EKSAKTA
Journal of Sciences and Data Analysis

E-ISSN: 2720-9326
P-ISSN: 2716-0459
DOI : 10.20885/EKSAKTA.vol7.iss1.art3

Negative Binomial Regression Analysis of Factors Influencing Tuberculosis Cases in West Java Indonesia

Aufa Fikriya[1], Shalfa Salsabilla[2], Raisah Zharifah Labibah[3], Sri Winarni[4]*, Defi Yusti Faidah[5], Anindya Apriliyanti Pravitasari[6], Triyani Hendrawati[7], Irlandia Ginanjar[8]

[1,2,3] Bachelor Programme of Statistics, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Indonesia
[4,5,6,7,8] Department of Statistics, Faculty of Mathematics and Natural Sciences, Padjadjaran University, Indonesia

* *Corresponding author: sri.winarni@unpad.ac.id*

Abstract: Tuberculosis (TB) remains a major public health problem globally, with West Java reporting the highest number of TB cases among all provinces in Indonesia in 2023. This study aims to identify key factors influencing TB incidence across districts and cities in West Java in 2024. The analysis focuses on healthy living behaviors, proper sanitation, HIV cases, and AIDS cases using a Negative Binomial Regression approach to address overdispersion in count data. The results show that proper sanitation has a significant negative association with TB incidence, while HIV and AIDS cases exhibit significant positive associations. The best-performing model includes these three variables, yielding a residual deviance of 27.615. These findings highlight the importance of integrated public health interventions that simultaneously improve sanitation and strengthen HIV/AIDS control programs to effectively reduce TB incidence in high-burden regions.

Keywords: Tuberculosis; Overdispersion; Negative Binomial Regression.

Introduction

Tuberculosis (TB) is recognized as a major global public health challenge and remains among the leading causes of death from infectious diseases [1]. The disease is caused by *Mycobacterium tuberculosis*, a bacterium that primarily infects the lungs but may also affect other organs of the body [2]. TB is transmitted through the air, and individuals can become infected by inhaling air contaminated with Mycobacterium tuberculosis [3]. The bacteria can be easily carried in airborne particles generated by coughing [4]. According to the Indonesian Ministry of Health, a single cough can release approximately 3,000 droplets containing up to 3,500 Mycobacterium tuberculosis bacteria, while a single sneeze can release between 4,500 and one million bacteria with an estimated 40,000 droplets [5]. The 2023 Global TB Report stated that the global increase in tuberculosis (TB) cases between 2020 and 2022 was largely attributed to Indonesia, Myanmar, and the Philippines, where the number of cases in these three countries was estimated to have risen by approximately 0.4 million [6]. Furthermore, Indonesia ranks second globally in terms

of the number of TB cases, with 1,060,000 new cases and 134,000 deaths each year, equivalent to 15 deaths every hour [7].

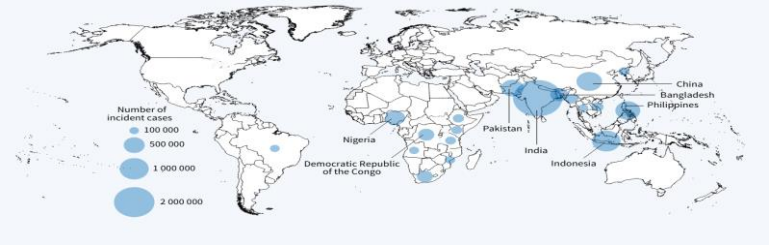

Figure 1. Estimated TB incidence in 2023 for countries with at least 100,000 new cases
Source: World Health Organization, Global Tuberculosis Report 2023

According to data from the Indonesian Ministry of Health, there was an increase in tuberculosis cases in Indonesia in 2023, reaching around 1,060,000 cases. This is the highest number ever recorded. Of the national TB cases, 57.9% were among men, while the remaining cases were among women [1]. At the provincial level, five provinces contributed to more than half of the national TB cases: West Java (105,794), East Java (71,791), Central Java (65,014), Jakarta (41,441), and North Sumatra (35,035). These regions are also among the most densely populated in Indonesia [8].

To accelerate TB eradication, Indonesia enacted Presidential Regulation No. 67 of 2021, targeting TB elimination by 2030. The regulation aims to reduce the number of cases to 65 per 100,000 population and mortality to 6 per 100,000 population [2]. Achieving these targets requires support through the enhancement of healthcare facilities, strengthening and expanding laboratory-based surveillance, tracking initial Lost to Follow Up (iLTFU) patients with drug-resistant TB (DR-TB) by establishing a "Tuberculosis Army," and developing a tuberculosis vaccine [9]. This endeavor is challenging and necessitates the cooperation of the community to reduce the risk factors associated with TB.

