\beta-1}\,dx$$

$$= \frac{1}{B(\alpha,\beta)} B\big((k+\alpha),\beta\big) = \frac{1}{\frac{\Gamma(\alpha)\,\Gamma(\beta)}{\Gamma(\alpha+\beta)}} \cdot \frac{\Gamma(k+\alpha)\Gamma(\beta)}{\Gamma\big((k+\alpha)+\beta\big)} = \frac{\Gamma(\alpha+\beta)\Gamma(k+\alpha)}{\Gamma(\alpha)\,\Gamma(k+\alpha+\beta)}, \tag{2}$$

thus, to search $\mu = E(X)$ can be used Eq. (2) with $k = 1$,

$$E(X) = \frac{\Gamma(\alpha+\beta)\Gamma(1+\alpha)}{\Gamma(\alpha)\,\Gamma(1+\alpha+\beta)} = \frac{\alpha}{\alpha+\beta}, \tag{3}$$

for X^2 with $k = 2$,

$$E(X^2) = \frac{\Gamma(\alpha+\beta)\Gamma(2+\alpha)}{\Gamma(\alpha)\,\Gamma(2+\alpha+\beta)} = \frac{\alpha(\alpha+1)}{(\alpha+\beta)\,(\alpha+\beta+1)}.$$

Thus, $Var(X)$ can be obtained as follows:

$$Var(X) = E\left[\big(X - E(X)\big)^2\right] = E(X^2) - \big(E(X)\big)^2 = \frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} - \left(\frac{\alpha}{\alpha+\beta}\right)^2$$
$$= \frac{1}{\alpha+\beta}\left(\frac{\alpha^2+\alpha}{\alpha+\beta+1} - \frac{\alpha^2}{\alpha+\beta}\right) = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}. \tag{4}$$

2.3 Beta Regression Model

The Beta distribution probability density function can be seen in Eq. (1), with parameters α and β. However, in modeling, it is often useful to model the average of the dependent variable. Additionally, it is important to define the model in such a way that it contains a dispersion parameter. To form a Beta regression model involving the average of the dependent variable along with the dispersion parameter, reparameterization of the probability density function of the Beta distribution is required. For example, $\mu = \frac{\alpha}{\alpha+\beta}$ and $\kappa = \alpha + \beta$. Thus, it is found that $\alpha = \mu\kappa$ and $\beta = (1-\mu)\kappa$. Based on Eqs. (3) and (4) then:

$$E(X) = \mu, \text{and } Var(X) = \frac{\mu(1-\mu)}{\kappa+1}.$$

It is known that μ is the average of the dependent variable and κ is the precision parameter. This means that for a certain μ, a larger value of κ will make the variance for the dependent variable smaller.

After going through the reparameterization process, the function involving the independent variable X and following the Beta distribution can be formulated as follows:

$$f(y_i;\mu,\kappa) = \frac{\Gamma(\kappa)}{\Gamma(\mu_i\kappa)\Gamma\big((1-\mu_i)\kappa\big)} {y_i}^{\mu_i\kappa-1}(1-y_i)^{(1-\mu_i)\kappa-1}, \quad 0 < y < 1. \tag{5}$$

This parameterization statement indicates that $0 < \mu < 1$ and $\kappa > 1$, and it can also be concluded that $\alpha = \mu\kappa > 0$ and $\beta = \kappa(1-\mu) > 0$. Thus, the Beta regression model allows researchers to link the expected value (μ_i) of the dependent variable with the independent variables (x_i) through a link function such as logit or probit, so that the mathematical relationship can be expressed as:

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

This formulation makes Beta regression a powerful approach for analyzing proportion data with non-constant variance (heteroscedasticity) or asymmetric distributions, which are often beyond the reach of classical linear regression.

2.3.1 Establishment of a Beta Regression Model

In forming the Beta regression model, a general linear model approach involving two link functions was used. The first function is for the location parameter μ, while the other function is used for the dispersion parameter κ. According to [14], these two functions are smooth, non-linear, and monotonous, mapping from an infinite linear free space into a sample space constrained to the interval (0,1). Suppose $\boldsymbol{X}$ and $\boldsymbol{W}$ are covariate matrices, with each matrix having the i-th row vector, namely $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$. Let also suppose that $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are each a vector of Beta regression coefficients. The general linear model for location parameters is:

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta}, \qquad i = 1, 2, \dots, n\,,$$

where g(·) is a monotone function, namely a link function that has derivatives. The following is a general form that shows the correlation between the average value and the variance [14]:

$$\sigma_i^2 = v(\mu_i)u(\kappa_i),$$

where v and u are functions that have no negative value. The precision parameter κ_i is assumed to be of a form that can be modeled as [14]:

$$h(\kappa_i) = \boldsymbol{w_i\delta}.$$

According to [15], the likelihood function for $\boldsymbol{\beta}$ when δ is constant will define the location submodel. On the other hand, if β is considered constant, then the likelihood function for $\boldsymbol{\delta}$ will form a dispersion submodel.

This research uses the logit link function because it can map $\mu \in (0,1)$ into a sample space that matches its distribution. Furthermore, the precision parameter κ is positive because the variance cannot have a negative value. The log link function that satisfies these properties is:

$$log(\kappa_i) = -\boldsymbol{w}_i\delta.$$

2.3.2 Parameter Estimation

Suppose $x_1, \dots, x_n$ are independent random variables, known for each $i = 1, \dots, n$, x_i following the distribution defined in Eq. (5) with mean μ_i and an unknown dispersion parameter κ which can be formulated as follows:

$$g(\mu_i) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \sum_{j=0}^{q} x_{ij}\beta_j = \eta_i, \qquad i = 1, 2, \dots, n, \tag{6}$$

with $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_q)$ is a vector of regression parameters, and $x_{i0}, x_{i1}, \dots, x_{iq}$ are observation data on q covariates, with $x_{i0} = 1$. The variance of the dependent variable y is a function of μ_i. Thus, it also depends on the value of the covariate [16].

