

An Analysis of Overdispersed Tuberculosis Case Data in Lampung Province Using Negative Binomial

Ma'rufah Hayati^{1*}, Reni Permata Sari², Rahmawati Azizah³, Mahfudz Hudori⁴, Kurnia Sari⁵

¹ Actuarial Science Study Program, Institute Teknologi Sumatera, Indonesia

^{2,4,5} Statistics Study Program, Nahdlatul Ulama University of Lampung, Indonesia

³ Al-Madani Institute of Economic, Indonesia

*Corresponding author: marufah.mt@at.itera.ac.id

*Submission date: 16 December 2025, Revision: 07 January 2026, Accepted: 07 April 2026

ABSTRACT

Tuberculosis (TB) continues to pose a serious challenge to public health, especially in low- and middle-income regions. Empirical analyses of TB incidence commonly in epidemiological studies rely on count data, which frequently deviate from the equidispersion assumption and thus require careful consideration in model specification. This study examines the distribution of TB cases in Lampung Province and investigates determinants associated with its occurrence through a Generalized Linear Model (GLM) framework. The analysis initially applies Poisson regression, followed by an assessment of the equidispersion assumption. Evidence of overdispersion is observed, indicating that the Poisson model is inadequate for the data structure. Consequently, negative binomial regression is employed to better accommodate the observed variability. Model comparison using the Akaike Information Criterion (AIC) demonstrates that the negative binomial model provides a superior fit relative to the Poisson specification. The results further reveal that the population living in poverty has a statistically significant positive association with TB incidence, whereas the availability of public hospitals and population density do not show significant effects. These findings highlight the importance of socioeconomic conditions in shaping TB transmission dynamics in Lampung Province. Overall, the study underscores the suitability of negative binomial regression for overdispersed TB count data and offers empirical evidence that may support policy formulation by emphasizing the integration of socioeconomic interventions alongside improvements in healthcare services.

KEYWORDS

Tuberculosis, Overdispersion, Generalized Linear Model (GLM), Negative Binomial Regression, Lampung Province.

1. INTRODUCTION

Tuberculosis (TB) is a communicable disease resulting from infection with *Mycobacterium tuberculosis*, which predominantly involves the pulmonary system but can also affect extrapulmonary organs. The spread of TB occurs via airborne particles expelled into the air, particularly by individuals with smear-positive pulmonary TB, who represent the most infectious group. It is estimated that a single smear-positive TB patient can transmit the infection to approximately 10–15 individuals annually [1]. This high transmission potential contributes to the persistence of TB as a long-term infectious condition of global concern. At the global level, TB remains among the infectious diseases associated with the highest mortality burden.

On a global scale, tuberculosis (TB) is recognized as one of the infectious diseases responsible for the highest number of deaths. Indonesia constitutes a significant contributor to this global disease burden [2]. Evidence from the national Tuberculosis Prevalence Survey indicates that bacteriologically confirmed pulmonary TB remains highly prevalent among the adult population in Indonesia [3]. In addition, TB has been reported as the foremost cause of mortality from infectious diseases and is positioned

among the leading causes of death nationally, following cardiovascular disorders and acute respiratory infections [4]. This epidemiological profile underscores the persistent complexity of TB prevention and control within the Indonesian health system.

Several social and behavioral factors are known to influence the incidence of TB. Prevalence surveys have shown that TB occurs more frequently in males than in females, which is suspected to be associated with treatment adherence and smoking behavior that is more common among men [5]. In addition, the role of the family is crucial in TB treatment and prevention, particularly in ensuring medication adherence and preventing transmission within households [6].

From a regional perspective, the epidemiological profile of tuberculosis in Indonesia indicates that Lampung Province remains among the areas with a substantial contribution to the national TB caseload. Reports on TB program performance indicate that TB case detection in Lampung Province has fluctuated and has not yet reached the nationally established targets. Several districts/cities report relatively high case detection rates, while others remain at low levels [7]. This situation reflects spatial variation in TB cases that requires further analysis to support more effective public health planning and policy formulation [8].

In epidemiological and public health studies, mathematical and statistical modeling has become an essential tool for understanding the dynamics of infectious diseases and supporting evidence-based decision making. Analyses involving disease incidence commonly rely on count-based outcomes, for which Poisson regression is frequently adopted as part of the Generalized Linear Model (GLM) framework. This modeling approach is built on the equidispersion premise, whereby the expected value and variability of the response are assumed to be identical. In practice, tuberculosis case data rarely conform to this requirement, as the observed variability often surpasses the mean level. Such a pattern reflects the presence of overdispersion, a phenomenon that has been consistently reported in TB-related studies [9, 10, 11, 12, 13].