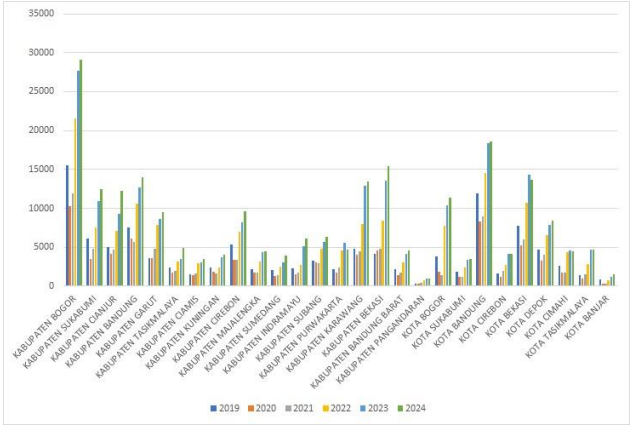

Figure 2. Number of TB cases by district/city in West Java in 2019-2024

Tuberculosis (TB) can be influenced by sanitation and residential conditions [10]. High population density and inadequate housing conditions that fail to meet health

standards contribute to poor air quality, facilitating the transmission of TB [11,12]. Individuals with TB have a higher likelihood of spreading the bacteria to those in close, confined spaces. Adequate ventilation in homes can help prevent TB. Poor lighting conditions, found in 46.2% of homes, are associated with a 1.056-fold increased risk of TB infection compared to homes with adequate lighting, which is present in 53.8% of cases. This is significantly related to air humidity; good lighting typically results in lower humidity, which inhibits bacterial growth. Mycobacterium tuberculosis thrives at temperatures between 31-37 degrees Celsius [13]. When a person is infected with Mycobacterium tuberculosis, the immune system attempts to combat the infection. However, if the individual also has another infection, such as HIV, which weakens the immune system, the risk of TB infection increases as the bacteria can more easily overcome the compromised immunity [14].

The primary objective of this study is to assess the impact of healthy living behaviors, proper sanitation, and HIV/AIDS cases on TB incidence in West Java province in 2024. By examining these factors, the study aims to assist the government and community in reducing TB cases and achieving reduction targets, particularly in West Java, which has the highest TB incidence in the country. This research provides valuable insights for the provincial government in addressing TB cases and builds on the limitations of previous studies [14]. Previous studies have examined the relationship between tuberculosis incidence and socio-environmental or comorbidity factors, including HIV/AIDS and housing conditions, at national or provincial levels, such as in East Java and other regions of Indonesia. However, most existing studies either focus on a limited set of explanatory variables or do not explicitly address overdispersion issues inherent in TB count data across heterogeneous regions. This study extends prior research by jointly examining sanitation, healthy living behaviors, and HIV/AIDS cases across all districts and cities in West Java using a negative binomial regression framework. By explicitly accounting for overdispersion, this study provides a more robust understanding of TB determinants, offering empirical evidence that complements and extends previous findings.

Materials and Methods

This section provides an overview of the Poisson Regression Model, overdispersion, the Negative Binomial Regression Model, and the Akaike Information Criterion (AIC). Additionally, it presents the research data and the methodology employed in the study.

Poisson Regression

The Poisson regression model is used to analyze count data and to describe the rate at which certain events occur. For a random variable Y, the Poisson distribution is defined by the probability mass function:

$$P(Y = y; \mu) = \frac{e^{-\mu}\mu^{y}}{y!} \quad \text{for } y = 0,1,2,\ldots \text{ and } \mu > 0 \qquad (1)$$

The distribution is characterized by a single parameter μ, which represents the mean frequency of events occurrence within a specified period. In this distribution, both the mean and variance are equal to μ:

$$E[y] = Var[y] = \mu \tag{2}$$

In Poisson regression, the dependent variable Y represents an observed count that follows the Poisson distribution. The mean rate μ is modeled as a function of several predictors $X = (X_1, X_2, \ldots, X_k)$ using the exponential link:

$$\mu = exp\{X\beta\} \tag{3}$$

Where β represents the vector of regression coefficients. Consequently, for observation i, the basic Poisson regression model is provided by [15]

$$P(Y_i = y_i|X_i, \beta) = \frac{e^{-exp\{X_i,\beta\}} exp\{X_i, \beta\}^{yi}}{yi!} \tag{4}$$

In other words, the categorical outcome for a given set of predictors follows a Poisson distribution with rate $exp\{X_i\beta\}$. The likelihood function for a sample size n in Poisson regression is:

$$L(\beta; y, X) = \prod_{i=1}^{n} \frac{e^{-exp\{X_i\beta\}} exp\{X_i\beta\}^{yi}}{yi!} \tag{5}$$

This yields the log likelihood:

$$l(\beta) = \sum_{i=1}^{n} y_i X_i \beta - \sum_{i=1}^{n} exp\{X_i\beta\} - \sum_{i=1}^{n} log(y_i!) \tag{6}$$

The logit function is defined as the natural logarithm of the odds that the dependent variable Y takes the value 1 rather than 0. Here, P refers to the probability that $Y = 1$ in a binary outcome model. The logit function is used as link function in logistic regression, which relates the probability of Y to the predictors. Specifically, the logit of Y is used in the regression model, rather than Y itself [16] :

$$ln\left(\frac{P}{1-P}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k \tag{7}$$

where $ln\left(\frac{P}{1-P}\right)$ is the log-odds of Y, and $\beta_0, \beta_1, \ldots, \beta_k$ are the regression coefficients.

