Journal of Mathematics Education and Application (JMEA)

Vol. 2, No 1, Februari 2023, pp. 01- 10 DOI: https://doi.org/10.30596/jmea.v2i1.12164

Binary Logistic Regression Analysis Using Stepwise Method on Tuberculosis Events

Rifan Halomoan Tua Sinaga[1*] , Open Darnius[2]

[1]Student of Mathematics Education, Universitas Sumatera Utara, Indonesia

[2]Lecturer In Mathematics Education, Universitas Sumatera Utara, Indonesia

*Corresponding Author E-mail: : rifansinaga@gmail.com

Article Info

Article History

Received : 06 November 2022

Accepted : 08 Desember 2022

Published : 28 Februari 2023

Keywords:

Akaike Information Criterion (AIC), Risk Factors, Stepwise Method, Tuberculosis

ABSTRACT

Tuberculosis is an infectious disease caused by the bacteria Mycobacterium tuberculosis. Among all the districts/cities of North Sumatra province, Medan has the highest cases of tuberculosis sufferers with a total of 12,105 cases in 2019. This study aims to determine the factors that significantly influence tuberculosis. The factors analyzed were age, gender, occupation, education, BCG immunization, history of diabetes mellitus and HIV infection. This study uses secondary data for the period January 2019 to December 2020 obtained from the Sentosa Baru Health Center. With the help of SPSS, this study uses a stepwise method with forward selection and backward elimination as the method for analysis. Akaike Information Criterion (AIC) is used to select the best model in the stepwise method. With the AIC criteria obtained, the best model is forward selection because the AIC value is lower at 28,527 compared to backward elimination at 41,664. Of the 7 variables studied, there are 3 factors that have a significant effect, namely age, history of diabetes mellitus, and HIV infection so that the model g(x) = 2.802 – 1.056 X_1 – 0.614 X_6 – 2.477 X_7.

To cite this article:

INTRODUCTION

Regression analysis is a statistical technique that is useful for examining and modeling the relationship between one response variable and one or more predictor variables. Regression analysis is useful for examining the relationship between two or more variables, especially in exploring the pattern of relationships in which the model is not yet fully known so that in its application it is more exploratory (A Agresti, 1990). Regression has various forms, including linear regression, dummy regression, panel data regression, and logistic regression. Logistic regression is a statistical analysis method to describe the relationship between response variables that have two or more categories with one or more explanatory variables on a category or interval scale (Aji,2014). The statistical method used in this study is binary logistic regression with the response variable in tuberculosis cases. Binary logistic regression is used to describe the relationship between the response variable and the predictor variable, where the response variable is a dichotomy that has two possible values. To perform variable selection, stepwise method is used. The inclusion and exclusion of variables with this method is completely based on statistical criteria, namely the pvalue. There are two versions of the stepwise method, namely forward selection and backward elimination(Mahmood, 2016). Tuberculosis is an infectious disease that greatly disrupts human activities so that this disease becomes one of the targets in sustainable health development. Indonesia is the second country with the highest number of tuberculosis sufferers. This encourages national tuberculosis control to continue to be carried out with intensification, acceleration, extensification and program innovation. In 2019, it was found that the number of tuberculosis cases was 33,779, an increase compared to all tuberculosis cases found in 2018, which was 26,418. In each district/city throughout North Sumatra, more cases occurred in men than women. Medan is the highest district/city with tuberculosis sufferers in North Sumatra with a total of 12,105 cases. There are so many risk factors that a person can be infected with tuberculosis. These risk factors often do not occur alone but interact. In this study, an analysis of the factors that are considered as triggers of tuberculosis will be carried out using the binary logistic regression method. In this study, the triggering factors used were age, gender, occupation, education, BCG immunization, history of Diabetes Mellitus and HIV infection.

