JHSP
CONSILIUM SANITATIS
JOURNAL OF HEALTH SCIENCE AND POLICY
Vol: 1, No. 2, April 2023, pp. 107 - 119
P-ISSN: 2985-9719 I E-ISSN: 2985-4628
Web: https://journals.eduped.org/index.php/JHSP

Modeling The Number of *Tuberculosis* Cases in West Java using The Negative Binomial Approach

Pemodelan Jumlah Kasus Tuberculosis di Provinsi Jawa Barat Menggunakan Pendekatan Binomial Negatif

Indira Ihnu Brilliant[1*], Deby Fakhriyana[2]

[1] Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
[2]Universitas Diponegoro, Semarang, Indonesia

*e-mail: indiraihnubrilliant@uny.ac.id

Abstract

Objective: This study aims to model the number of *Tuberculosis* cases in West Java Province in 2021 using the Negative Binomial Regression approach.

Methods: This study used quantitative analysis uses secondary data from the Central Bureau of Statistics website and the Health Office of West Java Province. 27 West Java districts/cities were studied. The number of *Tuberculosis* cases was assumed to be affected by population density, poverty, sanitation, and health complaints in the past month. Negative Binomial Regression was used to analyse data.

Results: The results showed that Poisson Regression caused overdispersion, which was solved using the Negative Binomial Regression approach. The Negative Binomial Regression model passed a detailed test. The partial test showed that only the variable percentage of low-income persons and the variable percentage of people with health concerns significantly affected the model with regression coefficients of 0,8755 and 1,0318, respectively. The final Negative Binomial Regression model with the lowest Akaike Information Criterion value of 491,9 is best for this investigation.

Conclusion: The most suitable model for modelling the number of *Tuberculosis* cases in West Java Province in 2021 is the Negative Binomial Regression model with independent variables that significantly influence the model, namely the percentage of poor people and the percentage of people who have had complaints recently.

Keywords: *Tuberculosis*, modeling, negative binomial.

Article History	Submitted	Revised	Accepted
	2023-04-07	2023-04-08	2023-04-09

Vol. 1, No. 2, April 2023, pp. 107 – 119
https://doi.org/10.56855/jhsp.v1i2.282

Introduction

Tuberculosis (TB) is the world's top infectious killer. Nearly 4,500 people die, and 30,000 falls ill from this disease daily. *Tuberculosis* is caused by bacteria (*Mycobacterium tuberculosis*) and most often attacks the lungs. *Tuberculosis* spreads through the air when people with pulmonary *Tuberculosis* cough, sneeze or spit. A person only needs to inhale a few germs to become infected. Most people who get TB live in low- and middle-income countries, but *Tuberculosis* is worldwide[1].

Based on the 2021 Indonesia Health Profile, Indonesia is ranked third with the highest number of *Tuberculosis* sufferers worldwide after India and China. In 2021, the most Tuberculosis cases were found in West Java Province, with 91,268 cases, followed by Central Java and East Java[2]. Therefore, the Ministry of Health is conducting a large-scale screening to be carried out starting in 2022 to find and treat these cases, thereby minimizing transmission and reducing the incidence of *Tuberculosis*. One of the steps taken by the Indonesian government to prevent *Tuberculosis* is to provide mandatory immunization with the BCG vaccine (*Bacillus Calmette-Guerin*), given before the baby is two months old[3].

The number of *Tuberculosis* cases that occur in a particular area is one of the events that follow the distribution of the Poisson distribution. A Poisson experiment in TB cases has characteristics including the number of TB incidents that occur in a certain area. Then the probability of TB occurrence in a certain area does not depend on the number of TB incidents outside the area[4]. Therefore, a Poisson regression analysis will be approached to analyze the factors that are thought to influence the number of *Tuberculosis* cases in West Java.

