Modeling of The Percentage of AIDS Sufferers in East Java Province using The Multipredictor Nonparametric Regression Approach Based on Spline Truncated Estimator

Nadia Murbarani[1], Yolanda Swaistika[1], Ananda Dwi[1], Baktiar Aris[1], and Nur Chamidah[2]

[1] Student of Study Program of Statistics, Department of Mathematics, Airlangga University, Surabaya, Indonesia
[2] Department of Mathematics, Faculty of Sciences and Technology, Airlangga University, Surabaya, Indonesia

nur-c@fst.unair.ac.id

Abstract. AIDS or Acquired Immune Deficiency Syndrome 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 R^2 value generated from the best model was 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.

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). The correlation pattern of factors that are thought to affect the number of AIDS sufferers has irregular data patterns or does not follow certain patterns, while the methods that can be used using nonparametric regression. Nonparametric regression of its function is assumed to be smooth, so that the nonparametric regression approach provides greater flexibility and forms of estimation of its regression function following the data pattern used (Hardle *et al*, 2004). The estimator that used in this study is a spline truncated. A spline truncated estimator is a segmented polynomial function that has better flexibility because there are knots that indicate changes in data behaviour patterns.

2. Literature Review

2.1. AIDS

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 caused by the HIV (Human Immunodeficiency Virus) (WHO, 2007). Transmission can occur through intimate contact (vaginal, anal, or oral), blood transfusion, contaminated needles, between mother and baby during pregnancy, childbirth, or breastfeeding, as well as other forms of contact with these body fluids.

2.2. Factors Related to AIDS

Here are some variables that are thought to affect the percentage of AIDS sufferers in East Java:

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).

2.3. Nonparametric Regression Analysis

Nonparametric regression is one method that is used to determine the pattern of the relationship between response variables and predictors where the function of the regression curve or the relationship pattern of the two variables is unknown. The regression model is generally as in the following equation:

$$y_i = f(x_i) + \varepsilon_i;\ i = 1,2,\dots,n \tag{1}$$

In nonparametric regression researchers look for the regression curves themselves without being influenced by the subjectivity factor. The nonparametric regression function is assumed to be smooth.

2.4. Spline Truncated Multipredictor Estimator

The spline truncated is one of the accesses to estimate the function $s(x_i)$ in nonparametric regression. A spline truncated is a polynomial cut with different polynomial segments that are joined together on the knots. Segmenting the spline truncated provides better flexibility than ordinary polynomials that allow the truncated spline regression model to adjust to the characteristics of the data. The spline function is truncated with one predictor variable, with the order (q) and the points of the knots $\tau_1, \tau_2, \tau_3, \ldots, \tau_m$ can be expressed in the following form:

$$s(x_i) = \beta_0 + \beta_1 x_i + \cdots + \beta_1 x_i^q + \sum_{k=1}^{m} \beta_{q+k}(x_i - \tau_i)_+^q \tag{2}$$

with $(x - \tau_k)_+^q = \begin{cases} (x - \tau_k)^q, x \geq \tau_k \\ 0 \qquad\quad , x < \tau_k \end{cases}$ (3)

β is a real constant (Eubank, 1999).

The extension of equation (2) with more than one predictor variable is called spline truncated multipredictor which is stated as follows:

$$s(x_i) = \sum_{j=1}^{p} \left(\sum_{h=0}^{q} \beta_{hj} x_{ji}^h + \sum_{k=1}^{m} \beta_{(q+k)} \left(x_{ji} - \tau_{jk} \right)_+^q \right) \tag{4}$$

with $(x_{ji} - \tau_{jk})_+^q = \begin{cases} (x_{ji} - \tau_{jk})_+^q, x_{ji} \geq \tau_{jk} \\ 0 \qquad\qquad , x_{ji} < \tau_{jk} \end{cases}$ (5)

Based on equation (4), the multipredictor linear truncated spline function with m point knots is as follows:

$$s(x_i) = \sum_{j=1}^{p} \left(\beta_{0j} + \beta_{1j} x_{ji} + \sum_{k=1}^{m_j} \beta_{(1+k)j} \left(x_{ji} - \tau_{jk} \right)_+ \right) \tag{6}$$

with $(x_{ji} - \tau_{jk})_+ = \begin{cases} (x_{ji} - \tau_{jk})_+, x_{ji} \geq \tau_{jk} \\ 0 \qquad\qquad , x_{ji} < \tau_{jk} \end{cases}$ (7)

2.5. Optimal Knot Point Selection

The best spline estimator is obtained using optimal knot points. Knot points are common fusion points where there is a change in behavior or curve behavior patterns. Optimal knot points can be obtained using the Generalized Cross Validation (GCV) method:

$$GCV(K_1, K_2, \ldots, K_r) = \frac{MSE(K_1, K_2, \ldots, K_r)}{(n^{-1} tr[I - H(K_1, K_2, \ldots, K_r)])^2} \tag{8}$$