Negative Binomial Regression

The Negative Binomial Regression model is less commonly used but is important in certain statistical contexts. While the statistical community often treats the negative binomial as a single model, similar to Poisson regression or logistic regression, it has its unique applications. The negative binomial distribution estimates the number of trials required to achieve a given number of successes while accounting for possible failures

before success occurs. Its probability mass function, where p represents the probability of success and $(1-p)$ the probability of failure, is expressed as [17]:

$$P(Y=y|r,p)=\begin{bmatrix}y-1\\r-1\end{bmatrix}p^r(1-p)^{y-r}\ \ for\ \ Y=r,r+1,r+2,\dots \tag{8}$$

Where r is the number of successes and y is the total number of trials. The expected value and variance of this distribution are given by:

$$E[y]=r\frac{(1-p)}{p}\ \ \text{and}\ \ Var[y]=r\frac{(1-p)}{p^2} \tag{9}$$

In the context of a mixed Poisson-Gamma distribution, where unobserved variance is modeled using a Gamma-distributed random effect with mean 1 and variance ϕ, the mean of the mixed distribution can be written as:

$$E(Y_i)=\tilde{\mu_i}=exp(X_i^T\beta+m_i)=exp(X_i^T\beta)\,exp(m_i)=\mu_i\delta_i \tag{10}$$

where δ_i represents the Gamma-distributed random effect, and μ_i is the mean of the Poisson component. With $\mu_i=exp(X_i^T\beta)$ representing the mean of the Poisson model, the probability density function for the mixed Poisson-Gamma distribution is expressed as:

$$f(y_i|x_i,\beta,\delta_i)=\frac{e^{-(\mu_i\delta_i)}(\mu_i\delta_i)^{y_i}}{y_i!} \tag{11}$$

To obtain the mixed Poisson-Gamma distribution, integrate the variable δ_i into the Poisson probability function:

$$f(y_i|x_i,\beta,\phi)=\int_0^\infty f(y_i|\phi,\delta_i)g(\delta_i)d\delta_i \tag{12}$$

where $f(y_i|\phi,\delta_i)$ is the Poisson probability function and $g(\delta_i)$ is the Gamma density function. This integration yields:

$$f(y_i|x_i,\beta,\phi)=\frac{\Gamma(y_i+\phi^{-1})}{\Gamma(\phi^{-1})y_i!}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i}\left(\frac{1}{1+\phi\mu_i}\right)^{\phi^{-1}} \tag{13}$$

Where $\phi>0, E[Y_i]=\mu_i$ and $Var[Y_i]=\mu_i+\phi\mu_i^2$. This formulation provides the mean and variance of the mixed Poisson-Gamma distribution, accounting for the additional variance introduced by the Gamma-distributed random effect.

Using the maximum likelihood estimation approach, the parameters of the Negative Binomial Regression model are obtained from the likelihood function:

$$L(\beta,\phi|y,x)=\prod_{i=1}^{n}\left\{\frac{\Gamma(y_i+\phi^{-1})}{\Gamma(\phi^{-1})y_i!}\left(\frac{\phi\mu_i}{1+\phi\mu_i}\right)^{y_i}\left(\frac{1}{1+\phi\mu_i}\right)^{\phi^{-1}}\right\} \tag{14}$$

Taking the natural logarithm of the likelihood function yields:

$$lnL(\beta,\phi|y,x)=\sum_{i=1}^{n}\left\{ln\left[\frac{\Gamma(y_i+\phi^{-1})}{\Gamma(\phi^{-1})\Gamma(y_i+1)}\right]-(y_i+\phi^{-1})ln(1+\phi\mu_i)+y_i ln(\phi\mu_i)\right\} \tag{15}$$

This logarithmic form simplifies the process of parameter estimation by converting the product of probabilities into a sum of logarithms.

Overdispersion

In the analysis of discrete data, the concept of overdispersion is essential. Overdispersion occurs when the observed variability within the data exceeds the level expected under the assumed probability distribution. This additional variability is not accounted for by the standard Generalized Linear Models (GLMs), as the mean and variance in GLMs are both determined by the same parameters through the predictors. Unlike linear regression, where the variability is typically well-modeled, overdispersion is a critical consideration in models for count data. In ordinary linear regression, the model is expressed as:

$$y_i \sim N(X_i^T \beta, \sigma^2) \tag{16}$$

where N denotes the normal distribution with mean $X_i^T \beta$ and variance σ^2. Overdispersion is not a concern in this context because the assumptions of normality and constant variance are typically satisfied.

In ordinary linear regression, the variance σ^2 is estimated independently of the mean function $X_i^T \beta$. However, with discrete response variables, the possibility of overdispersion arises because the commonly used distributions often specify a particular relationship between the variance and the mean. This is particularly relevant for Poisson regression, where the variance is assumed to equal the mean. When overdispersion occurs, the observed variance of the response exceeds what is predicted by the model. To test for overdispersion, the following test statistic is used:

$$\hat{\phi} = \frac{2\sum_{i=1}^{n}\left\{y_i log\frac{y_i}{\mu_i} - (y_i - \mu_i)\right\}}{n-p} \tag{17}$$

where:

$\hat{\phi}$: Dispersion parameter

y_i : Observed response value for the i-th observation

μ_i : Estimated mean for the i-th case

n : Number of observations

p : Number of parameters in model

The test statistic $\hat{\phi}$ helps determine if the observed variability in the data exceeds what is expected under the model assumptions.

Normality Test and Multicollinearity Test

Assessing whether data follows a normal distribution is crucial in many statistical analyses. The Shapiro-Wilk test is one of the most commonly used methods for assessing normality, particularly suitable for datasets with small sample sizes. This test evaluates if the sample data deviates significantly from a normal distribution. The Shapiro-Wilk test statistic is calculated using the following formula:

$$T_3 = \frac{1}{D}\left[\sum_{i=1}^{k} a_i (X_{n-i+1} - X_i)\right]^2 \quad (18)$$

Where D is the coefficient for the Shapiro-Wilk test, X_{n-i+1} is the $(n-i+1)$-th ordered value, X_i is the i-th ordered observation in the dataset, a_i represents the weights assigned in the Shapiro-Wilk test.

