



Available online at <http://scik.org>

Commun. Math. Biol. Neurosci. 2026, 2026:99

<https://doi.org/10.28919/cmbn/10206>

ISSN: 2052-2541

GENERALIZED ESTIMATING EQUATIONS FOR LONGITUDINAL ZERO-TRUNCATED NEGATIVE BINOMIAL DATA: ANALYSIS OF TUBERCULOSIS CASES IN CENTRAL SULAWESI

KHAIRUL PURQON*, NIRWAN, ANNA ISLAMIYATI

Department of Statistics, Faculty of Mathematics and Natural Sciences, Hasanuddin University, Makassar, 90245,
Indonesia

Copyright © 2026 the author(s). This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract: This research employs a ZTNB-GEE framework to examine positive annual tuberculosis counts recorded across 13 districts and cities in Central Sulawesi between 2020 and 2024. The data comprise 65 district/city-year observations and six predictors representing population density, inadequate housing, poverty, drinking-water quality, access to adequate sanitation, and public facilities that meet environmental health standards. The deviance and Pearson dispersion ratios were 106.228 and 111.963, respectively, indicating substantial overdispersion. ZTNB-GEE models were fitted using independence, exchangeable, unstructured, and first-order autoregressive working correlations. The corresponding QIC values were 84.47478, 96.29610, 269.64114, and 95.01389, so the independence structure was selected. Under the preferred correlation specification, only and satisfied the 5% significance criterion, with estimated coefficients of 0.00085832 and 0.02060237, respectively. Both coefficient estimates were positive. With the other covariates held fixed, the fitted mean was multiplied by 1.0008587 for every one-person increase in population density and by 1.0208161 for every additional thousand people living below the poverty line. These ratios are equivalent to

*Corresponding author

E-mail address: khairulpurqon10@gmail.com

Received August 6, 2026

changes of approximately 0.08587% and 2.08161%, respectively.

Keywords: generalized estimating equations; zero-truncated negative binomial; longitudinal data; tuberculosis; QIC.

2020 AMS Subject Classification: 62J12.

1. INTRODUCTION

Infection by *Mycobacterium tuberculosis* gives rise to tuberculosis, a communicable disease that is most frequently manifested in the lungs but may also develop in extrapulmonary organs. Its distribution is not uniform across areas because population, socioeconomic, housing, and environmental conditions vary between districts and cities [1]. These differences support the use of area-level analysis that also follows changes over time.

Indonesia continues to carry a high tuberculosis burden. The health profile cited in this study reported an estimated 1,090,000 incident cases in Indonesia and 10,084 incident cases in Central Sulawesi in 2024. Several tuberculosis-control indicators in Central Sulawesi had not reached their national targets as of March 20, 2025, including case detection, treatment success, contact investigation, and tuberculosis preventive therapy [2]. This situation motivates an assessment of factors associated with variation in tuberculosis case counts across districts, cities, and years.

The explanatory variables in this study reflect population, housing, socioeconomic, and environmental-health characteristics. Population density has been reported to have a positive relationship with tuberculosis incidence [3]. Housing conditions, including ventilation, humidity, and lighting, are related to the potential for transmission and provide a basis for using the number of inadequate houses as an area-level indicator [4]. Poverty, drinking-water quality, sanitation access, and public facilities are included as contextual indicators of economic and environmental-health conditions; they are not interpreted as direct mechanisms of tuberculosis transmission [5].

The response analysed in this research consists of annual integer-valued case totals. A Poisson formulation would require the conditional first and second moments to have an identical value. The observed variation, however, was substantially greater than the level implied by this restriction. Therefore, a Negative Binomial formulation was adopted because its additional parameter allows

the variance to increase independently of the expected count [6].

The observations in this study are longitudinal because each district or city was measured repeatedly from 2020 through 2024. Measurements from different districts or cities may be treated as independent, while repeated measurements within the same district or city may be correlated. Generalized Estimating Equations (GEE) provide population-averaged estimates for repeated or panel data and incorporate within-unit association through a working correlation matrix [7].

Every tuberculosis count observed in this study was greater than zero, which means that the empirical support begins at one. A ZTNB response is appropriate under this condition because it combines positive-only support with a dispersion parameter capable of representing variability beyond the Poisson level [9]. The truncation procedure does not modify the recorded observations; instead, the probability originally assigned to zero is redistributed across the possible positive outcomes [10].

Previous studies have separately used Negative Binomial regression for tuberculosis data and GEE for longitudinal count responses. A Negative Binomial analysis of tuberculosis cases in West Java identified health-related population characteristics as significant predictors [11], while a GEE study of longitudinal flood frequencies selected the independence working correlation using the minimum QIC criterion [8]. However, the simultaneous use of GEE and a ZTNB response for longitudinal tuberculosis counts at the district/city level remains limited in the literature reviewed for this study. Therefore, this research applies ZTNB-GEE to Central Sulawesi tuberculosis data and compares independence, exchangeable, unstructured, and AR-1 working correlations using QIC.