The model formed from Poisson regression is a non-linear regression model derived from the Poisson distribution, which is usually used to analyze data with the dependent variable as a discrete variable. A feature of the Poisson distribution is equidispersion, a condition where the mean and variance of the dependent variable are the same. However, in reality, conditions are often found in the field where the variance value of the data is greater than the mean value (overdispersion occurs). One method that can be used to overcome overdispersion is negative binomial regression[5]. In addition, there is also a possibility that underdispersion will occur, namely, a situation where the

Vol. 1, No. 2, April 2023, pp. 107 – 119
https://doi.org/10.56855/jhsp.v1i2.282

variance value is smaller than the mean[6]. One approach that can analyze when the underdispersion phenomenon occurs is Conway-Maxwell-Poisson (COM-Poisson)[7].

The factors that have been analyzed in several previous studies related to the incidence of *Tuberculosis* are divided into two types, namely host factors and social and environmental factors. Host factors include genetics, gender, immunity, smoking, malnutrition, HIV co-infection or other immunosuppressive diseases such as asthma and diabetes[8–11]. Social and environmental factors include close physical contact, density, indoor pollution, housing conditions, lifestyle, education, and socio-economic status[12–14]. Based on the description above, this study aims to model the number of *Tuberculosis* incidents in West Java Province in 2021 using the Negative Binomial approach.

Methods

This research method is included in the quantitative research category using secondary data obtained from the publication of the Central Bureau of Statistics of West Java Province and the West Java Provincial Health Office in 2021. The units of observation or subjects in this study are all districts/cities in West Java Province, consisting of 18 regencies and nine cities. At the same time, the object of this research is the number of *Tuberculosis* incidents in West Java Province in 2021. Then, the variables used in this study consist of response variables (dependent) and predictor variables (independent) with a more detailed explanation as follows:

Table 1. Research Variables and its Descriptions

Variable Status	Variable Symbol	Variable	Unit	Source
Response (Y)	TBC	Number of *Tuberculosis* incidents	Person	West Java Provincial Health Office
Predictor (X_1)	PoPP	Percentage of Poor Population	%	Central Bureau of Statistics of West Java Province
Predictor (X_2)	PD	Population Density	Life/km2	Central Bureau of Statistics of West Java Province

Predictor (X_3)	PH2APS	Percentage of Households Having Access to Proper Sanitation	%	Central Bureau of Statistics of West Java Province
Predictor (X_4)	P2H2CLM	Percentage of Population Having Health Complaints in the Last Month	%	Central Bureau of Statistics of West Java Province

Analysis Method

The analytical method used to model the number of *Tuberculosis* incidents in West Java in 2021 is to use the Negative Binomial regression approach. The steps taken to answer the objectives of this study are as follows:

1. Descriptive analysis
2. Multicollinearity testing
3. Determine the estimated parameters of the Poisson Regression model
4. Testing the equidispersion assumption
5. Determine the estimated parameters of the Negative Binomial Regression model.
6. Overall testing of the Negative Binomial Regression model
7. Partial testing of the Negative Binomial Regression model
8. Determination of the best model with Akaike Information Criterion.

Results

In this section, the research results will be presented, starting from descriptive statistics to testing hypotheses to obtain an appropriate model. The first stage to start the analysis is to present descriptive statistics. In this study, researchers used data on the

Vol. 1, No. 2, April 2023, pp. 107 – 119
https://doi.org/10.56855/jhsp.v1i2.282

number of cases or incidence of *Tuberculosis* in West Java Province, Indonesia, in 2021. This is presented in Figure 1. Then the correlation graph is presented in Figure 2.

