World Applied Sciences Journal 35 (11): 2456-2461, 2017
ISSN 1818-4952

DOI: 10.5829/idosi.wasj.2017.2456.2461

# Generalized Additive Model (GAM) Smoothing Penalized Piecewise Linear Basis

*Rosalinda Nainggolan, Yudhie Andriyana and Achmad Bachrudin*

Department of StatisticsUniversitas Padjadjaran, Bandung. 40132, Indonesia

**Abstract:** Mean Years of Schooling (MYS) is an important measurement to determine educational achievement in one region. From six provinces in Java Island, MYS in Central Java has the lowest level. This lowest achievement impacts the condition of socio-economic of an individual and also a region. Knowing the factors influencing the MYS is interesting to be studied. In this research, we proposed to use Generalized Additive Model (GAM) introduced by Hastie and Tibshirani in 1986. This method can be applied to data set where the relationship between response and predictor variables are not linear and more flexible to the distribution of the response as long as a member of the exponential family. In GAM can also control the smooth nature of a curve and hence the curve is protected from rigid and over-fitting properties.

**Key words:** Mean Years Of School · Exponential Family · Smoothing Technique

## INTRODUCTION

Mean years of schooling (MYS) as one of the construction indicators of the Human Development Index (HDI) can reflect the educational attainment of a region, whether it is from infrastructure, access, to the quality of education. MYS is the average number of years of schooling aged over 25 years who have completed formal education (Excluding repeated years) [1]. As an effort to increase the number of MYS in a region, the factors effecting MYS should be acknowledged so the right target policy can be created in order to increase the number of MYS such as the dependency ratio, the ratio of high school students, the ratio of high school students, sex ratio, GDP per capita and percentage of poverty.

From six provinces in Java Island, the HDI construction dimension of Central Java Province shows Mean years of Schooling (MYS) and Expected Years of Schooling (EYS) hits the lowest rank, Life expectancy reaches the 2nd rank and expenditure per-capita is in the 5th rank. The above explanation shows that the dimension of education in Central Java is placed in the lowest level compared to the other five provinces in Java Island. It also means that the government of Central Java Province needs to work hard to overtake the condition specially to meet the standard of MYS suggested by UNDP for 15 years. The data of MYS in Central Java in 2016 towards predictor variables illustrates the different spreading. Based on pre-specification by using scatter plot, the spreading dependency burden ratio towards MYS has declines which indicate that the low level of dependency burden has caused the percentage of MYS became higher. The spreading of teacher-student ratio, students ratio and sex ratio towards MYS shows a dispersed pattern, whereas the GDP per capita accumulates on lower left side. With such data pattern, the smoothing function is needed to make the curve follow the pattern of the existing data.

The pattern of data distribution in such a way is a problem that makes ordinary linear methods insufficient to model MYS data in Central Java in 2016. There should be a method that can involve the smoothing function for the predictors. One of the methods that can control this problem is by using the Generalized Additive Models (GAM) method. According to the formulation of the problem, the purpose of this study is to implement Generalized Additive Models (GAM) to model the Means of Years Schooling (MYS) in 2016 in Central Java Province. The benefit of this implementation is the gaining of statistical knowledge on Generalized Additive Models (GAM) in modelling the relationship of the response variable and some predictors that are strongly presumed to have a nonlinear relationship.

**Corresponding Author:** Rosalinda Nainggolan, Magister Statistika Terapan, Universitas Padjajaran, Jl. Dipatiukur Gd 4. 40132 Bandung, Indonesia.

## MATERIALS AND METHODS

**Generalized Linear Model (GLM):** In some real situation, the distribution of response is not always normally distributed. Nelden and Weddernburn [2] developed a linear model known as Generalized Linear Model (GLM). This model has more distribution alternatives than NLM. Which assumes that exponentially distributed family is considered as a normal distribution [2].