The multicollinearity test assesses whether independent variables in a regression model are correlated with each other. Detecting multicollinearity is important since strong correlations among predictors can produce unstable and unreliable estimates of regression coefficients. Two common methods for detecting multicollinearity are the Tolerance and Variance Inflation Factor (VIF). The VIF is calculated using the formula [18]:

$$VIF = \frac{1}{1-R^2} \quad (19)$$

where R^2 is the coefficient of determination obtained from regressing a particular independent variable against all other independent variables. To assess multicollinearity in a regression model, several criteria are used. Multicollinearity is generally considered absent if VIF is below 10 or if the tolerance value is above 0.01. Conversely, if the VIF exceeds 10 or the tolerance value falls below 0.01, it indicates the presence of multicollinearity. Additionally, multicollinearity is also identified if the correlation coefficient between any pair of independent variables exceeds 0.8. If the correlation coefficient is 0.8 or lower, multicollinearity is not typically viewed as a significant issue. These criteria help determine whether the independent variables in a regression model are too highly correlated, which could affect the reliability of the model's coefficients.

Simultaneous and Partial Test

It should be noted that the concepts of Sum of Squares for Regression (SSR) and Sum of Squares for Error (SSE) originate from ordinary linear regression. In the context of generalized linear models, including negative binomial regression, model significance is primarily assessed using likelihood-based statistics rather than variance decomposition. Therefore, the F-test formulation presented here is used as a conceptual reference, while inference in the negative binomial regression model relies mainly on likelihood ratio tests and Wald statistics for partial effects.

In regression analysis, it is important to evaluate both the overall and individual effects of independent variables on the dependent variable. The F-test serves as a primary method for assessing whether all independent variables, when considered together, significantly affect the dependent variable. This test examines whether the variation explained by the model is significantly greater than the variation due to random error. The F-test evaluates the joint significance of all predictors in the model. The significance level is typically set at 5%. When the p-value of the F-test is below 0.05, it indicates that the independent variables collectively have a statistically significant effect on the dependent variable. The F statistic is calculated using the formula [19]:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{(n-k-1)}} \tag{20}$$

where SSR denotes the Sum of Squares for Regression, SSE represents the Sum of Squares for Error, k is the number of predictor variables in the model, and n is the total number of observations in the sample. This formula compares the model's explained variance to the unexplained variance, adjusted for the number of predictors and observations, to determine the overall significance of the independent variables.

The F-test assesses the significance of all independent variables collectively. If the significance value of the F-test is less than 0.05, the null hypothesis (H_0) is rejected, implying that at least one independent variable significantly influences the dependent variable. Conversely, if the significance value is greater than 0.05, the null hypothesis is accepted, meaning that the independent variables do not significantly affect the dependent variable.

In regression analysis, it is essential to evaluate the significance of each independent variable in explaining the dependent variable. While the F test examines the overall model significance, the t test focuses on the contribution of individual predictors. This test evaluates whether a single predictor has a statistically significant effect on the dependent variable after accounting for the influence of other variables in the model. The formula for the t statistic is given as follows [16]:

$$t = \frac{\hat{\beta}_k}{se(\hat{\beta}_k)} \tag{21}$$

where $\hat{\beta}_k$ is the estimated value of the k-th regression coefficient and $se(\hat{\beta}_k)$ is the standard error of the k-th coefficient estimate.

Hypotheses are tested as follows: $H_0 : \beta_i = 0$ (indicating no significant effect), and $H_1 : \beta_i \neq 0$ (indicating a significant effect). The null hypothesis is rejected if the t statistic exceeds the critical value t_{table}, or if the p_{value} is less than 0.05. This indicates that the corresponding independent variable significantly affects the dependent variable. Otherwise, the null hypothesis is not rejected.

Akaike Information Criterion (AIC)

The Akaike Information Criterion (AIC) is a valuable tool for model selection, particularly in time series analysis where recent data is often more informative but can be limited in validation and test sets. Unlike traditional methods that split data into training, validation, and test sets, AIC allows for model selection by evaluating all available data, thus potentially improving model performance. AIC aims to balance the goodness of fit with model complexity, seeking the lowest possible AIC value to achieve the best trade-off between fit and generalizability. This balance is crucial for ensuring that the model performs well on new, unseen data. The formula for AIC is [20]:

$$AIC = -2\ ln\ (\hat{\theta}_{MLE}|y) + 2K \quad (22)$$

where $\hat{\theta}_{MLE}$ represents the maximum likelihood estimate of the model parameters, and K is the number of parameters in the model. AIC is particularly recommended when the number of observations K is less than $\frac{N}{40}$.

Research Procedure

The analytical procedure applied in this study consists of the following steps.

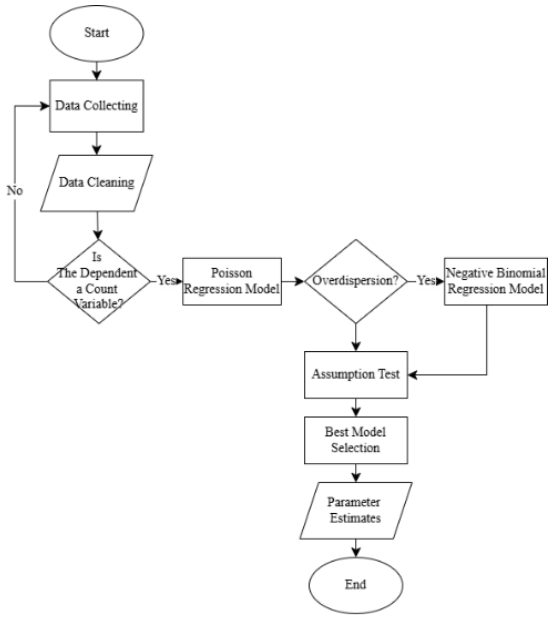

Figure 3. Flowchart Analysis

Reffering to **Figure 3**, the analytical procedure are outlined below:

1. Data related to the research variables were collected for subsequent analysis.
2. Data cleaning was performed to ensure that the dataset contained no missing values and was suitable for analysis.
3. Poisson regression analysis was applied when the dependent variable was identified as discrete count data.
4. An overdispersion test was conducted to prevent invalid inference in count data analysis and to avoid underestimated standard errors in Poisson regression when overdispersion is present.
5. When overdispersion was detected, a negative binomial regression model was employed as an alternative to the Poisson model.
6. Model assumption and diagnostic tests, including normality, multicollinearity, and both simultaneous and partial significance tests, were performed to ensure the reliability, validity, and interpretability of the estimated model.

7. The most appropriate model was selected based on the Akaike Information Criterion (AIC).
8. The selected best-fitting model was subsequently used to estimate the final model parameters.

Result

Descriptive Analysis

The distribution of tuberculosis (TB) cases across West Java, as well as the number of cases reported for each district and city, is illustrated in **Figure 4** and **Figure 5**. These figures provide an overview of the spatial variation and prevalence of TB within the province, highlighting areas with higher case counts and offering insights into the geographic distribution of the disease.

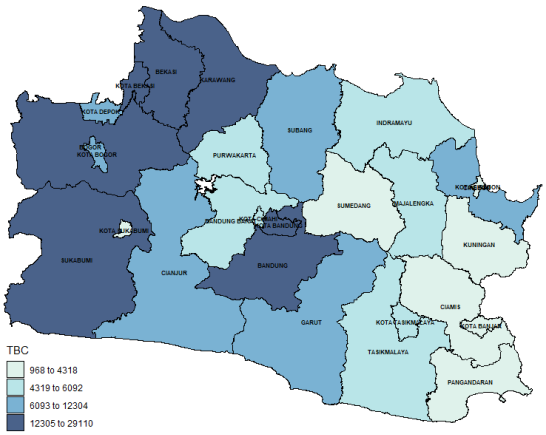

Figure 4. Distribution of TBC Cases in West Java

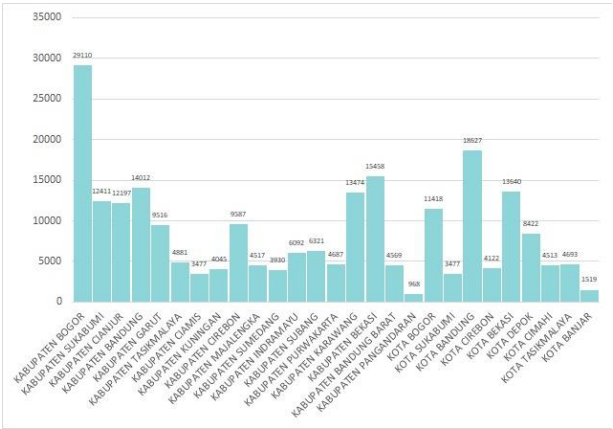

Figure 5. TBC Cases by District/City in West Java

It was observed that the highest number of tuberculosis (TB) cases in West Java is concentrated in Bogor District, which reported over 20,000 cases, followed closely by Bandung City. Conversely, Pangandaran District recorded the lowest number of TB cases in the province for 2024. The visual representation indicates that TB cases are

predominantly concentrated in the western part of West Java, with these areas showing darker shades on the map, signifying higher case counts. Notably, Sukabumi City, despite being surrounded by regions with high TB case numbers, has a relatively low total count of TB cases.

Table 1. Descriptive Statistics

Variabel	Min	Mean	Max
TB Cases (Y)	968	8507	29110
Percentage of Households Practicing Clean and Healthy Behaviors (X_1)	45,48	67,87	84,67
Percentage of Proper Sanitation (X_2)	45,88	75,70	99,08
HIV Cases (X_3)	51	394	1400
AIDS Cases (X_4)	13	113,9	373

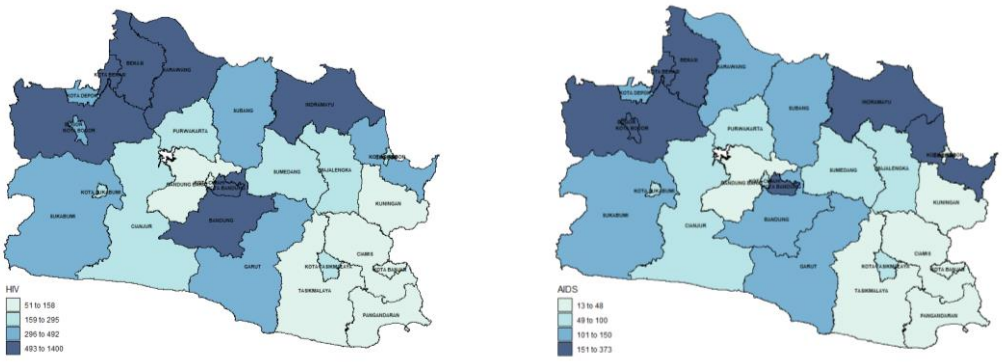

Figure 5. Distribution of HIV and AIDS Cases in West Java

Conversely, HIV cases exhibit a higher average compared to AIDS cases, making HIV more prevalent in West Java. Both HIV and AIDS cases are more uniform across the region. For instance the eastern Priangan region, such as the districts of Ciamis and Tasikmalaya, report a low number of both cases. Similarly, neighboring districts have HIV and AIDS prevalence rates that do not differ significantly between districts and surrounding cities. In both HIV and AIDS cases, cities and districts such as Bogor, Bekasi, and Bandung report higher incidences. Notably, these region also have high incidence of tuberculosis. Additionally, the data reveal that the percentage of proper sanitation is generally higher than the percentage of households practicing clean and healthy behaviors. This suggests that households with clean and healthy behaviors remain less common compared to those with proper sanitation in West Java.