When applied to data characterized by excess variability, the Poisson specification may produce distorted parameter estimates and compromise the reliability of statistical conclusions. Under such circumstances, the use of a more adaptable modeling framework becomes necessary, with negative binomial regression serving as a suitable alternative. Unlike the Poisson model, this approach allows the variance to exceed the mean, aligning more closely with the empirical properties of overdispersed count data. From a distributional perspective, the negative binomial model can be viewed as a Poisson–gamma mixture, a formulation that enables it to capture heterogeneity and additional variation present in the observed counts [14, 15].

In light of these methodological considerations, the present study focuses on modeling tuberculosis case counts in Lampung Province through the application of negative binomial regression and evaluating its performance relative to Poisson regression within the Generalized Linear Model framework. By adopting this comparative approach, the analysis seeks to better capture the underlying distributional features of TB incidence data and to generate evidence that may inform epidemiological assessment and the formulation of regionally responsive TB control strategies.

2. LITERATURE REVIEW

2.1 Generalized Linear Model (GLM)

The Generalized Linear Model (GLM) represents a generalization of classical linear regression, specifically developed to handle response variables whose probability distributions extend beyond the normal assumption. In contrast to ordinary linear regression, which directly links the mean of the response variable to the predictors through a linear form, the GLM framework introduces flexibility by incorporating a link function and specifying a response distribution drawn from the exponential family [16, 17].

Conceptually, a GLM considers a collection of independent random variables $Y_1, Y_2, \dots, Y_n$ each assumed to follow a distribution belonging to the exponential family. The model is structured around three fundamental elements. The first is the random component, which characterizes the distribution of each response variable Y_i , through an associated canonical θ_i . The second element is the linear predictor, defined as a linear combination of explanatory variables and regression coefficients, commonly written as $x_i^T \beta$. Where β represents the design matrix containing observed covariate values. The third element is the link function, which establishes a systematic relationship between the expected value of the response variable and the linear predictor, thereby connecting the stochastic and systematic components of the model. This relationship can be expressed

mathematically as

$$\mu_i = E(Y_i) \quad (1)$$

$$g(\mu_i) = x_i^T \beta \quad (2)$$

In this formulation, $g(\mu_i)$ denotes a monotonic link function, where μ_i represent the expected value of response variable and $x_i^T \beta$ corresponds to the linear predictor. The term Generalized Linear Model (GLM) refers to a broad class of statistical models introduced and formalized by McCullagh and Nelder (1989). Under this framework, the response variable is assumed to arise from a distribution belonging to exponential family, with its mean linked to a linear combination of explanatory variables. Although the functional relationship between the mean response and the covariates may be nonlinear, McCullagh and Nelder (1983) highlighted that linearity is preserved with respect to the model parameters, since the covariates enter the model through the linear component $x^T \beta$ [18].

Beyond logistic regression, Poisson regression constitutes one of the most frequently applied GLM variants. This approach is designed for modeling discrete count outcomes involving one or more predictor variables, where the response takes nonnegative integer values rather than binary outcomes. In epidemiology and public health research, Poisson regression is widely utilized to examine the frequency of disease events and to assess associations between incidence counts and relevant risk factors.

2.2 Poisson Regression

Regression analysis provides a formal statistical framework for characterizing the association between a response variable and a set of explanatory variables. Within this framework, the influence of each predictor on the response is quantified through parameters known as regression coefficients, commonly denoted by β . For response variables expressed as discrete counts, Poisson regression constitutes an appropriate modeling choice. This model is widely used for event data that occur sporadically over a given period and are recorded as non-negative integers [19]. The application of Poisson regression assumes that the response variable follows a Poisson distribution, where the data represent non-negative count outcomes, the observations are independent, and the mean of the distribution is equal to its variance (equidispersion). These assumptions form the basis for modeling the relationship between the response variable and the explanatory variables.

Let μ_i represent the expected number of events observed during an exposure period t_i . Under the assumptions of independence across observations and a constant mean structure, μ_i can be modeled as a function of several predictor variables. It can be formulated as a function of multiple covariates. In the Generalized Linear Model (GLM) setting, this relationship is established through a link function that connects the expected response to a linear predictor, expressed as

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (3)$$

Poisson regression adopts the logarithmic function as its link, ensuring that the modeled mean of the response variable remains strictly non-negative, in accordance with the nature of count data. Consequently, the association between the response and the predictors can be written as

$$\ln E(Y_i | X_i) = \ln(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \quad (4)$$

which may equivalently be represented in exponential form as

$$\mu_i = \exp(x_i^T \beta) = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \quad (5)$$

2.3 Negative Binomial Regression

Negative binomial regression constitutes a nonlinear modeling approach developed from a Poisson–Gamma mixture, in which the Poisson component governs the event-generating process while the Gamma component captures unobserved heterogeneity. As part of the Generalized Linear Model (GLM) family, this method is particularly suited for examining relationships between a count response variable and one or more explanatory variables when the observed variance exceeds the mean, a common feature of empirical count data [20].

