J Int Comp & He Inf.
Vol. 1, No. 2, September 2020

pISSN: 2715-6923, eISSN: 2721-9186
https://jurnal.unimus.ac.id/index.php/ichi

Journal of Intelligent Computing and Health Informatics

Modeling of Tuberculosis Case in Central Java 2018 with Three Knot Point

Research Article

Dina Fristantiningtyas Wiliyani Hapsari[1*], Laelatul Khikmah[2]

Department of Statistics, Akademi Ilmu Statistika Muhammadiyah Semarang, Semarang 50185, Indonesia

*dfristawh@gmail.com
aisyah.salsabila17@gmail.com

Article history :
Received : 24 Sep 2020
Accepted : 25 Sep 2020
Available online : 30 Sept 2020

ABSTRACT

Tuberculosis is an infectious disease caused by infection with the bacteria Mycobacterium Tuberculosis or known as Acid-Resistant Bacteria (BTA). In 2018, Central Java is one of the provinces with a high number of tuberculosis cases in Indonesia, which ranks 2nd with 67,941 cases after West Java. Many variables affect the number of TB cases. In this study we propose a modeling to determine the variables that affect the number of tuberculosis cases in Central Java. Based on data obtained from the Central Java Provincial Health Office in 2018, it shows that the pattern between the number of tuberculosis cases and the variables that are suspected of having an effect is not linear, then a spline regression approach is carried out. The results of this study indicate that the best spline regression model is to use a three-point node with significant variables, namely population density and malnutrition. The value of $R2$ obtained was 54.6%.

Keywords : Central java, knot point, spline regression, tuberculosis.

1. INTRODUCTION

Tuberculosis is an infectious disease caused by infection with the bacteria Mycobacterium Tuberculosis or known as Acid-Resistant Bacteria (BTA) (Kementerian Kesehatan RI, 2018). In 2016, Tuberculosis was ranked the tenth leading cause of death in the world with 10.4 million cases. In the same year, Tuberculosis was ranked fifth as the cause of death in countries with lower-middle-income status or middle-income countries, one of which was Indonesia (WHO, 2018). Indonesia is in the second position with the highest TB burden in the world. In 2017, the number of TB cases found in Indonesia was 446,732 cases (RI, 2018). Meanwhile, in 2018 it has increased to 566,623 cases of tuberculosis (Kementrian Kesehatan Republik Indonesia, 2019).

In 2017, there were three top provinces in Indonesia that had the most cases of tuberculosis. The province is West Java which is in the first rank with the number of Tuberculosis cases for all types of 78,698 cases, in the second rank is East Java Province with the number of Tuberculosis cases for all types as many as 48,323 cases, and Central Java is in the top three ranks in the TB cases with the number Tuberculosis cases of all types were 42,272 cases (RI, 2018). Meanwhile, in 2018, Central Java rose to second place after West Java with 67,941 cases of all types of TB (Kementrian Kesehatan Republik Indonesia, 2019).

Tuberculosis cases are increasing in Central Java Province, so various ways have been done by the Central Java Provincial Government to reduce the number of Tuberculosis cases. One way is to conduct research to determine the variables that affect the number of TB cases. Based on data on the number of Tuberculosis cases obtained from the Central Java Provincial Health Office in 2018, the pattern of the relationship between the number of tuberculosis cases and the variables suspected to be influential tends to be linearly not related (Kementrian Kesehatan Republik Indonesia, 2019). This problem can be overcome with a nonparametric approach that is often used in estimating curve shapes, namely Spline Nonparametric Regression (Wulandari et al., 2017).

The scope of this research is data on the number of tuberculosis cases in districts or cities in Central Java, which consists of 29 districts and 6 cities. The variables used in this study were the number of tuberculosis cases as a response variable, population density and malnutrition as a predictor variable. The statistical approach used was spline nonparametric regression. The purpose of this study was to determine the modeling of the number of Tuberculosis cases in Central Java 2018 with a spline nonparametric regression approach and to find out what variables were significant to the number of Tuberculosis cases in Central Java 2018.

2. THE MATERIAL AND METHOD

2.1. Linearity Testing

Linearity testing is useful for knowing whether two variables have a linear relationship or not, usually based on a scatterplot pattern between the response variable and the predictor variable.

2.2. Nonparametric Regression

Nonparametric regression is a statistical method used to determine the effect of dependent variables on independent variables in which the shape of the regression curve is not linear. In general, the nonparametric regression model is as follows Eq. 1 (Eubank, 1999).

$$y_i = f(x_i) + \varepsilon_i, i = 1,2,3,\dots,n \quad (1)$$

Where:

y_i : dependent variables
x_i : independent variables
$f(x_i)$: regression function of unknown shape nearity testing
ε : error

Nonparametric regression does not require certain assumptions like parametric regression does (Suparti et al., 2019).

