JSDS: JOURNAL OF STATISTICS AND DATA SCIENCE
VOLUME 1, No 2, October 2022
e-ISSN: 2828-9986
https://ejournal.unib.ac.id/index.php/jsds/index

Modeling of Tuberculosis Cases in Sumatra Region using Poisson Inverse Gaussian Regression

She Asa Handarzeni[1]*
[1] Bank Indonesia, Jakarta.

* Corresponding Author: She_asa@bi.go.id

Article Info

Article History:
Received: October 29, 2022
Revised: October 30, 2022
Accepted: October 30, 2022
Available Online: October 31, 2022

Key Words:
Poisson
PIG Regresssion
Overdispersi
Tuberculosis
Modelling

Abstract

In the Sumatra Region, tuberculosis (TB) is a disease that needs special attention because it tends to increase every year. Based on health theory, there are many factors that cause TB, but it is not easy to determine which factors have a significant effect. Therefore, in this study an analysis was carried out that could model, predict, and determine the factors causing TB disease in the Sumatra Region. The data used is data on TB cases in the Sumatra Region in 2018 taken from the Publication of the Central Statistics Agency. Poisson regression is an analysis that is suitable for modeling count data such as TB disease data. The assumption of Poisson regression is that the mean and variance of the response variables must be equal (equidispersion). However, the TB case data in the Sumatra Region in 2018 has an average value that is smaller than the variance (overdispersion) so it cannot be solved by Poisson regression. To overcome this problem, we need a method that can overcome overdispersion, namely Poisson Inverse Gaussian (PIG) regression. From the results of the analysis using PIG regression, it can be concluded that the factors that have a significant effect on TB cases in the Sumatra Region are the percentage of the male population (X_1), the percentage of the productive age population (X_2), the percentage of households with a floor area of ≤ $19m^2$ (X_3), and the percentage of households that have access to proper sanitation (X_4), where the model formed is

$$\hat{\mu} = exp\ (18.97511 - 0.15742X_1 - 0.08825X_2 + 0.13871X_3 + 0.01159X_4)$$

Based on the model, the predicted results of TB cases in the Sumatra Region had an average of 596.04178 where the lowest cases occurred in Pringsewu of 154.8943 and the highest cases occurred in Bukittinggi of 2719.59400.

1. INTRODUCTION

The total population of Indonesia in 2019 was recorded at 268,074,600 people. This figure represents 3.52% of the world's population [1]. A large population can cause complex problems, one of which is health problems. Currently, there are several population health problems that need serious attention, one of which is Tuberculosis (TB). TB is an infectious disease caused by the bacterium Mycobacterium tuberculosis, which attacks various organs, especially the lungs. This disease if left untreated or incomplete treatment can lead to complications and even death [2]. Sumatra is the area with the 2nd most TB cases in Indonesia after Java. To find out the factors that have a significant effect on TB, it is necessary to analyze the case.

1.1 Tree Regression

Poisson regression is a form of regression analysis that is used to model data whose response variable is in the form of count (sum), which assumes that the Y variable has a Poisson distribution [3]. The Poisson regression model is written as follows:

$$Y_i = \mu_i + \varepsilon_i \qquad i = 1,2, \ldots,$$

where Y_i is the number of events and μ_i is the average number of events and μ_i is assumed to be unchanged from data to data [4].

1.2 *Poisson Inverse Gaussian Regression*

PIG regression is a form of regression analysis of mixed distribution between Poisson and Inverse Gaussian. PIG regression is designed for data with a response variable in the form of counts that experience overdispersion cases so that it cannot be solved by Poisson regression. PIG is determined by two parameters, namely the average (μ) as the location parameter and the dispersion parameter (τ) as the shape parameter [5].

