M – 18

Regression Spline Truncated Curve in Nonparametric Regression

Syisliawati [1], Wahyu Wibowo[1], I Nyoman Budiantara[1]
[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia
syisliailasamad123@gmail.com

***Abstract*—**Nonparametric regression is used when regression curve is not known. Nonparametric regression curve is simply assumed to be smooth in the sense of space contained within a particular function. Data expected to find the shape of its estimation without being influenced by subjective factor of the researcher. Thus, the nonparametric regression approach has high flexibility. Nonparametric regression does not assume the shape of regression curve, regression curve is assumed to be contained within a particular function space. Infant deaths are deaths that occur in infants at an interval between time after birth until baby has not been exactly one year old. Magnitude which stated the possibility of babies die after birth until the age of one year per thousand live births is called Infant Mortality Rate (IMR). IMR data based on scatterplot with each predictor exhibits patterns that tend not follow a particular pattern (linear or a certain degree polynomial), so that the corresponding regression model approach is nonparametric spline regression model. Knot used is 1 knot, 2 knots and 3 knots and can be concluded that nonparametric regression model spline best produced is combination knot where significant variable are variable (X_1) The Percentage Of Aid Last Deliveries Performed By Non-Medical Assistance, (X_2) Marriage Percentages In Which Age of Women whom Married is Less Than 17 Years and (X_3) The Average Number Of Household Expenditure Per Capita Per Month.

Keywords: *nonparametrics, spline, knots, IMR*

I. INTRODUCTION

Regression modeling methods are divided into three, parametric regression, nonparametric regression and semiparametric regression. Semiparametric regression is used when one of the regression curve is not known, while others are known [1]. Nonparametric regression curve is only assumed to be smooth in the sense of space contained within a particular function. Data expected to find its shape of estimation, without being influenced by subjective factor of the researcher. Thus, the nonparametric regression approach has high flexibility [2]. Nonparametric regression does not assume the shape of regression curve, regression curve is assumed to be contained within a particular function space e.g. Sobolev Space (Eubank, 1988). In reality, not all of data model can be predicted with parametric regression approach in the absence of complete information about the regression curve shape. In such circumstances, it can be used nonparametric regression approach [3]. Nonparametric regression with spline approach is a method often used. In this study, we will estimate spline regression to model the function of Infant Mortality Rate (IMR) in Indonesia.

Infant deaths are deaths that occur in infants at an interval between the time after birth until the baby has not been exactly one year old. Magnitude which stated the possibility of babies die after birth until the age of one year per thousand live births is called Infant Mortality Rate (IMR). Based on the Indonesian Demographic and Health Survey, the IMR in Indonesia reached 32/1000 infant live births, this number is quite high compared to the standard of the Millennium Development Goals (MDGs), 23/1000 for IMR. Indonesian government is expected to reduce the value of IMR through programs initiated or to identify and resolve the factors that significantly affect the high value of IMR. East Java Province in the same year has IMR up to 32.43 deaths per 1,000 live births. In 2009 the IMR number is decreased to 31.41, while in 2010 the IMR became 29.99 deaths per 1,000 live births. This fact shows that East Java has not been able to reach the target of MDG's [4]. Data IMR based on scatter diagram with each predictor reveal patterns that tend not follow a particular pattern (linear or a certain degree polynomial), so that the corresponding regression model approach is nonparametric regression model. The characteristics of data IMR pattern is partially not have pattern so that the spline method is used to model the data pattern.

II. METHOD

A. Regression Analysis

Regression analysis is statistical analysis used to model the pattern of relationship between predictor variables and response variable. Parametric regression approach is used if the regression curve shapes are known. If relation of pattern data form linear pattern then used parametric linear regression approach. If relation of pattern data form a squares pattern then used quadratic regression approach, and others [2]. Relation of shape pattern can be identified based on past information or scatterplot of data [3].

B. Nonparametric Regression

Non-parametric regression is regression method approach in which the curve of regression function is unknown. The curve function assumed to be contained within a particular function space [2]. Nonparametric regression model is given by the following equation.

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

with y_i is the response, variable, x_i is the predictor variables, $f(x_i)$ is the regression function, where the shape of the curve is unknown and ε_i are normally distributed random error, with zero mean and variance σ^2.

