World Applied Sciences Journal 36 (3): 450-455, 2018
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2018.450.455

Poverty Modelling in Indonesia Using Generalized Additive Models B-Splines Regression

Arifin M. Kahar, Florencia M. Timothy, Yudhie Andriyana and Neneng Sunengsih

Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia

Abstract: Poverty is a multidimensional problem that is related to social, economic, cultural and other aspects. We then need to know some indicators which can reduce the poverty. In this study, we propose a Generalized Additive Models (GAM) B-Splines technique to modelling poverty in Indonesia using unemployment rate and per capita Gross Regional Domestic Product (GRDP). The result shows that the unemployment rate gives a more variability impact to the poverty, while per capita GRDP has a constant effect to the poverty until 70 million rupiah per year and then decreasing constantly.

Key words: Poverty · Generalized Additive Models · B-Splines

INTRODUCTION

In Indonesia, poverty reduction efforts have been carried out since the beginning of independence. The government has a great attention to the equality and prosperous society as contained in the fourth paragraph of the 1945 Constitution. Development programs implemented to improve the welfare society [1].

Macro poverty in Indonesia is calculated using basic needs approach conducted by Central Bureau of Statistics Indonesia (*BPS*). The basic needs components used by *BPS* consist of food and nonfood items according to urban and rural areas taken based on the results of National Socio-Economic Survey. Beginning 1998 basic needs approach used by *BPS* has been made improvements, where the number of components of basic needs consists of 52 types of food commodities and 51 nonfood commodities in urban areas and 47 commodities in rural areas. With this approach, poverty is shown the economic inability to gain the basic needs of food and nonfood as measured by the expenditure, then the threshold of this expenditure is called poverty line. People are classified as poor people when they have average per capita expenditure per month below the poverty line [1].

Currently, *BPS* reported that the number of poor people in Indonesia about 28.01 million people (10.86 percent) in March 2016. If it compares to 2007, the number of poor people has decreased by 9.16 million people from 37.17 million people (16.58 percent) [1]. Although in the last decade the government succeeded in reducing the poverty but in absolute numbers still quite a lot. This indicates that poverty is still a serious problem in Indonesia. Therefore, the government needs to continue reducing the number of poor people. Thus they can live in equality and prosperously by the mandate of 1945 Constitution. To create the formulation of poverty alleviation policy required indicators that truly caused poverty in Indonesia.

Several studies have examined the effect of some indicators of poverty in some regions of Indonesia using parametric regression. Wijayanto explained that unemployment rate has a significant impact on the percentage of poor people in Central Java using panel regression [2]. Then using linear regression, Wirawan and Arka found that per capita GRDP, mean years of schooling and unemployment rate significantly affect the percentage of poor people in Bali [3]. Based on the results of these study, we propose to modelling percentage of poor people in Indonesia using nonparametric regression. We use unemployment rate (X_1) and GRDP per capita (X_2) as predictors while the percentage of poor people is a response variable (Y).

MATERIALS AND METHODS

Linear Regression: Linear regression is one of the analytical techniques used to determine relationship between response variable expressed $Y_i = (y_1, y_2,...,y_n)$ and predictors $X_i = (x_1 x_2,...x_p)$ expressed as follows [4]:

Corresponding Author: Arifin M. Kahar, Department of Statistics, Universitas Padjadjaran, Bandung, Indonesia.

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \quad (1)$$

with ε_i is random error on the i^{th} observation assumed that ε_i i.i.d. $N(0,\sigma^2)$ and $(\beta_1, \beta_2,...,\beta_p)$ are unknown parameters. In matrix notation is written as follows:

$$y = X\beta + \varepsilon \quad (2)$$

with y is vector of response variable Y, X is matrix predictor and β is unknown parameter vector.

Generalized Linear Models (GLM): GLM is the development of linear regression which can solve several distributions, i.e., distributions that following to exponential familie such as Normal, Binomial, Poisson, Exponential, Gamma, Gaussian and Inverse Gaussian. The GLM can be formulated as follows [5]:

$$Y_i \overset{\text{i.i.d}}{\sim} \text{exponential family}(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i = \beta_0 + \beta_i x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i;\ i = 1,2,...,n \quad (3)$$

with $Y_i = (y_1, y_2, ..., y_n)^T$ and $E(Y_i) = \mu_i$ for I=1,2,...,n with linear predictor $g(\mu_i) = \eta_i$, where $g(.)$ is link function. Then ϕ is a known scale parameter. In matrix notation, can be formulated as follows [5]:

$$\eta = g(\mu) = X\beta + \varepsilon \quad (4)$$

However, GLM only able to modelling response variable derived from exponential family distributions with linear predictors.