GLM consists of three components, namely:

- The random component, determines the conditional distribution of the response variable, where $Y_i$ is independent. The distribution of $Y_i$ is a member of an exponential family, such as Normal, Binomial, Poisson, Gamma and Inverse-Gaussian.
- The linear predictor, $x_{ij}$ $i$ = 1,..., $n$ and $j$ = 1, ..., $p$ connects the parameter $\eta_i$ corresponds $E(Y_i)$ to the explanatory variable $x_{ij}$ using a linear combination, i.e., $\eta_i = \sum_{j=1}^{p} \beta_j x_{ij}, i = 1,...,n$.
- The link function is the function that transforms the expectations of the response variable $\mu_i = E(Y_i)$, $i$ = 1,...,$n$ with the linear predictors. Hence the GLM model can be written as follows:

$$Y_i \overset{iid}{\sim} exponential\ family(\mu_i, \phi) \quad (1)$$

where $g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}$; $i$ = 1,..., $n$ and $\phi$ is the scale parameter.

Then model (1) can be rewritten in vector form:

$$Y \overset{iid}{\sim} exponential\ family(\mu, \phi)$$
$$\eta = g(\mu) = X\beta$$

**Generalized Additive Model (GAM):** GAM was first developed by Hastie and Tibshirani [3]. GAM is an extension of an additive model. This method accommodates a nonlinearity of predictors. That nonlinear effect can be approximated by smoothing techniques.

GAM is also an extension of GLM assuming that response variable, $Y$, is a member of the following exponential family:

$$f(y_i;\theta_i;\phi) = exp\left(\frac{y_i\theta_i - c(\theta_i)}{a(\phi)} + d(y_i,\phi)\right)$$

where $\theta_i$ is called a canonical parameter and φ is the scale parameter. If $a(\phi) = 1$, $d(y_i, \phi( = d(y_i)$, then the natural form of exponential family distribution is $f(y_i, \theta_i) = \exp [y_i\theta_i - c(\theta_i) + d(y_i)]$, this represents an important relation between $\theta_i$ and $\mu_i$ which is $\theta_i = (c')^{-1} (\mu_i) = g(\mu_i)$ then it is connected to the predictors through the link function, $g(\mu_i)$ and hence the general form of GAM is formulated as follows:

$$Y_i \overset{iid}{\sim} exponential\ family(\mu_i, \phi)$$
$$g(\mu_i) = \eta_i = \beta_0 + f_1(x_{i1}) + \ldots + f_p(x_{ip}) + \epsilon_i; i = 1,2,\ldots n \quad (2)$$

where $f(.)$ is an unknown smooth function of the covariate $x_j$ for $j$ = 1,...,$p$. Equation (2) can be also written, in general, as follows:

$$g(\mu) = \eta = \beta_0 + \sum_{j=1}^{p} f_j(x_j) + \varepsilon$$

**The Piecewise Linear Basis:** Given a univariate variabel, $x$, a basis for piecewise linear functions will be constituted from the linear pieces that is continued in one point called knot. These knots are denoted as $x_m^*$: $m$ = 1,2,...,$k$ where $x_m^* > x^*_{m-1}$. Thus, the formulation of piecewise linear basis is Wood [4]:

$$b_m(x) = \begin{cases} (x - x^*_{m-1})/(x^*_m - x^*_{m-1}) & x^*_{m-1} < x \le x^*_m \\ (x^*_{m+1} - x)/(x^*_{m+1} - x^*_m) & x^*_m < x < x^*_{m+1} \\ 0 & \text{otherwise} \end{cases}$$

for $m = 2,\ldots,k-1$

while

$$b_1(x) = \begin{cases} (x^*_2 - x)/(x^*_2 - x^*_1) & x \le x^*_2 \\ 0 & \text{otherwise} \end{cases}$$

and

$$b_k(x) = \begin{cases} (x - x^*_{k-1})/(x^*_k - x^*_{k-1}) & x \le x^*_{k-1} \\ 0 & \text{otherwise} \end{cases}$$

$b_m(x)$ is zero everywhere, except over the interval between the knots immediately to either side of $x^*_m$. $b_m(x)$ increases linearly from 0 at $x^*_m$ -1 to 1 at $x^*_m$ and then decreases linearly to 0 at $x^*_{m+1}$.