The average parameter limits lie in the open interval $(0,1)$. Therefore, there is a need for a link function that can map parameters from intervals into real number space. The logit link function has a role as a canonical link function, which will produce parameter estimates in the form of log odds ratios. The following is the definition of the logit link function $g(\mu_i)$:

$$g(\mu_i) = logit(\mu_i) = ln\left(\frac{\mu_i}{1-\mu_i}\right) = \boldsymbol{x}_i^T\boldsymbol{\beta} = \mu_i = \frac{exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}{1 + exp(\boldsymbol{x}_i^T\boldsymbol{\beta})}. \tag{7}$$

For $y_i \sim B(\mu_i, \kappa)$, the density function y_i has a Beta distribution defined as follows:

$$f(y_i; \boldsymbol{\beta}, \kappa) = \frac{\Gamma(\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)} y_i^{\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa-1}(1-y_i)^{(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa-1}. \tag{8}$$

The following is the likelihood function of the density function y_i in Eq. (8):

$$L(\beta,\kappa) = \prod_{i=1}^{n} f(y_i; \boldsymbol{\beta}, \kappa) = \prod_{i=1}^{n}\left[\frac{\Gamma(\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}\right]\left(\prod_{i=1}^{n} y_i^{\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa-1}\right)\left(\prod_{i=1}^{n}(1-y_i)^{(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa-1}\right).$$

Thus, the log-likelihood function obtained from the likelihood function is as follows:

$$\begin{aligned}\ln L(\boldsymbol{\beta},\kappa) &= \sum_{i=1}^{n} \ln f(y_i; \boldsymbol{\beta}, \kappa) \\ &= \sum_{i=1}^{n} \ln\left[\Gamma(\kappa)\left(\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\right)^{-1}\left(\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\right)^{-1}\right] + \sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa - 1)\, ln\, y_i \\ &\quad + \sum_{i=1}^{n}\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa - 1\right)\ln(1-y_i).\end{aligned}$$

Estimates for vectors $\boldsymbol{\beta}$ and κ are obtained by calculating the derivative of the log-likelihood function with respect to the corresponding parameters:

1. The log-likelihood function is derived against β_j with $j = 0, 1, 2, \dots, q$, then:

$$\frac{\partial}{\partial \beta_j} \ln L\,(\boldsymbol{\beta}, \kappa) - \hat{\kappa} \sum_{i=1}^{n} x_{ij} \frac{\frac{\partial}{\partial \beta_j}\ \Gamma\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \hat{\kappa}\right)}{\Gamma\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \hat{\kappa}\right)} + \hat{\kappa} \sum_{i=1}^{n} x_{ij} \frac{\frac{\partial}{\partial \beta_j}\ \Gamma\left(\left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)\hat{\kappa}\right)}{\Gamma\left(\left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)\hat{\kappa}\right)}$$
$$+\hat{\kappa} \sum_{i=1}^{n} x_{ij} \ln y_i - \hat{\kappa} \sum_{i=1}^{n} x_{ij} \ln(1 - y_i) = 0. \tag{9}$$

2. If the log-likelihood function is derived against κ, then:

$$\frac{n}{\Gamma(\hat{\kappa})} \Gamma'(\hat{\kappa}) - \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \kappa}\ \Gamma\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \hat{\kappa}\right)}{\Gamma\left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}} \hat{\kappa}\right)} \left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right) - \sum_{i=1}^{n} \frac{\frac{\partial}{\partial \kappa}\ \Gamma\left(\left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)\hat{\kappa}\right)}{\Gamma\left(\left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)\hat{\kappa}\right)} \left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right)$$
$$+ \sum_{i=1}^{n} \left(\boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right) \ln y_i + \sum_{i=1}^{n} \left(1 - \boldsymbol{x}_i^T \widehat{\boldsymbol{\beta}}\right) \ln(1 - y_i) = 0. \tag{10}$$

An iterative approach using the Newton-Raphson iteration method is applied to find solutions of Eqs. (9) and (10).

2.4 Newton-Raphson Method for Beta Regression Models

The general formula for applying the Newton-Raphson method in the context of Beta model regression can be formulated as follows:

$$\boldsymbol{u}_{(i)} = \begin{pmatrix} \hat{\beta}_0^{(i)} \\ \hat{\beta}_1^{(i)} \\ \vdots \\ \hat{\beta}_q^{(i)} \\ \hat{\kappa}^{(i)} \end{pmatrix};\ f\left(\boldsymbol{u}_{(i)}\right) = \begin{pmatrix} \frac{\partial}{\partial \beta_0}\ \ln L\,(\boldsymbol{\beta}, \kappa)^{(i)} \\ \frac{\partial}{\partial \beta_1}\ \ln L\,(\boldsymbol{\beta}, \kappa)^{(i)} \\ \vdots \\ \frac{\partial}{\partial \beta_q}\ \ln L\,(\boldsymbol{\beta}, \kappa)^{(i)} \\ \frac{\partial}{\partial \kappa}\ \ln L\,(\boldsymbol{\beta}, \kappa)^{(i)} \end{pmatrix}.$$

It is known that $\boldsymbol{u}_{(i)}$ is a vector of estimated values at the i-th iteration for parameters $\boldsymbol{\beta}$ and κ, while $f\left(\boldsymbol{u}_{(i)}\right)$ is a vector of scores, which is the result of the first derivative of the log-likelihood function for parameters $\boldsymbol{\beta}$ and κ. The Hessian matrix is symbolized as $H_{\ln L(\boldsymbol{\beta},\kappa)}$ a matrix whose elements come from the second derivative of the log-likelihood function with respect to the parameters $\boldsymbol{\beta}$ and κ, namely:

$$H_{\ln L(\boldsymbol{\beta},\kappa)} = \begin{bmatrix} \frac{\partial^2}{(\partial \beta_0)^2}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_0 \partial \beta_1}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \cdots & \frac{\partial^2}{\partial \beta_0 \partial \beta_q}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_0 \partial \kappa}\ \ln L\,(\boldsymbol{\beta}, \kappa) \\ \frac{\partial^2}{\partial \beta_1 \partial \beta_0}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{(\partial \beta_1)^2}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \cdots & \frac{\partial^2}{\partial \beta_1 \partial \beta_q}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_1 \partial \kappa}\ \ln L\,(\boldsymbol{\beta}, \kappa) \\ \frac{\partial^2}{\partial \beta_2 \partial \beta_0}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_2 \partial \beta_1}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \cdots & \frac{\partial^2}{\partial \beta_2 \partial \beta_q}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_2 \partial \kappa}\ \ln L\,(\boldsymbol{\beta}, \kappa) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \frac{\partial^2}{\partial \beta_q \partial \beta_0}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_q \partial \beta_1}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \cdots & \frac{\partial^2}{(\partial \beta_q)^2}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \beta_q \partial \kappa}\ \ln L\,(\boldsymbol{\beta}, \kappa) \\ \frac{\partial^2}{\partial \kappa \partial \beta_0}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{\partial \kappa \partial \beta_1}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \cdots & \frac{\partial^2}{\partial \kappa \partial \beta_q}\ \ln L\,(\boldsymbol{\beta}, \kappa) & \frac{\partial^2}{(\partial \kappa)^2}\ \ln L\,(\boldsymbol{\beta}, \kappa) \end{bmatrix}.$$

The following are the values of each element of the Hessian matrix, written mathematically, the main diagonal elements with values $s = 0, 1, \ldots, q$ and $r = 0, 1, \ldots, q$:

1. $\frac{\partial^2}{\partial\beta_s\partial\beta_r} \ln L(\boldsymbol{\beta}, \kappa) = -k^2 \sum_{i=1}^{n} \left(x_{is}x_{ir} \left[\frac{\frac{\partial^2}{\partial\beta_s\partial\beta_r} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} - \frac{\frac{\partial}{\partial\beta_r} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\frac{\partial}{\partial\beta_s} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\left(\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\right)^2} \right] \right)$

$$-k^2 \sum_{i=1}^{n} \left(x_{is}x_{ir} \left[\frac{\frac{\partial^2}{\partial\beta_s\partial\beta_r} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}{\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)} - \frac{\frac{\partial}{\partial\beta_r} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\frac{\partial}{\partial\beta_s} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}{\left[\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\right]^2} \right] \right).$$