Generalized Additive Models (GAM): GAM was first developed by Hastie and Tibshirani [6]. GAM is an additive model which is extension of GLM has assuming that response variable is following exponential family. This method accommodates a nonlinear effect of predictors without we have to know the explicit effect. That nonlinear effect can be approximated by smoothing techniques. GAM can be formulated as follows [7]:

$$Y_i \overset{\text{i.i.d}}{\sim} \text{exponential family}(\mu_i, \phi)$$

$$g(\mu_i) = \eta_i = \alpha + f_1(x_{i1}) + f_2(x_{i2}) + ... + f_p(x_{ip}) + \varepsilon_i^i \quad i = 1,2,..., n \quad (5)$$

with $f_j(.)$ is nonparametric smoothing function on predictor X_j, for j = 1,2, ..., p. In general, GAM function can be expressed as follows:

$$g(\mu) = \eta = \alpha + \sum_{j=1}^{p} f_j(X_j) + \varepsilon \quad (6)$$

B-Splines: B-Splines is a segmented polynomial function on predictors interval were conduct by knot points (Piecewise polynomial) which is locally estimated based on the degree of splines. The degree of splines in B-Splines denoted v, with v = 1 (Linear), v = 2(Quadratic), v = 3(Cubic) and so on. Moreover, domain $x_{\min} - x_{\max}$ divided by u + 1 (t_0, ..., t_u) knots. Total knots of B-Splines are $u + 2v + 1$ and the number of Basis B-Splines are $m = u + v$. Hence, k^{th} B-Splines with degree of splines v and u+1 knots can be expressed as follows [8]:

$$B_k(x;v) = \frac{x - t_k}{t_{k+v-1} - t_k} B_k(x;v-1) + \left(1 - \frac{x - t_{k+1}}{t_{k+v} - t_{k+1}}\right) B_{k+1}(x;v-1) \quad (7)$$

with $B_k(x;v) = \begin{cases} 1, & \text{if } t_k \le x \le t_{k+1} \\ 0, & \text{otherwise} \end{cases}$

In general, if we have the model $Y = f(X) + \varepsilon$, then B-Splines function can be formulated as follows:

$$f(x) \approx \sum_{k=1}^{m} \alpha_k B_k(x;v) \quad (8)$$

with $\alpha = (\alpha_1, ..., \alpha_m)$ is coefficient vector of B-Splines and $B_k(.;v)$ for k = 1, 2, … , $u + v = m$ are the number of basis B-Splines with order v and u + 1 equidistant knots. Objective Function of B-Splines can be expressed as follows [8]:

$$\hat{\alpha} = \text{argmin}_{\alpha} \left\{ \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{m} \alpha_k B_k(x_i;v))^2 \right\} \quad (9)$$

Generalized Additive Models B-Splines (GAM B-Splines): GAM B-Splines is a GAM involving additive functions on a B-Splines basis which can be formulated as follows [9]:

$$g(\mu) = \eta = \alpha + \sum_{k=1}^{m_1} \alpha_k B_k(x_1;v_1) + ... + \sum_{k=1}^{m_j} \alpha_k B_k(x_j;v_j) + \varepsilon;\ j = 1,2, ..., p = \alpha + \sum_{j=1}^{p} \sum_{k=1}^{m_j} \alpha_{jk} B_{jk}(x_j;v_j) + \varepsilon \quad (10)$$

In general, GAM B-Splines is $g(\mu) = \eta = \beta\alpha$ where $\mathbf{B} = (\mathbf{1}, \mathbf{B_1}, ..., \mathbf{B_p})$ is a regressed matrix with size $n \times (1 + \sum_{j=1}^{p} m_j)$ and $\boldsymbol{\alpha}(\boldsymbol{\alpha}, \boldsymbol{\alpha_1}, ..., \boldsymbol{\alpha_p})^T$ is coefficient vector to

be estimated, where basis matrix on each *j* predictor, $\boldsymbol{\alpha}_j$ are coefficient vectors related to $\mathbf{B}_j$ to be estimated, with.