2. PRELIMINARIES

Longitudinal Data

A data set has a longitudinal organization when observations from the same unit are collected at several measurement occasions. Such an arrangement preserves information about temporal changes within each unit as well as variation across different units [20].

Repeated measurements from one district or city are generally dependent because observations recorded at adjacent or different years may still share the same underlying regional conditions [21]. For district or city i , the response and marginal-mean vectors are written as follows [7].

$$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \dots, Y_{iT_i})^T \quad (1)$$

Equation (1) defines the repeated-response vector for district or city i . Its components are the tuberculosis counts observed at the successive measurement times.

$$\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \dots, \mu_{iT_i})^T \quad (2)$$

Equation (2) defines the corresponding marginal-mean vector. Each component represents the expected tuberculosis count for district or city i at time t .

The number of repeated observations is denoted by T_i . In this study, every district or city contributes five annual measurements, so the within-unit association must be considered when the regression parameters and their standard errors are estimated.

Generalized Linear Model

A Generalized Linear Model extends regression analysis to responses whose probability behaviour is not adequately represented by a normal distribution. The response distribution is selected from the exponential family by considering the measurement scale and empirical characteristics of the outcome [16].

Specification of a GLM requires a probability model for the response, a systematic component formed from the explanatory variables and coefficients, and a function that relates this systematic component to the expected value of the response [13].

$$g(\mu_{it}) = \eta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} \quad (3)$$

Equation (3) establishes the relationship between the marginal expectation and the systematic component of the model. Through $g(\cdot)$, the mean μ_{it} is represented on the scale of the linear predictor $\mathbf{x}_{it}^T \boldsymbol{\beta}$.

$$\mu_{it} = g^{-1}(\eta_{it}) \quad (4)$$

Equation (4) reverses the transformation introduced by the link function. The inverse function converts the linear predictor back to the scale used for the expected response.

Here, μ_{it} denotes the expected response, x_{it} is the vector of explanatory variables, β is the regression-parameter vector, and $g(\cdot)$ is the selected link function. For count responses, a logarithmic link is commonly used because its inverse always produces a positive fitted mean.

Overdispersion

Under the Poisson assumption, the amount of conditional variation is completely determined by the expected response because both quantities are required to coincide [18]. The data are regarded as overdispersed when their observed variation is greater than the variability implied by the fitted Poisson expectation [17].

If this excess variability is ignored, uncertainty surrounding the coefficient estimates may be understated, which can lead to overly strong conclusions regarding parameter significance [17]. The condition and the two dispersion diagnostics used in this study can be expressed as follows [12].

$$\text{Var}(Y) > E(Y) \quad (5)$$

Equation (5) presents the basic condition used to recognize overdispersion. Equality between the two quantities is consistent with Poisson variation, while a larger variance indicates that the observed response is more heterogeneous than the Poisson model permits.

$$D = 2 \sum_{j=1}^N \left[y_j \log \left(\frac{y_j}{\hat{\mu}_j} \right) - (y_j - \hat{\mu}_j) \right] \quad (6)$$

Equation (6) summarizes the overall lack of agreement between the recorded counts and the values predicted by the Poisson model. Each observation contributes according to the difference between y_j and $\hat{\mu}_j$, after which all contributions are accumulated.

$$\hat{\phi}_D = \frac{D}{N-p} \quad (7)$$

Equation (7) expresses the deviance relative to the number of residual degrees of freedom. A ratio near unity provides little evidence against the Poisson variance assumption, whereas a considerably larger ratio indicates excess variability.

$$X_P^2 = \sum_{j=1}^N \frac{(y_j - \hat{\mu}_j)^2}{v(\hat{\mu}_j)} \quad (8)$$

Equation (8) defines the Pearson chi-square statistic. For each observation, the squared difference between the observed count y_j and fitted mean μ_j is divided by the variance function $V(\mu_j)$, and the resulting contributions are summed over all N observations.

$$\hat{\phi}_P = \frac{X_P^2}{N-p} \quad (9)$$

Equation (9) scales the Pearson statistic by the available residual degrees of freedom. Values markedly greater than one provide additional evidence that the count response is overdispersed.

The direct mean–variance comparison in Equation (5), the deviance and its ratio in Equations (6) and (7), and the Pearson statistic and its ratio in Equations (8) and (9) jointly provide the basis for evaluating whether the Poisson variance assumption is adequate [12].

Negative Binomial Regression

In Negative Binomial regression, unexplained heterogeneity is represented through an additional parameter rather than forcing the response variation to follow the Poisson restriction [17]. This specification allows the conditional variance to increase more rapidly than the conditional expectation [6].