2.3. Spline Regression

Spline is a nonparametric method. A spline is a piecewise polynomial that is segmented continuously. The segmented nature of the spline provides higher flexibility than ordinary polynomials. According to Hardle, spline has the advantage of overcoming data patterns that show sharp rises / falls with the help of knot points, and the resulting curve is relatively smooth (Härdle, 1992). The knot point is a point that shows changes in the behavior pattern of a different spline interval function.

In general, the m-order spline functions are as follows Eq. 2.

$$f(x_i) = \beta_0 + \beta_1 x_1^1 + \cdots + \beta_m x_i^m + \sum_{q=1}^{p} \beta(m + q)^{(x_i + k_q)_+^m} \quad (2)$$

Where:

β : constant
x_i : independent variables
k_q : q knots on the independent variable

2.4. Data Analysis Stages

The stages of this research are as follows:

Stage 1. Linearity testing: the first step is to create a scatterplot between the response variables and each predictor variable to determine the pattern of the relationship that occurs. This aims to determine the data patterns that are formed. If the plot shows a non-linear data pattern, then the Nonparametric Regression approach is used.

Modeling of Tuberculosis Case in Central Java 2018 with Three Knot Point
pISSN: 2715-6923, eISSN: 2721-9186, Vol. 1, No. 2, September 2020, pp. 56-60

Stage 2. Data modelling using spline regression by following steps as follows:

1. *determining the number of knot points*;

   This research uses one point knot, two point knot, and three point knot.

2. *optimal knot point selection*;

   The best Spline Nonparametric Regression Model is generated from the optimal selection of knot points. In selecting the optimal knot points with the Spline Nonparametric Regression approach in the number of Tuberculosis cases in Central Java in 2018, it was carried out based on the smallest GVC value at one knot point, two knot point, and three knot point.

3. *data modelling*;

   The best spline regression model on the number of Tuberculosis cases in Central Java in 2018 was obtained from optimal knots at three knot points. Then model the data according to the optimal knot point.

4. *testing the significance of parameters in the spline nonparamateric regression model*;

   The regression model parameter test was conducted to determine the predictor variables that had an effect on the response variable. In the spline nonparametric regression, the regression model parameter test was carried out after obtaining the regression model with the optimal knot point based on the minimum GCV. There are two stages of parameter testing, namely testing simultaneously and partially. Simultaneous testing of model parameters is a simultaneous regression curve parameter test using the F test. While partial testing is carried out to determine which individual parameters are significant in the model and to the response variable.

5. *interpreting the spline nonparametric regression model.*

   From the spline nonparametric regression model obtained, it can be interpreted that the variables that significantly influence the number of tuberculosis cases that occur in Central Java Province.

3. RESULT AND DISCUSSION

3.1. Analysis of Relationship Patterns of Variables Suspected of Affecting the Number of Tuberculosis in Central Java Province with a Scatter Plot

The scatter plot shows the shape of the relationship pattern between the response variable and the predictor variable. The following is a scatterplot between variables that are thought to affect the number of tuberculosis in Central Java Province.

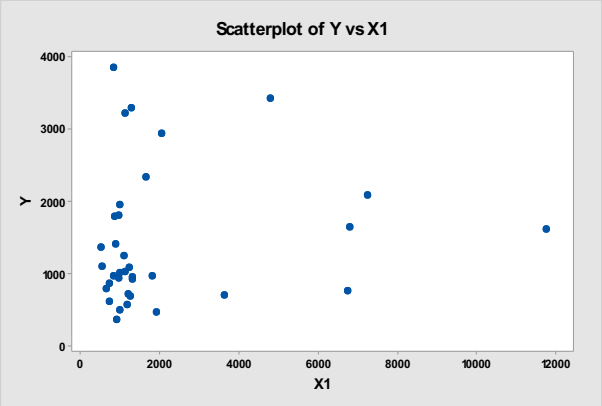

Fig. 1. Pattern of relationship between population density and number of tuberculosis cases.

Based on the scatterplot between Population Density (X1) and the Number of Tuberculosis Cases (Y) which can be seen in Figure 1, it is known that the variable Population Density (X1) and the Number of Tuberculosis Cases (Y) shows a non-linear relationship pattern, so the estimation model uses Nonparametric Regression.