2. $\frac{\partial^2}{\partial\beta_s\partial\kappa} \ln L(\boldsymbol{\beta}, \kappa) = \frac{\partial}{\partial\beta_s} \left[\frac{\partial}{\partial\beta_k} \ln L(\boldsymbol{\beta}, \kappa)\right]$

$$= -\sum_{i=1}^{n} x_{is} \left(\frac{\frac{\partial}{\partial\kappa} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} + (\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa) \left[\frac{\frac{\partial^2}{\partial\beta_s\partial\kappa} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} - \frac{\frac{\partial}{\partial\beta_r} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\frac{\partial}{\partial\beta_s} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\left(\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\right)^2} \right] \right)$$

$$+\sum_{i=1}^{n} x_{is} \left(\frac{\frac{\partial}{\partial\kappa} \Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa}{\Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa} + (1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa \left[\frac{\frac{\partial^2}{\partial\beta_s\partial\beta_r} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}{\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)} - \frac{\frac{\partial}{\partial\beta_r} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\frac{\partial}{\partial\beta_s} \Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}{\left[\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\right]^2} \right] \right)$$

$$+\sum_{i=1}^{n} x_{is} \ln y_i \sum_{i=1}^{n} x_{is} \ln(1-y_i).$$

3. $\frac{\partial^2}{\partial k^2} \ln L(\boldsymbol{\beta}, \kappa) = \frac{\partial}{\partial\beta_k} \left[\frac{\partial}{\partial\beta_k} \ln L(\boldsymbol{\beta}, \kappa)\right]$

$$= n \left[\frac{\Gamma''_{(\kappa)}}{\Gamma(\kappa)} - \left(\frac{\Gamma'_{(\kappa)}}{\Gamma(\kappa)}\right)^2 \right] - \sum_{i=1}^{n} (\boldsymbol{x}_i^T\boldsymbol{\beta})^2 \left(\frac{\frac{\partial^2}{\partial^2\kappa^2} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} - \left(\frac{\frac{\partial}{\partial\kappa} \Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} \right)^2 \right)$$

$$-\sum_{i=1}^{n} (\boldsymbol{x}_i^T\boldsymbol{\beta})^2 \left(\frac{\frac{\partial^2}{\partial^2\kappa^2} \Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa}{\Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa} - \left(\frac{\frac{\partial}{\partial\kappa} \Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa}{\Gamma(1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa} \right)^2 \right).$$

Noted that $\psi(u) = \frac{d}{du} \ln\Gamma(u) = \frac{\frac{\partial}{\partial u}\Gamma(u)}{\Gamma(u)}$ is the digamma function, and $\psi'(u) = \frac{d}{du}\psi(u)$ is the trigamma function. For example, $u = \boldsymbol{x}_i^T\boldsymbol{\beta}\kappa$ and $v = (1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa$, then:

1. $\frac{\frac{\partial}{\partial\beta_q}\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} = \psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(x_{iq}\kappa).$
2. $\frac{\frac{\partial}{\partial\kappa}\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)}{\Gamma(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)} = \psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(\boldsymbol{x}_i^T\boldsymbol{\beta}).$
3. $\frac{\frac{\partial}{\partial\beta_q}\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)}{\Gamma\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)} = \psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(-x_{iq}\kappa).$

So, Eqs. (9) and (10) can be written as follows:

$$\frac{\partial}{\partial\beta_q} \ln L(\boldsymbol{\beta}, \kappa) = 0 \Leftrightarrow -\hat{\kappa}^2 \sum_{i=1}^{n} \psi(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\hat{\kappa})(x_{iq})^2 - \hat{\kappa}^2 \sum_{i=1}^{n} \psi\left((1-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\hat{\kappa}\right)(x_{iq})^2$$

$$+\hat{\kappa} \sum_{i=1}^{n} x_{iq} \ln y_i - \hat{\kappa} \sum_{i=1}^{n} x_{iq} \ln(1-y_i) = 0,$$

and

$$\frac{\partial}{\partial\kappa} \ln L(\boldsymbol{\beta}, \kappa) = 0 \Leftrightarrow n\psi(\hat{\kappa}) - \sum_{i=1}^{n} \psi(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}\hat{\kappa})(\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2 - \sum_{i=1}^{n} \psi\left((1-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})\hat{\kappa}\right)(1-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}})^2$$

$$+\sum_{i=1}^{n} (\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) \ln y_i + \sum_{i=1}^{n} (1-\boldsymbol{x}_i^T\widehat{\boldsymbol{\beta}}) \ln(1-y_i) = 0.$$

Therefore, the new elements of the Hessian matrix can be written as follows:

1. $\frac{\partial^2}{\partial\beta_s\partial\beta_r} \ln L(\boldsymbol{\beta},\kappa) = \frac{\partial}{\partial\beta_s}\left[\frac{\partial}{\partial\beta_r}\ln L(\boldsymbol{\beta},\kappa)\right]$

$$= -\kappa^3\sum_{i=1}^{n}\psi'(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(x_{is})(x_{ir})^2 - \kappa^3\sum_{i=1}^{n}\psi'\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(-x_{is})\left(x_{iq}\right)^2.$$

2. $\frac{\partial^2}{\partial\beta_s\partial\kappa}\ln L(\boldsymbol{\beta},\kappa) = \frac{\partial}{\partial\beta_s}\left[\frac{\partial}{\partial\kappa}\ln L(\boldsymbol{\beta},\kappa)\right]$

$$= -\sum_{i=1}^{n}\left[\psi'(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(x_{is})\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right)^2 + \psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)2(\boldsymbol{x}_i^T\boldsymbol{\beta})(x_{is})\right]$$
$$-\sum_{i=1}^{n}\left[\psi'\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(-x_{is})(1-\boldsymbol{x}_i^T\boldsymbol{\beta})^2 + \psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)2\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(-x_{is})\right]$$
$$+\sum_{i=1}^{n}(x_{is})\ln y_i - \sum_{i=1}^{n}(-x_{is})\ln(1-y_i).$$

3. $\frac{\partial^2}{(\partial\kappa)^2} \ln L(\boldsymbol{\beta},\kappa) = \frac{\partial}{\partial\kappa}\left[\frac{\partial}{\partial\kappa}\ln L(\boldsymbol{\beta},\kappa)\right]$

$$= n\psi'(\kappa) - \sum_{i=1}^{n}\psi'(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(\boldsymbol{x}_i^T\boldsymbol{\beta})^3 - \sum_{i=1}^{n}\psi'\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(1-\boldsymbol{x}_i^T\boldsymbol{\beta})^3.$$