From a distributional standpoint, the negative binomial formulation arises by allowing the Poisson mean to vary according to a Gamma distribution, thereby introducing additional variability into the model. This mechanism relaxes the equidispersion

constraint inherent in Poisson regression and provides a natural way to address overdispersion in practice. As a result, negative binomial regression offers greater flexibility and is widely regarded as a more appropriate choice for highly variable count outcomes [21, 17].

Within the GLM framework, the systematic component of the negative binomial model links the expected value of the response variable to a linear predictor, which may be written as:

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \beta_0 + \sum_{j=1}^k \beta_j x_{ij}, i = 1, 2, \dots, n \quad (6)$$

In most applications, the logarithmic function is adopted as the link, such that, $g(\mu_i) = \ln g(\mu_i)$. This choice guarantees a positive mean structure and facilitates parameter estimation through maximum likelihood procedures [20]. Owing to these properties, negative binomial regression has become the standard alternative to Poisson regression for analyzing overdispersed count data, particularly in epidemiological and public health research settings [12, 15, 22].

3. METHODOLOGY

3.1 Data Sources

This study employs secondary data for the year 2021 obtained from the Lampung Provincial Health Office and publications of Statistics Indonesia (Badan Pusat Statistik/BPS). The unit of observation consists of 15 regencies/municipalities in Lampung Province. An overview of the variables incorporated in the analysis is provided in **Table 1**.

Table 1. Research Variables

No.	Variable	Variable Description	Unit/Measurement Scale
1	Y	Number of tuberculosis (TB) cases	Number of cases (persons)
2	X_1	Number of people living below the poverty line	Number of cases (persons)
3	X_2	Number of general hospitals	Number of units (hospitals)
4	X_3	Population density	Persons per km ²

3.2 Overview of the Analytical Process

The analytical process in this study was implemented in a structured and sequential manner, encompassing several key stages:

1. Data Acquisition

Information on the response variable (Y) and the set of explanatory variables (X) was obtained from authoritative secondary sources, specifically the Lampung Provincial Health Office and Statistics Indonesia (BPS).

2. Descriptive Analysis

An initial exploration of the data was conducted using descriptive statistical techniques. Summary measures, including the mean, variance, minimum, and maximum values, were used to characterize the distribution of each variable, complemented by graphical representations to aid interpretation.

3. Poisson Regression Specification

To examine the association between tuberculosis (TB) case counts and the explanatory variables, a Poisson regression model was first specified. The response variable was assumed to follow a Poisson distribution:

$$Y_i \sim \text{Poisson}(\mu_i)$$

with a logarithmic link function defined as:

$$\log(\mu_i) = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} \quad (7)$$

where:

Y_i : denotes the number of TB cases in the i -th regency or municipality

x_{1i} : number of people living below the poverty line

x_{2i} : number of general hospitals

x_{3i} : population density.

4. Assessment of Overdispersion

The adequacy of the Poisson model was evaluated by testing the equidispersion assumption. The null hypothesis (H_0) stated that the variance of Y_i equals its mean, whereas the alternative hypothesis (H_1) indicated the presence of overdispersion, defined by a variance exceeding the mean. This assessment relied on the ratio of the deviance or Pearson chi-square statistic to the corresponding degrees of freedom .

5. Negative Binomial Regression Modeling

When evidence of overdispersion was identified, the analysis proceeded with a negative binomial regression model as a more flexible alternative. In this formulation:

$$Y_i \sim \text{NB}(\mu_i, \alpha)$$

where α denotes the dispersion parameter. The systematic component retained the same logarithmic link function in equation (7)

6. Identification of Relevant Covariates.

The influence of each explanatory variable on TB incidence was evaluated through hypothesis testing of the estimated regression coefficients, allowing the identification of statistically meaningful predictors.

7. Model Comparison and Selection

To determine the most suitable specification, the Poisson and negative binomial models were compared using the Akaike Information Criterion (AIC). Preference was given to the model yielding the lowest AIC value.

8. Final Model Determination and Interpretation

The selected model was established as the final analytical framework based on both statistical performance and research objectives. The estimated parameters were then interpreted to elucidate the effects of the explanatory variables on the distribution of TB cases.

4. RESULT & DISCUSSION

This section outlines the empirical findings derived from the analytical procedures applied in the study, encompassing descriptive statistical techniques as well as Poisson and negative binomial regression models. Particular attention is given to evaluating the extent of overdispersion in tuberculosis case counts and to assessing the suitability of alternative modeling approaches for capturing the association between TB incidence and the selected explanatory variables. The discussion subsequently interprets the estimated parameters and considers their relevance for understanding the determinants of tuberculosis occurrence in Lampung Province.