Using λ as the mean before truncation and θ as the dispersion parameter, the probability mass function is [6].

$$P(Y = y) = \frac{\Gamma(y+\theta)}{\Gamma(\theta)\Gamma(y+1)} \left(\frac{\theta}{\theta+\lambda}\right)^\theta \left(\frac{\lambda}{\theta+\lambda}\right)^y \quad (10)$$

Equation (10) gives the distribution assigned to the underlying response before conditioning on positive observations. The quantity λ governs its location, θ regulates extra-Poisson variation, and $\Gamma(\cdot)$ is the Gamma function. At this stage the support is $y = 0, 1, 2, \dots$ because the zero outcome has not yet been removed.

The first two moments of the distribution show how the dispersion parameter produces variability beyond the poisson assumption [6]

$$E(Y) = \lambda \quad (11)$$

Equation (11) identifies as the expectation of the underlying Negative Binomial variable. It therefore refers to the average count before conditioning the response on positive values.

$$\text{Var}(Y) = \lambda + \frac{\lambda^2}{\theta} \quad (12)$$

As indicated by Equation (12), the response variance consists of λ together with the quadratic contribution $\frac{\lambda^2}{\theta}$. The presence of this second component causes the variance to exceed the underlying mean.

These distributional properties explain why the Negative Binomial model is appropriate for count data that display variability beyond the Poisson assumption.

Negative Binomial regression generally uses a log link to connect the mean to the explanatory variables [18].

$$\log(\lambda_{it}) = \eta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} \quad (13)$$

Equation (13) specifies the logarithmic link for the base Negative Binomial mean. It connects $\log(\lambda_{it})$ to the linear predictor $\eta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta}$, allowing the effects of the explanatory variables to be represented on an additive scale.

$$\lambda_{it} = \exp(\mathbf{x}_{it}^T \boldsymbol{\beta}) \quad (14)$$

Equation (14) is the inverse of the logarithmic link. Exponentiating the linear predictor produces λ_{it} and guarantees that the fitted base mean is strictly positive.

The initial Negative Binomial component is estimated by maximum likelihood. In this research, that initial fit is used to obtain the dispersion estimate required by the subsequent ZTNB-GEE variance specification.

Quasi-Likelihood

The quasi-likelihood approach bases estimation on the first two moments of the response rather than on a fully specified probability model [19]. For each observation, the estimating contribution is formed from its residual and is weighted inversely according to the variance associated with the fitted mean [20].

$$\frac{\partial Q(y, \mu)}{\partial \mu} = \frac{y - \mu}{V(\mu)} \quad (15)$$

Equation (15) indicates that the residual contribution is adjusted according to the variance expected at a given mean. Observations with different levels of model-based variability therefore

do not contribute equally to the estimating process.

This formulation provides the basis for a quasi-score equation. In longitudinal analysis, the individual contributions are combined and weighted by a covariance matrix that represents the repeated measurements within each observational unit [20].

Generalized Estimating Equations

The GEE approach develops a marginal regression model for observations that may be dependent within the same unit. Its coefficients describe changes in the population-average response rather than unit-specific effects [7]. Parameter estimates are obtained by finding the solution of the estimating equation presented below.

$$U(\beta) = \sum_{i=1}^n D_i^T V_i^{-1} (Y_i - \mu_i) = 0 \quad (16)$$

Equation (16) combines the estimating contributions from all districts and cities. Within each unit, the difference between the observed and expected response vectors is adjusted using the inverse covariance matrix and the sensitivity matrix D_i . The estimated coefficient vector is obtained when the combined contribution across units reaches zero.

The matrix D_i is the derivative of μ_i with respect to β , and V_i is the covariance matrix used in the calculations. The robust covariance of the estimator has the sandwich form [14].

$$\widehat{\text{Var}}_{\text{rob}}(\hat{\beta}) = M^{-1} B M^{-1} \quad (17)$$

Equation (17) gives the robust or sandwich covariance estimator of β . The matrices M^{-1} on both sides form the bread of the estimator, while B summarizes the empirical variability of the estimating functions. Its diagonal elements are used to obtain robust standard errors.

Provided that the marginal mean is correctly specified, the robust estimator protects the inference against moderate misspecification of the working correlation structure [14].

Working Correlation Structures

Dependence among yearly observations from the same area is represented through a working correlation matrix. Different matrix forms can produce different estimation efficiencies, while the robust covariance estimator limits the effect of an incorrectly chosen working pattern [7].

Independence. Under independence, only the diagonal elements are nonzero; therefore, no

working association is assigned between two different measurement years [7].

$$\mathbf{R}_i(\alpha) = \mathbf{I}_{T_i} \quad (18)$$

Equation (18) represents the independence working correlation. The matrix is an identity matrix, so every off-diagonal element equals zero and the repeated observations are treated as uncorrelated in the working model.

Exchangeable. The exchangeable form uses one common value of for all pairs of annual observations belonging to a given area [7].

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \cdots & \alpha \\ \alpha & 1 & \cdots & \alpha \\ \vdots & \vdots & \ddots & \vdots \\ \alpha & \alpha & \cdots & 1 \end{pmatrix} \quad (19)$$

Equation (19) represents the exchangeable working correlation. All diagonal elements equal one, while every off-diagonal element has the same value α . Thus, each pair of measurements within one district or city is assumed to have an equal correlation.