3. Materials and Methods

3.1. Source of 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.

3.2. Research Variables

The research variables used in the study are as follows:

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

3.3. Analysis Step

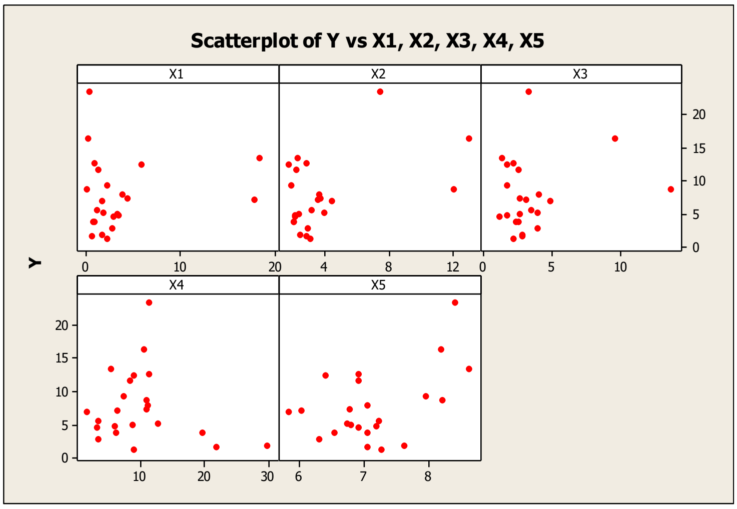
The analysis steps used in this study are as follows:

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 spline truncated method.
2. Analyzing and interpreting the best model for estimating the percentage of AIDS sufferers based on the truncated spline method.

4. 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 with Predictor Variables



Based on Figure 1. it can be seen that the pattern of the relationship between response variables with predictor variables forms an irregular pattern or spreads, so it tends not to form a particular pattern. So, a suitable approach for this type of modeling is a nonparametric regression approach. The estimator used in this study is the spline truncated because it can provide a better flexibility to the characteristics of a function or data, and is able to handle data characters or functions that are smooth (smooth).

4.1. The Selection of Optimum Knot Points

The Optimal knot point selection is done by determining the minimum GCV value produced. The results of the number of orders 1 and the optimum percentage of knot points as follows:

Table 2. The Optimum Knot Points.

Variable	Number of Knot Points	Knot Points	Minimum GCV
X_1	1	0,2484797	32,51998
	2	0,2484797 1,24848	**25,62689**
	3	0,2484797 2,24848 6,24848	27,09358
X_2	1	7,849093	22,66473
	2	4,849093 6,849093	22,55686
	3	1,849093 2,849093 7,849093	**19,59441**
X_3	**1**	1,119843	**5162,281**
	2	1,119843 12,119843	4076,318
	3	1,119843 11,119843 13,119843	4441,53
X_4	1	10,83847	4937,051
	2	11,83847 12,83847	**5229,726**
	3	10,83847 11,83847 12,83847	3847,046
X_5	**1**	7,860509	**4018,469**
	2	5,860509 7,860509	3795,763
	3	5,860509 6,860509 7,860509	4175,714

And the next step to analyze this research used a combination of the number of knot points obtained and have a minimum GCV value. The results of the combination of the optimum number of knot points are as follows:

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

Variable	Number of Knot Points	Knot Points	Minimum GCV
X_1	2	0,2484797 and 1,24848	22,29351
X_2	3	1,849093; 2,849093 and 7,849093	
X_3	1	1,119843	
X_4	2	11,83847 and 12,83847	
X_5	1	7,860509	

4.2. The Best Model Estimation

Based on the results of the combination of optimum knots, the best model in this study was obtained as follows:

$$\begin{aligned}\hat{y} = {} & 7{,}28 + 2{,}11x_1 + 0{,}30(x_1 - 0{,}25)_+^1 - 2{,}57(x_1 - 1{,}25)_+^1 + 5{,}55x_2 \\ & -7{,}91(x_2 - 1{,}85)_+^1 + 4{,}88(x_2 - 2{,}85)_+^1 - 1{,}97(x_2 - 7{,}85)_+^1 \\ & +3{,}47x_3 - 4{,}67(x_3 - 1{,}12)_+^1 + 0{,}62x_4 - 5{,}57(x_4 - 11{,}84)_+^1 \\ & +4{,}96(x_4 - 12{,}84)_+^1 - 2{,}81x_5 + 18{,}82(x_5 - 7{,}86)_+^1 \end{aligned} \qquad (9)$$

From this model, the MSE value was 1.685404 and the R^2 value was 93.84%, which means that the five predictor variables were able to explain 93.84% of the percentage of AIDS sufferers in East Java in 2016.