4.1 Descriptive Statistical Analysis

The dataset examined comprises tuberculosis (TB) case counts recorded across 15 regencies and municipalities in Lampung Province. The response variable corresponds to the number of TB cases (Y), while the explanatory variables include the population living below the poverty line (X_1), the number of general hospitals in Lampung Province (X_2), and population density (X_3). The variables were chosen based on their theoretical relevance to tuberculosis incidence reported in previous epidemiological studies and the availability of consistent data at the district level [1, 3, 5, 6, 7, 8]. Additionally, considering the relatively limited number of observations, the model specification was intentionally kept parsimonious to avoid overparameterization. To provide an initial characterization of the data structure and variability, numerical summary statistics for each variable are reported in **Table 2**, supported by graphical representations where appropriate.

Table 2. Numerical Summaries of the Study Variables

Variable	Max	Min	Average	Variance
Y	51.40	23.60	37.59	152.0957
X_1	159.79	15.24	72.26	850.3830
X_2	12.000	1.000	39.33	203.6499
X_3	6464.18	54.76	812.67	1500.9945

As shown in **Table 2**, tuberculosis (TB) case counts in Lampung Province display limited dispersion, reflected in an average of 37.59 cases accompanied by relatively low variability. This condition suggests that the distribution of TB cases tends to decrease and remains relatively well controlled. The variable representing the number of people living below the poverty line (X_1) ranges from 15.24 to 159.79 persons, with an average of 72.26 persons. A lower number of people living in poverty may reduce the risk of TB transmission. The availability of healthcare facilities constitutes an essential component in efforts to prevent and manage tuberculosis. Adequate availability of healthcare facilities plays an important role in the management and prevention of TB. Meanwhile, population density (X_3) shows a wide range of variation, with a minimum value of 54.76 persons and a maximum reaching 6,464.18 persons, as well as an average of 812.67 persons. High population density may increase the likelihood of TB transmission. Visual inspection of the data structure is conducted by applying density-based plots to the response variable Y and and bar-chart representations to the explanatory variables X .

Figure 1 presents the estimated density distribution of variable Y , which exhibits a pattern resembling a normal distribution. The density curve appears relatively symmetric with a single main peak around the value of 40, indicating that most observations are concentrated near this value. The fairly wide range of the data suggests variability across regions; however, no extreme deviations are observed. Overall, this visualization indicates that the distribution of TB cases shows moderate dispersion and is centered around the mean value.

The observed distribution indicates an absence of pronounced extreme values in the TB case counts, suggesting that the data generally conform to the basic features of count outcomes commonly analyzed using Poisson-based models. However, a distribution that appears approximately symmetric does not in itself guarantee the suitability of Poisson regression, as this approach fundamentally relies on the equidispersion condition, whereby the mean and variance are assumed to be equal. Accordingly, a formal assessment of overdispersion is required to evaluate model validity. When empirical evidence shows that variability exceeds the mean, negative binomial regression offers a more suitable modeling framework, as it explicitly accounts for excess dispersion and yields more reliable parameter estimates than the Poisson specification.

As illustrated in **Figure 2**, bar-chart visualizations highlight the distributional patterns of both the response and explanatory variables across administrative areas in Lampung Province. The visualization of the bar chart of the response variable (Y) exhibits considerable regional differences in TB case numbers, suggesting spatial heterogeneity in disease occurrence. The distribution of the number of people living below the poverty line (X_1) varies considerably across regions, suggesting differences in socioeconomic conditions that may contribute to disparities in TB cases. The bar chart of the number of general hospitals (X_2) indicates unequal availability of healthcare facilities, which may affect access to diagnosis and treatment services. Meanwhile, population density (X_3) exhibits substantial variation, with several regions showing markedly higher density levels, a condition that may increase the risk of TB transmission. Overall, these results highlight heterogeneity in both the response and explanatory variables across regions, supporting the application of count data regression models within the Generalized Linear Model