RESEARCH METHOD

This type of research is quantitativeresearch. With the help of SPSS, this study used stepwise methods with forward selection and backward elimination as methods to analyze.The data used in this study were taken from secondary data, namely data on tuberculosis patients from January 2019 to December 2020 Puskesmas Sentosa Baru.

This study was conducted from August 2021 to October 2021 at the New Sentosa Health Center located on Jalan Sentosa Baru No. 1. 22, Sei Kera Hilir I, Medan Perjuangan Subdistrict, Medan City, North Sumatra, Indonesia.The study variable used in this study was the response variable denoted Y and the predictor variable denoted X.

RESULTS AND DISCUSSION

A.Multicollinearity Test

The multicollinearity test aims to see whether or not there is a high correlation between variables. A regression model is said to be free of multicollinearity if it has a VIF value of not more than 10 and has a tolerance number of not less than 0.10.

Table 2. Multicollinearity Test

Variable	*Tolerance*	Nilai VIF
Age	0.787	1.270
Gender	0.923	1.083
Jobs	0.773	1.294
Education	0.898	1.114
BCGImmunization	0.917	1.090
History of Diabetes Mellitus	0.956	1.046
HIV Infection	0.964	1.037

Based on the table above, there is no multicollinearity between predictor variables. It can be seen that the tolerance number is not less than 0.10 and the VIF value is not more than 10. So the tuberculosis data at the Sentosa Baru Health Center can be used.

B. Parameter Estimation of Binary Logistics Regression Model

Determination of the binary logistic regression model using Newton Raphson. Determining the value of $\hat{\beta}_0$,using the Ordinary Least Square (OLS) method with predictor variables and response variables used are tuberculosis data at the Sentosa Baru.

Iterasi 1

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{1})} = \boldsymbol{\beta}^{(\mathbf{0})} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{(\mathbf{0})}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x})^{(\mathbf{0})})$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{1})} = \begin{bmatrix} 0{,}903 \\ -0{,}156 \\ -0{,}004 \\ \vdots \\ -0{,}599 \end{bmatrix} + \left[\begin{bmatrix} 0{,}720 & -0{,}025 & \ldots & -0{,}596 \\ -0{,}025 & 0{,}108 & \cdots & 0{,}001 \\ -0{,}060 & -0{,}003 & \cdots & 0{,}019 \\ \vdots & \vdots & \ddots & \vdots \\ -0{,}596 & 0{,}001 & \cdots & 0{,}616 \end{bmatrix} \begin{bmatrix} -40{,}184 \\ -23{,}484 \\ -16{,}388 \\ \vdots \\ -40{,}752 \end{bmatrix}\right]$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{1})} = \begin{bmatrix} 1{,}707 \\ -0{,}631 \\ -0{,}021 \\ -0{,}340 \\ -0{,}069 \\ 0{,}563 \\ -0{,}526 \\ -1{,}543 \end{bmatrix}$$

Iterasi 2

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{2})} = \boldsymbol{\beta}^{(\mathbf{1})} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{(\mathbf{1})}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x})^{(\mathbf{1})})$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{2})} = \begin{bmatrix} 1{,}707 \\ -0{,}631 \\ -0{,}021 \\ \vdots \\ -1{,}543 \end{bmatrix} + \left[\begin{bmatrix} 0{,}844 & -0{,}041 & \ldots & -0{,}707 \\ -0{,}041 & 0{,}120 & \cdots & 0{,}017 \\ 0{,}068 & 0{,}002 & \cdots & 0{,}023 \\ \vdots & \vdots & \ddots & \vdots \\ -0{,}707 & 0{,}017 & \cdots & 0{,}722 \end{bmatrix} \begin{bmatrix} -1{,}045 \\ -1{,}576 \\ -0{,}473 \\ \vdots \\ -1{,}138 \end{bmatrix}\right]$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{2})} = \begin{bmatrix} 1{,}875 \\ -0{,}726 \\ -0{,}016 \\ -0{,}396 \\ -0{,}079 \\ 0{,}622 \\ -0{,}581 \\ -1{,}662 \end{bmatrix}$$

