CAUCHY – Jurnal Matematika Murni dan Aplikasi
Volume 11 (1) (2026), Pages 1270-1285
p-ISSN: 2086-0382; e-ISSN: 2477-3344

Binary Logistic Regression Modeling using Bayesian method: Analysis and Simulation on Poverty Status of Districts in East Java

Achmad Efendi*, Restilia A. Sari, Samingun Handoyo, Nur S. Rahmi, and Friansyah Gani

Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Brawijaya, Malang, Indonesia

Abstract

This study aims to model poverty status using a Bayesian binary logistic regression approach and to identify factors influencing whether districts/cities in East Java are classified as having above-average or below-average poverty levels. The response variable is defined as a binary outcome based on the provincial average poverty rate, where observations are categorized into two groups: above-average and below-average poverty status. Bayesian estimation is employed to address potential instability in parameter estimation, particularly under small sample conditions. Parameter estimation is performed using Markov Chain Monte Carlo (MCMC) methods, specifically a Metropolis-Hastings algorithm within a Gibbs sampling framework, due to the intractability of the posterior distribution. Informative normal priors are specified based on empirical estimates obtained from maximum likelihood estimation. Convergence diagnostics indicate that the Markov chains reach stationarity after 266,000 iterations with a burn-in period of 60,000 and thinning of 10. The results show that the Human Development Index (HDI), Life Expectancy, and Gini Ratio significantly influence poverty status. The model demonstrates adequate fit based on the residual deviance test and achieves an in-sample classification accuracy of 84.2%. However, this accuracy may overestimate predictive performance since it is evaluated on the same dataset used for model estimation. Simulation studies with varying sample sizes indicate that Bayesian estimation produces estimates closer to the true parameter values compared to classical likelihood-based methods, particularly in small samples. The comparison is evaluated using bias and mean squared error (MSE), confirming the robustness of the Bayesian approach.

Keywords: Bayesian; likelihood; logistic regression; poverty level; simulation

1. Introduction

East Java has the highest poverty rate in Indonesia, yet existing models struggle with small-sample bias in district-level analysis [1]. We are trying to get a robust model, particularly with flexible sample sizes. Hence, we come up with the idea of using Bayesian estimation method used for logistic regression. In linear regression, the response variable is generally continuous, but many response variables are found to be categorical. Cases like this can be analyzed using a non-linear regression model, namely the logistic regression model. This regression model is used to determine the relationship between dichotomous or polychotomous response variables with one or more categorical or continuous predictor variables [2]. The application of logistic

*Corresponding author. E-mail: a_efendi@ub.ac.id

Submitted: February 24, 2026 Reviewed: March 14, 2026 Accepted: May 21, 2026
DOI: https://doi.org/10.18860/cauchy.v11i1.41326

regression has developed rapidly, starting from being used in research in the field of epidemiology to now being commonly used in other fields such as biomedicine, manufacturing, business and finance, and other fields.

The response variable in logistic regression can be categorical, can be nominal or ordinal [3]. Logistic regression models can be obtained by classical/likelihood or Bayesian estimation approaches [4]. In the classical approach, parameters are considered as constants whose values are not yet known. The classical approach commonly used is the Maximum Likelihood Estimation (MLE) method where the estimation process is only based on information from samples obtained from the population so that the sample size greatly influences the estimation results. Such as research by [5] using a binary logistic regression model on the open unemployment rate in West Sulawesi Province in 2017 with the Maximum Likelihood Estimation (MLE) method, it was found that gender and field of work variables affected the open unemployment rate. Binary logistic regression not only uses the classical method, but can also use the Bayesian method. The main advantage of the Bayesian method is the use of posterior probability [6]. Parameter estimation in the Bayesian method is not expressed as a point estimate, but is a statistical distribution, so it can be said that in this method the parameter is a variable that has a distribution. In the posterior, there is no longer a need to perform test statistics in inference, whereas in the classical likelihood-based method, test statistics must be used.

In the Bayesian estimation, parameter is estimated by combining the likelihood function of the sample data with the prior distribution to produce a posterior distribution. The posterior distribution, when applied to simple situations, can be solved analytically, but in complex cases, integration through simulation methods is needed. The use of binary logistic regression analysis with the Bayesian method has been applied in several studies, such as those conducted [7][8]. Furthermore, binary logistic regression research with the Bayesian method can be used in various fields, one of which is in the social and economic fields. Along with the development of the era, both developed and developing countries have serious problems that must be faced and resolved by the government in the social and economic fields, namely poverty. Poverty occurs in both developed and developing countries, especially in countries with high and dense populations such as Indonesia.

Based on information from Central Statistics Agency, poverty is the powerlessness of people to meet their life needs, both primary and secondary needs. The number of poor people in Indonesia in March 2023 reached 25.90 million people. The largest number of poor people is in East Java with a record of 4.18 million poor people or around 10.35% of the total population. To reduce the poverty rate in Indonesia, it is necessary to know what factors influence the high and low poverty rates in Indonesia so that in the future effective policies can be determined to reduce the poverty rate in Indonesia. Research using binary logistic analysis was conducted by [9] using the response variable, namely the percentage of poor people. While the predictor variables are the human development index, Gini ratio, GRDP growth rate, and population growth rate [10]. The results of this study indicate that there are two predictor variables that influence the poverty rate in Indonesia in 2021, namely the HDI and Gini ratio variables. Based on the description, this study will use binary logistic regression with the Bayesian method to determine the model formed and the factors that influence the poverty rate in East Java in 2022. All Bayesian inference is conducted through the use of Bayes' theorem [11–17]. The simulation result from estimating logistic regression parameters should also be used for confirmation of the beneficial part of using Bayesian estimation, particularly for the case of percentage of poverty rates. Specifically, the objectives of this study are 1) to identify the binary logistic regression model with the Bayesian method, and 2) to identify factors that significantly affect the percentage of poverty rates in cities/regencies in East Java in 2022, and 3) to conduct simulations to confirm the goodness of the Bayesian binary logistic regression model.