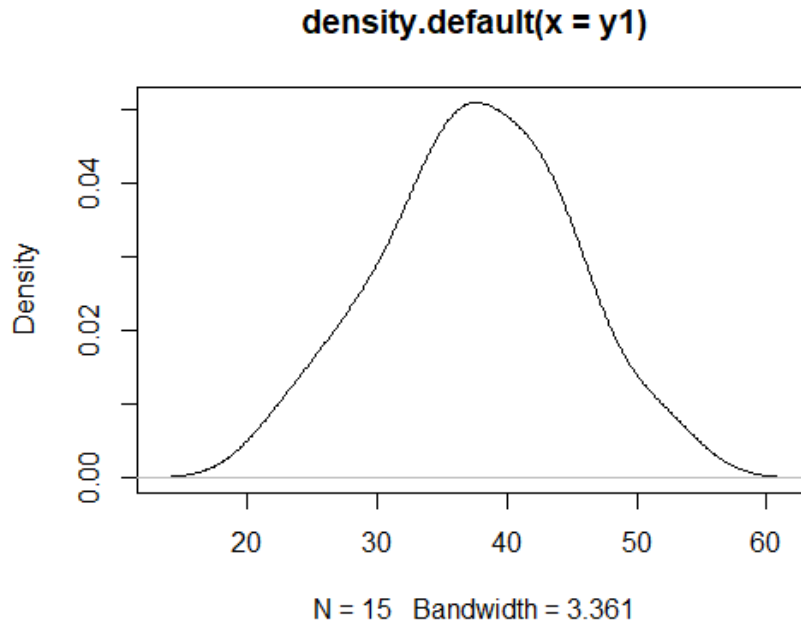


Figure 1. Density Plot of Y

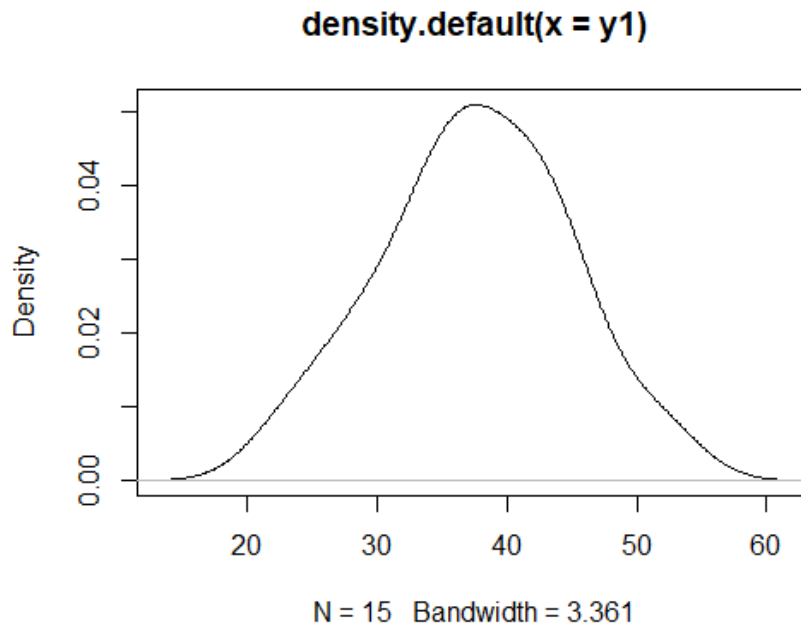


Figure 2. Bar-chart representation of the response variable Y and X_1, X_2, X_1

(GLM) framework in assessing factors related to tuberculosis cases in Lampung Province.

4.2 Modeling Tuberculosis Incidence with a Poisson Framework

A Poisson regression approach was employed to quantify the relationship between tuberculosis (TB) incidence and the selected explanatory variables in Lampung Province. Model parameters were estimated using the MASS library within the R computing

environment, and the corresponding results are reported in Table 4.2.

Table 3. Poisson Regression Coefficient Estimates for TB Case Counts

Parameter	Estimate	Z-Value	Pr(> z)
Intercept	3.109	739.013	$< 2 \times 10^{-16}***$
β_1	0.0131	18.367	$< 2 \times 10^{-16}***$
β_2	0.01058	-0.518	0.605
β_3	0.00004127	1.245	0.213

The estimated coefficients from the Poisson regression model for tuberculosis (TB) case counts in Lampung Province are reported in Table 4.2. The results show that the population living below the poverty line (X_1) is positively and significantly associated with TB incidence, indicating that increases in poverty levels correspond to higher expected numbers of TB cases. Conversely, the coefficients for the number of general hospitals (X_2) and population density (X_3) are not statistically significant at the 5% level, suggesting that their effects are not distinguishable from zero within this model specification. Based on the parameter estimates, the fitted Poisson regression model can be written as:

$$\log(\mu_i) = 3.109 + 0.0131X_{1i} - 0.01058X_{2i} + 0.00004127X_{3i}$$

or, equivalently, in exponential form,

$$\mu_i = \exp(3.109 + 0.0131X_{1i} - 0.01058X_{2i} + 0.00004127X_{3i})$$

Where μ_i represents the expected number of TB cases in the i -th regency or municipality.

4.3 Equidispersion Assumption and Overdispersion Testing

A fundamental premise of Poisson regression is the equidispersion condition, whereby the variance of the response variable is assumed to coincide with its expected value. In empirical analyses of count data, this requirement is frequently unmet, particularly in situations characterized by substantial variability or an excess of zero observations. Under such circumstances, the data tend to display overdispersion, reflected by a variance that surpasses the mean.

Consequently, once the Poisson model has been estimated, it is necessary to assess the validity of the equidispersion assumption through a formal overdispersion test. When evidence of overdispersion is present, reliance on the Poisson specification may result in inefficient estimation and unreliable inference. In such cases, negative binomial regression provides a more suitable modeling alternative due to its ability to accommodate additional dispersion in the data.

