Advances in Social Science, Education and Humanities Research, volume 474
Proceedings of the 2nd International Seminar on Science and Technology (ISSTEC 2019)

Modelling the Number of Tuberculosis (TB) Cases in Indonesia using Poisson Regression and Negative Binomial Regression

Rahmadi Yotenka[1,*], Alfazrin Banapon[1]

[1]*Statistics Department, Universitas Islam Indonesia*
Corresponding author:rahmadi.yotenka@uii.ac.id

ABSTRACT
Tuberculosis (TB) is still one of the world's health problems which continues to be overcome today. Indonesia is a country that accounts for 8% of the number of TB cases in the world in the third highest position after India and China. Tuberculosis is one of the lower respiratory and infectious diseases caused by the bacterium Mycrobacterium Tuberculosis. In this research, modelling the number of Tuberculosis cases was carried out to suppress the increase in the number of TB cases in the Indonesia with the Negative Binomial regression approach. Data on the number of TB cases is the count data so the analysis used to model the count data is Poisson regression. However, in analysis there is often an overdispersion phenomenon which will cause the estimation results to be biased. So that needs to be overcome by Negative Binomial regression, the results obtained are factors that have an influence on the number of TB cases in Indonesia are the percentage of districts/cities in Indonesia with a healthy environmental quality, the number of HIV (Positive) cases, the number of AIDS cases, and the percentage of households which has access to adequate drinking water sources with Pseudo-R 2 of 64.13% and the AIC of the Negative Binomial regression model is 724.33.
Keywords: *Negative Binomial Regression, Poisson Regression, Tuberculosis*.

1. INTRODUCTION

Tuberculosis (TB) is still a health problem, especially in developing countries including Indonesia. In Indonesia at least 300 people died from this TB disease (Intan, 2019). Globally, the number of TB cases is 6.4 million cases in 2018, equivalent to 64% of TB cases (10 million cases). Meanwhile, the number of TB cases in Indonesia was 511,873 cases in 2018 (data as of 31 January 2019) (Ministry of Health, Indonesian Ministry of Health Data and Information Center, 2018).

Factors that influence the occurrence of Tuberculosis are the socio-economic conditions of the community, namely nutritional status and environmental sanitation. The lower nutritional status and environmental sanitation cause low endurance so that it is easily infected with Tuberculosis when sick (Ministry of Health, 2011). Environmental factors play an important role in the transmission of disease, especially the home environment that does not meet the requirements. The home environment is one of the factors that has a major influence on the health status of its inhabitants (Ministry of Health, 2012). Furthermore, Tuberculosis (TB) and HIV/AIDS infection had a strong correlation and the presence of HIV/AIDS infection causes the increasing of TB disease incidence. People living with HIV/AIDS had 20 times increasing the risk of developing TB disease (Muna and Cahyati, 2019).

The number of TB cases can be reduced if the factors that influence the number of TB are known. The relationship of the number of TB cases with the factors that influence it can be determined by regression analysis. Regression analysis for census data is Poisson regression because the number of TB cases is census data (count data) that follows the Poisson distribution, the assumption that must be met is equidispersion or conditions where the expected price and variance are the same, but often in reality often occur in the field overdispersion symptoms or variance values greater than expected prices (McCullagh et al., 1989). When this happens Poisson regression can cause underestimation of the standard error (Hinde et al, 1998).

One method used in overcoming the problem of overdispersion in Poisson regression is Negative Binomial regression, so the purpose of this study is to look at the factors that influence the number of TB cases in Indonesia in 2018 with Negative Binomial regression.

2. MATERIALS AND METHODS

Poisson regression and negative binomial regression are the methods used in this study. First, the research data were analyzed using poisson regression. Because there is overdispersion in the poisson regression analysis, it is continued using the negative binomial regression approach. Negative binomial regression is used when the variance is greater than the mean made by poisson regression. The negative binomial distribution is an extension of the poisson-gamma distribution that contains the dispersion parameter θ.