The Result of Parameter Estimation

The parameter estimation was performed using Poisson regression within the Generalized Linear Model (GLM) framework, employing the 'log' link function to address the non-linear relationship between the dependent and independent variables. The estimation process utilized the maximum likelihood method (MLE) and involved four iterations of Fisher scoring. The results indicate that all four predictor variables were found to be statistically significant in influencing the response variable, as detailed in Table 2.

Table 2. Estimation of Poisson Regression Model

Variable	Parameter	Estimated Parameter	p_{value}	Decision
Intercept	β_0	$8.867e+00$	$< 2e-16$	Significant
X_1	β_1	$-3.788e-03$	$< 2e-16$	Significant
X_2	β_2	$-4.548e-03$	$< 2e-16$	Significant
X_3	β_3	$5.106e-04$	$< 2e-16$	Significant
X_4	β_4	$3.808e-03$	$< 2e-16$	Significant

Based on Table 2, the four predictor variables are found to be significant, and the estimated Poisson regression model is given by:

$$\mu_i = exp(8.867 - 0.003788x_1 - 0.004548x_2 + 0.0005106x_3 + 0.003808x_4).$$

Following this, classical assumption tests were conducted on the model. These tests include assessments of normality, as well as partial and simultaneous effects, and multicollinearity. Normality was tested using the Shapiro-Wilk test, as described in Equation (18). Simultaneous and partial effects were evaluated according to Equation (20), and the resulting p-values are presented in Table 3.

Table 3. Results of Normality Test and Simultaneous Test

Assumption	p_{value}	Decision
Normality	0.2916	Fulfilled
Simultaneity	$1.972336e-07$	Fulfilled

The simultaneous and partial tests, as shown in Table 2, indicate that the model is both feasible and linear. Additionally, the absence of high correlation among the predictors is confirmed by the VIF values, all of which are less than 5, as detailed in Table 4.

Table 4. Multicollinearity Test Results

X_1	X_2	X_3	X_4

1.483245	1.445590	3.589280	3.580089

Overdispersion Test

To ensure that the Poisson regression model assumptions are met, an overdispersion test is conducted to determine if the variance of the response variable exceeds the mean. Overdispersion occurs when the variability in the data is greater than what is assumed by the Poisson distribution, leading to biased parameter estimates with standard errors that are too low. The overdispersion test is performed using Equation (17), and the results are presented in Table 5.

Table 5. Overdispersion Test Results

Dispersion	p_{value}	Decision
980.7331	2.535e-05	Overdispersion

As shown in Table 5, the p_{value} for the dispersion test is 0.00002535, which indicates that there is significant overdispersion in the model. Since the dispersion value is greater than 1, this confirms the presence of overdispersion. To address this issue, a negative binomial regression model will be used as an alternative approach to better account for the overdispersed data.

The Result of Negative Binomial Regression

To address the overdispersion identified in the Poisson regression model, a negative binomial regression analysis was conducted. This approach, which follows the methodology outlined in Equation (15), is employed to better handle the variance issues observed. The results of the negative binomial regression model are presented in Table 6.

Table 6. Negative Binomial Regression Model Estimation

Variable	Parameter	Estimated Parameter	p_{value}	Decision
Intercept	β_0	8.7579843	$< 2e-16$	Significant
X_1	β_1	0.0010598	0.90370	Not Significant
X_2	β_2	−0.0104592	0.04132	Significant
X_3	β_3	0.0011019	0.00868	Significant
X_4	β_4	0.0034270	0.02163	Significant

As illustrated in Table 6, the analysis reveals that the variables for households with clean and healthy living behaviors do not have a significant effect on the occurrence of TB cases. However, the percentage of proper sanitation, HIV cases, and AIDS cases significantly affects TB incidence, as evidenced by a p-value, which is less than the alpha level of 0.05.

Following this, the model was re-evaluated to ensure that all necessary assumptions were met, as outlined in Equations (18) and (20). The results of these additional tests are presented in Table 7.

Table 7. Assumption Test on Significant Negative Binomial Model

Assumption	p_{value}	Decision
Normality	0.1875	Fulfilled
Simultaneity	5.081805e-08	Fulfilled

Table 7 indicates that the model meets the assumptions of normality and simultaneity. Additionally, as shown in Table 6, the model also satisfies the partial assumptions. Consequently, the negative binomial regression model is deemed appropriate for use. To ensure the robustness of the model, a comparison will be made with the previously used Poisson regression model. This comparison will focus on the AIC value to determine the most suitable model.

Best Model Selection

Selecting the best model involves evaluating and choosing the model that best fits the data. This process is often guided by criteria such as the AIC, which balances model fit and complexity. The model with the lowest AIC value is generally considered the best. The AIC values for each model are calculated using Equation (22) and presented in Table 8.