$$B_j = \begin{bmatrix} b_1(x_{1j}) & b_2(x_{1j}) & \cdots & b_{mj}(x_{1j}) \\ b_1(x_{2j}) & b_2(x_{2j}) & \cdots & b_{mj}(x_{2j}) \\ \vdots & \vdots & \ddots & \vdots \\ b_1(x_{nj}) & b_2(x_{nj}) & \cdots & b_{mj}(x_{nj}) \end{bmatrix}$$

and

$$\alpha_j = \begin{bmatrix} a_{j1} \\ a_{j2} \\ \vdots \\ a_{jm_j} \end{bmatrix}$$

$f(x)$ which is defined in equation (8) on each predictor $j = 1,2, ..., p$ can seen in equation (10) i.e. $f_1(x_1) = \sum_{k=1}^{m_1} \alpha_k B_k(x_1;v_1);\ldots;f_p(x_p) = \sum_{k=1}^{m_1} \alpha_k B_k(x_p;v_p)$. This function for each predictor e.g. x_1 can be defined $f_1(x_1) = f_1 = [f_1(x_{11}),\ldots,f_1(x_{1n})]^T$, thus it has $\mathbf{f}_1 = \mathbf{B}_1\boldsymbol{\alpha}_1$. Furthermore, to estimating, we need a contstraints to tackling the identifiability problem on each predictor [10].

$$\sum_{i=1}^{n} f_1(x_i) = 0 \text{ or equal to } 1^{\mathrm{T}}\mathrm{f}_1 = 0$$

where **1** is a vector $n \times 1$ with element 1. To apply a constraint $\mathbf{1}^{\mathrm{T}}\mathbf{f}_1 = \mathbf{1}^{\mathrm{T}}\mathbf{B}_1\boldsymbol{\alpha}_1 = 0$ for all $\boldsymbol{\alpha}_1$ is equivalent to applying $\mathbf{1}^{\mathrm{T}}\mathbf{B}_1 = 0$. Thus, the element of $\mathrm{B}_1\alpha_1$ reduced by each column average. Furthermore, we obtain central matrix column which is defined as $\mathbf{B}_1^* = \mathbf{B}_1 - \mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{B}_1/n$, Thus it will have a set of $\mathbf{f}_1^* = \mathbf{B}_1^*\boldsymbol{\alpha}_1 = \mathbf{B}_1\boldsymbol{\alpha}_1 - \mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{B}_1\boldsymbol{\alpha}_1/n = \mathbf{B}_1\boldsymbol{\alpha}_1 - \mathbf{1}c = \mathbf{f}_1 - c_1$ where $c = \mathbf{1}^{\mathrm{T}}\mathbf{B}_1/n$ is a scalar. The constraints applied will reduce rank of $\mathbf{B}_1^*$ to be m_1-1, i.e. it will have m_1-1 elements that are uniquely estimated [10].

Optimum Knots: In GAM, to obtain optimum knots we can use Generalized Cross Validation (GCV). The optimum knots are obtained when GCV score is minimum. GCV can be formulated as follows [11]:

$$GCV = (k_1,\ldots,k_u) = \frac{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\left(1 - \frac{1}{n}trace(A(k_1,\ldots,k_u))\right)^2} \quad (11)$$

with $A(k_1,\ldots,k_u)$ is matrix that involving knot points $(k_1,\ldots,k_u)$.

Goodness of Fit Model: Akaike's Information Criterion (AIC) found by Akaike which can be use to performing model comparisons on the same data. AIC is a useful method to get the best model. The AIC formula is [12]:

$$AIC = -2l(y;\alpha) + 2trace(H) \quad (12)$$

with $l(y;\alpha)$ is a log likelihood function. A model or distribution is best if it has smallest AIC.

In addition, Eubank mentioned that performance of regression estimator function can be determined by Mean Square Error (MSE) which is formulated as follows [12]:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (13)$$

Furthermore, to know the quality of a model, we can use coefficient of determination (R^2). R^2 score shows the model is able to explain data variability. The higher of R^2 can explain better quality of the model. R^2 score obtained by formula [12]:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (14)$$

RESULTS

The first step in regression analysis is to conduct pre-specification model. The result in Fig 1 shows that there exists a correlation between an unemployment rate to the percentage of poor people and there exists a correlation of per capita GRDP to the percentage of poor people. It is shown by the plot of data that tends to decrease, but the pattern cannot be ascertained by the shape of the relationship. Thus, we can use the nonparametric method to modelling data.