2.1. Population and Sample

The population in this study were all TB patients in Indonesia in 2018. The number of study sites used was 34 provinces in Indonesia.

2.2. Data and Operational Variables

The data used in this research is secondary data obtained through the publication of data on the Kementerian Kesehatan Republik Indonesia in 2018 and Badan Pusat Statistik (BPS) through www.bps.go.id.

This study uses ten variables consisting of nine independent variables and one dependent variable, the dependent variable in this study is Case Tuberculosis (tb_case), while independent variables are poor population (poor_perc), healthy area program (health_perc), smoker (smook_perc), seedy RT (seedy_perc), healthy environment (env_perc), cases of HIV (hiv_perc), cases of AIDS (aids_perc), healthy place to produce food (place_perc), and the source of drinking water (source_perc).

Table 1. Definition of Operational Variables

Variables	Definition	Scale
tb_case	The number of TB sufferers are male and female in each province in Indonesia in 2018	Ratio
poor_perc	Percentage of poor population in each province in Indonesia in 2018	Ratio
health_perc	Percentage of districts/cities in Indonesia held a healthy region order in 2018	Ratio
smook_perc	Percentage of smokers in Indonesia in 2018	Ratio
seedy_perc	Percentage of seedy households in Indonesia in 2018	Ratio
env_perc	Percentage of districts/cities in Indonesia with a healthy environmental quality in 2018	Ratio
hiv_perc	Percentage of HIV (Positive) cases in Indonesia in 2018	Ratio
aids_perc	Percentage of AIDS cases in Indonesia in 2018	Ratio
place_perc	Percentage of food management places that meet health requirements in Indonesia in 2018	Ratio
source_perc	Percentage of households which has access to adequate drinking water sources in Indonesia in 2018	Ratio

2.3. Multicollinearity

One of the conditions that must be fulfilled in forming the regression model is the absence of multicollinearity between independent variables, to detect the presence of multicollinearity in the regression model can use the VIF (Variance Inflation Factor) value. If the VIF value > 10 indicates that multicollinearity between independent variables. VIF values can be written as follows.

$$VIF_j = \frac{1}{1-R^2{}_j} \quad (1)$$

With $R^2{}_j$ is the coefficient of determination between independent variables (Hocking, 1996).

2.4. Poisson Regression

Poisson regression is a nonlinear regression model that is often used to model data *counts* (Agresti, 2002). If the discrete random variable (Y) is a Poisson distribution with the parameter μ then the probability function of the Poisson distribution can be stated as follows.

$$f(y,\mu) = \frac{e^{-\mu}\mu^y}{y!}; y = 0,1,2,...,n \quad (2)$$

With μ is the average of the Poisson distributed response variable where the average value and variance of Y have a value of more than 0. The equation of poisson regression model can be written as follows.

$$\mu_i = \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi}) \quad (3)$$

With μi is the average number of events that occur in a certain time interval.

2.5. Poisson Regression Model Parameter Estimation

One of method used for estimating Poisson regression parameters is the *Maximum Likelihood Estimation* (MLE) method. The function of *likelihood* is formulated as follows.

$$L(\beta) = \frac{\exp\left(-\sum_{t=1}^{n}\exp\left(x_i^T\beta\right)\right)\left(\exp\left(\sum_{t=1}^{n} y_t x_i^T \beta\right)\right)}{\prod_{t=1}^{n} y_t!} \quad (4)$$

where,

$\beta = \left[\beta_0\ \beta_1\ ....\beta_p\right]^T; x_i = \left[1\ x_{1i}\ ....\ x_{pi}\right]^T$

2.6. Testing Poisson Regression Model Parameters

Significance test simultaneously using the *Maximum Likelihood Ratio Test* (MLRT) with the following hypothesis (Mc Cullagh et al, 1989)