C. Nonparametric Spline Regression

Spline in nonparametric regression has high flexibility and the ability to estimate the behavior of data which tend to be different at different intervals [2]. This ability to estimate the behavior of data is indicated by the truncated function (pieces) which are attached to the estimator and pieces of so-called point knots. Knot point is the point of fusion joint that show changes in behavior patterns functions of different interval. Spline is one kind of piecewise polynomial, polynomial which has segmented nature. The segmented nature provides more flexibility than ordinary polynomials, thus allowing it to adapt more effectively to the local characteristics of a function or data. Spline function of degree p is any function that can generally presented in the following forms:

$$f(x_i) = \sum_{j=0}^{p} \beta_j x_i^j + \sum_{j=1}^{k} \beta_{j+p} (x_i - k_j)_+^p$$

with β_j are real constants, and

$$(x_i - k_j)_+^p = \begin{matrix} (x_i - k_j)^p & ; x \geq k_j \\ 0 & ; x < k_j \end{matrix}$$

If p = 1, 2, and 3 respectively spline linear, quadratic and cubic spline spline and k_j is point knots.

Assuming an error ε_i independent normal distribution with average zero and variance σ^2, then y_i in regression models also normally distributed with an average $f(x_i)$ and variance σ^2. Estimates for the parameter β by using the least squares method, namely by minimizing the sum of squared errors is as follows,

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_j - f(x_i)\right)^2$$

β estimator can be obtained from minimizing

$$Q(\beta) = (y - X\beta)^T (y - X\beta)$$

To obtain the best spline regression estimator, it is necessary to select knots optimal point. Method used is Generalized Cross Validation (GCV). GCV functions as follows:

$$GCV(K_1, K_2, \dots, K_r) = \frac{MSE(K_1, K_2, \dots, K_r)}{\left(n^{-1} tr\left[I - A\left((K_1, K_2, \dots, K_r)\right)\right]\right)^2}$$

Value $(K_1, K_2, \dots, K_r)$ is the point of knots, while the matrix $A(K_1, K_2, \dots, K_r)$ obtained from the equation $\hat{y} = A(K_1, K_2, \dots, K_r)\tilde{y}$ and $MSE(K_1, K_2, \dots, K_r) = n^{-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

III. PROCEDURES

This study uses secondary data from the Central Bureau of Statistics (BPS) in 2011. Observation unit in this study was 38 district/city in East Java in 2011. The variables used are Infant Mortality Rate (IMR), the percentage of aid last deliveries performed by non-medical assistance, marriage percentages in which age of women whom married is less than 17 years, the average number of household expenditure per capita per month, the percentage of infants aged 0-11 months were given breast milk, and the number of health facilities (hospitals and health centers). Step analysis are carried out as follows.

1. Creating a scatter plot between response variable with each predictor variable.
2. Modeling data using spline estimation with various knots (one knot, two knots and three knots).
3. Choosing the optimal knots point with GCV method
4. Establish the best models from the smallest GCV value.
5. Calculate the MSE and R^2 of data.

IV. RESULT AND DISCUSSION

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command

A. Characteristics of Infant Mortality and Factors Affecting

Data infant mortality rate and factors that influence described using mean, standard deviation, minimum and maximum values. Here are the characteristics of infant mortality data and factors that influence are presented in Table 1.

TABEL 1. CHARACTERISTICS OF INFANT MORTALITY AND FACTORS AFFECTING

Variable	Mean	Standard Deviation	Minimal	Maximal
Y	34,18	12,68	20,02	64,19
X_1	8,31	11,19	0,00	44,99
X_2	27,00	13,00	10,07	59,09
X_3	239322	105660	122240	558590
X_4	93,952	4,640	83,920	100,000
X_5	90,11	39,94	13,00	189,00

B. Scatterplot Data Infant Mortality and Factors Affecting

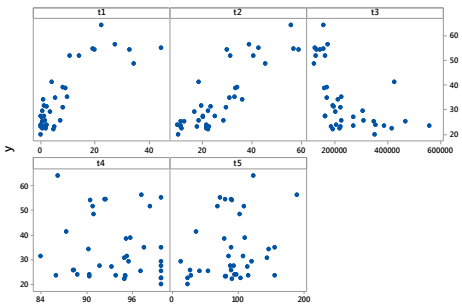

FIGURE 1. SCATTERPLOT DATA

Based on Figure 1, it shows that the relationship between the variables in infant mortality with each factor does not follow specific pattern because plot data is scattered randomly. Thus, the independent variable is nonparametric component.