However, such approach is problematic, since the selection of knot location becomes very significant to the fitted the regression model. A procedure to overcome this problem is by controlling the model's smoothness by

adding a 'wiggliness' penalty to the least squares fitting objective function. For example, rather than fitting the model by minimizing

$$\|\boldsymbol{y}-\boldsymbol{X\beta}\|^2$$

It could be fitted by minimizing

$$\boldsymbol{y}-\boldsymbol{X\beta}^2+\lambda\sum_{m=2}^{k-1}\left\{f\left(x_{m-1}^*\right)-2f\left(x_m^*\right)+f(x_{m+1}^*)\right\}^2 \quad (3)$$

where $\lambda \sum_{m=2}^{k-1} \{f(x_{m-1}^*)\ 2f(x_m^*)\ f(x_{m+1}^*)\}^2$ controls wiggliness, λ is a the smoothing parameter.

The coefficients of $f(.)$ are the function values at the knots, i.e., $\beta_m = f(x^*_m)$. Through equation (3), we state that

$$\begin{bmatrix}\beta_1-2\beta_2+\beta_3\\ \beta_2-2\beta_3+\beta_4\\ \beta_3-2\beta_4+\beta_5\\ \vdots\\ \vdots\end{bmatrix}-\begin{bmatrix}1&-2&1&0&.&.\\0&1&-2&1&0&.\\0&0&1&-2&1&0\\.&.&.&.&.&.\\.&.&.&.&.&.\end{bmatrix}\begin{bmatrix}\beta_1\\\beta_2\\\beta_3\\.\\.\end{bmatrix}$$

where

$$\boldsymbol{D}=\begin{bmatrix}1&-2&1&0&.&.\\0&1&-2&1&0&.\\0&0&1&-2&1&0\\.&.&.&.&.&.\\.&.&.&.&.&.\end{bmatrix}$$

hence

$$\sum_{m=2}^{k-1}(\beta_{m-1}-2\beta_m+\beta_{m+1})^2=\boldsymbol{\beta^T D^T D\beta}=\boldsymbol{\beta^T S\beta}$$

where $S = D^TD$.

**Fitting GAM with Piecewise Linear Estimator:** The problem of fitting Generalized Additive Model in Equation (1) by adding piecewise linear basis onto variable *x* is solved by maximization the following penalized likelihood function [5]:

$$l^*=l(\boldsymbol{y};\boldsymbol{\beta})-\tfrac{1}{2}\textstyle\sum_{j=1}^{p}\lambda_j\boldsymbol{\beta}_j^T\boldsymbol{S}_j\boldsymbol{\beta}_j \quad (4)$$

which can be achieved by penalized iterative least squares (PIRLS) technique, as follows:

1. Given the current linear predictor estimate, $\hat{\eta}^{[0]}$ and estimated mean response vector, $\hat{\eta}^{[0]}$,
2. Calculate $W^{[0]}$ and $Z^{[0]}$ with its elements

$$w_i^{[0]}\propto\frac{1}{V(\hat{\mu}_i^{[0]})g'(\hat{\mu}_i^{[0]})^2}$$

and

$$\hat{z}_i^{[0]}=g'(\hat{\mu}_i^{[0]})(y_i-\hat{\mu}_i^{[0]})+\hat{\eta}_i^{[0]}$$

where $var(Yi) = V(\mu_i)\phi$, $g$ is the link function and $W$ is the diagonal matrix such that Wii = wi.