Based on the description above, this study employs Bayesian binary logistic regression to model poverty status in districts/cities in East Java. The response variable is defined as a binary

outcome, where regions are classified into two categories: above-average and below-average poverty status based on the provincial mean. Therefore, this study aims to: (1) identify the Bayesian binary logistic regression model for poverty status; (2) determine factors that significantly affect poverty status in districts/cities in East Java in 2022, and; (3) conduct simulation studies to evaluate the performance of the Bayesian approach compared to the classical likelihood-based method under different sample sizes.

2. Methods

This section presents the methodological framework used in this study. The discussion begins with the formulation of the binary logistic regression model as the baseline framework, followed by the Bayesian estimation approach, inference procedure, and simulation design. Each component is described systematically to ensure clarity in model construction and interpretation.

2.1. Binary Logistic Regression

The binary logistic regression model is used to model the relationship between a binary response variable $Y_i \in \{0, 1\}$ and a set of predictor variables $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$, where $i = 1, 2, \ldots, n$. The model is defined through a linear predictor:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} = x_i^T \beta, \tag{1}$$

where $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ is the vector of regression parameters.

The conditional probability of success is modeled using the logit link function:

$$\pi_i = P(Y_i = 1 \mid x_i, \beta) = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)}. \tag{2}$$

Hence, the distribution of the response variable follows a Bernoulli distribution:

$$Y_i \mid x_i, \beta \sim \text{Bernoulli}(\pi_i). \tag{3}$$

The likelihood function is constructed based on the assumption of independent observations:

$$L(\beta \mid y, X) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1 - y_i}, \tag{4}$$

where $y = (y_1, y_2, \ldots, y_n)$ and $X = (x_1, x_2, \ldots, x_n)^T$. The corresponding log-likelihood function is given by:

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i \log(\pi_i) + (1 - y_i) \log(1 - \pi_i) \right]. \tag{5}$$

The score function (first derivative of the log-likelihood) is expressed as:

$$\frac{\partial \ell(\beta)}{\partial \beta} = \sum_{i=1}^{n} x_i (y_i - \pi_i). \tag{6}$$

Because the score equations are nonlinear in β, numerical optimization methods such as Newton-Raphson are required to obtain the maximum likelihood estimator.

2.2. Bayesian Method

Bayesian inference treats the parameter vector β as a random variable. The inference is based on Bayes' theorem, which combines the likelihood function and the prior distribution to obtain the posterior distribution.

Let $p(\beta)$ denote the prior distribution and $p(y \mid \beta, X)$ denote the likelihood function. The posterior distribution is given by:

$$p(\beta \mid y, X) = \frac{p(y \mid \beta, X)\, p(\beta)}{p(y \mid X)}. \tag{7}$$

where the marginal likelihood is:

$$p(y \mid X) = \int p(y \mid \beta, X)\, p(\beta)\, d\beta. \tag{8}$$

Since the marginal likelihood does not depend on β, the posterior distribution is commonly written in proportional form:

$$p(\beta \mid y, X) \propto p(y \mid \beta, X)\, p(\beta). \tag{9}$$

2.3. Prior Distribution, Likelihood, and Posterior Distribution

The prior distribution represents the uncertainty about the model parameters before observing the data. In this study, each regression parameter is assigned an independent normal prior distribution:

$$\beta_j \sim \mathcal{N}(\mu_j, \sigma_j^2), \quad j = 0, 1, \ldots, p. \tag{10}$$

Under the assumption of prior independence, the joint prior distribution is expressed as:

$$p(\beta) = \prod_{j=0}^{p} \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{(\beta_j - \mu_j)^2}{2\sigma_j^2}\right). \tag{11}$$

The hyperparameters μ_j and σ_j^2 are specified using an empirical Bayes approach based on maximum likelihood estimates (MLE). This strategy is adopted to incorporate data-driven prior information and to improve numerical stability of posterior estimation, particularly in finite-sample settings. However, since the prior is constructed using the same dataset as the likelihood, this introduces a potential dependence between prior specification and observed data. To mitigate its impact, the prior is interpreted as a weakly informative empirical prior rather than a fully independent prior.

The likelihood function for the binary logistic regression model is defined in observation-wise form as:

$$p(y \mid \beta, X) = \prod_{i=1}^{n} \left(\frac{\exp(x_i^T \beta)}{1 + \exp(x_i^T \beta)}\right)^{y_i} \left(\frac{1}{1 + \exp(x_i^T \beta)}\right)^{1-y_i}. \tag{12}$$

The posterior distribution is obtained by combining the likelihood and prior via Bayes' theorem:

$$p(\beta \mid y, X) \propto p(y \mid \beta, X)\, p(\beta). \tag{13}$$

Due to the non-conjugacy between the Bernoulli likelihood and the normal prior, the posterior distribution does not have a closed-form expression. Therefore, exact analytical inference is not feasible, and posterior approximation is required.

In this study, posterior inference is conducted using Markov Chain Monte Carlo (MCMC) methods. Specifically, a Metropolis-Hastings step is embedded within a Gibbs sampling framework to handle non-standard full conditional distributions. The full conditional distribution for each parameter β_j is given by:

$$p(\beta_j \mid \beta_{-j}, y, X) \propto p(y \mid \beta, X)\, p(\beta_j). \tag{14}$$

where β_{-j} denotes the vector of all parameters excluding β_j. Because this conditional distribution does not belong to a standard parametric family, direct sampling is not possible. Therefore, a Metropolis-Hastings update is applied within each Gibbs iteration to generate samples from the target posterior distribution.

2.3.1. Markov Chain Monte Carlo (MCMC)

Markov Chain Monte Carlo (MCMC) is a simulation-based method used to generate samples from a target posterior distribution when analytical solutions are not available. In this study, MCMC is used to approximate the posterior distribution of the parameter vector β in the Bayesian binary logistic regression model. The goal is to construct a Markov chain whose stationary distribution corresponds to $p(\beta \mid y)$. In binary logistic regression, the posterior distribution is analytically intractable due to the non-conjugacy between the Bernoulli likelihood and the normal prior. Therefore, numerical simulation methods are required to obtain posterior samples.