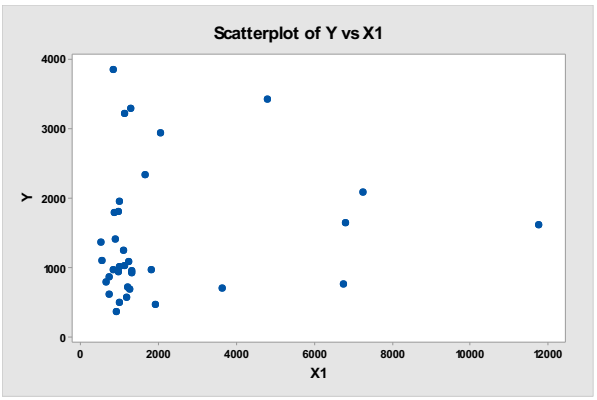

Fig. 2. Pattern of relationship between malnutrition and number of tuberculosis cases.

Based on the Scatter plot between Malnutrition (X2) and the Number of Tuberculosis Cases (Y) as seen in Figure 2, it is known that the variable Malnutrition (X2) and the Number of Tuberculosis Cases (Y) shows that the relationship pattern is not linear, so that the estimation of the model is correct. is Nonparametric Regression. Because the data has a non-linear relationship, this study uses nonparametric regression, namely spline regression to perform modeling. The best Spline Nonparametric Regression Model is generated from the optimal selection of knot points.

3.2. Optimal Knot Point Selection

The knot point is the point of changing data behavior at certain sub-intervals. The best spline nonparametric regression model is obtained from the optimal knot point. To get the optimal knot point, the Generalize Cross Validation (GCV) method is used. The optimal knot point is taken from the minimum GCV value. The following is

the selection of optimal knot points with one knot point, two knot point, three knot point for each variable that is thought to have an effect on the number of tuberculosis cases in Central Java in 2018.

3.2.1. Optimal knot point selection use one knot point

In modeling with one knot point, the minimum GCV value for the one-point knot spline regression which is located on the 13th data is obtained at 0.822659. It was found that the optimal knot point for the Population Density variable (X1) was at the knot point 1.109344 while in the Malnutrition variable (X2) the optimal knot point was at the knot point 0.791262, which was obtained based on the 13th data on each result one knot point per predictor variable.

3.2.2. Optimal knot point selection use two knot point

In the two knot point modeling, the minimum GCV value for the two-point knot spline regression located on the 365th data is 0.789645. It was found that the optimal knot point for the Population Density variable (X1) was at the knot point 1.109344 and 1.243871 while in the Malnutrition variable (X2) the optimal knot point was at the knot point 0.791262 and 0.937259, which were obtained based on data -365 on each result of two point knots for each predictor variable.

3.2.3. Optimal knot point selection use three one knot point

In the three-point knot modeling, the minimum GCV value for the two-point knot spline regression which is located in the 4317th data is 0.674327. It is found that the optimal knot point for the Population Density variable (X1) is at the knot point 1.243871; 1.378397 and 1.512924 while in the Malnutrition variable (X2) the optimal knot point is located at the knot point 0.937259; 1.083256 and 1.229253, which are obtained based on the 4317 data on each of the three points of knots for each predictor variable.

3.2.4. Optimal knot point selection

The best knot point on the spline regression is obtained from the minimum GCV value. Table 1 shows the comparison of the minimum GCV value from one point knot, two point knot, and three point knot. Based on the table, it is known that the minimum GCV value is three point knots of 0.674327.

Table 1. Comparison of GCV value.

Number of Knot Points	GCV
1	0.822659
2	0.789645
3	0.674327

3.2.5. Best spline regression model using optimum knot points

The best spline regression model is obtained from the optimum knot point. The parameter estimates in the best spline regression models are presented in Table 2.

Table 2.Estimated parameters for the best model.

Variables	Parameter	Estimation
Constanta	$\hat{\beta}_0$	0.8144912
x_1	$\hat{\beta}_1$	10.6000112
	$\hat{\beta}_2$	-17.2184020
	$\hat{\beta}_3$	13.5696348
	$\hat{\beta}_4$	-6.8637804
x_2	$\hat{\beta}_5$	-3.8041059
	$\hat{\beta}_6$	19.6146120
	$\hat{\beta}_7$	-27.8858553
	$\hat{\beta}_8$	12.8423556

3.2.6. Testing significance of parameters in spline regression model (current testing)

The purpose of testing simultaneously is to determine the significance of the parameters in the model as a whole. The following is an analysis of the variance of the nonparametric regression models presented in Table 3.

Table 3. ANOVA table of spline regression model.