Iterasi 3

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{3})} = \boldsymbol{\beta}^{(\mathbf{2})} + (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{V}^{(\mathbf{2})}\boldsymbol{X})^{-\mathbf{1}}\boldsymbol{X}^{\boldsymbol{T}}(\boldsymbol{Y} - \boldsymbol{\pi}(\boldsymbol{x})^{(\mathbf{2})})$$

$$\widehat{\boldsymbol{\beta}}^{(\mathbf{3})} = \begin{bmatrix} 1{,}875 \\ -0{,}726 \\ 0{,}016 \\ \vdots \\ -1{,}662 \end{bmatrix} + \left[\begin{bmatrix} 0{,}878 & -0{,}046 & \ldots & -0{,}763 \\ -0{,}046 & 0{,}124 & \cdots & 0{,}021 \\ -0{,}070 & -0{,}002 & \cdots & 0{,}024 \\ \vdots & \vdots & \ddots & \vdots \\ -0{,}736 & 0{,}021 & \cdots & 0{,}745 \end{bmatrix} \begin{bmatrix} 0{,}018 \\ -0{,}023 \\ 0{,}009 \\ \vdots \\ 0{,}012 \end{bmatrix}\right]$$

$$\hat{\boldsymbol{\beta}}^{(3)} = \begin{bmatrix} 1{,}880 \\ -0{,}729 \\ -0{,}015 \\ -0{,}396 \\ -0{,}078 \\ 0{,}624 \\ -0{,}583 \\ -1{,}667 \end{bmatrix}$$

The iteration process stops at the 3rd iteration with an estimate of $\hat{\boldsymbol{\beta}}^{(3)}$. These results are in accordance with the results of data processing using SPSS 25.

C. Forward Selection

The results of variable selection with forward selection can be seen as follows:

Outlier Data Identification

The outlier data resulted in the model being less good so it had to be removed from the research model. The following is the identification of outlier data in this study:

Table 3. Identification of Otulier Data in Forward Selection
Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		Y			Resid	ZResid
58	S	E**	,839	P	-,839	-2,281

The table above shows that there are outlier data, namely the 58th data. For further analysis, the 58th data were excluded from the research model. After the 58th data is removed in the research model, there are no more outliers, which means the data is good.

Simultaneous Test

Simultaneous testing aims to determine the relationship of the predictor variables to the overall response variable.

Table 4. Simultaneous Test on Forward Selection
Omnibus Tests of Model Coefficients

	Chi-square	df	Sig.
Step	3.938	1	.047
Block	21.611	3	.000
Model	21.611	3	.000

The Chi-Square table value with a df of 3 and a significance level of 5% is 7.815. If the calculated Chi-Square value is compared with the table Chi-Square value, then 21,611 > 7,815. So the decision taken is to reject H_0 which means that there is a coefficient β that has a significant simultaneous effect on the response variable.

Partial Test

Partial testing is carried out to determine the significance of each parameter on the response variable.

Table 5. Partial Test On Forward Selection

Variables in the Equation

		B	Wald	df	Sig.
Step 3[c]	Umur	-1.056	10.727	1	.001
	Riwayat_Diabetes_Mellitus	-.614	3.919	1	.048
	Infeksi_HIV	-2.477	4.861	1	.027
	Constant	2.802	5.975	1	.015

The results of the Wald test carried out on each predictor variable showed that there were 3 variables that had a significant effect on the response variable because the Wald test value was > 3.841 and the p-value was < 0.05. It was concluded that age, history of diabetes mellitus, and HIV infection had a significant influence on tuberculosis at the Sentosa Baru Health Center.

Model Fit Test

This test is carried out to test whether the model formed is feasible.