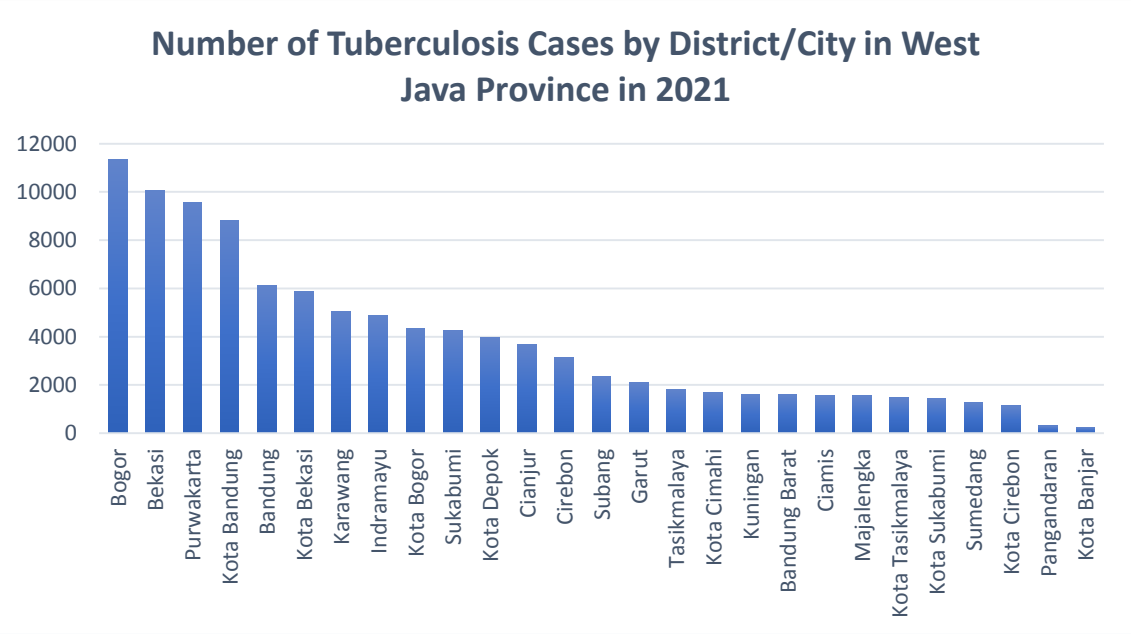

Figure 1. Graph of the number of *Tuberculosis* cases in West Java Province in 2021

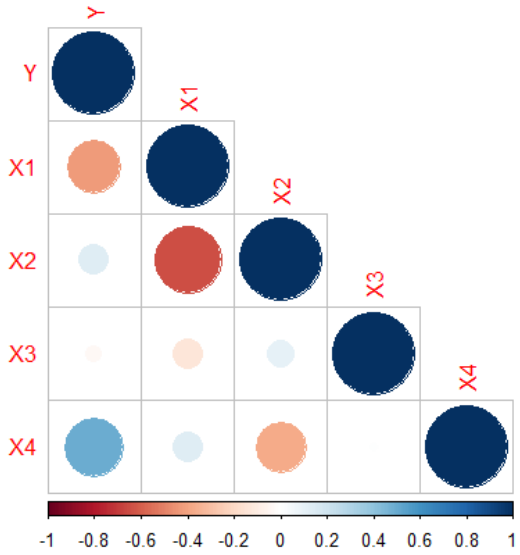

Figure 2. Graph of Correlation between variables

The next step is to conduct a test of multicollinearity to determine whether or not a model has been established and whether or not there is a perfect link between the variables that are independent. Table 2 displays the findings obtained from doing the test on multicollinearity.

Table 2. Multicollinearity Test Results

Variable	VIF
X_1 (PoPP)	1,7756
X_2 (PD)	2,0245
X_3 (PH2APS)	1,0213
X_4 (P2H2CLM)	1,1974

After going through multicollinearity testing, then looking for parameter estimates to form the initial model of Poisson Regression. Parameter estimates for the Poisson Regression model are presented in Table 3.