3. Calculate new estimate $\hat{\beta}$ by minimize

$$\|\sqrt{\boldsymbol{W}}\boldsymbol{z}-\sqrt{\boldsymbol{W}}\boldsymbol{X\beta}\|^2+\lambda\,\boldsymbol{\beta^T S\beta}$$

w.r.t. $\beta$ and hence updated estimates $\hat{\boldsymbol{\eta}}^{[t+1]}=\boldsymbol{X}\hat{\boldsymbol{\beta}}^{[t+1]}$ and $\hat{\mu}_i^{[t+1]}=g^{-1}(\hat{\eta}_i^{[t+1]})$. Hence

$$\hat{\boldsymbol{\beta}}^{[t+1]}=\left\{\left[\boldsymbol{X}^T\widehat{\boldsymbol{W}}^{[t]}\boldsymbol{X}-\lambda\boldsymbol{S}\right]^{-1}\boldsymbol{X}^T\widehat{\boldsymbol{W}}^{[t]}\hat{\boldsymbol{z}}^{[t]}\right\}$$

4. Iterate the equation above started from $t = 0$ until convergent, with $|\hat{\boldsymbol{\beta}}^{[t+1]}-\hat{\boldsymbol{\beta}}^{[t+1]}|<\delta$, where δ is the smallest number.

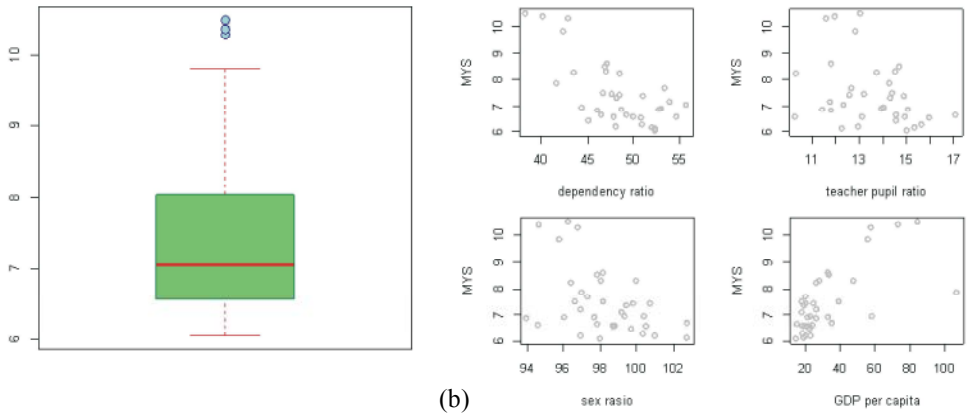
**Choosing the Smoothing Parameter, λ:** The smoothing parameter, λ, controls the trade-off between smoothness of the estimated $f(.)$ and fidelity to the data. If $\lambda$ is too high then the data will be over-smoothed and if it is too low then the data will be under-smoothed. Ideally, it would be good to choose $\lambda$ so that $\hat{f}(.)$ is as close as possible to $f(.)$. A suitable criterion might be to choose $\lambda$ to minimize generalized cross validation score

$$\mathcal{V}_g=\frac{n\sum_{i=1}^{n}(y_i-\hat{f}_i)^2}{[n-tr(\boldsymbol{A})]^2}$$

where $A$ is the influence (hat) matrix, $\boldsymbol{A}=\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{S})^{-1}\boldsymbol{X}^T$

## RESULTS

**Mean Years of School:** The data used in this study is obtained from Central Bureau of Statistics Indonesia (BPS) in 2016. An observation unit in this study focuses on the Regency/City in Central Java Province for 35 districts/cities. MYS is considered as the response variable ($Y$) and supporting factors such as dependency ratio, teacher-pupil ratio at high school, sex ratio and GDP per capita districts/cities in Java Province Central are




(a) (b)

Fig. 1: a) *Boxplot* Mean Years of School; b) *scatter plot* the relationship between explanatory variable and response variable

considered as predictor variables ($X$). The research of MYS has been also studied, for example, by Hudoyo *et al* [6].