Unstructured. In the unstructured form, correlations are estimated separately for the different pairs of measurement occasions. This avoids imposing a common pattern but requires a larger set of association parameters [7].

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha_{12} & \cdots & \alpha_{1T_i} \\ \alpha_{12} & 1 & \cdots & \alpha_{2T_i} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{1T_i} & \alpha_{2T_i} & \cdots & 1 \end{pmatrix} \quad (20)$$

Equation (20) represents the unstructured working correlation. Each pair of observation times has its own parameter α_{ts} , so the model does not impose a common or distance-based correlation pattern. This flexibility requires more correlation parameters to be estimated.

First-order autoregressive. The AR-1 structure assigns a correlation that decreases geometrically as the time lag between two measurements increases [7].

$$\mathbf{R}_i(\alpha) = \begin{pmatrix} 1 & \alpha & \alpha^2 & \cdots & \alpha^{T_i-1} \\ \alpha & 1 & \alpha & \cdots & \alpha^{T_i-2} \\ \alpha^2 & \alpha & 1 & \cdots & \alpha^{T_i-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \alpha^{T_i-1} & \alpha^{T_i-2} & \alpha^{T_i-3} & \cdots & 1 \end{pmatrix} \quad (21)$$

Equation (21) represents the first-order autoregressive, or AR-1, working correlation. The correlation between measurements at times t and s is α raised to the distance $|t - s|$. Therefore, observations recorded closer in time are assumed to have a stronger correlation.

Zero-Truncated Negative Binomial Distribution

The observed support is truncated at zero because the analysed sample contains no zero-valued tuberculosis count. Accordingly, the distribution is conditioned on a positive response, and the probability weights for $1, 2, \dots$ are recalculated to form a valid distribution [9]. Without this conditioning step, the fitted model would continue assigning probability to zero even though zero does not occur in the analysed data [10].

Let the latent untruncated response follow the Negative Binomial distribution in Equation (10). Its probability of zero is [9].

$$p_{0,it} = P(Z_{it} = 0) = \left(\frac{\theta}{\theta + \lambda_{it}} \right)^\theta \quad (22)$$

Equation (22) gives the probability that the latent Negative Binomial response equals zero. This probability, denoted by $p_{0,it}$, is removed when the distribution is conditioned on observing only positive tuberculosis counts.

After the zero outcome is excluded, the probability of each observed positive count is obtained by dividing its ordinary Negative Binomial probability by the probability of a positive response [9].

$$P(Y_{it} = y \mid Z_{it} > 0) = \frac{P(Z_{it}=y)}{1-p_{0,it}} \quad (23)$$

Equation (23) gives the Zero-Truncated Negative Binomial probability mass function. The ordinary Negative Binomial probability for a positive count is divided by $1 - p_{0,it}$ so that the remaining probabilities sum to one. The support of the observed response is $y = 1, 2, \dots$ because the zero category has been excluded..

Because the distribution is conditioned on a positive outcome, its observed marginal mean is not equal to the base mean λ_{it} . The conditional mean is [9].

$$\mu_{it} = E(Y_{it} \mid Z_{it} > 0) = \frac{\lambda_{it}}{1-p_{0,it}} \quad (24)$$

Equation (24) gives the conditional mean of the observed ZTNB response. Because zero is excluded, the base mean λ_{it} is divided by the probability of a positive response, $1 - p_{0,it}$. Hence, μ_{it} is the mean after truncation and differs from λ_{it} .

The corresponding conditional variance contains both the Negative Binomial dispersion and the adjustment caused by removing zero from the support [9].

$$\text{Var}(Y_{it} | Z_{it} > 0) = \frac{\lambda_{it} + \lambda_{it}^2(1 + \theta^{-1})}{1 - p_{0,it}} - \left(\frac{\lambda_{it}}{1 - p_{0,it}} \right)^2 \quad (25)$$

Equation (25) gives the conditional variance of the ZTNB response. The first part changes the second moment of the Negative Binomial to account for the chance of seeing a positive number, and the second part takes away the square of the average value given the conditions.

ZTNB-GEE Mean and Covariance Specification

In the parameterization used in this research, the logarithmic link is applied to the base Negative Binomial mean λ_{it} rather than directly to the conditional ZTNB mean μ_{it} [9]. For the six explanatory variables, the linear predictor is

$$\log(\lambda_{it}) = \beta_0 + \sum_{k=1}^6 \beta_k X_{k,it} \quad (26)$$

Equation (26) defines the linear predictor used in this study. The logarithm of the base mean is expressed as an intercept plus the contributions of the six explanatory variables measured for district or city i at time t .

After the inverse log link has produced λ_{it} , the zero-truncation correction gives the conditional mean [9].

$$\mu_{it} = \frac{\lambda_{it}}{1 - \left(\frac{\theta}{\theta + \lambda_{it}} \right)^\theta} \quad (27)$$

Equation (27) converts the base mean λ_{it} into the conditional ZTNB mean μ_{it} . The denominator is the probability that the latent count is positive; therefore, zero truncation is handled through a probability correction rather than by changing the recorded observations.