This study employs a Metropolis-Hastings within Gibbs sampling framework (Metropolis-within-Gibbs), since the full conditional distributions of the regression parameters do not have standard closed-form expressions. The general MCMC procedure for posterior sampling is described as follows [16]:

Algorithm 1 MCMC Posterior Sampling Procedure

1: Initialize parameter values $\beta^{(0)}$
2: For iteration $t = 1, 2, \ldots, T$, update each parameter sequentially using Metropolis-within-Gibbs steps
3: For each parameter β_j, generate candidate value β_j^* from proposal distribution

$$\beta_j^* \sim N(\beta_j^{(t-1)}, \tau_j^2) \tag{15}$$

4: Compute acceptance probability

$$\alpha = \min\left(1, \frac{p(\beta_j^* \mid \beta_{-j}, y)}{p(\beta_j^{(t-1)} \mid \beta_{-j}, y)}\right) \tag{16}$$

5: Accept or reject β_j^* based on α
6: Repeat until all parameters are updated for iteration t
7: Repeat steps until T iterations are completed
8: Discard the first B iterations as burn-in period
9: Use remaining samples $(\beta^{(B+1)}, \ldots, \beta^{(T)})$ for posterior inference
10: Summarize posterior results (mean, median, standard deviation, Monte Carlo error)

It should be noted that the use of priors derived from maximum likelihood estimates (MLE) obtained from the same dataset may introduce potential double use of data. In this study, these priors are treated as empirical priors to improve numerical stability, particularly under small sample conditions. In this context, the posterior distribution $p(\beta \mid y)$ cannot be obtained in closed form due to the non-conjugacy between the Bernoulli likelihood and the normal prior distribution. Therefore, MCMC sampling is required to approximate the posterior distribution.

2.3.2. Metropolis-within-Gibbs Sampling Algorithm

In this study, the parameter estimation of the Bayesian binary logistic regression model is conducted using a Metropolis-within-Gibbs sampling algorithm. The full conditional distribution of each parameter β_j is given by:

$$p(\beta_j \mid \beta_{-j}, y) \propto p(y \mid \beta)\, p(\beta_j). \tag{17}$$

At iteration t, a candidate value is generated from:

$$\beta_j^* \sim \mathcal{N}(\beta_j^{(t-1)}, \tau_j^2). \tag{18}$$

The acceptance probability is defined as:

$$\alpha = \min\left(1, \frac{p(\beta_j^* \mid \beta_{-j}, y)}{p(\beta_j^{(t-1)} \mid \beta_{-j}, y)}\right). \tag{19}$$

The updating rule is:

$$\beta_j^{(t)} = \begin{cases} \beta_j^*, & \text{if accepted with probability } \alpha, \\ \beta_j^{(t-1)}, & \text{otherwise.} \end{cases} \tag{20}$$

This procedure is repeated sequentially for all parameters within each iteration until convergence is achieved.

Convergence Algorithm and Parameter Significance Test

The convergence checking aims to assess whether the generated samples adequately represent the posterior distribution. This can be evaluated using trace plots, MC error, autocorrelation plots, and kernel density plots [16][18].

Parameter significance in the Bayesian framework is determined using credible intervals. For a posterior density $f(\beta|x)$, the $100(1-\alpha)\%$ credible interval $[C, D]$ satisfies:

$$P(\beta \in [C, D] \mid x) = \int_C^D f(\beta|x)\, d\beta = 1 - \alpha. \tag{21}$$

with endpoints determined by:

$$\int_{-\infty}^{C} f(\beta|x)\, d\beta = \alpha/2, \quad \int_D^{\infty} f(\beta|x)\, d\beta = \alpha/2. \tag{22}$$

2.4. Goodness of fit test and Model Classification Accuracy

The model suitability test aims to determine the suitability of the model in logistic regression. The test used is the residual deviance test [2]. The hypotheses used: $H_0 : y_k = \hat{y}_k$ (the model is appropriate); and $H_1 : y_k \neq \hat{y}_k$ (the model is not appropriate). The residual deviance test statistic can be written in the equation $D = -2\ln(f(y|\hat{\beta})) \sim \chi^2_{(n-p-1)}$. The model suitability test is carried out by comparing the residual deviance test statistic with the critical point of chi square.

The decision criterion in this test is to reject H_0 if $D > X^2_{\alpha,(n-p-1)}$, then the models obtained are not appropriate. The measure of the suitability of the logistic regression model has a high probability of classification accuracy or a low probability of misclassification [2]. To determine the suitability of the model, it is necessary to measure the goodness of the model. In binary logistic regression classification there are only two classes, namely 'Yes' or 'No'. Table 1 shows the predicted classification and actual classification, where: TP: actual class and predicted class show 'Yes'; FP: actual class shows 'No' but predicted class shows 'Yes'; FN: actual class shows 'Yes' but predicted class shows 'No'; TN: actual class and predicted class show 'No'. Determining accuracy of the classification can be assessed from the sensitivity $\frac{TP}{(TP+FN)}$, specificity $\frac{TN}{(FP+TN)}$, and accuracy $\frac{(TP+TN)}{(TP+FP+TN+FN)}$ [19].

Table 1: Cross Tabulation of Predicted Class and Actual Class

		Actual class	
		Yes	**No**
Prediction class	Yes	True Positive (TP)	False Positive (FP)
	No	False Negative (FN)	True Negative (TN)

2.5. Data and Research Variable

This study uses secondary data obtained from the Central Statistics Agency (BPS). The data used consist of the percentage of poverty rates in East Java and the factors that influence the percentage of poverty rates. The average percentage of poverty rates in East Java in 2022 was 10.38% [20].

The poverty status of cities/regencies is categorized into two groups, namely below-average and above-average poverty status based on the provincial average poverty rate in East Java. The grouping of poverty percentage levels is presented in Table 2.