C. Nonparametric Spline Truncated Regression Model

Nonparametric regression model, truncated spline, with a single point of knots in general are as follows.

$$y_i = \gamma_0 + \gamma_1 t_{i1} + \gamma_2 (t_{i1} - k_1)_+ + \gamma_3 t_{i2} + \gamma_4 (t_{i2} - k_1)_+ + \gamma_5 t_{i3} + \gamma_6 (t_{i3} - k_1)_+ + \gamma_7 t_{i4} + \gamma_8 (t_{i4} - k_1)_+ + \gamma_9 t_{i5} + \gamma_{10} (t_{i5} - k_1)_+ + \varepsilon_i$$

Here are iteration results of the best knot point value based on GCV minimum.

TABLE 2. SINGLE POINT OF KNOT

Knot X_1	Knot X_2	Knot X_3	Knot X_4	Knot X_5	GCV
0.92	11.07	131145.10	84.25	16.59	36.39
1.84	12.07	140050.20	84.58	20.18	36.66
2.75	13.07	148955.31	84.90	23.78	37.06
10.10	21.07	220196.12	87.53	52.51	37.50
11.02	22.07	229101.22	87.86	56.10	37.70

In Table 2, GCV smallest value obtained was 36.39. The value of knot point for variable X_1 is 0.92, variabel X_2 is 11.07, variable X_3 is 131145.10, variable X_4 is 84.25 and variable X_5 is 16.59.

Truncated spline nonparametric regression model with two point knots in general are as follows.

$$y_i = \gamma_0 + \gamma_1 t_{i1} + \gamma_2 (t_{i1} - k_1)_+ + \gamma_3 (t_{i1} - k_2)_+ + \gamma_4 t_{i2} + \gamma_5 (t_{i2} - k_1)_+ + \gamma_6 (t_{i2} - k_2)_+ + \gamma_7 t_{i3} + \gamma_8 (t_{i3} - k_1)_+ + \gamma_9 (t_{i3} - k_2)_+ + \gamma_{10} t_{i4} + \gamma_{11} (t_{i4} - k_1)_+ + \gamma_{12} (t_{i4} - k_2)_+ + \gamma_{13} t_{i5} + \gamma_{14} (t_{i5} - k_1)_+ + \gamma_{15} (t_{i5} - k_2)_+ + \varepsilon_i$$

Here are iterations results of two points based on the best knots GCV minimum.

TABLE 3. TWO POINT OF KNOT

X_1		X_2		X_3		X_4		X_5		GCV
K_1	K_2	K_1	K_2	K_1	K_2	K_1	K_2	K_1	K_2	
11,9	42,2	23,1	56,1	238006,3	531874,7	88,2	99,0	59,7	178,2	31,9
11,0	42,2	22,1	56,1	229101,2	531874,7	87,9	99,0	56,1	178,2	32,0
11,9	43,2	23,1	57,1	238006,3	540779,8	88,2	99,3	59,7	181,8	32,0
11,0	43,2	22,1	57,1	229101,2	540779,8	87,9	99,3	56,1	181,8	32,1
11,9	44,1	23,1	58,1	238006,3	549684,9	88,2	99,7	59,7	185,4	32,1

Based on Table 3, GCV smallest value obtained was 31.9. The value of knot point for variable X_1 are $k1$ amounted to 11,9 and k2 amounted to 42,2. Knots point value for variable X_2 are $k1$ amounted to 23.1 and k2 amounted to 56.1 $k2$. For variable X_3 are $k1$ amounted to 238006.3 and $k2$ amounted to 531874.7. For variable X_4 are $k1$ amounted to 88,2 and $k2$ amounted to 99.0, while for the variable X_5 are $k1$ amounted to 59.7 and $k2$ amounted to 178.2.

Nonparametric regression model with a three-point spline truncated knots in general are as follows..

$$y_i = \gamma_0 + \gamma_1 t_{i1} + \gamma_2 (t_{i1} - k_1)_+ + \gamma_3 (t_{i1} - k_2)_+ + \gamma_4 (t_{i1} - k_3)_+ + \gamma_5 t_{i2} + \gamma_6 (t_{i2} - k_1)_+ + \gamma_7 (t_{i2} - k_2)_+ + \gamma_8 (t_{i2} - k_3)_+ + \gamma_9 t_{i3} + \gamma_{10} (t_{i3} - k_1)_+ + \gamma_{11} (t_{i3} - k_2)_+ + \gamma_{12} (t_{i3} - k_3)_+ + \gamma_{13} t_{i4} + \gamma_{14} (t_{i4} - k_1)_+ + \gamma_{15} (t_{i4} - k_2)_+ + \gamma_{16} (t_{i4} - k_3)_+ + \gamma_{17} t_{i5} + \gamma_{18} (t_{i5} - k_1)_+ + \gamma_{19} (t_{i5} - k_2)_+ + \gamma_{20} (t_{i5} - k_3)_+ + \varepsilon_i$$

Here are the result of three-point iteration based on the best knots GCV value minimum.