Modelling: Based on the pattern in Fig 1, we propose GAM B-Splines as a nonparametric method to modelling data. Modelling is preceded by identifying the most appropriate exponential family distribution for response variable. The identification can conduct by modelling Generalized Additive Models Location and Shape (GAMLSS), then comparing the AIC score of each distribution [13]. Distribution with the smallest AIC is better to modelling GAM B-Splines. The corresponding distributions to the response variable in this study are

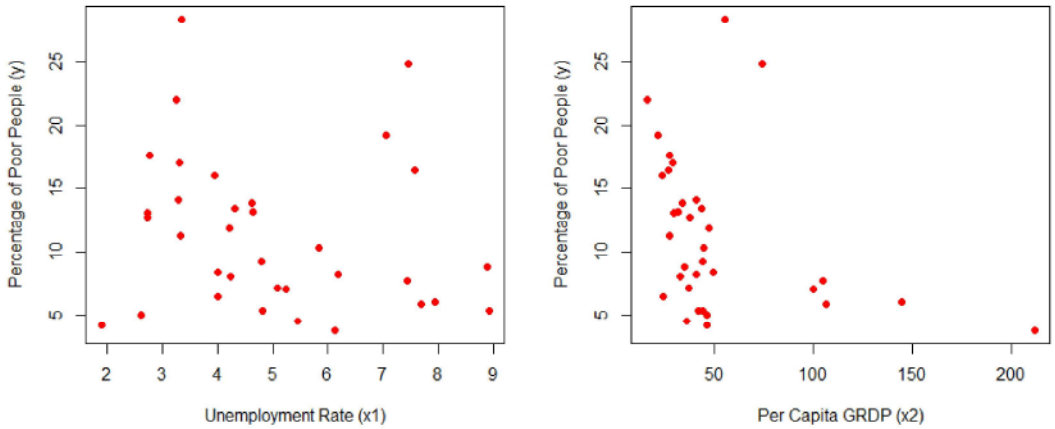

Fig. 1: Scatter plot the relationship between predictor and response variable

Table 1: AIC of corresponding distributions to the response variable

Distributions	AIC
Invers Gaussian	208.145
Log Normal	209.235
Invers Gamma	209.842
Gamma	210.916
Generalized Invers Gamma	211.190
Generalized Gamma	211.195
Weibull 1	215.641
Weibull 3	215.641
Weibull 2	215.649
Normal	222.453
Log Normal Type II	222.460
Eksponential	237.574
Pareto	239.832

Table 2: AIC, MSE and R-Square of GAM B-Splines based on degree of splines

Degree of Splines	AIC	MSE	R-Square
Cubic	193.811	0.010	0.522
Quadratic	198.962	0.011	0.543
Linear	201.758	0.015	0.425

[0, ∞]. Table 1 shows that the Inverse Gaussian distribution is better to modelling GAM B-Splines because it has smallest AIC, i.e., 208.145.

In addition, we propose Anderson Darling Test to testing the distribution was chosen to ensure that it is feasible to use. The null hypothesis is data of the response variable following the Inverse Gaussian distribution against the alternative hypothesis that data of the response variable is not following the Inverse Gaussian distribution [14]. The test will be based on the level of significance α (0, 05). The results obtained p-value 0.695, which means that the response variable following Inverse Gaussian distribution.

Modelling GAM B-Splines based on Gaussian Inverse distribution is conduct by different degree of splines i.e. v = 1(Linear), v = 2 (Quadratic) and v = 3 (Cubic). These models are compared based on AIC, MSE, dan R-Square as below:

Table 2 shows that at degree of splines 3, the model yields the smallest AIC 193, 811 and MSE 0.010, Thus it can be said that GAM B-Splines with v = 3 (Cubic) is the best model. This model produces 14 knot points optimum with a minimum GCV score 0.018. Furthermore, the number of basis obtained are 9 for each predictor. The model can be expressed as follows:

$$\begin{aligned}\hat{y}_i = {} & 0.0120 + 0.0131(B_1(x_1;3)) - 0.0313(B_2(x_1;3)) - \\ & 0.0296(B_3(x_1;3)) - 0.0185(B_4(x_1;3)) - 0.0050(B_5(x_1;3)) - \\ & 0.0200(B_6(x_1;3)) - 0.0182(B_7(x_1;3)) + 0.0039(B_8(x_1;3)) + \\ & 0.0238(B_9(x_1;3)) - 0.0009(B_1(x_2;3)) + 0.0011(B_2(x_2;3)) - \\ & 0.0034(B_3(x_2;3)) + 0.0046(B_4(x_2;3)) + 0.0147(B_5(x_2;3)) + \\ & 0.0255(B_6(x_2;3)) + 0.0372(B_7(x_2;3)) - 0.0493(B_8(x_2;3)) + \\ & 0.0613(B_9(x_2;3))\end{aligned} \tag{15}$$

In Fig 2, the contribution of each predictor can be explained implicitly, but the function of each predictor cannot be explained explicitly. The contribution given by each predictor should be adjusted to the model shape, related to link function applied in the model. Each predictor function will explain the expected value of the percentage of poor people in opposite directions.