If you want to know the relationship between a response variable Y and k predictor variables X_1, X_2,….,X_k, the multiple regression model to describe the relationship is as follows:

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i \quad i = 1,2,\ldots,$

when expressed in vector is as follows :

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta} + \varepsilon_i$$

and the expected value is

$$\begin{aligned} \mu_i &= E(Y_i) \\ &= E(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) + E(\varepsilon_i) \\ &= (\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}) + 0 \\ &= \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \end{aligned}$$

or when expressed in vector becomes

$$\mu_i = \boldsymbol{X}_i^T \boldsymbol{\beta} \tag{1}$$

However, the model in equation (1.2.1) is not suitable when applied to the response variable with a PIG distribution. The response variable of the model can be a real number in the interval (-∞,∞), while the value of the response variable in the PIG model is a non-negative integer. To overcome this problem, a natural log (*ln*) connecting function is used on the average using a linear model, so that the relationship between the response variable and the linear combination of predictor variables is:

$$\ln(\mu_i) = \boldsymbol{X}_i^T \boldsymbol{\beta} \text{ or } \mu_i = e^{\boldsymbol{X}_i^T \boldsymbol{\beta}}$$

1.3 Parameter Testing

Parameter testing was conducted to determine whether or not the influence of the predictor variable on the response variable was present. Parameter testing in the PIG regression model is carried out using simultaneous hypothesis testing on parameter and partial testing of parameters and.

1. Simultaneous testing

   Simultaneous testing includes all parameters together with the following hypotheses:

   $$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$$

   (Predictor variables simultaneously have no significant effect on the response variable)

   H_1: at least one of $\beta_i \neq 0, i = 1,2,\ldots,k$

   (at least one predictor variable has a significant effect on the response variable)

2. Partial testing

   Individual or partial testing is carried out with the following hypotheses:

   $$H_0: \beta_j = 0\,, \quad j = 1,2,\ldots,k$$

   (The j-th predictor variable has no significant effect on the response variable)

   $$H_1: \beta_j \neq 0\,, \quad j = 1,2,\ldots,k$$

   (The j-th predictor variable has a significant effect on the response variable)

   Parameter testing on τ:

   $$H_0: \tau = 0$$
   $$H_1: \tau \neq 0$$

1.4 Assumption Testing

Some assumption tests that must be carried out in the PIG Regression analysis are as follows:

1. Correlation test

   The following is the Pearson correlation test hypothesis:

   H_0: There is no correlation between variables

   H_1: There is a correlation between variables

2. Multicollinearity test

   One way to detect the occurrence of multicollinearity is to look at the value of the Variance Inflation Factor (VIF), which is a value that describes the increase in variance of the estimated parameters between predictor variables. VIF value > 10 indicates the presence of multicollinearity [6]. VIF can be written as follows:

$$VIF = \frac{1}{(1 - R^2_{Yjl})}$$

3. Overdispersion test

   The Poisson regression model requires equidispersion, which is a condition when the mean and variance of the response variables are equal. However, sometimes there is an overdispersion phenomenon in the data modeled with the Poisson distribution. Overdispersion means the variance is greater than the mean. This indicates that the Poisson regression model is not suitable for these data [4]. One way that can be done to detect the presence or absence of overdispersion in the response variable to be studied is the Pearson chi-square test statistic as follows:

$$VT = \sum_{i=1}^{n} \frac{(y_i - \bar{y})^2}{\bar{y}}$$

   If the dispersion index value is less than 1, it can be said that underdispersion occurs, on the other hand, overdispersion occurs when the dispersion index value is more than 1 [7]. The hypothesis used is as follows:
   H_0: There is no overdispersion on the data.
   H_1: There is an overdispersion on the data.

1.5 Selection of The Best Model with The Akaike Information Criteria (AIC) Method

The AIC method is a method that can be used to select the best regression model found by Akaike and Schwarz. The method is based on the maximum likelihood estimation method. The AIC value can be calculated as follows [8].

$$AIC = e^{\frac{2k}{n}} \frac{\sum_{i=1}^{n} \hat{u}_i{}^2}{n}$$

2. METHOD

This study will use the PIG Regression model to model TB cases in the Sumatra Region in 2018. The data are taken from the publications of the Central Statistics Agency, entitled "Provinsi Dalam Angka 2019" and " Statistik Kesejahteraan Provinsi Tahun 2018" with district/city observation units in the Sumatra Region. The data consists of 1 response variable, namely the number of TBs and 6 predictor variables, namely the percentage of the male population, the percentage of the productive age population, the percentage of households with a floor area of 19m^2, the percentage of households that have access to proper sanitation, and the percentage of households that have access to proper sanitation. have access to adequate drinking water sources, and Percentage of poor people.