4.3. The Best Interpretation Model Percentage of AIDS Sufferers

Assuming several values from other predictor variables, the best model can be interpreted to other results, as follows:

1. Assuming a variable other than 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:

$$\hat{y} = \begin{cases} 7{,}28 + 2{,}11x_1 & ; \quad x_1 < 0{,}25 \\ 7{,}205 + 2{,}41x_1 & ; \quad 0{,}25 \le x_1 < 1{,}25 \\ 10{,}4175 - 0{,}16x_1 & ; \quad x_1 \ge 1{,}25 \end{cases}$$

From the model obtained, can be interpreted as follows:

When 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. Assuming variables other than 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:

$$\hat{y} = \begin{cases} 7{,}28 + 5{,}55x_2 & ; \quad x_2 < 1{,}85 \\ 21{,}9135 - 2{,}36x_2 & ; \quad 1{,}85 \le x_2 < 2{,}85 \\ 8{,}0055 + 2{,}52x_2 & ; \quad 2{,}85 \le x_2 < 7{,}85 \\ 23{,}47 + 0{,}55x_2 & ; \quad x_2 \ge 7{,}85 \end{cases}$$

From the model obtained, can be interpreted as follows:

When the percentage of blood donors (X_2) is less than 1.85%, if every one percent increase in blood donors in a year, the percentage of AIDS sufferers will increase by 5.55%. The percentage of blood donors is between 1.85% to 2.85%, if every one percent increase in blood donors in a year, the percentage of AIDS sufferers will decrease by 2.36%. The percentage of blood donors is between 2.85% to 7.85%, if every one percent increase in blood donors in a year, the percentage of AIDS sufferers will increase by 2.52%. The percentage of blood donors is more than 7.85%, if every one percent increase in blood donors in 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. Assuming variables other than 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:

$$\hat{y} = \begin{cases} 7{,}28 + 3{,}47x_3 & ; \quad x_3 < 1{,}12 \\ 12{,}5104 - 1{,}2x_3 & ; \quad x_3 \ge 1{,}12 \end{cases}$$

From the model obtained, can be interpreted as follows:

When the percentage of health workers (*X₃*) is less than 1.12%, if every one percent increase in health workers in a year, the percentage of AIDS sufferers will increase by 3.47%. The percentage of health workers is more than 1.12%, 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:

$$\hat{y} = \begin{cases} 7{,}28 + 0{,}62x_4 & ; \quad x_4 < 11{,}84 \\ 73{,}2288 - 4{,}95x_4 & ; \quad 11{,}84 \le x_4 < 12{,}84 \\ 9{,}5424 + 0{,}01x_4 & ; \quad x_4 \ge 12{,}84 \end{cases}$$

From the model obtained, can be interpreted as follows:
When the percentage of condom users (X_4) is less than 11.84%, if every one percent increase in condom users in a year, the percentage of AIDS sufferers will decrease by 0.62%. The percentage of condom users is between 11.84% to 12.84%, if every one percent increase in condom users in a year, the percentage of AIDS sufferers will decrease by 4.95%. The percentage of condom users is more than 12.84%, 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. Assuming variables other than 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:

$$\hat{y} = \begin{cases} 7{,}28 - 2{,}81x_5 & ; \quad x_5 < 7{,}86 \\ -140{,}6452 + 16{,}01x_5 & ; \quad x_5 \ge 7{,}86 \end{cases}$$

From the model obtained, can be interpreted as follows:
When the percentage of the population aged 25-29 years (X_5) is less than 7.86%, if every one percent of the population is aged 25-29 years a year, the percentage of AIDS sufferers will decrease by 2.81%. The percentage of the population aged 25-29 years is more than 7.86%, 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.

5. Conclusion

Based on the results of the analysis that has been carried out, the conclusions that can be drawn are the best models and the R^2 values generated from the best model of 93.94%. This shows that the five predictor variables are able to explain 93.94% of the percentage of AIDS sufferers in East Java Province in 2016.

References

[1] Coovadia H and Hadingham J 2005 *Global Trends, Global Funds and Delivery Bottlenecks* Globalization and Health *J.J. HIV/AIDS*, 1, 1–10.

[2] East Java Provincial Health Office 2016 *Health Profile of East Java Province 2016* (Surabaya: East Java Provincial Health Office)

[3] East Java Provincial Health Office 2015 *Health Profile of East Java Province 2015* (Surabaya: East Java Provincial Health Office

[4] Eubank R 1988 *Spline Smoothing and Nonparametric Regression* (New York: Marcel Dekker)

[5] Eubank R 1999 *Nonparametric Regression and Spline Smoothing 2nd Edition* (New York: Marcel Deker)

[6] Hardle W, Muller M, Sperlich S, and Werwatz A 2004 *Nonparametric and Semiparametric*

Models (New York: Springer)

[7] Minister of National Education 2007 *Regulation of the Minister of National Education* (Jakarta: University of Indonesia)

[8] 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)

[9] WHO 2007 *Technical working groups for the development of an HIV/ AIDS DIAGNOSTIC support Toolkit*:p. 2