Thus, the general formula used in the Newton-Raphson Method for the Beta regression model case is as follows:

$$\begin{pmatrix}\hat{\beta}_0^{(i+1)}\\ \hat{\beta}_1^{(i+1)}\\ \vdots\\ \hat{\beta}_q^{(i+1)}\\ \hat{\kappa}^{(i+1)}\end{pmatrix} = \begin{pmatrix}\hat{\beta}_0^{(i)}\\ \hat{\beta}_1^{(i)}\\ \vdots\\ \hat{\beta}_q^{(i)}\\ \hat{\kappa}^{(i)}\end{pmatrix} - H^{-1}_{\ln L(x_{(i)})}\begin{pmatrix}\frac{\partial}{\partial\beta_0}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \frac{\partial}{\partial\beta_1}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \vdots\\ \frac{\partial}{\partial\beta_q}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \frac{\partial}{\partial\kappa}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\end{pmatrix}, \text{ with } \begin{pmatrix}\frac{\partial}{\partial\beta_0}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \frac{\partial}{\partial\beta_1}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \vdots\\ \frac{\partial}{\partial\beta_q}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\\ \frac{\partial}{\partial\kappa}\ln L(\boldsymbol{\beta},\kappa)^{(i)}\end{pmatrix}$$

are as follows:

$$= \begin{pmatrix} \left(-\kappa^2\sum_{i=1}^{n}\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(x_{i0})^2 - \kappa^2\sum_{i=1}^{n}\psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(x_{i0})^2 + (\kappa x_{i0})\sum_{i=1}^{n}\ln y_i - (\kappa x_{i0})\sum_{i=1}^{n}\ln(1-y_i)\right)^{(i)} \\ \left(-\kappa^2\sum_{i=1}^{n}\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(x_{i1})^2 - \kappa^2\sum_{i=1}^{n}\psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(x_{i1})^2 + (\kappa x_{i0})\sum_{i=1}^{n}\ln y_i - (\kappa x_{i1})\sum_{i=1}^{n}\ln(1-y_i)\right)^{(i)} \\ \vdots \\ \left(-\kappa^2\sum_{i=1}^{n}\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)\left(x_{iq}\right)^2 - \kappa^2\sum_{i=1}^{n}\psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)\left(x_{iq}\right)^2 + \left(\kappa x_{iq}\right)\sum_{i=1}^{n}\ln y_i - \left(\kappa x_{iq}\right)\sum_{i=1}^{n}\ln(1-y_i)\right)^{(i)} \\ \left(n\psi(\kappa) - \sum_{i=1}^{n}\psi(\boldsymbol{x}_i^T\boldsymbol{\beta}\kappa)(\boldsymbol{x}_i^T\boldsymbol{\beta})^2 - \sum_{i=1}^{n}\psi\left((1-\boldsymbol{x}_i^T\boldsymbol{\beta})\kappa\right)(1-\boldsymbol{x}_i^T\boldsymbol{\beta})^2 + (\boldsymbol{x}_i^T\boldsymbol{\beta})\sum_{i=1}^{n}\ln y_i + (1-\boldsymbol{x}_i^T\boldsymbol{\beta})\sum_{i=1}^{n}\ln(1-y_i)\right)^{(i)} \end{pmatrix}$$

2.5 Parameter Testing

The parameter significance test aims to evaluate whether the independent variable has a significant influence on the model. The simultaneous test in this study uses a partial likelihood ratio test. The test statistic in the simultaneous test is $G = -2[\ln L_R - \ln L_F]$, it is known that $L_R =$ Likelihood of the reduced model and $L_F =$ Likelihood of the full model, with the rejection criteria, namely reject H_0 if the $p\text{-}value < \alpha$, meaning that there is at least one β_j that is not equal to 0, while the partial test in this study uses a T test with the test statistic used in the partial test is $t = \left(\frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}\right)$, with the rejection criteria, namely, reject H_0 if the $p\text{-}value < \alpha$, meaning that variable β_j does not affect the model for $j = 1,2,3,\cdots,q$.

2.6 Test Assumptions

2.6.1 Normality Test

Normality checks are carried out to determine whether the residuals from the model are normally distributed or not. According to [15], several causes of residuals not being normally distributed are the presence of outliers in the data and errors in collecting data. To overcome this problem, one way that can be done is to carry out transformations such as logarithms or square roots on the data.

According to [17], there are several normality tests, namely the Jarque-Bera test, Kolmogorov-Smirnov test, Cramer-Von Mises test, Lilliefors test, Shapiro-Wilk test, Anderson-Darling test, D'Agostino-Pearson test, and Chi-Square test. Yap *et al.* [17] conducted research comparing eight normality tests on seizure data using Monte Carlo simulation. The results obtained showed that the Shapiro-Wilk test was the best test for generated data. Therefore, normality testing in this study used the Shapiro-Wilk test. Although the comparative study [17] was conducted using Monte Carlo simulation data, the results remain relevant because such simulation approaches are commonly used to assess the reliability and efficiency of various statistical tests under various data conditions. The consistent superiority of the Shapiro-Wilk test across various simulation scenarios indicates that this test has good performance in general, including when applied to empirical data.

Data normality testing was performed using the Shapiro-Wilk test [18]. The hypotheses used were H_0: residuals are normally distributed and H_1: The residuals are not normally distributed. The Shapiro-Wilk test statistic is calculated using the formula:

$$W = \frac{\left[\sum_{i=1}^{k} a_i \left(x_{[n-i+1]} - x_{[i]}\right)\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

The symbol x_i represents the value of the i-th observation of the sample data, while $\bar{x}$ is the average of the entire sample. The notations $x_{[i]}$ and $x_{[n-i+1]}$ each represents the data values that have been sorted from smallest to largest, namely order statistics, with $x_{[1]}$ being the smallest value and $x_{[n]}$ the largest value. Furthermore, a_i is the coefficient obtained based on the expected value and covariance of the order statistics of the standard normal distribution, which have been tabulated or calculated previously. The rejection region is determined based on the $p\text{-}value$. If the $p\text{-}value < \alpha$, then H_0 is rejected, meaning the residuals are not normally distributed. If the $p\text{-}value \geq \alpha$, then H_0 is accepted, meaning the residuals are normally distributed.

2.6.2 Multicollinearity Detection

Multicollinearity is a condition in which two or more independent variables in a regression model are strongly correlated with each other, meaning that there is a strong linear relationship between these variables. This can cause problems in regression analysis, such as making coefficient estimates unstable or reducing the accuracy of model predictions [19]. According to [20], multicollinearity is caused by several factors, namely the implementation of data collection, limitations contained in the model, or differences in the sample population.

Several ways to detect multicollinearity are:

1. Correlation Analysis: The correlation coefficient between independent variables should be low (below 0.8). If there is a strong relationship, it can cause multicollinearity problems [21].
2. Variance Inflation Factor (VIF): A value that describes the increase in variance of estimated parameters for all independent variables. If $VIF > 10$, then multicollinearity occurs, whereas if $VIF < 10$, then multicollinearity does not occur [19].