Table 3. Parameter Estimation Results of the Poisson Regression Model

Variable	Est. Coefficient	Std. Error	z-value	p-value
Intercept	$9{,}166 \times 10^{1}$	$2{,}639 \times 10^{-2}$	347,368	0,0000*
X_1 (PoPP)	$-1{,}461 \times 10^{-1}$	$1{,}596 \times 10^{-3}$	-91,565	0,0000*
X_2 (PD)	$-3{,}312 \times 10^{-6}$	$9{,}730 \times 10^{-7}$	-3,404	0,0006*
X_3 (PH2APS)	$-5{,}665 \times 10^{-3}$	$2{,}083 \times 10^{-4}$	-27,202	0,0000*
X_4 (P2H2CLM)	$2{,}700 \times 10^{-2}$	$2{,}208 \times 10^{-4}$	122,272	0,0000*

*significant at α = 5%
AIC = 33.901

After obtaining the parameter estimates for the Poisson Regression model, then the equidispersion assumption is tested to determine whether there is overdispersion in the model. The null hypothesis of the equidispersion test is that there is no overdispersion in the Poisson Regression, while the alternative hypothesis is that there is overdispersion in the Poisson Regression. The significance level used is 5% with the ϕ test statistic, and the rejection criterion is to reject H_0 when $\phi > 1$ [15]. Based on the results of statistical calculations of the ϕ test, a result of 1,510.33 was obtained, so a decision was made that the null hypothesis was rejected (overdispersion occurred). Then estimate the parameters for the Negative Binomial Regression model shown in Table 4.

Vol. 1, No. 2, April 2023, pp. 107 – 119
https://doi.org/10.56855/jhsp.v1i2.282

Table 4. Parameter Estimation Results of the Negative Binomial Regression Model

Variable	Est. Coefficient	Std. Error	z-value	p-value
Intercept	8,747 x 10^{1}	1,017 x 10^{1}	8,601	0,0000*
X_1 (PoPP)	-1,341 x 10^{-1}	6,053 x 10^{-2}	-2,216	0,0267*
X_2 (PD)	-1,819 x 10^{-6}	4,079 x 10^{-5}	-0,045	0,9644
X_3 (PH2APS)	-3,019 x 10^{-3}	8,200 x 10^{-3}	-0,368	0,7127
X_4 (P2H2CLM)	3,153 x 10^{-2}	1,324 x 10^{-2}	2,382	0,0172*

*significant at α = 5%
AIC = 495,77

After obtaining the parameter estimates of the Negative Binomial Regression model, proceed to test the suitability of the Negative Binomial Regression model, an overall test is carried out systematically as follows[15]:

1. Hypothesis

   $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ (Negative Binomial Regression model cannot be used as a model).

   H_1 : there is at least one $\beta_j \neq 0,\ j = 1,2,3,4$ (Negative Binomial Regression models can be used as models).

2. Significance Level

   $\alpha = 0{,}05$

3. Statistics Test

$$D_{hit} = 2\sum_{i=1}^{n}\left\{y_i\, ln\left(\frac{y_i}{\mu_i}\right) - \left(\frac{1}{k} + y_i\right) ln\left(\frac{1 + ky_i}{1 + k\mu_i}\right)\right\}$$

4. Rejection Criteria

   Reject H_0 if $D_{hit} > \chi^2_{\alpha;p}$

5. Decision

   Because D_{hit} (28,98447) > $\chi^2_{(0,05;4)}$ (9,488), then H_0 is rejected.

6. Conclusion

   By using a 95% confidence level, the existing data does not support the null hypothesis, or the existing data indicates that the Negative Binomial Regression model can be used as a model.

Vol. 1, No. 2, April 2023, pp. 107 – 119
https://doi.org/10.56855/jhsp.v1i2.282

In the overall test, it is concluded that the null hypothesis is rejected, or it can also be said that there is at least one $\beta_j \neq 0$ with $j = 1,2,3,4$. To find out which β_j is not equal to zero, a partial (individual) test for the formed Negative Binomial Regression model can be carried out.

1. Partial Test for X_1

   Systematically, hypothesis testing can be written as follows:

   a. Hyphotesis

      $H_0 : \beta_1 = 0$ (the regression coefficient X_1 has no significant effect)

      $H_1 : \beta_1 \neq 0$ (the regression coefficient X_1 has a significant effect)

   b. Significance Level

      $\alpha = 0{,}05$

   c. Statistics Test

      $$Z_{hit} = \frac{\hat{\beta}_1}{\sqrt{Var(\hat{\beta}_1)}}$$

   d. Rejection Criteria

      Reject H_0 if $|Z_{hit}| \geq Z_{\alpha/2}$ or p-*value* < α

   e. Decision

      Because $|Z_{hit}|$ (2,216) > $Z_{0{,}05/2}$ (1,96) or p-*value* (0,0267) < α (0,05), then H_0 is rejected

   f. Conclusion

      By using a 95% confidence level, the existing data does not support the null hypothesis or the existing data states that the regression coefficient X_1 has a significant effect.