Table 4. Equidispersion Assumption and Overdispersion Testing

Deviance Value	Df (Degree of Freedom)	Deviance Value / Df
31.793	11	2.890273

Referring to the results in **Table 4**, the Poisson model yields a deviance statistic of 31.793 with 11 degrees of freedom, resulting in a deviance-to-degree-of-freedom ratio of 2.89. A ratio exceeding unity indicates that the observed variability in the data is greater than would be expected under the equidispersion assumption, thereby signaling a violation of this fundamental Poisson requirement. These findings confirm the presence of overdispersion in the tuberculosis case counts. Under such conditions, reliance on the Poisson specification may lead to inefficient coefficient estimates and distorted standard errors. Accordingly, a negative binomial model is considered a more suitable alternative, as it explicitly accommodates excess dispersion and offers more stable inferential results.

Table 5. Parameter Estimates for the Negative Binomial Regression of TB Case Counts

Parameter	Estimate	Z-Value	Pr(> z)
Intercept	3.069×10^0	27.845	$< 2 \times 10^{-16}^{**}$
X_1	1.423×10^{-2}	12.375	$< 2 \times 10^{-16}^{**}$
X_2	-2.852×10^{-2}	-0.813	0.416
X_3	6.073×10^{-5}	1.069	0.285

4.4 Modeling TB Case Counts Under a Negative Binomial Framework

In response to the overdispersion identified in the Poisson framework, the analysis proceeded with the application of negative binomial regression. This approach relaxes the restriction of equal mean and variance, allowing for a more flexible representation of count data. The estimated parameters from the negative binomial model are reported in **Table 5**. Using the estimated coefficients, the fitted negative binomial regression model is written as

$$\mu_i = \exp(3.069 + 0.01423X_{1i} - 0.02852X_{2i} + 0.00006073X_{3i})$$

Where μ_i represents the expected number of tuberculosis (TB) cases in the i -th regency or municipality. The results indicate that the intercept is statistically significant, representing the baseline expected log-count of TB cases when all explanatory variables are held at their reference values. Among the covariates considered, X_1 which captures the number of individuals living below the poverty line, exhibits a positive and statistically significant relationship with TB incidence, indicating that higher poverty levels are associated with increased expected case counts when other factors are held constant. By contrast, the coefficients for the number of general hospitals X_2 and X_3 (population density) are not statistically significant at the 5% level. This suggests that, after accounting for excess dispersion in the data, there is limited empirical support for a direct effect of these variables on TB case counts in Lampung Province under the negative binomial specification.

Taken together, these results demonstrate that the negative binomial model offers a more robust and reliable representation of the tuberculosis case data than the Poisson model, particularly in settings where overdispersion is present.

4.5 Assessing Covariate Impacts (Marginal Effect) via Incidence Rate Ratios (IRR)

For ease of interpretation within the negative binomial framework, the estimated regression coefficients (β) are re-expressed in terms of Incidence Rate Ratios (IRR) by applying the exponential transformation, such that.

$$IRR = \exp(\beta)$$

This transformation allows the coefficients to be interpreted on a multiplicative scale. Specifically, the IRR quantifies the proportional change in the expected mean number of tuberculosis (TB) cases resulting from a one-unit increase in a given explanatory variable, while all other covariates are held constant.

Among the explanatory variables considered, the size of the population living below the poverty line (X_1) emerges as the strongest and only statistically significant determinant of tuberculosis incidence. The estimated Incidence Rate Ratio (IRR) of 1.0143 implies that a one-unit increase in the poor population corresponds to a 1.43% increase in the expected TB incidence rate, assuming other factors remain unchanged. This result reinforces the importance of socioeconomic vulnerability as a key underlying factor in the spread of tuberculosis.

The coefficient associated with the number of general hospitals (X_2) shows a negative association with TB cases, suggesting that increased availability of healthcare facilities may help reduce TB incidence. However, this effect is not statistically significant, indicating that the mere quantity of healthcare facilities may be insufficient without considering service quality, accessibility, and effectiveness.

Population density (X_3) This finding suggests that population concentration, in isolation, is insufficient to explain variations in TB transmission once socioeconomic disparities are explicitly accounted for in the model.

Based on the negative binomial regression results and IRR interpretation, it can be concluded that socioeconomic factors—particularly the size of the poor population—are the primary determinants of increased tuberculosis (TB) cases in Lampung Province. These findings emphasize that TB control strategies should not focus solely on medical interventions but must be integrated with broader social and economic policies.

Table 6. Interpretation of IRR from the Negative Binomial Regression

Variable	Estimated β	IRR = $\exp(\beta)$	Interpretation
Intercept	3.069	21.53	Baseline expected mean of TB cases under the reference condition.
X_1 (Poor population)	0.01423	1.0143	A one-person increase in the poor population increases the expected TB incidence by 1.43%.
X_2 (General hospitals)	-0.02852	0.972	An increase of one general hospital is associated with an estimated 2.8% reduction in the expected tuberculosis incidence; however, this effect does not achieve statistical significance.
X_3 (Population density)	0.00006073	1.00006	Higher population density is linked to a marginal increase of approximately 0.006% in tuberculosis incidence, although the estimated effect is not statistically significant.