Table 2: Grouping of Poverty Status

Information	**Condition**	**Code**
Below average poverty status	$\leq 10.38\%$	0
Above average poverty status	$> 10.38\%$	1

The structure of the data used in this study is presented in Table 3. The data structure to

Table 3: Data Structure

i	Y_i	X_{1i}	X_{2i}	$\cdots$	X_{pi}
1	Y_1	X_{11}	X_{21}	$\cdots$	X_{p1}
2	Y_2	X_{12}	X_{22}	$\cdots$	X_{p2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	Y_n	X_{1n}	X_{2n}	$\cdots$	X_{pn}

be used in this study uses binary logistic regression analysis as shown in Table 3, where: n is the number of responses; p is the number of predictor variables; $i = 1, 2, \ldots, n$; $j = 1, 2, \ldots, p$; Y_i is the value of the i-th response of the response variable; X_{ji} is the value of the j-th predictor variable of the i-th response. This study identifies factors that influence the poverty rate of cities/regencies in East Java with binary logistic regression using the Bayesian method with one response variable and four predictor variables presented in Table 4.

Table 4: Response and Predictor Variables

Variable	**Note**	**Category**	**Scale**
Y	Poverty status	0: Below average; 1: Above average	Nominal
X_1	Human Development Index (HDI)	-	Ratio
X_2	Life expectancy	0: ≤ 71.74; 1: > 71.74	Nominal
X_3	Open unemployment	0: $\leq 5.49\%$; 1: $> 5.49\%$	Nominal
X_4	Gini ratio	-	Ratio

Interpretation of logistic regression is based on the odds ratio (OR). The OR value (ψ) is obtained from the probability of an event occurring compared to the probability of the event not occurring [21]. This is expressed through the odds ratio, which represents a comparison of the probability of an event between two groups. A high odds ratio value indicates that the probability for a particular category is much greater than that of the reference group.

2.6. Data Analysis and Simulation

Data processing is conducted using RStudio. The simulation study is performed with 1000 replications for each sample size to evaluate the performance of the estimators under repeated

sampling. The performance of the estimators is assessed using bias and mean squared error (MSE), defined as:

$$\text{Bias} = \mathbb{E}(\hat{\beta} - \beta), \quad \text{MSE} = \mathbb{E}[(\hat{\beta} - \beta)^2]$$

In this study, simulation data are generated based on a predefined logistic regression model using parameter values obtained from empirical estimates. Sample sizes considered are $n = 10$, $n = 38$, and $n = 100$. It should be noted that the current simulation design assumes fixed true parameter values and a correctly specified model structure. Therefore, the simulation primarily evaluates estimator performance under an ideal model setting. This may lead to optimistic performance results, especially for the Bayesian estimator.

To address this limitation, future studies are recommended to consider additional scenarios such as model misspecification, different signal-to-noise ratios, and alternative prior specifications. These extensions are expected to provide a more comprehensive assessment of estimator robustness under more realistic conditions.

Algorithm 2 Bayesian Binary Logistic Regression Procedure

1: Perform descriptive statistical analysis
2: Check multicollinearity using VIF
3: Specify prior, likelihood, and posterior distributions
4: Estimate parameters using MCMC (Metropolis-Hastings within Gibbs)
5: Assess convergence using trace plots, MC error, autocorrelation, and density plots
6: Evaluate parameter significance using credible intervals
7: Construct the Bayesian logistic regression model
8: Assess model fit using residual deviance
9: Evaluate classification performance (accuracy, sensitivity, specificity)
10: Interpret model using odds ratios

3. Results and Discussion

This section presents the empirical results obtained from the Bayesian binary logistic regression model applied to poverty status in districts/cities in East Java. The discussion begins with descriptive statistics, followed by model estimation results, parameter inference, model evaluation, and simulation outcomes.

3.1. Decriptive Statistics

Descriptive statistics are used to determine the general description of the data to be analyzed. This study uses data from East Java Province consisting of 38 cities/regencies. It is known that cities/regencies in East Java are categorized into two categories, namely poverty rates above the average of 45% and poverty rates below the average of 55%. Based on these data, it shows that cities/regencies in East Java have more cities/regencies that are in the category of poverty rates below the provincial average. Predictor variables include the human development index, life expectancy, open unemployment rate, and Gini ratio. It is known that most cities/regencies in East Java are included in the category of life expectancy of more than 71.74 years, which is 71%. While cities/regencies in the category of life expectancy of less than 71.74 years are 29%. It can be seen that the percentage of the category of open unemployment rate of more than 5.49% is lower than the category of open unemployment rate of less than 5.49%, which is 39%. While the category of open unemployment rate is less than 5.49% which is 61%. The number of cities/regencies that have an open unemployment rate category of more than 5.49% is 15 cities/regencies, while the category of open unemployment rate is less than 5.49% is 23 cities/regencies.

Table 5: Decriptive Statistics

Variable	Minimum	Maximum	Average	Std. Error
X_1 (HDI)	63.39	82.74	72.969	5.077
X_4 (Gini Ratio)	0.26	0.42	0.334	0.038

Based on the results of descriptive statistics in Table 5, it can be seen that the average Human Development Index (X_1), in each city/district in East Java has an average of 72.969. The highest Human Development Index is 82.74 in Surabaya City. The lowest Human Development Index is 63.39 in Sampang Regency. The average difference in the Human Development Index of each city/district in East Java to the average is 5.077. The Gini ratio (X_4) in each city/district in East Java has an average of 0.334. The highest Gini ratio is 0.421 in Malang City. The lowest Gini ratio is 0.266 in Sumenep Regency. The average difference in the Gini ratio of each city/district in East Java to the average is 0.038. Multicollinearity test is conducted to evaluate whether there is a relationship between predictor variables. The way to find out the presence of multicollinearity is by looking at the VIF (Variance Inflation Factor) value. The assumption of non-multicollinearity will not be met if the VIF value is > 10. The VIF values are presented in Table 6. Based on Table 6, it is found that the VIF values for all variables are no more than 10, so the assumption of non-multicollinearity is met.