Table 8. Best Model Selection

Model	AIC
Poisson Regression	26958.08
Negative Binomial Regression	510.7046
Negative Binomial Regression with Significant Variable	508.7172

As shown in Table 8, the model with significant variables in the negative binomial regression has the lowest AIC value compared to both the Poisson regression model and the full negative binomial regression model. This indicates that the negative binomial regression model with significant predictor variables provides a better fit for the data. Therefore, the most appropriate and effective model is the negative binomial regression with significant variables. The parameter estimates for this best model are summarized in Table 9.

Table 9. Best Model Parameter Estimation

Variable	Parameter	Estimated Parameter	p_{value}	Decision

Intercept	β_0	8.8073690	$< 2e-16$	Significant
X_2	β_2	−0.0101465	0.01777	Significant
X_3	β_3	0.0010998	0.00798	Significant
X_4	β_4	0.0034248	0.02002	Significant

Additionally, the negative binomial regression model with significant predictor variables demonstrates a good level of fit. The residual deviance of 27.615, with 23 degrees of freedom is considerably smaller than the null deviance of 101.786, indicating that the model fits the data well.

Discussion

The presence of overdispersion confirms that the variability of TB cases across districts and cities in West Java cannot be adequately captured by a Poisson model. This finding reflects unobserved heterogeneity across regions, such as differences in population density, healthcare access, and socio-economic conditions. The negative binomial regression model effectively accommodates this excess variability, resulting in more reliable parameter estimates.

The significant positive effects of HIV and AIDS cases on TB incidence are consistent with previous studies, which emphasize immunosuppression as a major driver of TB vulnerability [14], [21]. Meanwhile, the negative association between proper sanitation and TB incidence highlights the critical role of environmental health interventions in reducing disease transmission. Compared to previous studies, this research provides a more detailed regional-level analysis while explicitly addressing overdispersion, thereby strengthening the empirical foundation for TB control policies in West Java.

The analysis of TB cases in West Java, which utilized discrete data from 2024, initially employed the Poisson regression model. This model was chosen due to its suitability for count data. However, the analysis uncovered overdispersion, as the variance in TB cases exceeded the mean, indicating that the Poisson model was not fully capturing the variability in the data. To address this issue, a negative binomial regression approach was adopted. The negative binomial regression revealed that the percentage of proper sanitation, HIV cases, and AIDS cases had a significant impact on the number of TB cases, with a p-value of 0.01777, 0.00798, and 0.02002, which is below the alpha level of 0.05. This result is consistent with previous research conducted in East Java [14], which also found a significant relationship between HIV and TB incidence. The significant effect of HIV and AIDS on TB cases underscores the critical need for targeted interventions to reduce TB incidence, particularly in regions like West Java. Additionally, the results of this study are consistent with previous research [21], indicating that the negative binomial regression model provides a better fit compared to the Poisson regression model.

These findings highlight the importance of public health authorities and policymakers to prioritize efforts aimed at managing HIV and AIDS as part of a broader strategy to control TB. Effective HIV and AIDS treatment and prevention programs could be crucial in reducing the incidence of TB and improving overall health outcomes in the region. In addition, the proper sanitation should not be underestimated, as maintaining a clean environment and ensuring access to proper sanitation facilities can substantially reduce the transmission of infectious diseases, including tuberculosis (TB). Efforts to enhance sanitation should be integrated with other public health interventions to achieve comprehensive and sustainable TB control.

Conclusion

The analysis of factors influencing TB cases in West Java for the year 2024 identified that among the four predictor variables examined, only the variables for households with clean and healthy living behaviors did not meet the necessary assumptions and demonstrated a significant impact on TB incidence in the region. The negative binomial regression approach was found to be more suitable compared to ordinary linear regression and Poisson regression methods, effectively addressing the issue of overdispersion present in the data. The best model, as determined from the negative binomial regression, has the parameter estimates represented by the equation:

$$\mu_i = exp(8.8073690 - 0.0101465x_2 + 0.0010998x_3 + 0.0034248x_4).$$

This indicates that for every one-unit increase in HIV cases and AIDS cases, the number of TB cases is expected to rise by 0.0010998 and 0.0034248. Additionally, for every one-unit increase in proper sanitation, the number of TB cases is estimated to decrease by 0.0101465. In the absence of proper sanitation, HIV cases, and AIDS cases, the baseline estimate for the number of TB cases is 8.8073690. This underscores the crucial role of managing proper sanitation, HIV, and AIDS to control and reduce TB incidence in the region, highlighting the need for integrated public health strategies.

This study demonstrates that negative binomial regression is an appropriate and robust approach for modeling TB incidence in West Java due to the presence of overdispersion in the data. The results indicate that improved sanitation significantly reduces TB cases, while increases in HIV and AIDS cases are associated with higher TB incidence. These findings underscore the importance of integrated public health strategies that simultaneously target environmental sanitation and HIV/AIDS management.

The novelty of this study lies in its application of negative binomial regression to district- and city-level TB data in West Java while explicitly addressing overdispersion and comparing model performance using AIC. This approach provides more reliable inference than conventional Poisson regression and offers valuable evidence to support targeted TB reduction policies in high-burden regions.

Acknowledgment

The research was supported by the Internal Matching Funds (IMF) grant provided by Universitas Padjadjaran, Indonesia, through the project *"Negative Binomial Regression Analysis of Factors Influencing Tuberculosis Cases in West Java, Indonesia"* (Contract No. 4356/UN6.D/PT.00/2025).