$H_0 = \beta_1 = \beta_2 = ... = \beta_p = 0$

$H_1 =$ there is at least one $\beta_k \neq 0; k = 1, 2, ..., p$

Test Statistics:

$$G^2 = -2\ln\left(\frac{L(\hat{\varpi})}{L(\hat{\Omega})}\right) \quad (5)$$

Reject H_0 if $G^2 > \chi^2_{(a,p)}$, it means that there is at least one independent variable that gives effect to the dependent variable. Then the parameters are tested partially with the following hypothesis

$H_0 = \beta_k = 0$

$H_1 = \beta_k \neq 0; k = 1, 2, ..., p$

Test Statistics:

$$Z_{count} = \frac{\hat{\beta}_k}{SE\left(\hat{\beta}_k\right)} \quad (6)$$

Reject H_0 if $|Z_{count}| > Z_{(a/2)}$ with a is the level of significance.

2.7. Overdispersion Poisson Regression

Poisson regression is said to overdispersion if the value of the variance is greater than the average value. If the data discrete happen overdispersion and keep using regression poisson as a method of settlement, it will be obtained a conclusion that is not valid because the value of *the standard error* becomes *underestimate*.

Overdispersion is the ratio value of deviance with degrees freely, when the ratio was > 1 then indicates happen overdispersion the poisson regression model (Famoye et al, 2004)

2.8. Negative Binomial Regression

The Negative Binomial regression model has the following pdf (Greene, 2008):

$$P(y, \mu, \theta) = \frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) y!}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu}\right)^{y} \quad (7)$$

Negative Binomial regression model can be used to model Poisson data that has overdispersion because Negative Binomial distribution is an extension of the Poisson-Gamma distribution that contains the dispersion parameter θ (Hilbe, 2011). According to Osgood (2000), Paternoster and Brame (1997) the Negative Binomial model is an alternative that is often used for cases of overdispersion in regression (Berk et al, 2008). The Negative Binomial regression model is stated as follows

$$\hat{y}_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_p X_{pi}) \quad (8)$$

Estimation of Negative Binomial Regression Parameters

The *Maximum Likelihood Estimation* (MLE) method is used to estimate parameters in the Negative Binomial regression. The *likelihood* function of the Negative Binomial regression is as follows.

$$L(y, \mu, \theta) = \prod_{i=1}^{n} \frac{\Gamma\left(y + \frac{1}{\theta}\right)}{\Gamma\left(\frac{1}{\theta}\right) y!}\left(\frac{1}{1+\theta\mu}\right)^{\frac{1}{\theta}}\left(\frac{\theta\mu}{1+\theta\mu}\right)^{y} \quad (9)$$

Negative Binomial Regression estimation uses *Newton Raphson's* iteration *method* to maximize the *likelihood* function.

2.9. Testing Negative Binomial Regression Parameters

Significance testing simultaneously for estimating the parameters of the Negative Binomial Regression model uses the deviance test with the following hypothesis (Hosmer, 1995).

$H_0 = \beta_1 = \beta_2 = ... = \beta_p = 0$

$H_1 =$ there is at least one $\beta_k \neq 0; k = 1, 2, ..., p$

Test Statistics:

$$G^2 = -2\ln\left(\frac{L(\hat{\varpi})}{L(\hat{\Omega})}\right) = 2\left(\ln L(\hat{\Omega}) - \ln L(\hat{\varpi})\right) \quad (10)$$

Reject H_0 if the test statistics $G^2 > \chi^2_{(a,p)}$

Partially significant test to determine which parameters have an influence on the model, with the following hypothesis.

$H_0 = \beta_k = 0$

$H_1 = \beta_k \neq 0; k = 1, 2, ..., p$

Test Statistics: p

$$W_k = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \quad (11)$$

H_o is rejected if the test statistic W or $t_{count} > t_{(n-k,\ a/2)}$, meaning that the k-th parameter is significant for the Negative Binomial regression model.