This two-stage specification is retained throughout estimation: the log link first produces λ_{it} , after which the zero-truncation adjustment produces μ_{it} [10].

Denote by A_i a diagonal matrix whose entries are the conditional ZTNB variances defined in

Equation (25). The covariance representation for the repeated responses is then written as

$$V_i = \phi A_i^{1/2} R_i(\alpha) A_i^{1/2} \quad (28)$$

Equation (28) combines marginal variability and within-area association in one matrix. The entries of A_i represent the ZTNB variances for the five annual measurements, while $R_i(\alpha)$ describes the selected correlation pattern. The parameter ϕ adjusts the overall magnitude of the covariance. The matrix $R_i(\alpha)$ is chosen from the four candidate structures, while ϕ is a scale parameter. Regression parameters are updated iteratively until the estimating equation reaches the specified convergence criterion.

Comparison of Candidate Correlation Structures

Since GEE parameter estimation does not rely on a fully specified likelihood, QIC is used to compare the proposed correlation assumptions [15]. All candidate models retain the same marginal mean formulation, so the correlation structure with the minimum QIC value is selected.

$$QIC = -2Q(\hat{\beta}; I) + 2\text{tr}(V_m^{-1} B V_e) \quad (29)$$

Equation (29) defines the QIC used to compare the candidate working correlations. The first term measures model fit through the quasi-likelihood under independence, while the trace term penalizes covariance complexity. The preferred structure is the one with the smallest QIC.

The QIC comparison is applied only after all candidate structures have been estimated using the same response distribution and mean specification [15].

Tuberculosis and the Study Context

Infection by *Mycobacterium tuberculosis* causes a communicable illness that predominantly affects the respiratory system. Nevertheless, clinical manifestations may also occur in organs other than the lungs [1]. Its continuing burden in Indonesia has encouraged statistical studies that examine differences in case occurrence across regions [22].

Individuals with pulmonary disease may release infectious particles into the air, creating a greater opportunity for exposure where many people occupy a limited space [23]. At the regional level, population density can therefore serve as an indicator of the intensity of potential contact [3].

Inadequate ventilation, humidity, and limited natural lighting may create housing conditions

that are less supportive of tuberculosis prevention [4]. Poverty, drinking-water quality, sanitation access, and environmentally compliant public facilities are treated in this study as contextual measures of socioeconomic and environmental-health conditions rather than as direct biological causes of infection [5].

Individual susceptibility, household exposure, preventive behaviour, and adherence to treatment also influence the chance that infection or transmission will occur within a family setting [24].

Research Method

Annual records supplied by the Central Sulawesi Provincial Health Office were organized as a balanced longitudinal data set covering 13 districts and cities over the 2020–2024 period [2]. Five measurements for each area resulted in 65 district/city-year observations. The response is the positive number of recorded tuberculosis cases, and six predictors describe population, housing, poverty, drinking-water quality, sanitation, and environmental-health conditions of public facilities.

Table 1. Variables and measurement units

Model function	Variable	Symbol	Definition used in this study	Unit
Response	Tuberculosis cases	Y	Annual number of tuberculosis cases included in the analysis	Cases
Predictor	Population density	X ₁	Population per square kilometre	People/km ²
Predictor	Inadequate houses	X ₂	Houses not meeting safety, health, and minimum-space criteria	Units
Predictor	Population living in poverty	X ₃	Population with expenditure below the poverty line	Thousand people
Predictor	Drinking-water quality	X ₄	Drinking water meeting appropriate and safe quality standards	%
Predictor	Access to adequate sanitation	X ₅	Households with access to adequate sanitation facilities	%
Predictor	Public facilities	X ₆	Public facilities meeting environmental-health requirements	%

The analysis began with a descriptive examination of the variables and an assessment of extra-Poisson variation. An initial ZTNB model was then used to obtain the dispersion estimate. Using

this estimate, four ZTNB-GEE specifications were fitted under independence, exchangeable, unstructured, and AR-1 working correlations. The final structure was determined from the smallest QIC value.

3. MAIN RESULTS

Descriptive Statistics and Overdispersion

The tuberculosis response ranged from 128 to 1,847 cases, with a mean of 488.82, a median of 408.00, and a standard deviation of 354.52. The mean exceeding the median indicates the presence of several relatively high tuberculosis counts. Among the explanatory variables, X_2 displayed the widest absolute spread. In contrast, X_5 had the lowest variability around its sample mean, with a standard deviation of 9.73.