The stages of the research carried out are as follows:

1. Descriptive statistics, including the calculation of the maximum, minimum, mean, variance, and standard deviation of each variable. Calculations are carried out using the R 3.5.1 . program.
2. Overdispersion test using AER package from software R 3.5.1.
1. Correlation test using R 3.5.1. program.
2. Multicollinearity test using VIF value.
3. PIG regression modeling which consists of parameter estimation and parameter significance testing either simultaneously or partially. This process is carried out using the gamlss package from the R3.5.1 software.
4. Selection of the best model based on the smallest AIC value.
5. Interpretation of results and drawing conclusions.

3. RESULTS AND DISCUSSION

3.1 Descriptive Statistics

In this study, modeling of the number of TB cases in the Sumatra region was carried out using PIG regression. The descriptive statistical value of each of these variables can be seen in the table below:

Tabel 1. Descriptive Statistics of Research Variables

Variable	Minimum	Maximum	Average	Variance
Y	0.000	5313.000	540.442	424543.974
X_1	47.812	57.973	50.617	1.345
X_2	55.770	73.030	65.653	7.666
X_3	0.000	11.690	2.251	4.817
X_4	7.400	99.210	64.795	439.368
X_5	6.080	95.460	53.040	497.785
X_6	2.390	27.790	11.155	26.283

Based on Table 1, it can be seen that the number of TB cases in the Sumatra Region has an average of 540.442 with a variant of 424543.974. The highest number of cases occurred in Palembang City as many as 5131 cases and the lowest cases occurred in West Lampung, South Lampung, Metro City, and Sibolga, namely 0 cases.

3.2 Overdispersion Test Result

The test criteria used are Reject H_0 if p.value < α with value of 5% or 0.05. From the test results using the AER package on Software R, the p.value is $0.00946 < \alpha = 0.05$ so that there is enough evidence to reject H_0. This shows that there is an overdispersion in the response variable so that the data analysis using PIG regression can be continued.

3.3 Multicollinearity Test Results

VIF value greater than 10 indicates the occurrence of multicollinearity in the data. From the test results using the R software, the VIF value is obtained as follows:

Table 2. VIF Value of Predictor Table

Varible	X_1	X_2	X_3	X_4	X_5	X_6
VIF	1.20080	1.36809	1.25978	1.40328	1.29250	1.45783

From Table 2 it can be seen that the VIF value for all predictor variables is less than 10, which means that there is no multicollinearity between predictor variables so that the analysis using PIG regression can be continued.

3.4 Correlation Test Results

The following are the results of correlation testing carried out using the R software:

Table 3. Correlation Between Variables X_i and Y

Variable	r_{X_iY}	***p-value***
X_1	-0.01160	0.88560
X_2	0.21623	0.00707
X_3	0.26451	0.00091
X_4	0.23746	0.00302
X_5	0.24715	0.00200
X_6	-0.12346	0.12720

Table 3 shows that with = 5% the predictor variables that have a relationship with the response variable are X_2, X_3, X_4, and X_5. However, the variables X_1 and X_2 are still included in the study because theoretically these two variables have a relationship with the response variable. there is no correlation with a value of 0 between the predictor variable and the response variable, so the analysis using PIG regression can be continued.

3.5 PIG Regression Modeling

PIG regression modeling can be done if the data meets all assumptions, namely that there is no multicollinearity and the data has overdispersion. PIG regression modeling on TB data in the Sumatra Region in 2018 was carried out using the GAMLSS package from the R software. All predictor variables used can produce several combinations of models, but in the discussion only convergent models are shown. The PIG regression model is said to be convergent if it reaches the deviation value or stable deviation at a certain iteration.

The following is a convergent PIG regression model:

$\hat{\mu} = exp(20.12895 - 0.15979X_1 - 0.09574X_2 + 0.12083X_3 + 0.00889X_4 + 0.00044X_5 - 0.03033X_6)$, $\tau = 1.12080$

$\hat{\mu} = exp(20.33460 - 0.16267X_1 - 0.09618X_2 + 0.12237X_3 + 0.00891X_4 - 0.03111X_6)$, $\tau = 1.12610$

When written in vector form it becomes:

$$\begin{bmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{bmatrix} = e^{[1 \quad X_1 \quad X_2 \quad X_3 \quad X_4 \quad X_5 \quad X_6]\begin{bmatrix}20.12895 & 20.33460\\ -0.15979 & -0.16267\\ -0.09574 & -0.09618\\ 0.120830 & 0.122370\\ 0.008890 & 0.008910\\ 0.000440 & 0.000000\\ -0.03033 & -0.03111\end{bmatrix}}$$

After obtaining the parameters of the possible models, parameter testing is carried out to see if the model formed is significant.