In this study, the VIF value was used to detect multicollinearity. The following is the VIF Eq. (11) [22]:

$$VIF_j = \frac{1}{\left(1 - R_j^2\right)} \qquad (11)$$

There are various ways to overcome multicollinearity, including:

1. Increase the number of observations
2. Remove independent variables that have a high correlation from the regression model
3. Carry out variable transformations, which can be done using the natural logarithms (ln) with the following formula:

$$Z^* = \ln(Z).$$

2.7 Selection of the Best Model

2.7.1 Akaike Information Criterion (AIC)

In this study, one criterion used to evaluate the performance of discrete time series data models is the Akaike Information Criterion (AIC) (Akaike, 1974). AIC is mathematically defined by the following formula:

$$AIC = -2\ln(L) + 2k. \tag{12}$$

The variable k represents the number of parameters estimated in the model, and L is the maximum log-likelihood value of the model. AIC balances model fit and model complexity. A lower AIC value indicates a better model, making AIC suitable for model selection.

2.7.2 Bayesian Information Criterion (BIC)

The model in this study also uses the Bayesian Information Criterion (BIC) as an indicator of model fit and complexity. The following formula mathematically defines the BIC:

$$BIC = -2\ln(L) + k\ln(n), \tag{13}$$

where n is the sample size, k is the number of estimated parameters, and L remains the maximum log-likelihood. The BIC also evaluates model fit. A lower BIC value indicates a more optimal model, especially for large data sets.

2.8 Poverty Concept

In general, there are two forms of economic conditions referred to as poverty, namely relative poverty and absolute poverty. Relative poverty refers to economic inequality, which is measured based on differences in income levels between groups of people with the lowest, middle, and highest incomes [23]. Absolute poverty refers to the application of the minimum standards necessary for a person to be able to adequately fulfill their basic life needs, including food and non-food needs. A person is considered to be poor if their average monthly per capita expenditure is below the poverty line [24].

2.8.1 Poverty Line

The approach applied by Statistics Indonesia regarding the concept of the poverty line is [2]:

1. The Poverty Line (PL) consists of a combination of the Food Poverty Line (FPL) and the Non-Food Poverty Line (NFPL). A person is considered a poor individual if their average monthly per capita expenditure is below the poverty line.
2. FPL involves expenditure required to meet minimum food requirements, which is measured by a daily calorie intake of 2100 kilocalories per capita. The details of the basic food needs commodity package include 52 types of goods.
3. NFPL refers to the minimum needs for housing, clothing, education, and health. The basic non-food goods package includes 51 types of goods in urban areas and 47 types of goods in rural areas.

The following is the formula for calculating the poverty line [2]:

$$PL = FPL + NFPL.$$

How to calculate PL:

1. The first stage involves identifying a reference population group, consisting of 20 percent of the population who are above the Temporary Poverty Line (TPL). This group is specifically defined as marginal-class residents.
2. FPL describes the total expenditure value for 52 basic food commodities consumed by the reference population, adjusted to the daily per capita standard of 2100 kilocalories. The basic formula used in calculating FPL is:

   $$FPL_j = \sum_{k=1}^{52} P_{jk} Q_{jk} = \sum_{k=1}^{52} V_{jk}.$$

   Then, FPL_j assess the balance with 2100 kilocalories by multiplying 2100 by the implicit average price of calories according to the region j of the reference population, so that:

$$\overline{HK_j} = \frac{\sum_{k=1}^{52} V_{jk}}{\sum_{k=1}^{52} K_{jk}},$$

$$F_j = \overline{HK_j} \times 2100.$$

3. NFPL is the total value of minimum needs for certain non-food commodities, including aspects such as housing, clothing, education, and health. Mathematically, the minimum non-food requirement value can be formulated as follows:

$$NF_m = \sum_{i=1}^{n} r_i \times V_i.$$

2.8.2 Poverty Indicators

Measuring absolute poverty using an income approach or the ability to meet basic needs from an expenditure perspective, produces three categories of poverty indicators, namely [2]:

1. The percentage of the population living below the poverty line is usually referred to as the *head-count index* (P_0).
2. The poverty gap index (P_1) which is a measure of the average difference between expenditure per poor individual and the poverty line.
3. The poverty severity index (P_2) which provides information about the distribution of expenditure among the population in poor conditions.

The poverty level of a region can be seen from the proportion of poor people, with a high percentage of poor people reflecting a high level of poverty in that region. The percentage of poor people (Headcount Index/P_0) is those living below the poverty line [2].

2.9 Factors Affecting Poverty

According to [4], the causes of poverty originate from household and individual characteristics.

Table 1. Main Determinants of Poverty

Characteristics	Information
Household	Employment and income structure, wages/income, type of work Education of household members and the average level of health Ownership of assets such as houses, land, equipment, and production tools Household size/number of household members Dependency ratio
Individual	Education Employment Status Health Status

Characteristics that influence poverty at the household and individual level can be grouped into three categories, namely economic, demographic, and social categories.

1. Demographic Category
   a. Household structure and size
   According to [4], socio-economic survey data in Cambodia in the 1993-1994 period showed that households with an average of around 6.6 members tended to have high levels of poverty.
   b. Dependency ratio
   This ratio is a comparison between the number of people of non-productive age, such as children and the elderly, to the population of productive age.

2. Economic Category
   a. Employment and Income
   The indicators analyzed include the number of household members working, duration of working time, type of work, wage/income level, and open unemployment rate.
   b. Ownership of assets such as land, agricultural land, living supplies, agricultural equipment, tools, house buildings, savings, and other financial assets.

3. Social Category
   a. The social category involves health aspects, such as nutritional status, infant mortality rate, level of morbidity, use of health services by households, ownership of Social Insurance Administration Organization, and life expectancy.
   b. Education is also a factor, including literacy level, average years of schooling, and availability and utilization of educational services (schools) by households.
   c. Housing aspects are also analyzed through indicators such as type of house, availability of clean water, access to communications, electricity, and level of environmental cleanliness.

Based on the theory originating from [4], this research will focus on household and individual characteristics, especially in the economic and social categories, as factors that influence population poverty. The following is a picture of the conceptual framework for determining factors that influence poverty in this research.