Table 2. Numerical summary of the study variables

Variable	Minimum	Maximum	Mean	Median	Standard deviation
Y	128.00	1847.00	488.82	408.00	354.52
X_1	12.00	1101.00	126.39	50.00	265.05
X_2	447.00	56720.00	11269.71	8842.00	8992.21
X_3	10.32	78.76	30.26	26.75	16.98
X_4	13.73	100.00	82.52	89.61	19.79
X_5	56.40	99.27	74.61	72.71	9.73
X_6	4.11	100.00	56.01	60.19	25.86

The Poisson-based deviance was 6,161.2 and the Pearson chi-square statistic was 6,493.854. With 58 residual degrees of freedom, the dispersion ratios were 106.228 and 111.963, respectively. Both ratios were much greater than one, confirming that the tuberculosis counts were overdispersed and supporting the use of a Negative Binomial response [13].

Initial Zero-Truncated Negative Binomial Model

The initial ZTNB model was fitted before the longitudinal correlation was introduced. Population density (X_1), the population living in poverty (X_3), drinking-water quality (X_4), and access to adequate sanitation (X_5) were significant at the 5% level. The number of inadequate houses (X_2) and public facilities (X_6) were not significant. The estimated dispersion parameter was

$\theta = 4.791893$ and was subsequently used in the ZTNB-GEE variance specification.

Table 3. Estimates obtained from the preliminary ZTNB fit

Parameter	Estimate	Std. error	z-value	p-value	Decision
β_0	3.59614102	0.66195615	5.4326	0.00000005	Significant
β_1	0.00085832	0.00025770	3.3308	0.00086610	Significant
β_2	-0.00001159	0.00000823	-1.4088	0.15889584	Not significant
β_3	0.02060243	0.00434234	4.7445	0.00000208	Significant
β_4	0.00618187	0.00300154	2.0596	0.03944020	Significant
β_5	0.01890018	0.00704844	2.6815	0.00732991	Significant
β_6	-0.00060673	0.00255130	-0.2378	0.81202761	Not significant

ZTNB-GEE Modeling under Alternative Working Correlations

The longitudinal ZTNB-GEE model was fitted under four candidate working correlations. The same mean specification and six predictors were retained so that the QIC comparison reflected differences in the working correlation structure.

Independence Structure

Table 4. ZTNB-GEE estimates under the independence structure

Parameter	Estimate	Robust SE	p-value	Decision
β_0	3.59614874	0.86787521	0.00003419	Significant
β_1	0.00085832	0.00022212	0.00011147	Significant
β_2	-0.00001159	0.00000661	0.07972640	Not significant
β_3	0.02060237	0.00375789	0.00000004	Significant
β_4	0.00618181	0.00374020	0.09837141	Not significant
β_5	0.01890015	0.01054256	0.07301320	Not significant
β_6	-0.00060677	0.00301656	0.84058480	Not significant

Under independence, population density (X_1) and the population living in poverty (X_3) were significant, with p-values of 0.00011147 and 0.00000004. Their coefficients were positive. The remaining four predictors had p-values greater than 0.05.

Exchangeable Structure

Table 5. ZTNB-GEE estimates under the exchangeable structure

Parameter	Estimate	Robust SE	p-value	Decision
β_0	4.77080698	0.97409223	0.00000097	Significant
β_1	0.00100271	0.00017913	0.00000002	Significant
β_2	-0.00001667	0.00000195	0.00000000	Significant
β_3	0.01603413	0.00516113	0.00189181	Significant
β_4	-0.00022119	0.00233037	0.92438019	Not significant
β_5	0.01096173	0.01187638	0.35601444	Not significant
β_6	0.00186914	0.00231311	0.41905306	Not significant

Under the exchangeable structure, X_1 , X_2 , and X_3 were statistically significant. The coefficients of X_1 and X_3 were positive, whereas the coefficient of X_2 was negative. X_4 , X_5 , and X_6 were not significant at the 5% level.

Unstructured Correlation

Table 6. ZTNB-GEE estimates under the unstructured correlation

Parameter	Estimate	Robust SE	p-value	Decision
β_0	-0.10414092	3.91186725	0.97876139	Not significant
β_1	0.00049151	0.00052778	0.35171396	Not significant
β_2	-0.00001732	0.00000890	0.05163697	Not significant
β_3	0.02831782	0.00781048	0.00028827	Significant
β_4	0.00410479	0.00326998	0.20937111	Not significant
β_5	0.07228420	0.05349844	0.17664869	Not significant
β_6	-0.00334122	0.00315884	0.29017615	Not significant

The unstructured model identified only X_3 as significant ($p = 0.00028827$). The p-value for X_2 was 0.05163697, which was close to but still above the 0.05 significance threshold. All other predictors were not significant.

First-Order Autoregressive Structure

Table 7. ZTNB-GEE estimates under the AR-1 structure

Parameter	Estimate	Robust SE	p-value	Decision
β_0	5.57242400	0.68090537	0.00000000	Significant
β_1	0.00117546	0.00018147	0.00000000	Significant
β_2	-0.00000619	0.00000146	0.00002235	Significant
β_3	0.01302055	0.00449661	0.00378392	Significant
β_4	-0.00060258	0.00197013	0.75971147	Not significant
β_5	0.00096969	0.00882607	0.91251479	Not significant
β_6	0.00046178	0.00201692	0.81890582	Not significant

The AR-1 structure also identified X_1 , X_2 , and X_3 as significant. X_1 and X_3 had positive coefficients, while X_2 had a negative coefficient. X_4 , X_5 , and X_6 were not statistically significant.