Table 6: VIF for each Predictor Variable

Variable	VIF
Human Development Index (HDI)	2.7982
Life expectancy	2.6340
Open unemployment	1.2413
Gini ratio	1.0904

3.2. Maximum Likelihood Estimation Approach

Parameter estimation for binary logistic regression is generally solved using the maximum likelihood estimation (MLE) approach. MLE is a parameter estimation method by maximizing the likelihood function. In the Bayesian approach, the MLE parameter estimation value can be used as a prior distribution when the initial information is limited and the model formed is complex. In this study, the MLE parameter estimation results contain relevant and significant initial information. So the MLE stimation results are used as the formation of priors for the average value and standard deviation value. The results of the MLE approach parameter estimation are presented in Table 7.

Table 7: Results of MLE Approach Parameter Estimation

Variable	Parameter	Coefficient	Standard Error
Intercept	β_0	52.2468	19.9746
HDI	β_1	-0.5850	0.2486
Life expectancy	β_2	-3.0466	1.5515
Open unemployment	β_3	-0.4509	1.0288
Gini ratio	β_4	-27.9093	18.2605

3.3. Bayesian Logistic Regression Parameter Estimation

The Bayesian method aims to obtain a posterior distribution obtained from the multiplication of the prior distribution and the likelihood function. The Bayesian method in determining the prior distribution does not yet know the parameter distribution to be used. The prior distribution in this study is based on the types of informative priors and non-conjugate priors. The use of informative priors is based on data information obtained from the results of the MLE approach parameter estimation, while the use of non-conjugate priors is based on the subjectivity of the

researcher. One of the criticisms of Bayesian, especially priors, is that Bayesian is subjective [22]. Researchers can provide initial confidence (prior) in parameters because of the assumption that the parameters are random variables. In this study, the non-conjugate prior used is the normal distribution [4]. Based on research [23], the prior distribution for the parameters of the Bayesian binary logistic regression model can follow the normal distribution. The probability density function of the normal distribution of each binary logistic regression parameter is written as follows

$$f(\beta_j) = \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(\beta_j - \mu_j)^2\right) \tag{23}$$

Table 8: Prior Distribution Parameter Values

Variable	**Parameter**	μ	σ
Intercept	β_0	52.2468	19.9746
HDI	β_1	-0.5850	0.2486
Life expectancy	β_2	-3.0466	1.5515
Open unemployment	β_3	-0.4509	1.0288
Gini ratio	β_4	-27.9093	18.2605

The value of the normal prior parameter is obtained from the estimation results with the MLE approach. Table 8 presents the value of the normal prior distribution parameters that will be used. The parameter estimation process for binary logistic regression with the Bayesian method begins with the formation of a likelihood function. The Likelihood function is obtained from a random sample and will be used to obtain the posterior distribution. The likelihood function for the binary logistic regression model is written as follows

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{\exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}\right)^{y_i} \left(\frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{p} \beta_j x_{ij})}\right)^{1-y_i} \tag{24}$$

The posterior distribution in Bayesian is obtained from the multiplication of the prior distribution and the likelihood function. If the likelihood function and the normal prior distribution are combined then the following equation will be formed. This equation is used in the formation of a full conditional distribution to generate samples. Sample generation is carried out using the Gibbs sampling algorithm which aims to update parameters.

$$f(\beta_j \mid x) \propto f(x \mid \beta_j)\, f(\beta_j)$$

where,

$$\begin{aligned} f(\beta_j|x) = \prod_{i=1}^{n} &\left(\frac{1}{1 + \exp\left(\sum_{i=1}^{n}\sum_{j=1}^{p} \beta_j x_{ji}\right)}\right) \left(\exp\left(\sum_{i=1}^{n}\sum_{j=1}^{p} \beta_j x_{ji}\right)\right)^{y_i} \\ &\times \frac{1}{\sqrt{2\pi\sigma_j^2}} \exp\left(-\frac{1}{2\sigma_j^2}(\beta_j - \mu_j)^2\right) \end{aligned} \tag{25}$$

Furthermore, convergence checks can be seen based on the results of the MCMC process using trace plots, autocorrelation plots, MC errors, and kernel density plots. The algorithm convergence check using trace plots forms an up and down pattern, so burn-in is needed. This aims to eliminate the influence of non-convergent initial values. In the MCMC simulation of the Gibbs sampling algorithm, iterations are carried out until the sample reaches convergence. In the trace plot with 266,000 iterations by setting 60,000 iterations as burn-in, it does not show a particular pattern, so that the simulated random sample has shown a convergent value. Convergence checks for random samples of the Gibbs sampling algorithm MCMC can also be done by examining the autocorrelation plot. If the autocorrelation value at each lag approaches zero, then the random sample is said to have converged. Based on the ACF plot at thin 10, it

shows that only lag 0 is close to 1 and the other lags are close to 0. This shows that the resulting sample has converged and does not have high autocorrelation. Kernel density plot is a picture of the posterior distribution, if the resulting distribution is in accordance with the prior form, then the random sample is said to be convergent.