4.6 Health Policy–Based Conclusions

Based on the negative binomial regression results and IRR interpretation, it can be concluded that socioeconomic factors—particularly the size of the poor population—are the primary determinants of increased tuberculosis (TB) cases in Lampung Province. These findings emphasize that TB control strategies should not focus solely on medical interventions but must be integrated with broader social and economic policies.

4.7 Policy Implications

4.7.1 Strengthening socioeconomic-based interventions

Local governments should prioritize poverty alleviation programs, improvements in housing quality, and enhanced access to adequate nutrition for low-income populations as integral components of TB control strategies.

4.7.2 Optimization of primary healthcare services

Although the number of hospitals does not show a significant effect, improving service quality, early case detection, and treatment adherence at primary healthcare facilities remains crucial for reducing TB transmission Risk-based regional approach Areas with high concentrations of impoverished populations should be prioritized for active surveillance, contact tracing, and community-based health education programs.

4.7.3 Cross-sectoral policy integration

Effective TB control requires collaboration across sectors, including health, social welfare, and regional planning agencies, to create environments that support the prevention and control of infectious diseases.,

4.8 Selection of the Best Model

Following the estimation of the negative binomial regression, model adequacy was assessed to determine the most appropriate specification for the data. Model comparison was conducted using the Akaike Information Criterion (AIC), a metric that evaluates the trade-off between explanatory power and model parsimony, with lower values indicating preferable performance. The AIC values obtained for the Poisson and negative binomial models are summarized in **Table 7**.

Table 7. AIC-based Comparison of Competing Regression Models

Model	AIC Value
Poisson Regression	127.3
Negative Binomial Regression	126.2

As shown in **Table 7**, the negative binomial model yields a smaller AIC than the Poisson alternative, indicating a superior balance between fit and complexity. This evidence supports the selection of the negative binomial regression as the preferred model for explaining variation in tuberculosis (TB) case counts across regencies and municipalities in Lampung Province.

5. CONCLUSION

This study examines tuberculosis (TB) incidence in Lampung Province using a modeling approach that explicitly addresses overdispersion in count data. The findings demonstrate that the Poisson regression framework fails to adequately meet the equidispersion requirement, limiting its ability to capture the observed variability in TB case counts. In contrast, negative binomial regression offers a more appropriate specification by accommodating excess dispersion and yielding more reliable parameter estimates. The empirical results identify the population living below the poverty line as a significant determinant of increased TB incidence, underscoring the importance of socioeconomic vulnerability in shaping TB transmission patterns. Although the estimated effects of public hospital availability and population density are directionally consistent with theoretical expectations, neither variable attains statistical significance within the fitted model. Taken together, these results confirm the superiority of the negative binomial model for analyzing TB case data in Lampung Province. From a policy perspective, the findings highlight the need for TB control strategies that extend beyond clinical interventions by incorporating poverty reduction measures and enhancing the quality and accessibility of healthcare services. Future research may benefit from integrating a broader set of social, environmental, and health system indicators to develop a more comprehensive and nuanced modeling framework.