Selection and Interpretation of the Final Model

Table 8. QIC values for the candidate working correlations

Working correlation	QIC
Independence	84.47478
Exchangeable	96.29610
Unstructured	269.64114
Autoregressive AR-1	95.01389

The independence structure produced the smallest QIC, 84.47478. It was therefore selected as the best working correlation among the four candidates. The selected structure does not change the longitudinal nature of the data; rather, it indicates that the independence working correlation provided the most favorable QIC for the specified marginal model.

Using the estimates in Table 4, the final model for the base Negative Binomial mean is

$$\begin{aligned} \log(\lambda_{it}) = & 3,59614874 + 0,00085832X_{1,it} - 0,00001159X_{2,it} \\ & + 0,02060237X_{3,it} + 0,00618181X_{4,it} \\ & + 0,01890015X_{5,it} - 0,00060677X_{6,it}. \end{aligned}$$

The exponentiated coefficient of X_1 is 1.0008587. Therefore, when the other explanatory variables remain unchanged, each additional person per square kilometre multiplies the fitted tuberculosis mean by 1.0008587. This ratio is equivalent to a relative change of approximately

0.08587%. Exponentiation of the X_3 estimate produces a mean ratio of 1.0208161. With all other predictors fixed, adding one thousand people below the poverty line multiplies the fitted mean by this value, corresponding to a relative change of approximately 2.08161%.

Densely populated areas may provide more opportunities for close contact [3]. Poverty may reflect socioeconomic conditions related to living environments and access to health services [5]. These findings represent population-averaged associations from an ecological longitudinal model and should not be interpreted as individual-level causal effects. The number of inadequate houses, drinking-water quality, access to adequate sanitation, and public facilities did not show statistically significant associations in the selected model.

4. CONCLUSION

This research analysed positive and highly dispersed tuberculosis counts reported annually for 13 districts and cities in Central Sulawesi. The ZTNB-GEE model was evaluated under independence, exchangeable, unstructured, and AR-1 correlation assumptions. Among these alternatives, independence produced the lowest QIC and was retained for final interpretation.

The selected model provides evidence that X_1 and X_3 are related to higher fitted tuberculosis counts. The estimated effects of X_2 , X_4 , X_5 , and X_6 did not satisfy the 5% significance criterion. These findings describe average relationships across districts and cities and do not establish causal effects at the individual level.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

- [1] S.S. Rajak, S. Ismail, R. Resmawan, Metode Conditional Autoregressive dalam Analisis Penyebaran Kasus Penyakit Tuberculosis, *Jambura J. Probab. Stat.* 2 (2021), 28-34.
<https://doi.org/10.34312/jjps.v2i1.9771>.
- [2] Dinas Kesehatan Provinsi Sulawesi Tengah, Profil Kesehatan Provinsi Sulawesi Tengah 2024, Dinas Kesehatan Provinsi Sulawesi Tengah, 2024.

- [3] D.N.A. Ningrum, A.R. Trisnaningsih, Tuberculosis Risk Mapping in Yogyakarta, Indonesia: an Ecological Study, *Unnes J. Public Health* 12 (2023), 1-13.
<https://journal.unnes.ac.id/sju/index.php/ujph/article/view/65193>.
- [4] A. Maharani, A. Siwiendrayanti, Analisis Potensi Penularan Tuberkulosis Paru di RW 10 Desa Kedung Waringin, Kecamatan Bojonggede, Kabupaten Bogor, *J. Kesehat. Masy.* 11 (2023), 481-491.
<https://doi.org/10.14710/jkm.v11i5.37692>.
- [5] A. Nurjanah, F.Y. Rahmalia, H.R. Paramesti, L.A. Laily, F.K. Pradani PH, et al., Determinan Sosial Tuberculosis di Indonesia, *J. Penelit. Pengemb. Kesehat. Masy. Indones.* 3 (2022), 71-82.
- [6] L. Rivas, M. Galea, On Estimation and Influence Measures for the Negative Binomial Regression Model Based on Q-Function, *Commun. Stat. Theory Methods* 51 (2022), 1954-1974.
<https://doi.org/10.1080/03610926.2021.1942493>.
- [7] S. Musthofa, L.H. Hasibuan, D.M. Putri, M. Jannah, I.D. Rianjaya, Analisis Data Longitudinal dengan Respon Biner Menggunakan Generalized Estimating Equation (GEE), *Math. Appl. J.* 5 (2023), 186-196.
<https://doi.org/10.15548/map.v5i2.7416>.
- [8] R. Kara, A. Yeşilova, Zero Truncated Models in Regression Analysis: an Examination of Their Advantages on Small Mean Values, *Yüzüncü Yıl Üniv. Fen Bilim. Enst. Derg.* 30 (2025), 102-112.
<https://doi.org/10.53433/yyufbed.1590611>.
- [9] L.C. Dennett, A. Overstall, D. Böhning, Zero-Truncated Modelling in a Meta-Analysis on Suicide Data after Bariatric Surgery, *Am. Stat.* 79 (2025), 492-499. <https://doi.org/10.1080/00031305.2025.2507380>.
- [10] I.I. Brilliant, D. Fakhriyana, Modeling the Number of Tuberculosis Cases in West Java Using the Negative Binomial Approach, *Consil. Sanit. J. Health Sci. Policy* 1 (2023), 107-119.
<https://doi.org/10.56855/jhsp.v1i2.282>.
- [11] M. Fadil, R. Raupong, N. Ilyas, Mengatasi Overdispersi Menggunakan Regresi Binomial Negatif dengan Penaksir Maksimum Likelihood pada Kasus Demam Berdarah di Kota Makassar, *Estim. J. Stat. Its Appl.* 5 (2024), 37-51. <https://doi.org/10.20956/ejsa.v5i1.14552>.
- [12] C.T. Fatmala, M. Hayati, R. Permatasari, M. Hudori, D.Y. Dalimunthe, Pemodelan Jumlah Kasus HIV/AIDS di Provinsi Lampung Menggunakan Regresi Binomial Negatif, *J. Math.: Theory Appl.* 6 (2024), 168-177. <https://doi.org/10.31605/jomta.v6i2.4069>.