Mean years of school (MYS) is an average number of completed years of education aged 25 years and older, excluding years spent for repeating individual grades [1]. The mean of MYS in Central Java is 7.45 years. It means that people aged 25 years and older in Central Java have an average length of schooling from 7 to 8 years. But there are three cities whose value is higher than other districts / cities. Those three cities are Magelang city, Surakarta city and Semarang city, in 10, 29, 10, 37 and 10.49 years.

**Modelling:** Generalized Additive Models (GAM) modelling is conducted firstly by knowing whether or not the response variable follows an exponential family distribution. In accordance with the stages of the research, the proper exponential family of the response variable is chosen via Akaike Information Criterion (AIC) [7] generated by a simple regression model where AIC is formulated as follows:

$$AIC = -2l(y;\beta) + 2\,tr(A)$$

where $l(y;\beta)$ is the likelihood of the model.

The results are shown in Table 1. It shows that the Inverse Gaussian gives the smallest AIC indicating that the response variable follows the Gaussian inverse distribution.

The next step is to determine the knots for each predictors. In this case the authors use equidistant knots is equal to 4 and then we optimize the smoothing parameter $\lambda$

Table 1: AIC criterion of the regression model based types of the response distributions of range [0, ∞]

| Distribusi | AIC |
|---|---|
| Inverse Gaussian | 81, 5428 |
| Weibull | 85, 5407 |
| Gamma | 86, 0569 |
| Log Normal | 86, 2892 |
| Inverse Gamma | 86, 5491 |
| Generalized Gamma | 87, 3044 |
| Generalized Inverse Gaussian | 88, 1313 |
| Normal | 91, 2495 |
| Exponential | 223, 9910 |

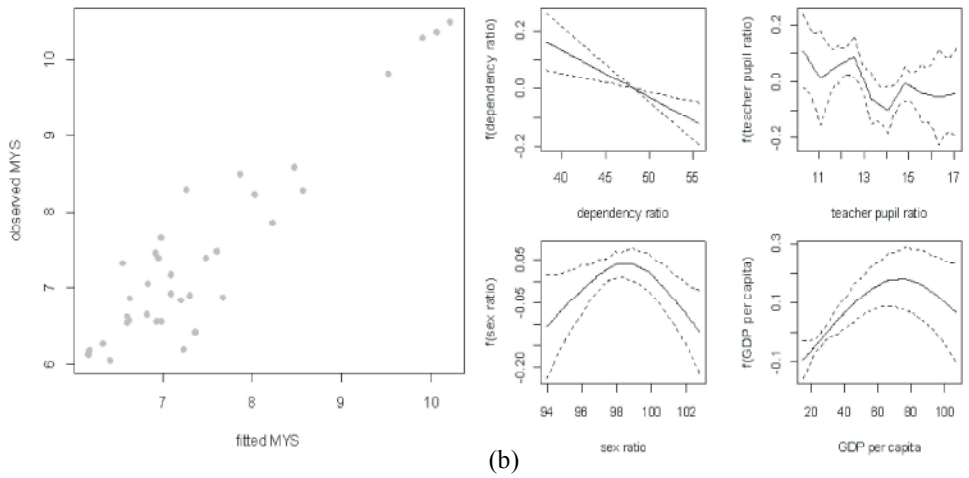
Table 2: MSE, R-Square and AIC of the model

| Metode | MSE | *R-Square* | AIC |
|---|---|---|---|
| (1) | (2) | (3) | (4) |
| LM | 0.5452575 | 0.6322593 | 90.0983 |
| GLM | 0.5369187 | 0.6654928 | 83.9893 |
| GAM | 0.203789 | 0.86255 | 74.0833 |

Besides applying Generalized Additive Models (GAM), we also apply Normal Linear Model (NLM) and Generalized Linear Model (GLM) to the data. The results are shown in the Table 2:

Table 2 demonstrates that GAM is the best model based on model selection criteria. The performances of the regression techniques can be compared via mean squared error (MSE). It appears that the smallest MSE is generated by using the Generalized Additive Model (GAM). R-Square is also higher than the other two methods which is 0, 8626 showing that the four variables included in the model are able to explain 86, 26% of the Mean Years of School.