Table 6. Results of the Model Suitability Test in Forward Selection

Chi-square	*Df*	*p-value*
0.879	2	0.644

The Chi-Square table value with a df of 2 and a significance level of 0.05 is 12,592. Based on Table 4.5 with Chi-Square value < 5.992 and p-value > 0.05, which means it failed to reject H_0. So it can be concluded that the model is appropriate or there is no significant difference between the observations and the possible predictions of the model.

Classification Accuracy T

he following are the results of the classification accuracy from the results is.

Table 7. Accuracy of Classification in Forward Selection

Observation		Prediction		Percentage True
		Extrapulmonary	Lung	
Tuberculosis	Extrapulmonary	104	18	85,2%
	Lung	52	32	38,1%

Total Percentage	66,0%

Table above shows that 104 patients with extrapulmonary tuberculosis were correctly classified and 32 patients were classified as having pulmonary tuberculosis. So that the accuracy of the modeling classification in binary logistic regression on tuberculosis at the Sentosa Baru Health Center is 66.0%.

D.Backward Elimination

The results of the variable selection with backward elimination can be seen as follows:

Identification of Outlier Data

There is an outlier data, namely the 58th data. For further analysis, the 58th data were excluded from the research model. After the 58th data was removed in the research model, there were still outliers as shown in the table.

Table 8. Identification of Data Outliers in Backward Elimination (Step 2)
Casewise List[b]

Case	Selected Status[a]	Observed	Predicted	Predicted Group	Temporary Variable	
		Y			Resid	ZResid
89	S	E**	.817	P	-.817	-2.113

The table above shows that there are data outliers, namely the 89th data. For further analysis, the 89th data were excluded from the research model. After the 89th data is issued in the research model, there are no more outliers, which means the data is good.

Simultaneous Test

Simultaneous testing aims to determine the relationship of the predictor variables to the overall response variable.

Table 9. Simultaneous Test on Backward Elimination
Omnibus Tests of Model Coefficients

	Chi-square	Df	Sig.
Step	-.654	1	.419
Block	29.658	4	.000
Model	29.658	4	.000

The Chi-Square table value with a df of 4 and a significance level of 5% is 9.488. If the calculated Chi-Square value is compared with the table Chi-Square value, then 29.658 > 9.488. If the p-value is compared to the value of α, then 0.000 < 0.05. So that the decision taken is to reject H_0 which means that there is a coefficient β that has a significant simultaneous effect on the response variable.

Partial Test

Partial testing was conducted to determine the significance of each parameter on the response variable.

Table 10. Simultaneous Test on Backward Elimination

	B	Wald	df	Sig.
Age(1)	-.963	8.648	1	.003
Immunization_BCG(1)	.659	4.555	1	.033
History Diabetes Mellitu(1)	-.649	4.120	1	.042
Infection_HIV(1)	-21.840	.000	1	.999
Constant	21.857	.000	1	.999

Based on Table above, it can be seen that the results of the Wald test carried out on each predictor variable showed that there were 3 variables that had a significant effect on the response variable because the Wald test value was > 3.841 and the p-value was < 0.05. It was concluded that age, BCG immunization, and history of diabetes mellitus had a significant influence on tuberculosis at the Sentosa Baru Health Center.

Model Fit Test

This test is carried out to test whether the model formed is feasible.

Table 11. Model Conformity Test Results on Backward Elimination

Chi-square	*df*	*p-value*
6.503	5	0.260

Based on Table above with Chi-Square value < 11.071 and p-value > 0.05, which means it failed to reject H0. So it can be concluded that the model is appropriate or there is no significant difference between the observations and the possible predictions of the model.

Classification Accuracy

The following are the results of the classification accuracy from the results is.

Table 12. Accuracy of Classification in Backward Selection

Observation		Prediction		Percentage True
		Extrapulmonary	Lung	
Tuberculosis	Extrapulmonary	82	39	67,8%
	Paru	33	51	60,7%
Total Percentage				64,9%

abel above showed that patients with extrapulmonary tuberculosis were correctly classified as 82 patients and 51 patients were correctly classified as having pulmonary tuberculosis. So that the accuracy of the modeling classification in binary logistic regression on tuberculosis at the Sentosa Baru Health Center is 64.9%.