References

[1] Kementerian Kesehatan Republik Indonesia. "Laporan Program Penanggulangan Tuberkulosis". 2024.

[2] Presiden Republik Indonesia, "*Peraturan Presiden Republik Indonesia Nomor 67 Tahun 2021 tentang Penanggulangan Tuberkulosis"*, Jakarta, 2021.

[3] C. R. Plumlee, F. J. Duffy, B. H. Gern, J. L. Delahaye, S. B. Cohen, C. R. Stoltzfus, et al., "Ultra-low dose aerosol infection of mice with *Mycobacterium tuberculosis* more closely models human tuberculosis," *Cell Host & Microbe*, vol. 29, pp. 68–82.e5, 2021, doi: 10.1016/j.chom.2020.10.003.

[4] T. L. Chiyaka, G. R. Nyawo, C. C. Naidoo, S. Moodley, J. C. Clemente, S. T. Malherbe, et al., "PneumoniaCheck, a novel aerosol collection device, permits capture of airborne *Mycobacterium tuberculosis* and characterisation of the cough aeromicrobiome in people with tuberculosis," *Annals of Clinical Microbiology and Antimicrobials*, vol. 23, 2024, doi: 10.1186/s12941-024-00735-x.

[5] R. Dhand and J. Li, "Coughs and sneezes: Their role in transmission of respiratory viral infections, including SARS-CoV-2," *American Journal of Respiratory and Critical Care Medicine*, vol. 202, pp. 651–659, 2020, doi: 10.1164/rccm.202004-1263PP.

[6] World Health Organization, "*Global Tuberculosis Report 2023"*, Geneva: WHO, 2023.

[7] M. Lengari, P. Weraman, Y. K. Syamruth, L. P. Ruliati, and A. A. Adu, "Spatial autocorrelation of population density, HIV/AIDS, and diabetes mellitus with pulmonary tuberculosis in Kupang, East Nusa Tenggara, Indonesia," *Indonesian Journal of Medicine*, vol. 10, pp. 93–104, 2025, doi: 10.26911/theijmed.2025.10.02.01.

[8] Ministry of Health Republic of Indonesia, "*Revised National Strategy of Tuberculosis Care and Prevention in Indonesia 2020–2024 and Interim Plan for 2025–2026"*, Jakarta: MoH RI.

[9] J. Stoklosa, R. V. Blakey, and F. K. C. Hui, "An overview of modern applications of negative binomial modelling in ecology and biodiversity," *Diversity*, vol. 14, 2022, doi: 10.3390/d14050320.

[10] L. B. Diantara, H. Hasyim, I. P. Septeria, D. T. Sari, G. T. Wahyuni, and R. Anliyanita, "Tuberkulosis masalah kesehatan dunia: Tinjauan literatur," *Jurnal 'Aisyiyah Medika*, vol. 7, 2022, doi: 10.36729/jam.v7i2.855.

[11] A. M. Liyew, A. C. A. Clements, T. Y. Akalu, B. Gilmour, and K. A. Alene, "Ecological-level factors associated with tuberculosis incidence and mortality: A systematic review and meta-analysis," *PLOS Global Public Health*, vol. 4, 2024, doi: 10.1371/journal.pgph.0003425.

[12] R. K. Mahato, K. M. Htike, A. B. Koro, R. K. Yadav, V. Sharma, A. Kafle, et al., "Spatial autocorrelation with environmental factors related to tuberculosis prevalence in Nepal, 2020–2023," *Infectious Diseases of Poverty*, vol. 14, 2025, doi: 10.1186/s40249-025-01283-y.

[13] N. Muna and W. H. Cahyati, "Determinan kejadian tuberkulosis pada orang dengan HIV/AIDS," *Higeia Journal of Public Health Research and Development*, 2019, doi: 10.15294/higeia/v3i2/24857.

[14] R. N. Hakim, "*Pengaruh Jumlah Kasus HIV/AIDS dan Cakupan Rumah Sehat terhadap Jumlah Kasus Tuberkulosis di Provinsi Jawa Timur"*, Skripsi, 2018.

[15] I. Pardoe, "*Applied Regression Modeling*, 3rd ed.", Hoboken, NJ: Wiley, 2021.

[16] D. N. Gujarati and D. C. Porter, "*Basic Econometrics*, 5th ed.", New York: McGraw-Hill, n.d.
[17] M. Al Haris and P. R. Arum, "Negative binomial regression and generalized Poisson regression models on the number of traffic accidents in Central Java," *Barekeng: Jurnal Ilmu Matematika dan Terapan*, vol. 16, pp. 471–482, 2022, doi: 10.30598/barekengvol16iss2pp471-482.
[18] N. Shrestha, "Detecting multicollinearity in regression analysis," *American Journal of Applied Mathematics and Statistics*, vol. 8, pp. 39–42, 2020, doi: 10.12691/ajams-8-2-1.
[19] R. A. Johnson and G. K. Bhattacharyya, "*Statistics: Principles and Methods*, 6th ed.", Hoboken, NJ: Wiley, n.d.
[20] S. Portet, "A primer on model selection using the Akaike Information Criterion," *Infectious Disease Modelling*, vol. 5, pp. 111–128, 2020, doi: 10.1016/j.idm.2019.12.010.
[21] M. O. Adenomon and G. S. Akinyemi, "Trend analysis of tuberculosis cases and the effect of HIV cases on tuberculosis cases in some West African countries using panel Poisson and negative binomial regression models," 2019, doi: 10.20944/preprints201910.0362.v1.