MODELING OF THE PERCENTAGE OF AIDS SUFFERERS IN EAST JAVA PROVINCE USING NONPARAMETRIC REGRESSION APPROACH BASED ON TRUNCATED SPLINE ESTIMATOR

Nadia Murbarani[*1], Yolanda Swastika[*2], Ananda Dwi[*3], Baktiar Aris[*4] and Nur Chamidah[5‡]

[*]Student of Study Program of Statistics, Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia
[1]nadiamurbarani@gmail.com, [2]yolandaswastika9c35@gmail.com, [3]putiananda134@gmail.com, [4]arisbaw.ba@gmail.com
[5]Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia, nur-c@fst.unair.ac.id
[‡]corresponding author

Abstract

Acquired Immune Deficiency Syndrome (AIDS) is a set of symptoms and infection or a syndrome that arise due to damage to the human immune system. AIDS is a health problem that often occurs in developing countries, including in Indonesia. East Java Province was ranked first in the highest number of AIDS sufferers in Indonesia ever reported from 1987-2016 as many as 16,911 people out of a total of 86,780 people. In order to overcome AIDS cases, it is necessary to know the factors that influence it. Data on the percentage of AIDS sufferers and their predictor variables have irregular data patterns or do not match in certain patterns, then the method that can solve these problems is by using the nonparametric regression based on spline truncated estimator. A spline truncated estimator is a segmented polynomial function that has better flexibility because there are knot points indicating changes in data behaviour patterns. The data that used in this study is a secondary data in 2016 obtained from the East Java Provincial Health Office. The results showed that the determination coefficient (R^2) based on the best model of 93.84%. This shows that the variables of health facilities, blood donors, health workers, condom users, and residents of 25-29 years are able to explain 93.84% of the percentage of AIDS sufferers in East Java Province in 2016.

Keywords: AIDS, truncated spline estimator, multi predictor nonparametric regression, east java.

1. Introduction

Acquired Immune Deficiency Syndrome (AIDS) is a set of symptoms and infections or a syndrome that arises due to damage to the human immune system due to infection from the HIV virus. AIDS is a health problem that often occurs in developing countries (Coovadia and Hadingham, 2005), including in Indonesia. The number of AIDS sufferers in Indonesia in 2016 reached 7,491 people, this number has increased from 2015 (AIDS Indonesia, 2016). East Java Province was ranked first in the highest number of AIDS sufferers in Indonesia ever reported from 1987-2016 as many as 16,911 people out of a total of 86,780 people (AIDS Indonesia, 2016).

However, to know relationship between percentage of AIDS suffers and factors that affect it we build statistical modeling with two approaches, i.e., parametric and nonparametric models. The parametric regression model can be used by assuming the form of regression function is known, for examples, linear, quadratic, cubic, etc. If we cannot assume particularly the form of regression function is unknown, for this case, we can use nonparametric regression model approach. In the nonparametric regression, its function is assumed to be smooth, so that the nonparametric regression model approach provides greater flexibility and forms of estimation of its regression function following the data pattern used (Hardle et al., 2004).

There are many researchers who used estimators to estimate the regression function in the nonparametric regression model. Aydin and Yilmaz (2018); Chamidah & Lestari (2016); Lestari et al. (2017); Chamidah et al. (2018); Lestari et al. (2018); Lestari et al. (2019a, 2019b) used spline estimator. Chamidah and Saifudin (2013), Chamidah et al. (2018), Chamidah et al. (2019) studied by using kernel, local linear and local polynomial estimators, respectively. The advantages of these estimators are able to estimate the function at each point such that the model closes to the real pattern, and also no need much data to estimate the model. But, they haven't discussed about relationship between percentage of AIDS suffers and factors that affect it by using nonparametric regression model approach.

Because of the pattern of factors (as predictor variables) that affect to the percentage of AIDS sufferers (as response variable) have irregular patterns or does not follow certain patterns, in this study we use nonparametric regression model approach. The estimator that used in this study is truncated spline. The truncated spline estimator is a segmented polynomial function that has better flexibility because there are knots that indicate changes in data behavior patterns. For choosing optimal knots, we use generalized cross validation (GCV) method and creating R-code for analyzing data.

2. Methodology

2.1   Factors Related to AIDS

Some variables that are thought to affect the percentage of AIDS sufferers in East Java are as follows:
1. Health Facilities
   HIV / AIDS health facilities are a place not only to get a treatment services, but also as a preventative measure to reduce the number of people living with HIV / AIDS. In addition, this health facility also provides insight and space for a

consultation for the community to find out more information about HIV / AIDS (Pratiwi, 2011).