E. Stepwise Method Comparison

After getting the results from the two versions of the stepwise method, namely forward selection and backward elimination, a comparison of the two models will be carried out to obtain the best model based on the AIC method.

Table 13. Stepwise Method Comparison

Models	1	2
Methods	Forward	Backward
-2 Log Likelihood	256,916	247,817
Classification Accuracy	66%	64,90%
Significant Variables	Age	Age
	History of Diabetes Mellitus	BCG Immunization
	HIV Infection	History of Diabetes Mellitus
AIC	28,527	41,664

Forward selection produces the best model with an AIC value of 28,527 compared to backward elimination with an AIC value of 41,664. Due to the forward selection method, the factors that significantly influence the occurrence of tuberculosis are age, history of diabetes mellitus, and HIV infection.

F. Binary Logistics Regression Model

The significant variables obtained based on the forward selection method are variables X1, X6, and X7, so that the regression coefficient values are obtained as follows:

Table 14. Regression Coefficient

Variabel	
Constant	2,802
age	-1,056
History of Diabetes Mellitus	-0,614
HIV Infection	-2,477

The logit model generated from binary logistic regression based on significant variables is as follows:

$$g(x) = 2{,}802 - 1{,}056\ X_1 - 0{,}614\ \ X_6 - 2{,}477\ X_7$$

Based on this model, the calculation of the probability function generated as follows:

$$\pi(x) = \frac{e^{2{,}802\ -\ 1{,}056\ X_1 -\ 0{,}614\ \ X_6 -\ 2{,}477\ X_7}}{1 + e^{2{,}802\ -\ 1{,}056\ X_1 -\ 0{,}614\ \ X_6 -\ 2{,}477\ X_7}}$$

Odds Ratio The value of the odds ratio (Exp(β)) on the factors that influence tuberculosis at the Sentosa Baru Health Center based on the model formed can be seen in tabel below.

Table 15. Odds Ratio

Variabel		Exp()
Age (X_1)	-1,056	0,348
History of Diabetes Mellitus(X_6)	-0,614	0,541
HIV Infection (X_7)	-2,477	0,084

Based on the table above, the age variable has an odds ratio of 0.348, which means non-productive age has a 0.348 times lower risk of developing pulmonary tuberculosis than productive age. The variable history of diabetes mellitus has an odds ratio of 0.541 which means that having no history of diabetes mellitus is 0.541 times lower than having a history of diabetes mellitus. The HIV Journal of Mathematics Technology and Education Vol. , No. , 2022 15 infection variable has an odds ratio of 0.084, which means that those who are not infected with HIV are 0.084 times less likely to have pulmonary tuberculosis than those who are infected with HIV.

Conclusions

Based on the results and previous discussions, the best model obtained from the stepwise method is forward selection with the lowest AIC value of 28,527. The binary logistic regression model obtained is

$$g(x) = 2{,}802 - 1{,}056\ X_1 - 0{,}614\ \ X_6 - 2{,}477\ X_7$$

From this model, only 3 of the 7 predictor variables that have a significant effect on tuberculosis are X_1 (age), X_6 (history of diabetes mellitus), and X_7 (HIV infection). The coefficient of age is -1.056 with an odds ratio of 0.348 which means non-productive age has a 0.348 times lower risk of developing pulmonary tuberculosis than productive age. The variable history of diabetes mellitus has a coefficient of -0.614 with an odds ratio of 0.541 which means that having no history of diabetes mellitus is 0.541 times less likely to have pulmonary tuberculosis than having a history of diabetes mellitus. The HIV infection variable has a coefficient of -2.477 with an odds ratio of 0.084, which means that those who are not infected with HIV are 0.084 times less likely to have pulmonary tuberculosis than those who are infected with HIV.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis.* New York: John Wiley and Sons.