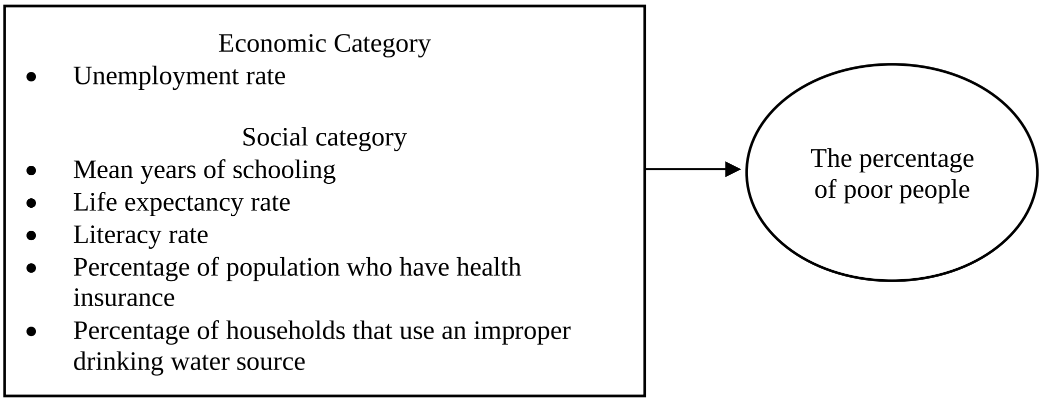

Figure 1. Research Conceptual Framework

3. RESULTS AND DISCUSSION

The type of data in this research is numerical data. Numerical data in this research are dependent variable data, such as the percentage of poor people (Y) and independent variables such as the unemployment rate (UR) (X_1), mean years of schooling (MYS) (X_2), percentage of the population who have health insurance (PPH) (X_3), life expectancy rate (LER) (X_4), literacy rate (LR) (X_5), percentage of households that use an improper drinking water source (IDW) (X_6). The research data source was taken from the publication of Statistics Indonesia. This research uses data from 2019-2022 per province in Indonesia, while the software used is R-Studio software.

3.1 Beta Regression Parameter Estimation

The parameter β_j in the model is an unknown parameter, and its value will be estimated using the Newton-Raphson iteration method. The following are influencing factors for the percentage of poor people in Indonesia:

Table 2. Beta Regression Model Parameter Estimation

Year	Variables	$\hat{\beta}_j$	Standard Error	*p-value*
	Intercept	7.50	1.98	0.0002
	UR	5.73	4.99	0.2505
	MYS	-0.05	0.12	0.6915
2019	PPH	1.58	0.53	0.0029
	LER	-0.08	0.03	0.0047
	LR	-3.76	1.84	0.0409
	IDW	-0.73	0.78	0.3534
	Intercept	7.24	1.97	0.0002
2020	UR	5,93	4.95	0.2310
	MYS	-0.01	0.13	0.9580
	PPH	1.34	0.53	0.0114

	LER	-0.08	0.03	0.0127
	LR	-4.68	1.87	0.0123
	IDW	-0.38	0.82	0.6402
2021	Intercept	7.02	2.03	0.0005
	UR	-3.69	5.51	0.5029
	MYS	0.09	0.13	0.5010
	PPH	0.85	0.57	0.1334
	LER	-0.08	0.03	0.0081
	LR	-4.71	1.95	0.0160
	IDW	-0.10	0.93	0.9166
2022	Intercept	8.42	2.04	0.0000
	UR	-2.58	5.83	0.6580
	MYS	0.07	0.12	0.5441
	PPH	0.97	0.52	0.0646
	LER	-0.09	0.03	0.0038
	LR	-5.62	2.01	0.0052
	IDW	-0.11	0.95	0.9087

Thus, the Beta regression model estimates for 2019-2022 are obtained as follows:

Year 2019:

$$log\left(\frac{\mu}{1-\mu}\right) = 7.50 + 5.73\text{UR} - 0.05\text{MYS} + 1.58\text{PPH} - 0.08\text{LER} - 3.76\text{LR} - 0.73\text{IDW} \quad (14)$$

Year 2020:

$$log\left(\frac{\mu}{1-\mu}\right) = 7.24 + 5.93\text{UR} - 0.01\text{MYS} + 1.34\text{PPH} - 0.07\text{LER} - 4.68\text{LR} - 0.38\text{IDW} \quad (15)$$

Year 2021:

$$log\left(\frac{\mu}{1-\mu}\right) = 7.02 - 3.69\text{UR} + 0.09\text{MYS} + 0.85\text{PPH} - 0.07\text{LER} - 4.71\text{LR} - 0.10\text{IDW} \quad (16)$$

Year 2022:

$$log\left(\frac{\mu}{1-\mu}\right) = 8.42 - 2.58\text{UR} + 0.07\text{MYS} + 0.97\text{PPH} - 0.09\text{LER} - 5.62\text{LR} - 0.11\text{IDW} \quad (17)$$

3.2 Testing Beta Regression Model Assumptions

3.2.1 Normality Test

Normality checks are carried out to determine whether the residuals from the model are normally distributed or not.

Table 3. Normality Test Results

Year	W Value	p-$value$	Decision	Conclusion
2019	0.966	0.361	Accepted H_0	Residues are normally distributed
2020	0.966	0.371	Accepted H_0	Residues are normally distributed
2021	0.985	0.918	Accepted H_0	Residues are normally distributed
2022	0.985	0.905	Accepted H_0	Residues are normally distributed

Table 3 shows that the data from 2019 to 2022 have $p\text{-}value \geq \alpha = 0.05$, meaning that all residuals in the data are normally distributed.

3.2.2 Multicollinearity Detection

Multicollinearity is a condition in which two or more independent variables in a regression model are strongly correlated with each other, meaning that there is a strong linear relationship between these variables. The regression model is considered good if there is no multicollinearity.

Table 4. Multicollinearity Detection Results

Year	Variables	VIF Value
2019	UR	1.584
	MYS	2.979
	PPH	1.263
	LER	1.448
	LR	2.380
	IDW	1.678
2020	UR	1.429
	MYS	3.007
	PPH	1.201
	LER	1.516
	LR	2.398
	IDW	1.654
2021	UR	1.936
	MYS	3.220
	PPH	1.290
	LER	1.391
	LR	2.300
	IDW	2.649
2022	UR	1.8934
	MYS	3.0102
	PPH	1.2477
	LER	1.4951
	LR	2.1575
	IDW	1.7513

Table 4 shows that the data from 2019 to 2022 has a VIF value < 10, meaning that all variables in the data from 2019 to 2022 do not have multicollinearity.

3.3 Beta Regression Parameter Testing

3.3.1 Simultaneous Parameter Testing

To find out whether all the variables in Eqs. (14), (15), (16), and (17) influence the model, the parameters are tested using the partial likelihood ratio test as follows:

Table 5. Simultaneous Parameter Testing Results

Year	*p-value*	Decision	Conclusion
2019	1.743×10^{-4}	Rejected H_0	There is at least one independent variable that influences the model
2020	6.082×10^{-4}	Rejected H_0	There is at least one independent variable that influences the model
2021	2.162×10^{-3}	Rejected H_0	There is at least one independent variable that influences the model
2022	1.207×10^{-4}	Rejected H_0	There is at least one independent variable that influences the model

Table 5 shows that the data from 2019 to 2022 have a $p\text{-}value < \alpha = 0.05$, indicating that at least one independent variable influences the model.