3. RESULTS AND DISCUSSION

3.1. Multicollinearity Checks

Table 2. VIF Value of the Independent Variable

Independent Variable	VIF	Independent Variable	VIF
poor_perc	2.84	hiv_perc	3.64
health_perc	2.35	aids_perc	1.31
smook_perc	1.27	place_perc	1.62
seedy_perc	2.92	source_perc	1.48
env_perc	3.96		

Based on the test results obtained the value of each independent variable < 10, meaning that multicollinity does not occur or between independent variables do not correlate with each other.

3.2. Poisson Regression

Data on the number of TB cases is assumed to be Poisson distributed because it is the data *count.* The following is a parameter estimate of the Poisson regressi on model

Table 3. Parameter Estimation of the Poisson Regression Model

Variable	Estimation	Z count	P value
Const	3.975	120.5	$2x10^{-16}$
poor_perc	$-1.115x10^{-1}$	-216.5	$2x10^{-16}$
health_perc	$3.198x10^{-2}$	301.1	$2x10^{-16}$
smook_perc	$1.976x10^{-1}$	272.8	$2x10^{-16}$
seedy_perc	$7.275x10^{-2}$	178.6	$2x10^{-16}$
env_perc	$-3.159x10^{-2}$	-463.7	$2x10^{-16}$
hiv_perc	$3.161x10^{-2}$	380.9	$2x10^{-16}$
aids_perc	$9.781x10^{-4}$	391.5	$2x10^{-16}$
place_perc	$-3.011x10^{-4}$	-155.8	$2x10^{-16}$
source_perc	$-3.421x10^{-2}$	-164.8	$2x10^{-16}$
Deviance = 543845		DF=20	
AIC = 544235			
$G^2 = 6.4283x10^5$			

Tests carried out with a significance level of 5%. Simultaneous testing parameters Poisson regression model aims to determine whether the independent variables simultaneously influence the dependent variable. The hypothesis used is as follows.

$H_0 = \beta_1 = \beta_2 = ... = \beta_6 = 0$

2.10. Pseudeo-R^2

Pseudeo-R^2 denoted R^2 meaning that the k-th parameter is significant for the Negative Binomial regression model.

$$R^2_p = 1 - \frac{L_p}{L_1} \quad (12)$$

Where L_p is the *likelihood* function value of the complete model while L_1 is the *likelihood* function value of the model that only contains *intercepts.*

One way to detect the occurrence of multicollinearity is to look at the VIF (Variance Influence Factor) value.

k

H = there is at least one $\beta \neq 0$; $k = 1,2,...,9$

1

Based on the test results with a significance level of 5% obtained $\chi^2_{(df=9, a=5\%)}$
= 16,91898 which means < G^2 =

$6{,}4283x10^5$ so reject H_0 which means at least one independent variable that affects the dependent variable. So it needs to proceed with partial testing with the following hypothesis

$H_0 = \beta_k = 0$

$H = \beta_k \neq 0;\ k = 1,2,...,9$

Based on the partial test results with a significance level of 5%, it was found that $|Z \text{ count}| > Z_{(0.05/2)} = 1$or *p value* < *a* = 5% which means that reject H_0 or any independent variable influence on the dependent variable.

3.3. Overdispersion Examination

Poisson regression has the expectation value is the same as the value of the variance or the so-called equidispersion, but in this case, the number of TB cases overdispersed or conditions where the expected price is greater than the variance. To detect the occurrence of overdispersion by looking at the deviance value in the Poisson regression model is 543845 and the degree of freedom 24, if the ratio of deviance and degree of freedom > 1, meaning that the number of TB cases have overdispersion cases.