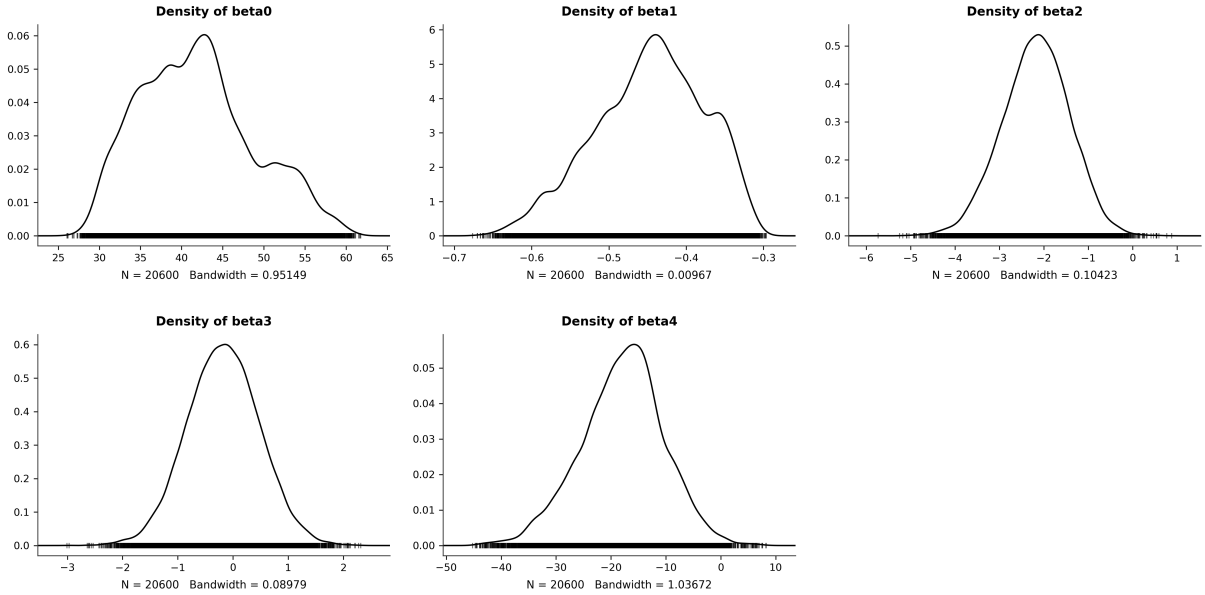

Fig. 1: Kernel Density Plot

Fig. 1 shows that the posterior distribution has resembled a normal distribution. The distribution shape of the kernel density plot is in accordance with the prior distribution. So, it can be said that the random sample generated from the MCMC simulation has converged. The next convergence check is the MC error. The convergence check with MC error can be seen in Table 9. Based on the table, it can be seen that the MC error results for each parameter are less than 1% standard deviation. This means that the convergence of the random sample MCMC Gibbs sampling algorithm for each parameter is met.

Table 9: MC Error

Variable	Parameter	SD	1%SD	MC Error
Intercept	β_0	0.223	0.0022	0.0016
HDI	β_1	0.009	0.0001	0.0001
Life expectancy	β_2	0.619	0.0062	0.0043
Open unemployment	β_3	0.684	0.0068	0.0048
Gini ratio	β_4	0.233	0.0023	0.0016

3.4. Parameter Significance Test

In the Bayesian method, the parameter significance test is carried out by examining the credible interval, where significant variables can be determined at a significance level of 5%. The values from the 2.5% to 97.5% quantile provide a credible interval value for each particular variable. A parameter is considered significant if the credible interval value does not contain zero. The credible interval value for each variable is presented in Table 10. It can be seen that of the four variables used, there are three variables that are proven to be significant, namely the Human Development Index, Life Expectancy, and Gini Ratio variables. Meanwhile, the Open Unemployment Rate variable is not significant based on the credible interval value.

Table 10: Credible Interval

Variable	Parameter	2.5%	97.5%	Note
Intercept	β_0	51.811	52.687	Sig.
HDI	β_1	-0.604	-0.567	Sig.
Life expectancy	β_2	-4.240	-1.834	Sig.
Open unemployment	β_3	-1.768	0.917	Not Sig.
Gini ratio	β_4	-28.364	-27.451	Sig.

3.5. Formation of Binary Logistic Regression Model with Bayesian Method

After testing the significance of the parameters, the next step is to form the Bayesian binary logistic regression model. The final model coefficients are obtained from the posterior mean of the MCMC simulation results using the Metropolis-within-Gibbs algorithm. The estimated classification results are presented in Table 11.

Model suitability is evaluated using the residual deviance test. The hypotheses are H_0: the model fits the data well, and H_1: the model does not fit the data well. The residual deviance value obtained is 30.5 with degrees of freedom equal to 33, following a chi-square distribution. At a significance level of 0.05, the critical value is $\chi^2_{0.05,33} = 47.399$. Since $D = 30.5 < 47.399$, the null hypothesis is not rejected, indicating that the model provides an adequate fit to the observed data.

The classification results based on the confusion matrix are shown in Table 11. Based on

Table 11: Confusion Matrix and Classification Performance of Bayesian Logistic Regression Model

Prediction	**Actual**		**Total**
	Above average (1)	**Below average** (0)	
Above average (1)	15	4	19
Below average (0)	2	17	19
Total	17	21	38

Table 11, the model achieves a sensitivity of 0.882, a specificity of 0.809, and an overall accuracy of 0.842. These results indicate that the model has a relatively good ability to classify observations in the dataset used in this study.

However, it should be noted that these evaluation metrics are calculated using the same dataset employed for model estimation (in-sample evaluation). Therefore, the reported accuracy may be optimistic and potentially overestimate the model's predictive performance on new or unseen data. This limitation is common in studies without independent test data or cross-validation procedures. Future research is recommended to apply out-of-sample validation techniques, such as cross-validation or data splitting, to obtain a more robust evaluation of predictive performance.

The interpretation of the model is based on odds ratios derived from posterior estimates. The results show that an increase in the Human Development Index (HDI) and life expectancy is associated with a decrease in the odds of a district/city being classified as having above-average poverty. The Gini Ratio also shows a significant relationship with poverty status, while open unemployment is not statistically significant in this model. These interpretations describe associations observed in the data and should not be interpreted as causal effects, as the analysis is based on observational cross-sectional data.

3.6. Simulation Result

The simulation design includes data generation based on a logistic model with predefined parameter values. For each scenario, repeated simulations are conducted to evaluate estimation performance. The comparison between methods is assessed using metrics such as bias and mean squared error (MSE).

In this section, a simulation is carried out with initial values obtained from the results of likelihood and Bayesian estimates on poverty data in East Java province in 2023. The simulation

results are presented in Table 12. It can also be seen in Figure **??** that the estimation results with Bayesian and likelihood methods vary across different sample sizes.