The curve generated by unemployment rate (X_1) indicates that there is an exist changed pattern on a particular interval. An unemployment rate at interval 1.9 to 3.2 percent shows the increasing pattern, then decreasing at interval 3.3 to 5.5 percent and increasing again at

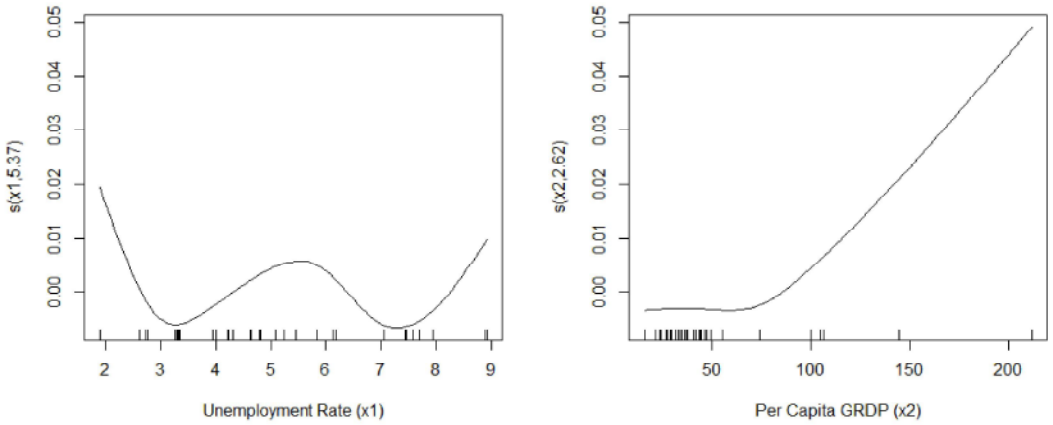

Fig. 2: GAM B-Splines Regression curve

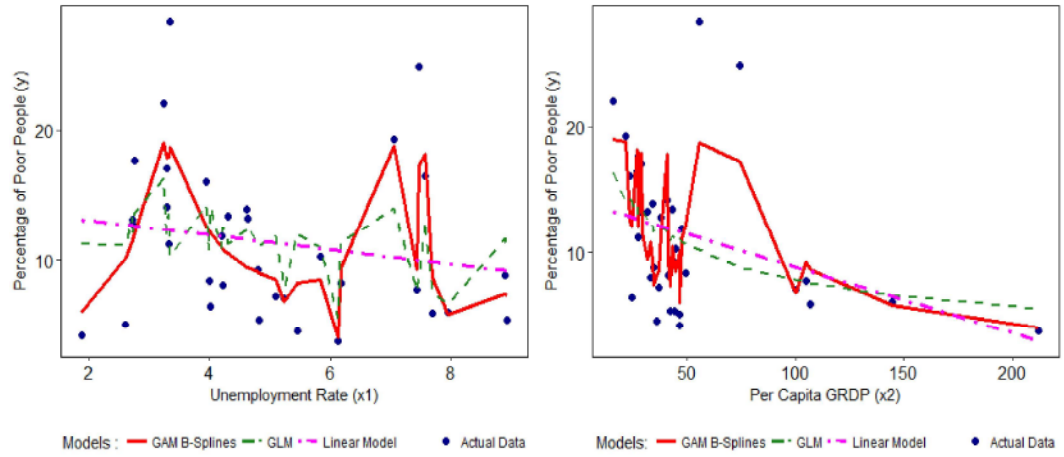

Fig. 3: Comparison of regression curves with actual data

Table 3: AIC, MSE and R-Square of the models

Models	AIC	MSE	R-Square
GAM B-Splines	193.811	0.010	0.522
GLM	209.647	0.023	0.174
Linear Model	222.453	32.123	0.119

interval 5.6 to 7.2 percent, further down again at interval 7.3 to 9 percent. It means that the unemployment rate in Indonesia gives a more variability impact to the percentage of poor people in Indonesia.