3.6 Parameter Test

Parameter testing in PIG regression was carried out twice, namely simultaneous testing and partial testing.

1. Simultaneous test

Simultaneous parameter testing includes all parameters in the possible models that are formed. This test is carried out using the statistical value G. The hypothesis used in the simultaneous test is:

$H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$

H_1: at least one $\beta_i \neq 0, i = 1,2,\ldots,k$

with the decision criteria reject H_0 if the value of Statistics $G > \chi^2_{(\alpha,v)}$

The following are the results of simultaneous parameter testing carried out using R software and Microsoft Office Excel:

Table 4. Simultaneous Parameter Testing

	Variables	G	V	$\chi^2_{(\alpha,v)}$	Conclusion
Model 1	$X_1, X_2, X_3, X_4, X_5, X_6$	2364.647	146	175.1976	Reject H_0
Model 2	X_1, X_2, X_3, X_4, X_6	2364.653	147	176.2938	Reject H_0

From Table 4 it can be seen that the value of the G statistic in the two possible models is greater than the value of $\chi^2_{(\alpha,v)}$. This shows that together the predictor variable values have a significant effect on the response variable.

2. Partial test

Individual parameter testing is carried out to see which predictor variables have a significant influence on the response variable. Individual parameter testing is carried out on parameters and with the following hypothesis:

a. Test on β
   $H_0: \beta_j = 0\,, \quad j = 1,2,\dots,k$
   $H_0: \beta_j \neq 0\,, \quad j = 1,2,\dots,k$
b. Test on τ
   $H_0: \tau = 0$
   $H_0: \tau \neq 0$

The test criteria used is to reject H_0 if p.value < α. The following are the results of individual parameter testing carried out using the R software:

Tabel 5. Parameter Testing Individually

	β_0	β_1	β_2	β_3	β_4	β_5	β_6	τ
Model 1	20.12895	-0.15979	-0.09574	0.12083	0.00889	0.00044	-0.0303	1.12080
pvalue	0.00014	0.10185	0.01987	0.01712	0.10027	0.93197	0.18339	0.00000
Model 2	20.3346	-0.16267	-0.09618	0.12237	0.00891	-	-0.0311	1.12610
pvalue	0.00000	0.08350	0.01870	0.01330	0.09740	-	0.13910	0.00000

By using α = 0.10, the result is that in the first model the parameter β_0 has a significant effect on the << 0.01 level, the β_2 parameter has a significant effect on the 1.987% level. The parameter β_3 has a significant effect at the level of 1.1712%, while the other parameters have no significant effect. In the second model, the β_0 parameter has a significant effect on the << 0.01 level, the β_1 parameter has a significant effect on the 8.35% level, the β_2 parameter has a significant effect on the 1.87% level, the β_3 parameter has a significant effect on 1.33% level, the parameter β_4 has a significant effect at the level of 9.74%, while the parameter β_6 has no significant effect. For the dispersion parameter (τ), both models have significant parameters at the level of << 0.01%. That is, both models are able to overcome the problem of overdispersion.

3.7 Best Model Selection

Parameter testing produces models with significant parameters. From these models, the best model is selected based on the smallest AIC value.