2. Partial Test for X_2

   Systematic hypothesis testing for X_2 can also be done in the same way as the partial test for X_1. However, the researcher briefly describes the partial tests X_2, X_3, and X_4. In the partial test for X_2, the null hypothesis is that the regression coefficient X_2 does not have a significant effect ($\beta_2 = 0$), In contrast, the alternative hypothesis is the regression coefficient X_2 has a significant effect ($\beta_2 \neq 0$). The significance level used is 5% ($\alpha = 0{,}05$). Then the test statistic used is the Z test with the calculated Z value of -0.045 with a p-value of 0.9644. Then the decision criteria reject H_0 if

$|Z_{hit}| \geq Z_{\alpha/2}$ or p-*value* < α, so that the decision H_0 fails to be rejected because $|Z_{hit}|$ (0,045) < $Z_{0,05/2}$ (1,96) or p-*value* (0,9644) > α (0,05). Therefore, it can be concluded that by using a 95% confidence level, the existing data support the null hypothesis, or the existing data shows that the regression coefficient X_2 has no significant effect.

3. Partial Test for X_3

   In the partial test for X_3, the null hypothesis is that the regression coefficient X_3 has no significant effect ($\beta_3 = 0$), while the alternative hypothesis is that the regression coefficient X_3 has a significant effect ($\beta_3 \neq 0$). The significance level used is 5% ($\alpha = 0{,}05$). Then the test statistic used is the Z test with the calculated Z value of -0.368 with a p-value of 0.7127. Then the decision criteria reject H_0 if $|Z_{hit}| \geq Z_{\alpha/2}$ or p-value < α, o that the decision H_0 fails to be rejected because $|Z_{hit}|$ (0,368) < $Z_{0,05/2}$ (1,96) or p-value (0,7127) > α (0,05). Therefore, it can be concluded that by using a 95% confidence level, the existing data support the null hypothesis, or the existing data shows that the regression coefficient X_3 has no significant effect.

4. Partial Test for X_4

   In the partial test for X_4, the null hypothesis is that the regression coefficient X_4 has no significant effect ($\beta_4 = 0$), while the alternative hypothesis is that the regression coefficient X_4 has a significant effect ($\beta_4 \neq 0$). The significance level used is 5% ($\alpha = 0{,}05$). Then the test statistic used is the Z test with the calculated Z value of 2.382 with a p-value of 0.0172. Then the decision criteria reject H_0 if $|Z_{hit}| \geq Z_{\alpha/2}$ or p-value < α, so that a decision H_0 is rejected because $|Z_{hit}|$ (2,382) > $Z_{0,05/2}$ (1,96) or p-value (0,0172) < α (0,05). Therefore, it can be concluded that by using a 95% confidence level, the existing data does not support the null hypothesis, or the existing data shows that the regression coefficient X_4 has a significant effect.

After conducting partial tests for all independent variables, the estimated parameters that are already significant are obtained in Table 5.

Table 5. Parameter Estimation Results of Negative (Significant) Binomial Regression Model

Variable	Est. Coefficient	Std. Error	z-value	p-value
Intercept	8,51727	0,49678	17,145	0,0000*