2. Blood Donors
Blood donors are people who donate their blood to help others who need it. Blood donation is usually needed by people who experience illness or accident (Ministry of National Education, 2007).

3. Health Workers
In UU Number 23 of 1992 concerning Health, the health worker is any person who is devoted to health, has knowledge and or skills through education in the health sector that requires authority in carrying out health services.

4. Condom Users
Condoms are contraceptives used during sexual intercourse, to avoid pregnancy. In addition to preventing pregnancy, the use of condoms also serves to avoid someone from sexually transmitted diseases, such as AIDS.

5. Residents Aged 25-29 Years
In the last few years the age group infected with HIV / AIDS is the age of 25-29 years which is a productive age. Most of the sexes of AIDS cases are male dominated and the most dominant age of active sexual groups, namely age 25-29 years (East Java Provincial Health Office, 2015).

2.2   Material and Data

The data that used in this study is data on the percentage of AIDS sufferers along with AIDS-related factors in East Java Province in 2016. The data used by researchers is secondary data sourced from the East Java Provincial Health Office. The observation unit used is 24 administrative areas in East Java Province.
The research variables used in the study are given in Table 1.

Table 1. Research Variables.

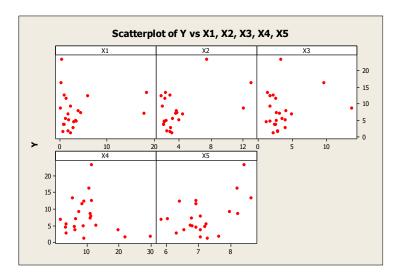| Variables	Explanation
Y	The Percentage of AIDS Sufferers
X_1	Percentage of Health Facilities
X_2	Percentage of Blood Donors
X_3	Percentage of Health Workers
X_4	Percentage of Condom Users
X_5	Percentage of Residents Aged 25-29 Years

2.3   Analysis Method

The analysis steps used in this study are as follows:

1. Estimating the best model for the percentage of AIDS sufferers based on predictor variables with GCV criteria on the order and optimal knot points using the truncated spline method.
2. Analyzing and interpreting the best model for estimating the percentage of AIDS sufferers based on the truncated spline method.

3.   Result and Discussion

Regression analysis is one of the statistical methods used to determine the relationship between response variables and predictor variables. Information about the correlation between response variables and predictor variables can be seen by creating scatterplot diagrams. The pattern of the correlation between the percentage of AIDS sufferers and predictor variables used in the study is as follows:

Figure 1. Scatterplot between Percentage of AIDS Sufferers
versus Predictor Variables

Figure 1 shows that the patterns of the relationship between response variables and predictor variables form an irregular pattern or spreads, so they do not form a particular pattern. So, a suitable approach for this case is a nonparametric regression approach. The estimator used in this study is the truncated spline because it can provide a better flexibility to the characteristics of a function or data, and can be able to handle data characters or functions that are smooth.

3.1   The Selection of Optimum Knot Points

The Optimum knot point selection is done by determining the minimum GCV value. The number of knot points and minimum GCV values for order one spline of 5 predictor variables are given in Table 2.

Table 2. Number of knot points and minimum GCV values

Variable	Number of Knot Points	Knot Points	Minimum GCV
X_1	1	0.248	32.520
	2	0.248 1.248	**25.627**
	3	0.248 2.248 6.248	27.093
X_2	1	7.849	22.665
	2	4.849 6.849	22.557
	3	1.849 2.849 7.849	**19.594**
X_3	**1**	1.120	**5162.281**
	2	1.120 12.120	4076.318
	3	1.120 11.120 13.120	4441.53
X_4	1	10.838	4937.051
	2	11.838 12.838	**5229.726**
	3	10.838 11.838 12.838	3847.046
X_5	**1**	7.860	**4018.469**
	2	5.860 7.860	3795.763
	3	5.860 6.860 7.860	4175.714

Next, based on Table 2, we get the combination of the optimum number of knot are given in Table 3.

Table 3. Combination of the optimum number of knot points.

Variable	Number of Knot Points	Knot Points	Minimum GCV
X_1	2	0.248 and 1.248	
X_2	3	1.849; 2.849 and 7.849	
X_3	1	1.120	22.293
X_4	2	11.838 and 12.838	
X_5	1	7.860	

3.2   The Best Model Estimation

Based on Table 3, we obtain the best model as follows:

$$\hat{y} = 7.28 + 2.11x_1 + 0.30(x_1 - 0.25)^1_+ - 2.57(x_1 - 1.25)^1_+ + 5.55x_2$$

$$-7.91(x_2 - 1.85)^1_+ + 4.88(x_2 - 2.85)^1_+ - 1.97(x_2 - 7.85)^1_+$$

$$+3.47x_3 - 4.67(x_3 - 1.12)^1_+ + 0.62x_4 - 5.57(x_4 - 11.84)^1_+$$

$$+4.96(x_4 - 12.84)^1_+ - 2.81x_5 + 18.82(x_5 - 7.86)^1_+$$

From this model, mean square error (MSE) value is 1.685404 and the R^2 value is 93.84%. It means that the five predictor variables are able to explain 93.84% of the percentage of AIDS sufferers in East Java in 2016.