3.3.2 Partial Parameter Testing

Partial parameter testing is carried out to find out which variables have a significant influence on the model by testing each variable with the t-test. The results of partial parameter testing using the t-test are as follows:

Table 6. Partial Parameter Testing Results with t-test

Year	Variable	$t_{Statistic}$	t_{table}	*p-value*	Decision
2019	UR	1.149	2.0595	0.2505	Accepted H_0
	MYS	-0.379	2.0595	0.6915	Accepted H_0
	PPH	2.973	2.0595	0.0029	Rejected H_0
	LER	-2.828	2.0595	0.0047	Rejected H_0
	LR	-2.045	2.0595	0.0409	Rejected H_0
	IDW	-0.928	2.0595	0.3534	Accepted H_0

2020	UR	1.198	2.0595	0.2310	Accepted H_0
	MYS	-0.053	2.0595	0.9580	Accepted H_0
	PPH	2.530	2.0595	0.0114	Rejected H_0
	LER	-2.493	2.0595	0.0127	Rejected H_0
	LR	-2.502	2.0595	0.0123	Rejected H_0
	IDW	-0.467	2.0595	0.6402	Accepted H_0
2021	UR	-0.670	2.0595	0.5029	Accepted H_0
	MYS	0.673	2.0595	0.5010	Accepted H_0
	PPH	1.501	2.0595	0.1334	Accepted H_0
	LER	-2.649	2.0595	0.0081	Rejected H_0
	LR	-2.409	2.0595	0.0160	Rejected H_0
	IDW	-0.105	2.0595	0.9166	Accepted H_0
2022	UR	-0.443	2.0595	0.6580	Accepted H_0
	MYS	0.607	2.0595	0.5441	Accepted H_0
	PPH	1.847	2.0595	0.0646	Accepted H_0
	LER	-2.892	2.0595	0.0038	Rejected H_0
	LR	-2.794	2.0595	0.0052	Rejected H_0
	IDW	-0.115	2.0595	0.9087	Accepted H_0

Based on Table 6, the t-test results show that in 2019 and 2020, three independent variables had a significant influence on the model at a significance level of 5%, namely the percentage of the population who have health insurance, and the literacy rate. The final models formed for these years are as follows:

Year 2019: $$log\left(\frac{\mu}{1-\mu}\right) = 7.66 + 1.29PPH - 0.09LER - 4.01LR$$

Year 2020: $$log\left(\frac{\mu}{1-\mu}\right) = 6.99 + 1.19PPH - 0.08LER - 4.31LR$$

The t-test results based on Table 6 show that in 2021 and 2022, four independent variables have a significant effect on the model at a significance level of 5%, namely the mean years of schooling, the percentage of the population who have health insurance, and the literacy rate. The final models formed are as follows:

Year 2021: $$log\left(\frac{\mu}{1-\mu}\right) = 7.88 - 0.09LER - 3.88LR$$

Year 2022: $$log\left(\frac{\mu}{1-\mu}\right) = 9.16 - 0.10LER - 4.73LR$$

3.4 Evaluation of Model Goodness

Model quality evaluation was conducted solely to determine which of the four models performed better. In this study, AIC and BIC were used to determine the best model. The lower the AIC and BIC values, the better the model's performance. The following table shows the AIC and BIC values for each model from 2019 to 2022:

Table 7. AIC and BIC Value Results

Year	AIC	BIC
2019	-123.732	-116.100
2020	-123.409	-115.777
2021	-120.576	-114.471
2022	-126.691	-120.585

Based on Table 7, it is known that the AIC and BIC values for the 2019 data model are -123.732 and -116.100, the 2020 data model are -123.409 and -115.777, the 2021 data model are -120.576 and -114.471, and the 2022 data model are -126.691 and -120.585. The AIC and BIC values indicate that life expectancy, literacy rates, and the percentage of the population with health insurance significantly affect the model for the 2019-2022 period. This finding is in line with previous research that emphasizes the role of education, health, and human development in improving social welfare [25].

4. CONCLUSION

The regression model equation from 2019 to 2022 shows the β_j value that influences the dependent variable. The factors that influence the percentage of the poor population using Beta regression in 2019 and 2020 are the percentage of the population with health insurance, the life expectancy variable, and the literacy rate variable. The factors that influence the percentage of the poor population in 2021 and 2022 are the life expectancy variable and the literacy rate variable. The overall conclusion of this study is that there are differences in the factors that influence the percentage of the poor population in Indonesia between 2019 and 2020, as well as between 2021 and 2022. In 2019 and 2020, there is one factor that is not present in the following year, namely, the percentage of the population with health insurance.

Author Contributions

Muhammad Arib Alwansyah: Problem Formulation, Theoretical Framework Development, Beta Regression Method Selection, And Poverty Data Analysis in Indonesia. Sigit Nugroho: Validation and Interpretation of the Research Results. Ramya Rachmawati: Literature Review, Methodology Refinement, and Conclusions. All authors were actively involved in manuscript revision, provided final approval, and are accountable for the overall content of this article.

Funding Statement

This research was conducted without any funding support from government agencies or the private sector. All research activities and article preparation were carried out independently by the authors without financial assistance from any party.

Acknowledgment

The authors would like to express their appreciation to the academic institutions that provided computing and data processing support throughout the research process. They also thank their fellow academics for their valuable input and discussions regarding the application of beta regression to data on the percentage of poor people in Indonesia. They also thank the reviewers for their constructive comments that helped improve the quality of this article.

Declarations

The authors declare that there are no conflicts of interest associated with this research.

Declaration of Generative AI and AI-Assisted Technologies

AI-assisted technology (e.g., ChatGPT) was used to support light paraphrasing and sentence restructuring for clarity. The authors confirm that the underlying ideas, arguments, data analyses, and conclusions are original and were not generated by AI. All AI-assisted edits were critically reviewed and validated by the authors.

REFERENCES

[1] W. Sari *et al.*, "KEBIJAKAN ANTI KEMISKINAN PROGRAM PEMERINTAH DALAM PENANGGULANGAN KEMISKINAN DI INDONESIA," *J. Inov. Penelit.*, vol. 4, no. 1, pp. 1–10, 2022.

[2] Imam Machdi, *STATISTIK INDONESIA STATISTICAL YEARBOOK OF INDONESIA 2023*, 1st ed., vol. 1101001. Jakarta: Jakarta: Badan Pusat Statistik, 2023.

[3] S. Martha, Y. Yundari, S. W. Rizki, and R. Tamtama, "PENERAPAN METODE GEOGRAPHICALLY WEIGHTED PANEL REGRESSION (GWPR) PADA KASUS KEMISKINAN DI INDONESIA," *BAREKENG: J. Ilmu Mat. dan Terap.*, vol. 15, no. 2, pp. 241–248, 2021. doi: https://doi.org/10.30598/barekengvol15iss2pp241-248

[4] J. Haughton and S. R. Khandker, *HANDBOOK ON POVERTY AND INEQUALITY*, First edit., no. January 2009. Washington DC: The World Bank, 2009.

[5] W. Chetrin, "PEMODELAN PERSENTASE KEMISKINAN DI JAWA TIMUR DENGAN PENDEKATAN REGRESI NONPARAMETRIKA ADITIF BERDASARKAN ESTIMATOR PENALIZED SPLINE," Departemen Matematika Fakultas Sains dan Teknologi, Universitas Airlangga, 2016.