REFERENCES

- [1] I. S. Budi, Y. Ardillah, I. P. Sari, and D. Septiawati, “Analisis Faktor Risiko Kejadian penyakit Tuberculosis Bagi Masyarakat Daerah Kumuh Kota Palembang,” *JKLI*, vol. 17, no. 2, p. 87, Oct. 2018.
- [2] H. I. Zebua and G. A. Harefa, “Structural Equation Model (SEM) dalam Pemodelan Kemiskinan di Pulau Sumatera,” *IJAS*, vol. 5, no. 1, p. 48, May 2022.
- [3] L. Pangaribuan, K. Kristina, D. Perwitasari, T. Tejayanti, and D. B. Lolong, “Faktor-Faktor yang Mempengaruhi Kejadian Tuberkulosis pada Umur 15 Tahun ke Atas di Indonesia,” *hsr*, vol. 23, no. 1, pp. 10–17, May 2020.
- [4] M. S. Tamunu, D. N. Pareta, H. Hariyadi, and F. A. Karauwan, “Skrining Fitokimia Dan Uji Aktivitas Antioksidan Ekstrak Daun Benalu Pada Kersen Dendrophloe pentandra (L.) Dengan Metode 2,2- diphenyl -1- Picrylhydrazyl (DPPH),” *J.Biofar.Trop*, vol. 5, no. 1, pp. 79–82, Apr. 2022.
- [5] E. Y. Muhammad, “Hubungan tingkat pendidikan terhadap kejadian tuberkulosis paru,” *Jurnal Ilmiah Kesehatan Sandi Husada*, vol. 8, no. 2, pp. 288–291, 2019.
- [6] D. Andriani and S. Sukardin, “Pengetahuan dan Sikap Keluarga Dengan Pencegahan Penularan Penyakit Tuberculosis (TBC) Di Wilayah Kerja Puskesmas Penana’e Kota Bima,” *JIIKI*, vol. 10, no. 03, pp. 72–80, Sep. 2020.
- [7] W. Aryawati, N. Indrawati, E. Yuliana, and H. F. Usfa, “Analisis kejadian kasus baru tuberkulosis dinas kesehatan kabupaten lampung tengah tahun 2022,” *Jurnal Pendidikan Dan Konseling (JPDK)*, vol. 4, no. 4, pp. 2276–2281, 2022.
- [8] N. I. Fawzi, “Analisis Program Dots Untuk Menurunkan Kasus Tuberculosis Di Sekitar Taman Nasional Gunung Palung, Kalimantan Barat,” *juke*, vol. 13, no. 1, pp. 25–30, Feb. 2020.
- [9] M. Hayati, A. H. Wigena, A. Djuraidah, and A. Kurnia, “A new approach to statistical downscaling using Tweedie compound Poisson gamma response and lasso regularization,” *Commun. Math. Biol. Neurosci.*, 2021.
- [10] M. Hayati, K. Sadik, and A. Kurnia, “Conwey-Maxwell Poisson Distribution: Approach for Over- and-Under-Dispersed Count Data Modelling,” *IOP Conf. Ser.: Earth Environ. Sci.*, vol. 187, p. 012039, Nov. 2018.

- [11] C. Tri Fatmala, M. Hayati, R. Permatasari, M. Hudori, and D. Yuliana Dalimunthe, “Pemodelan Jumlah Kasus HIV/AIDS di Provinsi Lampung Menggunakan Regresi Binomial Negatif,” *Jomta*, vol. 6, no. 2, pp. 168–177, Oct. 2024.
- [12] D. I. Purnama, “Comparison of Zero Inflated Poisson (ZIP) Regression, Zero Inflated Negative Binomial Regression (ZINB) and Binomial Negative Hurdle Regression (HNB) to Model Daily Cigarette Consumption Data for Adult Population in Indonesia,” *J*, vol. 17, no. 3, pp. 357–369, May 2021.
- [13] P. R. Arum, I. Manfaati Nur, A. Jihan Syafiqoh, and H. Rizky Utami, “Permodelan Jumlah Kasus Tuberkulosis Di Kabupaten Purbalingga Tahun 2022 Menggunakan Regresi Binomial Negatif: Modeling the Number of Tuberculosis Cases in Purbalingga Regency in 2022 Using Negative Binomial Regression,” *jodi*, vol. 1, no. 2, pp. 44–50, Dec. 2023.
- [14] M. Fathurahman, “Regresi Binomial Negatif untuk Memodelkan Kematian Bayi di Kalimantan Timur,” *J. Eksponensial*, vol. 13, no. 1, p. 79, Jun. 2022.
- [15] H. M. Winata, “Mengatasi Overdispersi Dengan Regresi Binomial Negatif Pada Angka Kematian Ibu Di Kota Bandung,” *J. Gauss*, vol. 11, no. 4, pp. 616–622, Feb. 2023.
- [16] M. Hayati and A. Muslim, “Generalized Linear Mixed Model and Lasso Regularization for Statistical Downscaling,” *ENTHUSIASTIC*, vol. 1, no. 01, pp. 36–52, Apr. 2021.
- [17] M. Hayati and R. Permatasari, “Comparison of Generalized Linear Model between Gamma and Tweedie Compound Response for Rainfall Prediction in Lampung Province,” *AJPAS*, vol. 26, no. 1, pp. 41–49, Jan. 2024.
- [18] P. McCullagh and J. Nelder, *Generalized Linear Models*, 2nd ed. Routledge, Jan. 2019.
- [19] S. Iin, “Regresi poisson dan penerapannya untuk memodelkan hubungan usia dan perilaku merokok terhadap jumlah kematian penderita penyakit kanker paru - paru,” *Mat. UNAID*, vol. 1, no. 1, pp. 71–76, 2022.
- [20] A. Agresti, *Foundations of linear and generalized linear models*, ser. Wiley series in probability and statistics. Hoboken, New Jersey: John Wiley & Sons Inc, 2015.
- [21] J. W. Hardin and J. M. Hilbe, *Generalized linear models and extensions*, fourth edition ed. College Station, Texas: Stata Press, 2018.
- [22] A. Safitri, I. Rahmi, and D. Devianto, “Penerapan regresi poisson dan binomial negatif dalam memodelkan jumlah kasus penderita aids di indonesia berdasarkan faktor sosiodemografi,” *J. Mat. UNAND*, vol. 3, no. 4, pp. 58–65, 2014.