3.3   The Interpretation of Model Percentage of AIDS Sufferers

Interpretations of the best model given in section 3.2 are as follows:

1. By assuming variables except X_1 is constant, the relationship between the percentage of health facilities (X_1) and the percentage of AIDS sufferers in East Java are as follows:

$$\hat{y} = \begin{cases} 7.28 + 2.11x_1 & ; \quad x_1 < 0.25 \\ 7.205 + 2.41x_1 & ; \quad 0.25 \le x_1 < 1.25 \\ 10.417 - 0.16x_1 & ; \quad x_1 \ge 1.25 \end{cases}$$

For the percentage of health facility availability (X_1) is less than 0.25%, if every one percent increase in health facilities in a year, the percentage of AIDS sufferers will increase by 2.11%. The percentage of health facilities is between 0.25% to 1.25%, if every one percent increase in health facilities in a year, the percentage of AIDS sufferers will increase by 2.41%. The percentage of health facilities is more than 1.25%, if every one percent increase in health facilities in a year, the percentage of AIDS sufferers will decrease by 0.16%. This happens because with the increase in health facilities, AIDS can be detected early, besides that the community becomes more aware of the dangers of AIDS.

2. By assuming variables except X_2 are constant, the relationship between the percentage of blood donors (X_2) and the percentage of AIDS sufferers in East Java are as follows:

$$\hat{y} = \begin{cases} 7.28 + 5.55x_2 & ; \quad x_2 < 1.85 \\ 21.913 - 2.36x_2 & ; \quad 1.85 \le x_2 < 2.85 \\ 8.005 + 2.52x_2 & ; \quad 2.85 \le x_2 < 7.85 \\ 23.47 + 0.55x_2 & ; \quad x_2 \ge 7.85 \end{cases}$$

For the percentage of blood donors (X_2) is less than 1.85%, then if every one percent increase in blood donors in a year, the percentage of AIDS sufferers will increase by 5.55%. For the percentage of blood donors is between 1.85% to 2.85%, then if every one percent increase in blood donors in a year, the

percentage of AIDS sufferers will decrease by 2.36%. For the percentage of blood donors is between 2.85% to 7.85%, then if every one percent increase in blood donors in a year, the percentage of AIDS sufferers will increase by 2.52%. For the percentage of blood donors is more than 7.85%, then if every one percent increases in blood donors a year, the percentage of AIDS sufferers will increase by 0.55%. This happens because blood donors can distribute the HIV virus if it is not processed properly. So the Indonesian Red Cross officer needs to do a more detailed examination of the blood filter process to prevent the spread of AIDS.

3. By assuming variables except X_3 are constant, the relationship between the percentage of health workers (X_3) and the percentage of AIDS sufferers in East Java are as follows:

$$\hat{y} = \begin{cases} 7.28 + 3.47x_3 & ; \quad x_3 < 1.12 \\ 12.510 - 1.2x_3 & ; \quad x_3 \geq 1.12 \end{cases}$$

For the percentage of health workers (X_3) is less than 1.12%, then if every one percent increase in health workers in a year, the percentage of AIDS sufferers will increase by 3.47%. For the percentage of health workers is more than 1.12%, then if every one percent increase in health workers in a year, the percentage of AIDS sufferers will decrease by 1.2%. This happens because if there are many health workers, then handling the AIDS sufferers will be better, and faster.

4. Assuming variables other than X_4 are constant, the relationship between the percentage of condom users (X_4) and the percentage of AIDS sufferers in East Java is as follows:

$$\hat{y} = \begin{cases} 7.28 + 0.62x_4 & ; \quad x_4 < 11.84 \\ 73.229 - 4.95x_4 & ; \quad 11.84 \leq x_4 < 12.84 \\ 9.542 + 0.01x_4 & ; \quad x_4 \geq 12.84 \end{cases}$$

For the percentage of condom users (X_4) is less than 11.84%, then if every one percent increase in condom users in a year, the percentage of AIDS sufferers will decrease by 0.62%. For the percentage of condom users is between 11.84% to 12.84%, then if every one percent increase in condom users in a year, the percentage of AIDS sufferers will decrease by 4.95%. For the percentage of condom users is more than 12.84%, then if every one percent increase in condom users in a year, the percentage of AIDS sufferers will increase by 0.01%. This happens because condoms can also prevent transmission of AIDS and there needs to be more detail information about the correct use of condoms to prevent the spread of AIDS so the AIDS sufferers can be decrease sharply.