[6] N. Zuhdiyaty and D. Kaluge, "ANALISIS FAKTOR - FAKTOR YANG MEMPENGARUHI KEMISKINAN DI INDONESIA SELAMA LIMA TAHUN TERAKHIR," *J. Ilm. Bisnis dan Ekon. Asia*, vol. 11, no. 2, pp. 27–31, 2018. doi: https://doi.org/10.32812/jibeka.v11i2.42

[7] D. Kusumaningrum, A. S. E. Hidayat, K. A. Notodiputro, A. Kurnia, B. Sartono, and I. M. Sumertajaya, "FACTORS AFFECTING INDONESIAN PADDY HARVEST FAILURE: A COMPARISON OF BETA REGRESSION, QUASI-BINOMIAL REGRESSION, AND BETA MIXED MODELS," *Barekeng*, vol. 18, no. 4, pp. 2611–2622, 2024. doi: https://doi.org/10.30598/barekengvol18iss4pp2611-2622

[8] M. Al Faruq, A. S. Almaida, K. Fajri, and A. N. N. F. Tyaga, "MODEL INDEKS KEKUATAN NEGARA DENGAN REGRESI BETA: ANALISIS PENGARUH STABILITAS EKONOMI, DEMOGRAFI, DAN INFRASTRUKTUR TEKNOLOGI TERHADAP KEMAMPUAN PERTAHANAN DAN KEAMANAN," *Pop. J. Penelit. Mhs.*, vol. 4, no. 1, pp. 247–259, 2025. doi: https://doi.org/10.58192/populer.v4i1.3005

[9] B. N. Khairunnisa and N. Hajarisman, "PENERAPAN MODEL REGRESI BETA PADA PROPORSI ANGKA KEMATIAN NEONTAL TINGKAT KABUPATEN/KOTA DI PROVINSI JAWA BARAT TAHUN 2020," *Bandung Conf. Ser. Stat.*, vol. 3, no. 2, pp. 236–245, 2023. doi: https://doi.org/10.29313/bcss.v3i2.7893

[10] U. Panitanarak *et al.*, "A NEW BETA DISTRIBUTION WITH INTERDISCIPLINARY DATA ANALYSIS," *AIMS Math.*, vol. 10, no. 4, pp. 8495–8527, 2025. doi: https://doi.org/10.3934/math.2025391

[11] A. B. Simas, W. Barreto-Souza, and A. V. Rocha, "IMPROVED ESTIMATORS FOR A GENERAL CLASS OF BETA REGRESSION MODELS," *Comput. Stat. Data Anal.*, vol. 54, no. 2, pp. 348–366, 2010. doi: https://doi.org/10.1016/j.csda.2009.08.017

[12] C. C. Lee, T. Roback, and H. S. Lim, "FACTORS IMPACTING WORK ENGAGEMENT OF GEN Z EMPLOYEES: A REGRESSION ANALYSIS," *J. Leadership, Account. Ethics*, vol. 18, no. 3, pp. 147–159, 2021. doi: https://doi.org/10.1016/j.csda.2009.08.017

[13] J. Fox, *APPLIED REGRESSION ANALYSIS AND GENERALIZED LINEAR MODELS*, 3rd ed. Thousand Oaks, CA, USA: Sage Publications, 2021.

[14] A. S. Hamad, M. Mahmood, and F. Hussein, "COMPARATIVE ANALYSIS OF BETA REGRESSION MODELS (BRM) WITH DIFFERENT LINK FUNCTION USING REAL AND SIMULATION DATA : IDENTIFYING THE OPTIMAL MODEL," vol. 15, no. 2, pp. 394–404, 2025. doi: https://doi.org/10.32894/1913-015-002-028

[15] G. K. Smyth, "ROYAL STATISTICAL SOCIETY WILEY GENERALIZED LINEAR MODELS WITH VARYING DISPERSION," vol. 51, no. 1, 1989. doi: https://doi.org/10.1111/j.2517-6161.1989.tb01747.x

[16] P. R. Sihombing, K. A. Notodiputro, and A. Kurnia, "BIBLIOMETRIC MAPPING AND TREND ANALYSIS OF BETA REGRESSION MODELING : A DECADE OF," *J. dan Penelit. Tek. Inform.*, vol. 9, no. 3, pp. 1062–1072, 2025. doi: https://doi.org/10.33395/sinkron.v9i3.14949

[17] B. W. Yap and C. H. Sim, "COMPARISONS OF VARIOUS TYPES OF NORMALITY TESTS," *J. Stat. Comput. Simul.*, vol. 81, no. 12, pp. 2141–2155, 2011. doi: https://doi.org/10.1080/00949655.2010.520163

[18] S. S. Shapiro and M. B. Wilk, "AN ANALYSIS OF VARIANCE TEST FOR NORMALITY (COMPLETE SAMPLES)," *Biometrika*, vol. 52, no. 3/4, p. 591, 1965. doi: https://doi.org/10.2307/2333709

[19] T. P. Ryan, *MODERN REGRESSION METHODS*, 2nd ed., vol. 34, no. 11. John Wiley & Sons. New York., 1997.

[20] R. Rachmawati, N. Afandi, and M. A. Alwansyah, "SURVIVAL ANALYSIS ON DATA OF STUDENTS NOT GRADUATING ON TIME USING WEIBULL REGRESSION, COX PROPORTIONAL HAZARDS REGRESSION, AND RANDOM SURVIVAL FOREST METHODS," *Barekeng: J. Math. & App.*, vol. 19, no. 3, pp. 2111–2126, 2025. doi: https://doi.org/10.30598/barekengvol19iss3pp2111-2126

[21] M. Ridwan and S. Sunendiari, "MENDETEKSI DAN MENGATASI MULTIKOLINIERITAS PADA DATA PENELITIAN DIABETES MELITUS WANITA SUKU INDIAN TAHUN 2018," *Pros. Stat.*, pp. 64–70, 2021.

[22] J. Fox and G. Monette, "GENERALIZED COLLINEARITY DIAGNOSTICS," *J. Am. Stat. Assoc.*, vol. 87, no. 417, pp. 178–183, 1992. doi: https://doi.org/10.1080/01621459.1992.10475190

[23] N. S. Rahayu, "GEOGRAPHICALLY WEIGHTED PANEL REGRESSION UNTUK PEMODELAN PERSENTASE PENDUDUK MISKIN DI PROVINSI JAWA TENGAH," Institut Teknologi Sepuluh Nopember, 2017.

[24] Parkit Handono, N. D. Santi, and M. Fitria, *PENGHITUNGAN DAN ANALISIS KEMISKINAN MAKRO INDONESIA 2023*, 1st ed., 2021. Jakarta: Badan Pusat Statistik, 2023.

[25] Luhgiatno *et al.*, *PENGANTAR EKONOMI PEMBANGUNAN*, First Edit. Kabupaten Purbalingga: Eureka Media Aksara.