Poisson regression is not suitable for overdispersion cases because it will produce a biased and inefficient parameter estimate. Distribution serine g used for overdispersion case is Negative Binomial. The initial step in this Negative Binomial regression modelling is to determine the value θ which aims to minimize the dispersion parameters so that it can overcome this problem. Initial θ can be done by *trial-and-error* to get the value of the ratio of deviance and degrees of freedom is 1. Here are the results of *trial-and-error* initial θ

Table 4. Initial Value θ

Initial	Deviance	Df	Deviance/df
1	23.749	24	0.9895
2	47.493	24	1.9789
1.5	35.622	24	1.4843
1.1	26.124	24	1.0885
1.09	25.886	24	1.0786
1.02	24.224	24	1.0093
1.011	24.010	24	1.0004
1.0101	23.989	24	0.9995
1.0105	23.999	24	0.99996
1.0106	24.001	24	1.000042
1.01057	24	24	1

Based on the *trial-and-error* initial θ obtained initial θ where the ratio of the value of the degree of free deviance with a value of 1 is equal to 1.01057 so do Negative Binomial regression modelling with initial θ at 1.01057.

3.4. Negative Binomial Regression Modelling

Testing concurrently significant parameters of the Negative Binomial regression model. The following are estimated parameters of the Negative Binomial regression model.

Table 5. Parameter Estimation of the Negative Binomial Regression Model

Variable	Estimation	t count	P value
Const	7.134	2.285	0.031
poor_perc	$-9.215x10^{-2}$	-1.923	0.066
health_perc	$1.607x10^{-2}$	1.477	0.153
smook_perc	$1.418x10^{-1}$	2.220	0.036
seedy_perc	$2.310x10^{-2}$	0.572	0.573
env_perc	$-3.303x10^{-2}$	-3.491	0.002
hiv_perc	$3.221x10^{-2}$	3.337	0.003
aids_perc	$1.175x10^{-3}$	3.015	0.006
place_perc	$1.671x10^{-6}$	0.007	0.995
source_perc	$-4.108x10^{-2}$	-1.923	0.066
Deviance = 24		*DF=24*	
AIC = 726.27			
$G^2 = 34.858$			

Simultaneously testing parameter significance Negative Binomial regression model aims to determine jointly whether the same independent variables influence the dependent variable. The hypothesis used is as follows:

$H_0 = \beta_1 = \beta_2 = ... = \beta_9 = 0$

$H_1 =$ there is at least one $\beta_k \neq 0; k = 1,2,...,9$

Based on the test results with a significance level of 5% obtained $\chi^2_{(df=9,a=5\%)} = 16.91898 < G^2 = 34.858$ so reject H_0 means that there is at least one independent variable that

gives effect to the dependent variable. So that it needs to be continued in partial testing with the following hypothesis.

$H_0 = \beta_k = 0$

$H_1 = \beta_k \neq 0; k = 1,2,...,9$

Based on the partial test results, there are a number of variables that do not influence the dependent variable, so one needs to be excluded one by one and the independent variables obtained that influence the dependent variable are as follows

Table 6. Parameter Estimation of the Negative Binomial Regression Model

Variable	Estimation	t count	P value
Const	11.660	7.794	$1.35x10^{-8}$
env_perc	-0.021	-2.220	0.034
hiv_perc	0.039	3.656	0.001
aids_perc	0.001	2.314	0.028
source_perc	-0.053	-2.473	0.019
Deviance = 29			DF=29
AIC = 724,33			
G^2 = 25,461			

The conclusion that the factors that give a statistical influence on the number of TB cases in Indonesia are the healthy environment (env_perc), cases of HIV (hiv_perc), cases of AIDS (aids_perc), and the source of drinking water (source_perc).

3.5. Pseudo R^2

In testing the goodness of the model or *goodness of fit* researchers use *Pseudo* R^2. The R^2_p value obtained was 0.6413 where the value was relatively large, so it can be said that 64.13% of the models used were good enough to describe or explain the number of TB cases. Whereas 35.87% is influenced by other factors not included in the model.

3.6. Selection and Formation of the Best Model

After testing with Poisson regression and Negative Binomial regression. Next is the selection of the best model between the two methods based on the AIC (Akaike Information Criterion) value. Poisson regression itself has an AIC value = 544.235, while the Negative Binomial regression with all significant parameters has an AIC value = 72.33, so it can be concluded that the method to be used to model the number of TB cases is the Negative Binomial regression because it has the smallest AIC value.