5. By assuming variables except X_5 are constant, the relationship between the percentage of the population aged 25-29 years (X_5) and the percentage of AIDS sufferers in East Java are as follows:

$$\hat{y} = \begin{cases} 7.28 - 2.81x_5 & ; \quad x_5 < 7.86 \\ -140.645 + 16.01x_5 & ; \quad x_5 \geq 7.86 \end{cases}$$

For the percentage of the population aged 25-29 years (X_5) is less than 7.86%, then if every one percent of the population is aged 25-29 years a year, the percentage of AIDS sufferers will decrease by 2.81%. For the percentage of the population aged 25-29 years is more than 7.86%, then if every one percent of the

population is aged 25-29 years a year, the percentage of AIDS sufferers will increase by 16.01%. This happens because at the age of 25-29 years is an age that is susceptible to aids disease, where the age is productive age.

4. Conclusion

Based on the results of the analysis that has been carried out, we can conclude that the best estimated model has R^2 value of 93.94%. It means that the five predictor variables are able to explain 93.94% of the percentage of AIDS sufferers in East Java Province in 2016.

References

Aydin, D., & Yilmaz, E. (2018). Modified spline regression based on randomly right-censored data: A comparative study. *Communications in Statistics-Simulation and Computation*, *47*(9), 2587-2611.

Chamidah, N., & Saifudin, T. (2013). Estimation of children growth curve based on kernel smoothing in multi-response nonparametric regression. *Applied Mathematical Sciences*, *7*(37), 1839-1847.

Chamidah, N., Tjahjono, E., Fadilah, A. R., & Lestari, B. (2018). Standard growth charts for weight of children in east java using local linear estimator. In *Journal of Physics: Conference Series* (Vol. 1097, No. 1, p. 012092). IOP Publishing.

Chamidah, N., Gusti, K. H., Tjahjono, E., & Lestari, B. (2019). Improving of classification accuracy of cyst and tumor using local polynomial estimator. *Telkomnika*, *17*(3), 1492-1500.

Chamidah, N., & Lestari, B. (2016). Spline estimator in homoscedastic multiresponse nonparametric regression model in case of unbalanced number of observations. *Far East Journal of Mathematical Sciences*, *100*(9), 1433-1453.

Coovadia, H. M., & Hadingham, J. (2005). HIV/AIDS: global trends, global funds and delivery bottlenecks. *Globalization and health*, *1*(1), 13.

East Java Provincial Health Office. (2015). *Health Profile of East Java Province 2015*. Surabaya: East Java Provincial Health Office.

East Java Provincial Health Office. (2016). *Health Profile of East Java Province 2016*. Surabaya: East Java Provincial Health Office.

Hardle W, Muller M, Sperlich S, and Werwatz A. (2004). *Nonparametric and Semiparametric Models*. New York: Springer.

Lestari, B., Chamidah, N., and Saifudin, T. (2019a). Estimasi fungsi regresi dalam model regresi nonparametrik birespon menggunakan estimator spline dan estimator kernel. *Jurnal Matematika, Statistika, dan Komputasi*, 15(2), 20-24.

Lestari, B., Fatmawati, and Budiantara, I., N. (2017). Estimasi fungsi regresi nonparametrik multirespon menggunakan reproducing kernel Hilbert space berdasarkan estimator smoothing spline. In *Proceeding of National Seminar on Mathematics and Its Aplications*, October 21, 2017, Airlangga University, Surabaya, Indonesia, 243-250.

Lestari, B., Fatmawati, Budiantara, I. N., & Chamidah, N. (2018). Estimation of regression function in multi-response nonparametric regression model using smoothing spline and kernel estimators. In *Journal of Physics: Conference Series* (Vol. 1097, No. 1, p. 012091). IOP Publishing.

Lestari, B., Fatmawati, Budiantara, I. N. (2019b), Spline estimator and its assymptotic properties in Multiresponse nonparametric regression. *Songklanakarin Journal of Science and Technology*, in press.

Ministry of National Education. (2007). *Regulation of the minister of national education*, Jakarta: University of Indonesia.

Pratiwi H. (2011). *Conditions and concepts of drought disaster management in central java articles are presented in the national disaster mitigation and resilience seminar*. Semarang: UNISSULA.