The curve generated by per capita GRDP (X_2) explains that when per capita GRDP is less than 70 million rupiah per year, the percentage of poor people has a constant effect. However, when the per capita GRDP is over 70 million rupiah per year, the percentage of poor people decreasing constantly.

Model Comparison: In this study, we propose to compare nonparametric GAMs B-Splines to the parametric model, i.e., GLM and Linear Model as shown Table 3.

Table 3 shows that GAM B-Splines has the smallest AIC and MSE score of 193.8108 and 0.0100 compared to the GLM and Linear models. Meanwhile, R-Square of GAM B-Splines is higher than others. R-square score 0.522 means that variation of the data can be explained in GAM B-Splines is 52.2 percent.

Furthermore, regression curve of the predictors in Fig 3 shows that GAM B-Splines more closely matches or tends to follow the actual data pattern. This result indicates that GAM B-Splines is better than GLM and Linear Model in modelling the percentage of poor people based on unemployment rate and per capita GRDP at 34 provinces in Indonesia 2016.

CONCLUSSION

The conclusions in this study are to modelling percentage of poor people in Indonesia based on unemployment rate and per capita GRDP in 2016 is more feasible using nonparametric GAM B-Splines regression

compared to parametric models such as GLM and Linear Regression. The variations of data can explain the model is 52.20 percent with AIC 193.811 and MSE 0.010. Furthermore, the unemployment rate gives a more variability impact to the poverty, while per capita GRDP has a constant effect to the poverty until 70 million rupiah per year and over that is decreasing constantly.

ACKNOWLEDGEMENTS

We would like to thanks to the Central Bureau of Statistics Indonesia and Master Program of Applied Statistics Universitas Padjadjaran for their supports.

REFERENCES

1. Badan Pusat Statistik, 2016. Penghitungan dan Analisis Kemiskinan Makro Indonesia 2016. Jakarta. Badan Pusat Statistik, pp: 1-13.
2. Wijayanto, R.D., 2010. Analsis Pengaruh PDRB, Pendidikan and Pengangguran terhadap Kemiskinan Kabupaten/Kota Jawa Tengah 2005-2008. Undergraduate Thesis. Universitas Diponegoro, Semarang.
3. Wirawan, I.M.T. and S. Arka, 2015. Analsis Pengaruh Pendidikan, PDRB per kapita, dan Pengangguran terhadap Persentase Penduduk Miskin di Bali. E-Jurnal Ekonomi Pembangunan, 4(5): 546-560, Universitas Udayana.
4. Yan, X. and X.G. Su, 2009. Linear Regression Analysis Theory and Computing. USA. World Scientific, pp: 10-12.
5. Nainggolan, R., Y. Andriyana and A. Bachrudin, 2017. Generalized Additive Model (GAM) Smoothing Penalized Piecewise Linear Basis. World Applied Sciences Journal, 35(11): 2456-2461.
6. Hastie, T.J. and Tibshirani, 1986. Generalized Linear Models. Statistical Science, 1(3): 297-318.
7. Beck, N. and S. Jackman, 1997. Getting The Mean Right is a Good Thing: GAMs. University of California, San Diego.
8. Boor, C.D., 1972. On Calculating B-Splines. Journal Of Approximation Theory 6. Academic Press, Inc, pp: 50-62.
9. Marx, B.D. and P.H.C. Eilers, 1998. Direct generalized additive modeling with penalized likelihood. Computational Statistics and Data Analysis, 28(2): 193-209.
10. Wood, S.N., 2006. Generalized Additive Models: an Introduction with R. Chapman and Hall/ CRC Press, London, pp: 141-155.
11. Krivobokova, T., 2006. Theoritical and Practical Aspects of Penalized Spline Smoothing. Dissertation. Bielefeld University, pp: 14-15.
12. Eubank, R., 1999. Nonparametric Regression and Spline Smoothing. New York: Marcel Dekker.
13. Stasinopoulos, D.M., R.A. Rigby, V. Voudouris, G. Heller and F.D. Bastiani, 2017. Flexible Regression and Smoothing Using GAMLSS in R. Chapman and Hall/CRC Press, London.
14. Folks, J.L. and A.S. Davis, 1981. Regression Models for the Inverse Gaussian Distribution. Statistical Distributions in Scientific Work. NATO Advanced Study Institutes Series (Series C: Mathematical and Physical Sciences), vol 79. Springer, Dordrecht.