- [13] A. Setiawan, Regresi Poisson Menggunakan Generalized Estimating Equation (Case Study: Flood Occurance Longitudinal Data in East Java Province Year 2011-2013), Masters Thesis, Institut Teknologi Sepuluh Nopember, Surabaya, 2017. <https://repository.its.ac.id/3904/>.
- [14] Y. Wilandari, S.H. Kartiko, A.R. Effendie, Estimasi Cadangan Klaim Menggunakan Generalized Linear Model (GLM) dan Copula, *J. Gaussian* 9 (2020), 411-420.
<https://doi.org/10.14710/j.gauss.v9i4.29260>.
- [15] H.M. Winata, Mengatasi Overdispersi dengan Regresi Binomial Negatif pada Angka Kematian Ibu di Kota Bandung, *J. Gaussian* 11 (2023), 616-622. <https://doi.org/10.14710/j.gauss.11.4.616-622>.
- [16] N.M. Dhahari, A. Sofro, Analisis Regresi Binomial Negatif untuk Pemodelan Angka Positif Penyakit Kusta di Jawa Timur, *MATHunesa J. Ilm. Mat.* 12 (2024), 49-56.
<https://doi.org/10.26740/mathunesa.v12n1.p49-56>.
- [17] M. Rezapour, S.J. Walker, D.A. Ornelles, P.M. McNutt, A. Atala, et al., Analysis of Gene Expression Dynamics and Differential Expression in Viral Infections Using Generalized Linear Models and Quasi-Likelihood Methods, *Front. Microbiol.* 15 (2024), 1342328.
<https://doi.org/10.3389/fmicb.2024.1342328>.
- [18] B. Tantular, D.Y. Faidah, F. Indrayatna, Quasi Likelihood on Linear Mixed Effect of Binary Response in Longitudinal Data, *Commun. Math. Biol. Neurosci.* 2025 (2025), 19.
<https://doi.org/10.28919/cmbn/9025>.
- [19] A. Fernandes, L.R. Rahmanda, Solimun, Analisis Model Jalur Longitudinal Berbasis GLMM pada Kasus Pasien Penderita Tuberkulosis Paru: Studi Simulasi, *J. Teknol. Inf. Ilmu Komput.* 12 (2025), 625-632. <https://doi.org/10.25126/jtiik.2025128415>.
- [20] R.B. Arie Salim, G. Anuraga, Pemodelan Kejadian Penyakit Tuberkulosis di Provinsi Jawa Barat Tahun 2023 Menggunakan Metode Geographically Weighted Negative Binomial Regression (GWNBR), *Mandalika Math. Educ. J.* 7 (2025), 891-902.
- [21] A.W. Muktiwijaya, J. Nugraha, The Negative Binomial Regression Approach Is More Appropriate for Identifying Tuberculosis Risk Factors (Data Analysis of Madiun and Ponorogo Districts), *J. Ris. Kesehat. Poltekkes Depkes Bandung* 17 (2025), 687-699.
<https://doi.org/10.34011/juriskesbdg.v17i2.2878>.

- [22] Wulandari, M.F. Abadi, N.W.D. Bintari, Hasil Pemeriksaan Mikroskopis BTA pada Sputum Anggota Keluarga yang Tinggal Serumah dengan Pasien BTA Positif di Wilayah Kerja UPT Puskesmas Gadung Sulawesi Tengah, Tentiro: J. Pendidik. Kesehat. 2 (2025), 40-55.
<https://doi.org/10.70420/tentiro.v2i1.151>.