The best model of Negative Binomial regression is:

$$\ln(\hat{\mu}_i) = 11.660 - 0.021X_5 + 0.039X_6 + 0.001X_7 - 0.053X_9$$

Based on the model, can be in the interpretation right that each

additional 1% districts/cities in Indonesia that meet the Environmental Quality Health, the average number of TB cases will be reduced by 1,021 ≈ 1 person, assuming other variables constant. Meanwhile, if there is an increase in one case of HIV (Positive) in Indonesia, then the number of TB cases will also increase by 1,0398 ≈ 1 person. Then if there is an increase in one AIDS case in Indonesia, then the number of TB cases will also increase by 1.001 ≈ 1 person and if an additional 1% of households in Indonesia have access to a reliable source of drinking water, the average number of TB cases will decrease by 1.054 ≈ 1 person.

4. CONCLUSION

With the Negative Binomial regression modelling, it is found that the factors that have a statistically influential effect on the number of TB cases in Indonesia are the percentage of districts/cities in Indonesia with a healthy environmental quality, the number of HIV (Positive) cases, the number of AIDS cases, and the percentage of households which has access to adequate drinking water sources. With *pseudo* R^2 is 64.13% of the variance of the counts on TB can be explained by the model.

REFERENCES

[1] Agresti, A. (2002). Categorical Data Analysis Second Edition. New York: John Wiley & Sons, Inc

[2] Berk, Richard. MacDonald, J.M. (2008). Overdipersion and Poisson Regression. Philadelphia: Springer

[3] Famoye, F., Wulu, J.T. and Singh, K.P. (2004). *On The Generalize Poisson Regression Model with an Application to Accident Data*. *Journal of Data Science 2*. 287-295

[4] Greene, W. (2008). Functional Forms for the Negative Binomial Model for Count Data, Foundation, and Trends in Ecometrics,99,585-590. New York: New York University.

[5] Intan, G. (2019, Maret 23). 300 Orang Per Hari Meninggal di Indonesia Akibat Penyakit TBC. Retrieve from voaindonesia:

voaindonesia.com/a/orang-per-hari- meninggal-di-indonesia-akibat- penyakit-tbc/4849081.html

[6] Hilbe, J. (2011). Negative Binomial Regression, Second Edition. New York: Cambridge University Press

[7] Hinde J, Dem'etrio CGB. (1998). *Overdispersion: Models and Estimation. Computational Statistics and Data Analysis* 27: 151-170

[8] Hocking, R.R. (1996). *Method and Applications of Linier Models.* New York: John Wiley and Sons,Inc

[9] Hosmer, David Watson and Lemeshow, Sticher. (1995). Applied Logistic Regression. New York: John Wiley and Sons Inc

[10] Kementerian Kesehatan, RI. (2011). Strategi Nasional Pengendalian Tuberculosis di Indonesia 2010-2014. Jakarta: Kementerian Kesehatan RI Direktorat Jenderal Pengendalian Penyakit dan Penyehatan Lingkungan

[11] Kemenkes. (2018). Data dan Informasi Profil Kesehatan Indonesia 2017. Jakarta: Kementerian Kesehatan RI

[12] Kemenkes. (2018). Pusat Data dan Informasi Kementrian Kesehatan RI. Jakarta: Kementerian Kesehatan

[13] Mc Cullagh p, Nelder JA. (1989). *Generalized Linear Models Second Edition*, London: Chapman and Hall

[14] Muna, N., and Cahyati, W. H. (2019). Determinan Kejadian Tuberkulosis pada Orang dengan HIV/AIDS. HIGEIA Journal Of Public Health Research And Development, 3(2), 2019

[15] WHO. (2018). Global, Tuberculosis Report. World Health Organization