Table 6. Comparison of AIC values

	AIC
Model 1	2382.579
Model 2	2378.905

In Table 6 it can be seen that the model that has the smaller AIC value is model 2, so the better model used in this study is model 2 with the following details:

Table 7. Estimated Parameters of The Selected Model (Model 2)

Parameter	**Estimated value**
β_0	18.97511
β_1	-0.15742
β_2	-0.08825
β_3	0.13871
β_4	0.01159

When written in the form of a model, it becomes :

$$\hat{\mu} = exp\ (18.97511 - 0.15742X_1 - 0.08825X_2 + 0.13871X_3 + 0.01159X_4)$$

The model above shows that every 1% addition of the percentage of the male population will double the average TB case by exp(-0.15742) or 0.85435 times the original TB case average if other variables are constant. Every 1% addition of the percentage of the productive age population will double the average TB case by exp(-0.08825) or 0.91553 times the original average if other variables remain constant. Each additional 1% of the percentage of households with a floor area of ≤ 19m^2 will multiply the average TB cases by exp(0.13871) or 1.14879 times the average TB cases from the original average if other variables remain constant. Each additional 1% of the percentage of households that have access to proper sanitation will double the average TB case by exp(0.01159) or 1.01166 times the previous average if other variables remain constant.

3.8 Prediction

The prediction results of TB cases in the Sumatra Region as a whole had an average of 596.04178 where the lowest cases occurred in Pringsewu at 154.8943 and the highest cases occurred in Bukittinggi at 2719.59400. The following is a comparison chart between the observed values and predictions:

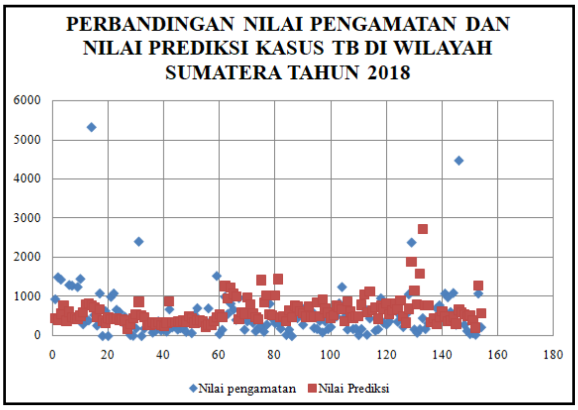

Figure 1. Comparison of observed values with predicted values

From Figure 1 it can be seen that the predicted value is close to the actual value because it is still around the same value.

4. CONCLUSION

The results of the analysis of TB cases in the Sumatra Region in 2018 showed that the factors that significantly influenced were the percentage of the male population (X_1), the percentage of the productive age population (X_2), the percentage of households with a floor area of ≤ 19m^2 (X_3), and the percentage of households that have access to proper sanitation (X_4). The model formed is as follows:

$$\hat{\mu} = exp\ (18.97511 - 0.15742X_1 - 0.08825X_2 + 0.13871X_3 + 0.01159X_4)$$

REFERENCES

[1] Anonim. 2020. Daftar Negara Menurut Jumlah Penduduk. Indonesia: Wikipedia Indonesia.

[2] Anonim. 2018. *Info Datin: Tuberkulosis*. Indonesia: Kementrian Kesehatan RI.

[3] Aulele, S N. 2011. *Model Geographically Weighted Poisson Regression dengan Pembobot Fungsi Kernel Gauss Studi Kasus: Jumlah Kematian Bayi di Jawa Timur Tahun 2007*. Jurnal Barekeng Vol 5 No 2.

[4] Cahyandari, R. 2014. *Pengujian Overdispersi pada Model Regresi Poisson (Studi Kasus: Laka Lantas Mobil Penumpang di Provinsi Jawa Barat)* . Jurnal Statistika Vol 14 No 2.

[5] Fuadyah, K. 2019. *Estimasi Parameter Regresi Poisson Inverse Gaussian (PIG) Menggunakan Metode MLE dan Penerapannya pada Data Kematian Ibu di Jawa Timur Tahun 2017.* Malang: Universitas Negeri Malang.

[6] Herindrawati, A Y., I. N. Latra dan Purhadi. 2017. *Pemodelan Regresi Poisson Inverse Gaussian Studi Kasus: Jumlah Kasus Baru HIV di Provinsi Jawa Tengah Tahun 2015*. Jurnal Sains Dan Seni ITS .Vol 6 No 1.

[7] Karlis, D dan E. Xekalaki. 2000. *A Simulation Comparison of Several Procedures for Testing the Poisson Assumption*. The Statistician, Vol. 49 No. 3.

[8] Akaike, H. 1978. *A Bayesian Analysis of The Minimum AIC Procedure*. Annals Institute of Statistical Mathematics. Part A 9-4.