TABLE 4. THREE POINT OF KNOT

		X_1	X_2	X_3	X_4	X_5	GCV
	K_1	0.92	11.07	131145.10	84.25	16.59	
1	K_2	11.94	23.08	238006.33	88.19	59.69	28.36858
	K_3	41.32	55.09	522969.59	98.69	174.63	
	K_1	0.92	11.07	131145.10	84.25	16.59	
2	K_2	11.94	23.08	238006.33	88.19	59.69	28.38476
	K_3	42.24	56.09	531874.69	99.02	178.22	
	K_1	0.92	11.07	131145.10	84.25	16.59	
3	K_2	11.94	23.08	238006.33	88.19	59.69	28.39802
	K_3	43.15	57.09	540779.80	99.34	181.82	
	K_1	0.92	11.07	131145.10	84.25	16.59	
4	K_2	11.94	23.08	238006.33	88.19	59.69	
	K_3	44.07	58.09	549684.90	99.67	185.41	28.40616
	K_1	0.92	11.07	131145.10	84.25	16.59	
5	K_2	22.04	34.08	335962.45	91.80	99.20	28.60887
	K_3	24.79	37.08	362677.76	92.78	109.98	

Based on Table 4, GCV smallest value obtained is 24.67.

D. Optimal Knot Point Selection

After getting knots value for each predictor variable, the next step is to compare GCV value of each model to choose which one is the best knots. Here are the smallest GCV value of each point knots results.

TABLE 5. SMALLES GCV VALUE FROM EACH MODELLING

Knot	GCV
1 Knot	36.4
2 Knots	31.9
3 Knots	28.4
Knot Combinations	27.3

Bold –Knot value which has smallest GCV

The smallest GCV value is modeling with knot combinations that is equal to 27,3.

E. Modelling Infant Mortality by Optimal Knot Point

The combination of knots optimum point is

$$\hat{y} = -0.00003 + 0.1569x_{i1} + 0.1524(x_{i1} - 0.9182)_+ + 0.2183x_{i2} + 0.2178(x_{i2} - 11.07)_+ \\ +0.00037x_{i3} - 0.000054(x_{i3} - 13145.1)_+ + 0.000024(x_{i3} - 238006.3)_+ \\ -0.000062(x_{i3} - 522969.6)_+ - 0.14976x_{i4} - 0.1463(x_{i4} - 84.25)_+ \\ -0.118x_{i5} + 0.2211(x_{i5} - 59.69)_+ + 0.0283(x_{i5} - 178.22)_+$$

R^2 value of this model is 89.38 percent. This means that 10.62 percent of IMR are able to be explained by variable percentage of aid last deliveries performed by non-medical assistance, marriage percentages in which age of women whom married is less than 17 years, the average number of household expenditure per capita per month, the percentage of infants aged 0-11 months were given breast milk, and the number of health facilities in Spline Regression Model Truncated with combination knot of optimum knots.

F. Parameter Regression Examination

There are two test parameter estimation to be performed, simultaneously testing and individual testing. Here are the results of simultaneously testing by using F-statistic test.

TABLE 6. ANALYSIS OF VARIANCE

Variance Source	db	*Sum Square*	*Mean Square*	F	P-*value*
Regresi	13	5308.908	408.3775	16.20971	8.23152e-09
Error	24	604.6414	25.19339		
Total	37	5913.549			

F-test value is 16.21 and the F degree of freedom table value for v_1 amounted to 13 and v_2 amounted to 24. Thus, the value of F-test > F_{tabel} so it can be concluded that we rejected H_0 which mean there are at least one predictor variable which has significant impact on the model. Next step is doing individual testing to determine what variables that has significant impact. The results of individual tests are presented in Table 7.