In Figure 2.a), we show the quality of the proposed method, where the diagonal pattern indicates that the estimators fit nicely to the real/observed variables. While in Figure 2.b) shows that in general, the pointwise




(a) (b)

Fig. 2: a) Fitted MYS v.s Observed MYS ; b) Piecewise linear estimator of each predictors.

confidence interval of predictors is always wider in the tails of the data. Besides that, we shows the existence of linear and nonlinear functions where the linearity is given by dependency ratio and the nonlinearities are given by the rest of the predictors.

## DISCUSSION

Generalized additive model is a flexible method for statistical modelling appropriate for nonliniear covariate in exponential family models. GAM has been widely used in many studies for example Generalized Additive Models for Pair-Copula Constructions [8] Generalized Additive Modelling of Sample Extremes [9] Generalized additive models as an alternative approach to the modelling of the tree height-diameter relationship [10] and the other research.

This study concludeds that the performance of Generalized Additive Models is more optimal than the Generalized Linear Model and Linear Model. The MSE Generalized Additive Models is smaller than MSE Generalized Linear Model and Linear Model. It can be happened because the Generalized Additive Model can overcome the data that has a non-linear pattern and the response variable follows a wider family of distribution (As long as a member of exponential family).

For further research, we suggest to compare some of the smoothing techniques to get the best model and also consider the influence of scale and shape parameters. In the other words, we suggest to involve the variance and the skewness as appear in Generalized Additive Model Location, Scale and Shape (GAMLSS) technique, introduced by Rigby and Stasinopoulos [11].

## ACKNOWLEDGEMENTS

We would like to acknowledge the support of Statistics Indonesia and Master Program in Applied Statistics of Padjadjaran University which help in terms of data and writing of this paper.

## REFERENCES

1. Badan Pusat Statistik, 2015. Indeks Pembangunan Manusia. [http://ipm.bps.go.id/page/ipm]. Accessed October 13, 2017
2. Nelder, J.A. and R.W.M. Wedderburn, 1972. Generalized linear models. Journal of the Royal Statistical Society, A135: 370-384.
3. Hastie, T.J. and Tibshirani, 1990. Generalized Additive Model 4th ed. Chapman and Hall CRC Press, London.
4. Wood, S.N., 2006. Generalized Additive Models: An Introduction with R, Second Edition. Chapman and Hall/ CRC Press, London.
5. Wood, S.N. and H. Augustin, Nicole, 2002. GAMs with integrated model selection using penalized regression splines and applications to environmental modelling. Ecological Modelling, 157: 157-177.
6. Hudoyo, L.P., Y. Andriyana and B. Handoko, 2017. Modeling relationship between mean years of schooling and household expenditure at Central Sulawesi using constrained B-splines (COBS) in quantile regression. American Institute of Physics, 1827: 020021.
7. 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.

8. Vatter, T. and T. Nagler, 2017. Generalized Additive Models for Pair-Copula Constructions. Preprint available at arXiv:1608.01593.
9. Chavez-Demoulin, V. and A.C. Davison, 2005. Generalized Additive Modelling of Sample Extremes. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54: Part 1: 207-222.
10. Adamec, Z. and K. Drápela, 2015. Generalized additive models as an alternative approach to the modelling of the tree height-diameter relationship. Journal of Forest Science, 61(6): 235-243.
11. Rigby, R.A. and D.M. Stasinopoulos, 2005. Generalized Additive Models for Location, Scale and Shape. Journal of the Royal Statistical Society: Series C (Applied Statistics), 54: 507-554.