Source	df	SS	MS	F-value	P-Value
Regression	8	18.570	2.321	3.911	0.0038
Error	26	15.430	0.593		
Total	34	34			

Based on Table 3, it is known that the test statistic using Fcount is 3.91 with a p-value of 0.0038. At a significant level (α) of 5%, the p-value is less than α, so that H0 is rejected or together the predictor variables have a significant effect on the number of tuberculosis cases in Central Java.

3.2.7. Testing significance of parameters in the spline regression model (partial testing)

The test results simultaneously show that at least one parameter of the spline regression model is significant. The results of individual tests with a significant level (α) of 5% can be seen in Table 4.

Table 4. Partial testing result.

Variables	Parameter	Estimation	t	P-Value
Constanta	$\hat{\beta}_0$	0.814	0.083	0.934
x_1	$\hat{\beta}_1$	10.600	2.069	0.048
	$\hat{\beta}_2$	-17.218	-2.178	0.038
	$\hat{\beta}_3$	13.570	1.968	0.059
	$\hat{\beta}_4$	-6.864	-1.701	0.1
x_2	$\hat{\beta}_5$	-3.804	-0.414	0.682
	$\hat{\beta}_6$	19.615	1.300	0.204
	$\hat{\beta}_7$	-27.886	-2.409	0.023
	$\hat{\beta}_8$	12.842	2.403	0.024

From Table 4 with a significant level (α) 5%, it is found that significant parameters are $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\beta}_7$ and $\hat{\beta}_8$ it can be concluded that the population density and malnutrition

variables significantly affect the number of tuberculosis cases. So that the spline regression model on the variables that affect the number of tuberculosis cases in Central Java based on the results of the partial test is as follows:

$$y = 0.814 + 10.600x_1 - 17.218(x_1 + 1.24)_+^1 - 3.804x_2 - 27.886(x_2 + 1.08)_+^1 + 12.842(x_2 + 1.23)_+^1$$

The model is obtained based on the results of significant parameters that have been partially tested and the results from the optimal knot point of each predictor variable based on the minimum GCV value. The optimal knot point for the Population Density variable ($X1$) is at the knot point 1.243871. Whereas in the Malnutrition variable ($X2$) the optimal knot points are located at the 1.083256 and 1.229253 knots, which are obtained based on the 4317th data on each of the three knot points for each predictor variable.

4. CONCLUTION

This research is only limited to three point knots. To get maximum results, it is necessary to experiment for knots of more than three or as many independent variables as used. So this research needs to be refined by adding the number of points of knots in the modeling.

REFERENCES

Eubank, R.L., 1999. Nonparametric Regressionand Spline Smoothing. CRC press.

Härdle, W. and Linton, O., 1994. Applied nonparametric methods. Handbook of econometrics, 4, pp.2295-2339.

Kementerian Kesehatan RI, 2018. InfoDatin Tuberculosis. Kementeri. Kesehat. RI 1.

Kementrian Kesehatan Republik Indonesia, 2019. Profil Kesehatan Indonesia 2018 Kemenkes RI. (2019). Profil Kesehatan Indonesia 2018 [Indonesia Health Profile 2018]. http://www.depkes.go.id/resources/download/pusdatin/profil-kesehatan-indonesia/Data-dan-Informasi_Profil-Kesehatan-Indonesia-2018.pdf[Ind.

RI, K.K., 2018. Provil Kesehatan Indonesia 2017. https://doi.org/10.1002/qj

Suparti, S., Prahutama, A., Rusgiyono, A., Sudargo, S., 2019. Modeling Central Java Inflation and Grdp Rate Using Spline Truncated Birespon Regression and Birespon Linear Model. Media Stat. 12, 129. https://doi.org/10.14710/medstat.12.2.129-139

WHO, 2018. WHO TB burden report 2018, Workplace Health and Safety. https://doi.org/10.1177/2165079915607875

Wulandari, H., Kurnia, A., Sumantri, B., Kusumaningrum, D., Waryanto, B., 2017. Penerapan Analisis Regresi Spline Untuk Menduga Harga Cabai Di Jakarta. Indones. J. Stat. Its Appl. 1, 1–12. https://doi.org/10.29244/ijsa.v1i1.47

Dina Fristantiningtyas Wiliyani Hapsari received the Diploma 3 (A.Md., Stat.) degree in Applied of Statistics from Akademi Statistika (AIS) Muhammadiyah Semarang in 2020. Her research interests include linear regression and spline regression.

Laelatul Khikmah received the Master of Science (M.Si) degree in Applied of Statistics from IPB University in 2017. Her research interests include linear regression and categorical data analysis.