Table 12: Simulation result with Likelihood and Bayesian

Parameter	Initial value	Likelihood			Bayesian		
		10	**38**	**100**	**10**	**38**	**100**
β_0	50	12.004	43.895	40.097	40.435	55.321	52.543
β_1	-0.5	-0.738	-0.634	-0.658	-0.683	-0.504	-0.578
β_2	-3.0	-0.283	-2.754	-5.037	-2.275	-2.875	-3.078
β_3	-0.5	-1.298	-0.478	-0.682	-0.421	-0.498	-0.512
β_4	-20	-10.670	-17.004	-18.894	-12.315	-18.874	-18.756

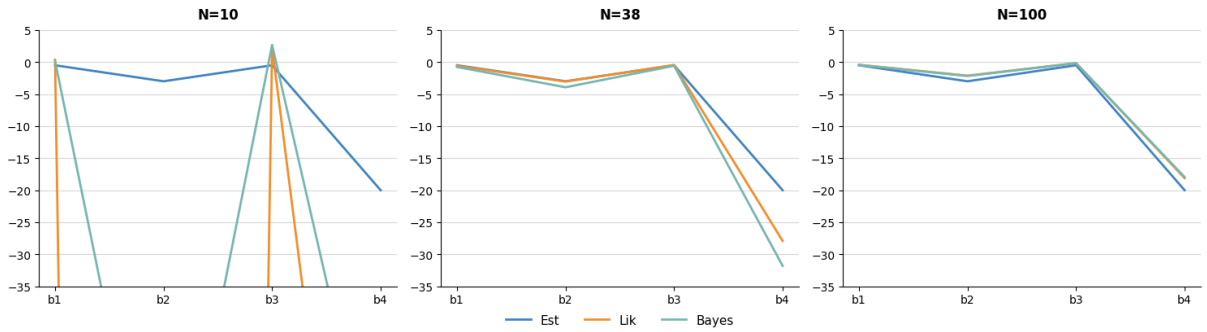

Figure 2: Simulation Results Comparing Bayesian and Likelihood Estimation

Fig. 2: Simulation Results Comparing Bayesian and Likelihood Estimation under Different Sample Sizes. The x-axis represents sample size (n = 10, 38, 100), while the y-axis represents parameter estimates. The horizontal line indicates the true parameter values used in the data-generating process.

It should be noted that the horizontal lines in Fig. 2 represent the true parameter values used in the data-generating process, while the x-axis corresponds to different sample sizes (n = 10, 38, and 100), and the y-axis represents the estimated parameter values under both likelihood and Bayesian approaches.

As shown in Table 12, the Bayesian estimates are generally closer to the initial parameter values compared to the likelihood estimates in several cases, particularly for smaller sample sizes such as $n = 10$. However, this pattern is specific to the simulation setting used in this study and should not be interpreted as a general superiority of one method over the other.

It is also important to note that the simulation results are influenced by the assumed data-generating process, including the chosen parameter values and model structure. In addition, the Bayesian performance is affected by the use of empirical priors derived from the same dataset, which may contribute to the observed behavior in small sample settings.

Overall, 2 shows that as the sample size increases, both likelihood and Bayesian estimates become closer to the initial parameter values, and the differences between the two methods tend to decrease.

Finally, the conclusions drawn from this simulation are conditional on the specific parameter setting and design of the study. Future research may consider alternative parameter configurations, prior sensitivity analysis, and different data-generating mechanisms to further evaluate the robustness of both estimation methods.

The simulation results suggest that Bayesian estimation provides more stable parameter recovery under small sample conditions, which is particularly relevant for district-level poverty modeling where data availability is often limited. This has important implications for policy evaluation, as more stable estimates allow for more reliable identification of high-poverty regions and improve the robustness of targeted intervention strategies.

4. Conclusion

Based on the results of this study, it can be concluded that Bayesian binary logistic regression provides a robust framework for modeling poverty status, defined as whether districts/cities fall above or below the provincial average. The results indicate that the Human Development Index (HDI), Life Expectancy, and Gini Ratio significantly influence poverty status, while the Open Unemployment Rate is not statistically significant.

Furthermore, although parameter estimates obtained from classical likelihood and Bayesian approaches are relatively similar, the simulation study demonstrates that the Bayesian method performs better in small sample conditions, as indicated by lower bias and mean squared error (MSE), reflecting more stable and reliable estimates.

This study has several limitations, including the relatively small sample size, the use of empirical priors derived from the same dataset, and the dichotomization of the response variable, which may lead to information loss. Future research is recommended to explore alternative prior specifications, incorporate spatial dependence, and extend the modeling framework to account for temporal dynamics. These extensions are expected to provide more comprehensive insights for poverty-related policy formulation.

CRediT Authorship Contribution Statement

Achmad Efendi: Conceptualization, Methodology, Formal Analysis, Investigation, Writing–Original Draft. **Restilia A. Sari:** Data Curation, Formal Analysis, Validation, Writing–Review & Editing. **Samingun Handoyo:** Software, Visualization, Validation. **Nur S. Rahmi:** Supervision, Project Administration, Writing–Review & Editing. **Friansyah Gani:** Methodology, Supervision, Writing–Review & Editing, Funding Acquisition.

Declaration of Generative AI and AI-assisted technologies

No generative AI or AI-assisted technologies were used during the preparation of this manuscript.

Declaration of Competing Interest

The authors declare no competing interests

Funding and Acknowledgments

We are grateful for the funding from Universitas Brawijaya, Hibah Penelitian Internal FMIPA 2024, Grant No. 2612.27/UN01.F09/PN/2024.

Data and Code Availability

The research data supporting the findings of this study are secondary data obtained from the Central Statistics Agency (Badan Pusat Statistik - BPS) of East Java Province. The dataset, which includes poverty rates and their influencing factors in East Java for the year 2022, is publicly accessible through the official BPS website at https://jatim.bps.go.id [20]. Specific calculation codes used in this study are available from the corresponding author upon reasonable request.