TABLE 7. INDIVIDUAL TESTING PARAMETER RESULTS

Variable	Parameter	Estimator	*t*	P-*value*
Konstan	γ_0	-3.532E-05	-0.285	0.77833
X_1	γ_1	1.569E-01	2.398	0.024632
	γ_2	1.524E-01	2.335	0.028209
X_2	γ_3	2.183E-01	3.776	0.000926
	γ_4	2.178E-01	3.774	0.000931
X_3	γ_5	3.727E-04	3.363	0.002583
	γ_6	-5.437E-04	-3.712	0.001087
	γ_7	2.447E-04	3.809	0.000852
	γ_8	-6.184E-04	-2.534	0.018221
X_4	γ_9	-1.497E-01	-1.563	0.131213
	γ_{10}	-1.463E-01	-1.602	0.122276
X_5	γ_{11}	-1.118E-01	-1.122	0.273009
	γ_{12}	2.211E-01	1.778	0.088093
	γ_{13}	2.829E-02	1.418	0.168963

Individual parameter test results of 14 parameters which are contained in nonparametric spline regression model, 6 parameters are not significant because p-value is greater than α. The significant parameters are $\gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6, \gamma_7$ and γ_8.

G. Testing Residual Assumption

Residual assumptions that must be fulfilled to find the best model are assumption of identical residual, independent and normally distributed. Testing assumptions using identical residual test presented Glejser role is in the following Table

TABLE 7. GLEJSER TESTING RESULTS

Sumber Variasi	db	*Sum Square*	*Mean Square*	F	P-*value*
Regresi	13	47.904	3.685	0.514	0.894
Error	24	172.184	7.174		
Total	37	220.088			

P-value of Glejser testing residual is 0.894, the value is greater than the value of α of 0.05, so can be concluded that H_0 failed to reject it means that heteroskedastisity is not occur in the model so that identical

residual assumption are met. Examination of independent residual assumption use ACF plot is presented in Figure 2 below.

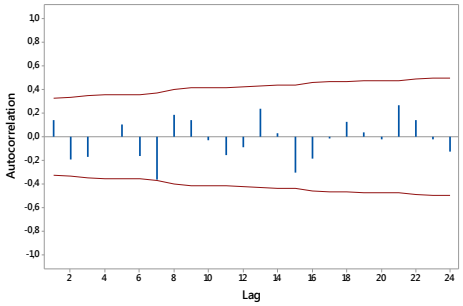

FIGURE 2. RESIDUAL ACF PLOT

Based on ACF plot for residual, there is no significant autocorrelation (ACF) values or out of the upper limit and lower limit. It can be concluded that the assumption of residuals independent are met and no autocorrelation between residuals. Here is residual normal distribution testing using Kolmogorov Smirnov test statistics that are presented in Figure 3.

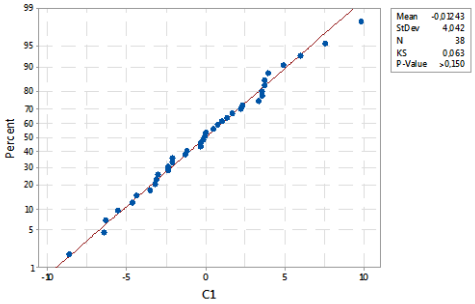

FIGURE 3. RESIDUAL NORMALITY PLOT

Based on normality plot model residual, obtained p-value is greater than 0.150 where the value is greater than the value of α which is equal to 0.05. Testing for normality using Kolmogorov Smirnov statistics with p-value > α can be concluded that H_0 failed to reject, which means residual has normal distribution. So, assumption of normal distribution are met. The test results for assumption of identical residual, independent, and normal distribution are met, then the model used is appropriate and can be done interpretation of the best model.

CONCLUSIONS

Based on modeling results of infant mortality rate in Indonesia using nonparametric regression splines can be concluded that truncated nonparametric spline regression model is the best produced knots with determination coefficient of 89.38 percent where significant variablea are variable (X_1) The Percentage Of Aid Last Deliveries Performed By Non-Medical Assistance, (X_2) Marriage Percentages In Which Age Of Women Whom Married Is Less Than 17 Years and (X_3) The Average Number Of Household Expenditure Per Capita Per Month.

REFERENCES

[1] Budiantara, I.N., (2006), *Model Spline dengan Knots Optimal*. Jurnal Natural FMIPA Universitas Jember, Jember.

[2] Eubank, R., (1988), *Spline Smoothing and Nonparametric Regression***,** Marcel Dekker, New York.

[3] Hardle, W., (1990), *Applied Nonparametric Regression*, Cambrige University Press, Boston.

[4] KemenKes, 2012, *Data Informasi Kesehatan Provinsi Jawa Timur*, Kementrian Kesehatan Republik Indonesia, Jakarta.