X_1 (PoPP)	-0,13292	0,04592	-2,895	0,0038*
X_4 (P2H2CLM)	0,03128	0,01223	2,558	0,0105*

*significant at α = 5%
AIC = 491,9

Discussion

In this section, the results of the research will be discussed in more detail. In Figure 1, it can be seen that the highest number of *Tuberculosis* cases by district/city in West Java Province in 2021 will occur in Bogor Regency, followed by Bekasi Regency and Purwakarta Regency, while the two districts/cities with the lowest incidence of *Tuberculosis* are Kota Banjar and Pangandaran Regency. Then, Figure 2 shows the correlation between variables. This study used 4 (four) variables to model the number of *Tuberculosis* cases. The correlation in Figure 2 is formed based on the magnitude of each correlation value between the two variables. Still, it focuses more on visually seeing the correlation between the independent and dependent variables. Based on Figure 2, it can be seen that population density and the percentage of people who have health complaints to the number of *Tuberculosis* cases have a positive correlation because they are coloured blue. In contrast, the percentage of poor people and the percentage of households that have access to proper sanitation have a negative correlation because they are coloured red.

The next step is checking multicollinearity, which is used to find out a model that is formed and whether there is a perfect relationship between the independent variables. A good model is one in which there is no perfect relationship between the independent variables but a perfect relationship between the independent and dependent variables. One way that can be used to detect multicollinearity is VIF (Variance-Inflating Factor). Multicollinearity occurs when the VIF value is > 10 [16]. Based on Table 2, it can be seen that the VIF values for all independent variables are less than 10, so it can be said that there is no multicollinearity.

Once it is known that there is no multicollinearity, then the independent variables look for parameter estimates to form a Poisson Regression model, the results of which are presented in Table 3. Based on Table 3, it can be seen that the parameter estimates of the regression coefficients have been obtained, so the initial model of the Poisson Regression formed are as follows:

$$\hat{\mu} = \exp(9{,}166 - 0{,}1461PoPP - 0{,}000003312PD - 0{,}005665PH2APS + 0{,}027P2H2CLM) \quad (1)$$

Equidispersion is one of the characteristics of Poisson, which states the equality of the average value and variance[5], so in this study, the assumption of equidispersion was tested. This test is carried out by dividing the Pearson Chi-Square value by the degree of freedom. The Pearson Chi-Square value was 33,227.27. Then the degree of freedom is 22, obtained from *n*-*p*-1 where *n* observations in this study are 27 and *p* is the number of independent variables, namely 4. The results from dividing the Pearson Chi-Square value with the degrees of freedom are 1,510.33. The results of the equidispersion test show that there is overdispersion. One of the analytical methods that can be used to overcome this problem is using the Negative Binomial Regression approach. Table 4 presents the results of parameter estimation for the Negative Binomial Regression coefficient so that the initial model of Negative Binomial Regression is obtained, namely:

$$\hat{\mu} = \exp(8{,}747 - 0{,}1341PoPP - 0{,}000001819PD - 0{,}003019PH2APS + 0{,}03153P2H2CLM) \quad (2)$$

Based on a series of hypothesis testing (overall test and partial test for each independent variable) that have been carried out, the final Negative Binomial Regression model is obtained which is suitable for modeling the number of *Tuberculosis* cases in West Java Province in 2021, namely as follows:

$$\hat{\mu} = \exp(8{,}51727 - 0{,}13292PoPP + 0{,}03128P2H2CLM) \quad (3)$$

Then to determine the best model of the models in this study, you can compare the value of the AIC (Akaike Information Criterion) for each model formed which is presented in Table 6.

Table 6. Best Model Selection Criteria

Regression Models	AIC
Poisson	33.901
Negative Binomial (initial)	495,77
Negative Binomial (final)	491,9

The best model has the smallest AIC value. Based on Table 6, it can be seen that the model that has the smallest AIC value is the final Negative Binomial Regression model. The model formed in equation[3] can be interpreted as follows:

1. For every 1% increase in the percentage of poor people, it will increase the number of *Tuberculosis* cases in West Java Province by 0,8755 times.

2. For every 1% increase in the percentage of the population who had health complaints in the past month, it will increase the number of *Tuberculosis* cases in West Java Province by 1,0318 times.