References

[1] Wayne Wobcke and Siti Mariyah. “Machine learning and data augmentation in the proxy means test for poverty targeting”. In: *Statistical Journal of the IAOS* 39.4 (2023), pp. 961–977. DOI: 10.3233/SJI-230033.

[2] David W Hosmer and Stanley Lemeshow. *Applied Logistic Regression*. 2nd ed. New York: John Wiley & Sons, Inc., 2000. DOI: 10.1002/0471722146.

[3] Syarifah Nur Aini et al. "Deteksi Nefropati Diabetik Pada Pasien Diabetes Melitus Menggunakan Regresi Logistik". In: *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer* 9.2 (2025), pp. 1–10. https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14478.

[4] William M Bolstad. *Introduction to Bayesian Statistics*. New Jersey, USA: John Wiley & Sons, 2004. DOI: 10.1002/0471654213.

[5] A Safitri, Sudarmin, and M Nusrang. "Model Regresi Logistik Biner pada Tingkat Pengangguran Terbuka di Provinsi Sulawesi Barat Tahun 2017". In: *VARIANSI: Journal of Statistics and Its Application on Teaching and Research* 1.2 (2019), pp. 1–6. DOI: 10.35580/variansiunm10620.

[6] David B Dunson. "Commentary: Practical advantages of Bayesian analysis of epidemiologic data". In: *American Journal of Epidemiology* 153.12 (2001), pp. 1222–1226. DOI: 10.1093/aje/153.12.1222.

[7] Desi Kurniawati and Hery T Sutanto. "Faktor-Faktor Yang Mempengaruhi Anemia Remaja Putri Dengan Menggunakan Bayesian Regresi Logistik Dan Algoritma Metropolis-Hasting". In: *Jurnal Ilmiah Matematika* 7.1 (2019), pp. 1–6. DOI: 10.26740/mathunesa.v7n1.p1-6.

[8] P A Lukman, S Abdullah, and A Rachman. "Bayesian logistic regression and its application for hypothyroid prediction in post-radiation nasopharyngeal cancer patients". In: *Journal of Physics: Conference Series* 1725.1 (2021), p. 012010. DOI: 10.1088/1742-6596/1725/1/012010.

[9] A I Nurrizqi, Erfiani, Indahwati, A Fitrianto, and R Amelia. "Pemodelan Regresi Logistik Berbasis Backward Elimination Untuk Mengetahui Faktor-Faktor yang Memengaruhi Tingkat Kemiskinan di Indonesia Tahun 2021". In: *Jurnal Statistika dan Aplikasinya* 6.2 (2022), pp. 160–170. DOI: 10.21009/JSA.06206.

[10] Friansyah Gani, Henny Pramoedyo, and Achmad Efendi. "Spatial Variation of HDI in East Java: A Tricube-Based Geographically Weighted Regression–Flower Pollination Algorithm Modeling Approach". In: *CAUCHY – Jurnal Matematika Murni dan Aplikasi* 11.1 (2026), pp. 239–254. DOI: 10.18860/ca.v11i1.XXXXX.

[11] S James Press. *Bayesian Statistics: Principles, Models and Applications*. New York: John Wiley & Sons, 1989. DOI: 10.1002/9780470316696.

[12] Jose M Bernardo and Adrian F M Smith. *Bayesian Theory*. Wiley, 2000. DOI: 10.1002/9780470316870.

[13] Michael D Lee. "A Bayesian analysis of retention functions". In: *Journal of Mathematical Psychology* 48.5 (2004), pp. 310–321. DOI: 10.1016/j.jmp.2004.06.002.

[14] Edward Greenberg. *Introduction to Bayesian Econometrics*. Cambridge University Press, 2012. DOI: 10.1017/CBO9781139150736.

[15] Zening Zhao, Wei Duan, Guojun Cai, Meng Wu, and Songyu Liu. "CPT-based fully probabilistic seismic liquefaction potential assessment to reduce uncertainty: Integrating XGBoost algorithm with Bayesian theorem". In: *Computers and Geotechnics* 149 (2022), p. 104868. DOI: 10.1016/j.compgeo.2022.104868.

[16] Ioannis Ntzoufras. *Bayesian Modelling Using WinBUGS*. New Jersey: John Wiley & Sons, Inc., 2009. DOI: 10.1002/9780470434550.

[17] Peyman Amirafshari and Athanasios Kolios. "Estimation of weld defects size distributions, rates and probability of detections in fabrication yards using a Bayesian theorem approach". In: *International Journal of Fatigue* 159 (2022), p. 106763. DOI: 10.1016/j.ijfatigue.2022.106763.

[18] Muhammad Yarahmadi and Amin Salehi. "A comparative Bayesian PINN–MCMC analysis of Barrow–Tsallis holographic dark energy with neutrinos: Toward resolving the Hubble tension". In: *Journal of High Energy Astrophysics* 50 (2026), p. 100498. DOI: 10.1016/j.jheap.2025.100498.

[19] Cagatay Catal. "Performance Evaluation Metrics for Software Fault Prediction Studies". In: *Acta Polytechnica Hungarica* 9.4 (2012), pp. 211–225. DOI: 10.12700/APH.9.4.2012.4.13.

[20] Badan Pusat Statistik Jawa Timur. "Badan Pusat Statistik Jawa Timur". In: *Statistik Sektoral Jawa Timur* (2023). Diakses pada 10 Desember 2023. https://www.jatim.bps.go.id.

[21] Alan Agresti. *Categorical Data Analysis*. New York: John Wiley & Sons, Inc., 2002. DOI: 10.1002/0471249688.

[22] Kevin Ross. *An Introduction to Bayesian Reasoning and Methods*. bookdown.org, 2022. https://bookdown.org/kevin_davisross/bayesian-reasoning-and-methods/.

[23] D A M Rohmah, A B Astuti, and A Efendi. "A Statistical Analytics of Migration Using Binary Bayesian Logistic Regression". In: *BAREKENG: Journal of Mathematics and Its Applications* 17.3 (2023), pp. 1725–1738. DOI: 10.30598/barekengvol17iss3pp1725-1738.