Aji, C. M. (2014). Analisis Faktor-faktor yang Mempengaruhi Laju Pertumbuhan Penduduk Kota Semarang Tahun 2011 Menggunakan Geographically Weighted Logistic Regression. *Jurnal Gaussian*, Vol 3 No 2: 161-171.

Fatah, K. S., & Mahmood, R. F. (2016). Parameter Estiamtion for Binary Logistic Regression Using Different Iterative Methods. *Journal of Zankoy Sulaimani,* Vol 19 No 2: 177-178.

Harlan, Johan. 2018. *Analisis Regresi Logistik*. Depok: Gunadarma.

.Hosmer, D. W., and Lemeshow, S. 2000.*Applied Logistic Regression*. New York: John Wiley and Sons, Inc.

Johnson, Richard A. dan Dean W. Wichern. 2007. *Applied Multivariate Statistical Analysis.* New Jersey: Pearson Education, Inc.

Kementerian Kesehatan RI. Profil Kesehatan Indonesia 2018. Jakarta: Pusat Data dan Informasi Kementrian Kesehatan RI; 2019.

Kim, J.S., and Dailey, R.J. 2008.*Biostatistic for Oral Healthcare*. USA: Blackwell Munksgaard.

Kuncoro, Mudrajad. 2004. *Metode Kuantitatif: Teori dan Aplikasi untuk Bisnis dan Ekonomi.* Yogyakarta: UPP AMP YKN.

Muaz, Fariz. Faktor-Faktor yang Mempengaruhi Kejadian Tuberkulosis Paru Basil Tahan Asam di Puskesmas Wilayah Kecamatan Serang Kota Serang tahun 2014. Jakarta: Universitas Syarif Hidayatullah.

Nirwana.S.R.A. 2015. Regresi Logistik Multinomial dan Penerapannya dalam Menentukan Faktor yang Berpengaruh pada Pemilihan Program Studi di Jurusan Matematika UNM.[Skripsi].Makassar:Universitas Negeri Makassar,Program Sarjana.

Peeters, B., Dewil, R., & Smets, I. Y. (2012). Improved Process Control of an Industrial Sludge Cemtrifuge-dryer Installation Through Binary Logistic Regression Modeling of The Fouling Issues. *Journal of Process Control*, Vol 22: 1390-1391.

Putri, E. I. S., Indriati, D. W., & Wahyunitisari, M. R. 2020. The Prevalence of Diabetes Mellitus among Hospitalized Tuberculosis Positive Cases in Hajj Hospital Surabaya. *Malaysian Journal of Medicine and Health Sciences.* 16(1): 235-239

Ranuh,I.G.N.2008. Pedoman Imunisasi di Indonesia. Edisi ketiga.Jakarta: Badan Penerbit Ikatan Dokter Anak Indonesia

RI, K. (2016). Peraturan Menteri Kesehatan Republik Indonesia Nomor 67 Tahun 2016 Tentang Penanggulangan Tuberkulosis. *Jakarta:Kementerian Kesehatan Republik Indonesia*.

Santoso, S. (2012).*Analisis SPSS pada Statistik Parametrik.*Jakarta: PT. ElexMedia Komputiondo.

Sugiarto,D.S.2006. *Metode statistika untuk bisnis dan ekonomi*.Jakarta: *Gramedia Pustaka Utama.*

Sumut, Dinkes. (2019). Profil Kesehatan Provinsi Sumatera Utara Tahun 2019. *Sumatera Utara.*

Suyo, J.2010, *Herbal Penyembuhan Gangguan Sistem Pernafasan*.,Yogyakarta: B First

Wijaya, I. M. K. (2013, December).Infeksi HIV (human immunodeficiency virus) pada penderita tuberkulosis.In *Prosiding Seminar Nasional MIPA*.

World Health Organization.Global Tuberculosis Report 2020. Geneva: World Health Organization; 2020.