The final Negative Binomial regression model in this study was formed from two significant independent variables: the percentage of poor people and the percentage of people who had health complaints in the last month. This is in accordance with several studies that have been done previously that socio-economic status (in this case, poverty) is a factor that influences the number of *Tuberculosis* cases. Then someone who has recently had health complaints can be a sign that the immune system in that person's body is decreasing or weakening. According to the Ministry of Health, a weak immune system can cause a person to be easily exposed to *Tuberculosis* bacteria. Therefore, the variable percentage of the population who had health complaints in the last month can significantly influence the modelling of the number of *Tuberculosis* cases (in this case West Java Province in 2021).

Conclusion

Based on the several steps that have been carried out to obtain the best model for modelling the number of *Tuberculosis* cases in West Java Province in 2021, it can be concluded that the most suitable model for this study is the Negative Binomial Regression model. Then the independent variables that significantly influence the model are the percentage of poor people and the percentage of people who have had complaints recently.

References

1. World Health Organization. Tuberculosis.
2. Kementrian Kesehatan Republik Indonesia. Pusat Data dan Informasi . Retrieved from Pusat Data dan Informasi Kementerian Kesehatan Republik Indonesia.
3. Kementrian Kesehatan Republik Indonesia. Stop Tuberculosis. Retrieved from Kementerian Kesehatann Jenderal Pelayanan Kesehatan.
4. Walpole R, Myers R. *Ilmu Peluang Dan Statistika Untuk Insinyur Dan Ilmuwan*. ITB; 1995.
5. Cameron A., Trivedi P. *Regression Analysis of Count Data*. Cambridge University Press; 1998.
6. Wang W, Famoye F. Modeling household fertility decisions with generalized Poisson regression. *J Popul Econ*. 1997;10(3):273-283. doi:10.1007/s001480050043
7. Shmueli G, Minka TP, Kadane JB, Borle S, Boatwright P. A useful distribution for fitting discrete data: Revival of the Conway-Maxwell-Poisson distribution. *J R Stat*

Soc Ser C Appl Stat. 2005;54(1):127-142. doi:10.1111/j.1467-9876.2005.00474.x
8. Diendéré E, Tiéno H, Bognounou R, Ouédraogo D, Simporé J, Ouédraogo-Traoré RDJ. Prevalence and risk factors associated with infection by human immunodeficiency virus, hepatitis B virus, syphilis and bacillary pulmonary tuberculosis in prisons in Burkina Faso. *Med Trop*. Published online 2011.
9. Du Preez K, Mandalakas AM, Kirchner HL, et al. Environmental tobacco smoke exposure increases Mycobacterium tuberculosis infection risk in children. *Int J Tuberc Lung Dis*. 2011;15(11):1490-1496. doi:10.5588/ijtld.10.0759
10. Aravindan P. Host genetics and tuberculosis: Theory of genetic polymorphism and tuberculosis. *Lung India*. 2019;36(3):41-46. doi:10.4103/lungindia.lungindia
11. Byrd RP, Mehta JB, Roy TM. Malnutrition and pulmonary tuberculosis. *Clin Infect Dis*. 2002;35(5):634-636. doi:10.1086/342314
12. Singh SK, Kashyap GC, Puri P. Potential effect of household environment on prevalence of tuberculosis in India: Evidence from the recent round of a cross-sectional survey. *BMC Pulm Med*. 2018;18(1). doi:10.1186/s12890-018-0627-3
13. Siroka A, Law I, Macinko J, et al. The effect of household poverty on tuberculosis. *Int J Tuberc Lung Dis*. 2016;20(12):1603-1608. doi:10.5588/ijtld.16.0386
14. Mohidem NA, Hashim Z, Osman M, Muharam FM, Elias SM, Shaharudin R. Environment as the risk factor for tuberculosis in Malaysia: A systematic review of the literature. *Rev Environ Health*. 2021;36(4):493-499. doi:10.1515/reveh-2020-0096
15 Caraka RE., Yasin H. *Geographically Weighted Regression (GWR) Sebuah Pendekatan Regresi Geografis.* Mobius; 2017.
16 Gujarati DN. *Basic Econometrics, Fourth Edition*. The McGraw-Hill Companies; 2004.