Application of generalized additive models with P-Spline basis for infant mortality rate data

D Ahya\textsuperscript{1,*}, Miftahuddin\textsuperscript{1}
\textsuperscript{1}Department of Statistics, Faculty of Mathematics and Sciences, Syiah Kuala University, Banda Aceh 23111, Indonesia

E-mail: desfira.ahya@students.stat.unsyiah.ac.id

Abstract. Infant Mortality Rate (IMR) is an indicator of health that reflects the state of the health status of a population. The infant mortality rate in Aceh province period of 2014 has the highest position compared to the previous 4 years that amounted to 15/1000 live births. Generalized Additive Models (GAM) is a method that can handle the condition of the data which response variable does not have to be normally distributed and related with the predictor variable also does not have to be linear. Fitting GAM models with P-spline base can produce a smoother curve and avoid the occurrence of misfitting models. Application of GAM P-spline bases in this research is used data number of IMR in Aceh province period of 2012-2015. The purpose of this research is to get the best fitting model of IMR with GAM models P-spline bases. The results show that the best model of GAM P-spline which can explain IMR in Aceh Province period of 2012-2015 with knots 7 and GCV is 0.0723.

1. Introduction
Infant mortality rate (IMR) is one of the health indicators that determine the level of public health. Reidpath and Allotey [1] stated that IMR is an essential indicator for the health of the entire population, reflecting that structural factors such as economic development, living conditions, social welfare, and environmental quality affecting the health of the entire population have an impact on the IMR. It is because the IMR is sensitive to changes in the level of health and welfare of the community. During 2012-2014, IMR in Aceh Province continues to increase. In 2014, IMR is 15/1000 live birth [2]. The increased of IMR indicates that the more babies who died. It cannot be left alone, because the child's survival determines the quality of human resources in the future.

A research in analyzing of the infant mortality rate in Aceh Province show that factors affecting of infant mortality were percentage of poor households who use clean water, the number of community health centers and health facilities, the latest birth attendants with the help of TBAs and family, and nutritional coverage of toddlers receiving care [3]. In addition, research [4] analyzing the risk factors for low birth weight associated with IMR using the logistic regression analysis method.

Based on the description above, then in this research analyzing IMR using Generalized Additive Models (GAM) method. The GAM model can be more flexible than the additive model because it can accommodate complex data issues, such as nonnormality errors, nonlinear relationships, and autocorrelation variables [5]. A smoothing method in the GAM model that can give better results is a smoothing spline [6]. Smoothing spline approaches are various, among others original spline, spline

*Corresponding author: desfira.ahya@students.stat.unsyiah.ac.id
type M, spline relaxed, weighted spline, B-spline, P-spline, spline truncated and others. In this research used one spline bases that is P-spline. P-spline is one of the most popular smoothing approaches due to its simplicity and flexibility [7]. The purpose of this research is to get the best fitting model of IMR in Aceh Province period 2012-2015 period with GAM models P-spline bases.

2. Literature Review

2.1. Infant Mortality Rate
Infant Mortality Rate (IMR) is the number of infant deaths under one year of age per 1000 live births in a given year [8]. Based on the cause, infant mortality is distinguished by endogenous and exogenous factors. Endogenous infant mortality (neonatal mortality) is the incidence of death occurring in the first month since the baby was born, generally caused by a factor brought by birth, inherited by the parents at the time of conception or obtained from his mother during pregnancy. While exogenous mortality (neonatal post-mortem) is infant mortality occurring between the age of one month or until one year, caused by factors related to environmental influences. The following formula can calculate the calculation of infant mortality.

\[
IMR = \frac{number\ of\ infant\ mortality\ below\ one\ year}{the \ number\ of\ live\ births\ in\ a\ given\ year} \times 1.000
\]

2.2. Generalized Additive Models
Generalized Additive Models (GAM) is the development of Generalized Linear Models (GLM) by replacing linear functions \( \sum_{j=1}^{p} \beta_j X_j \) with additive functions \( \sum_{j=1}^{p} f_j (X_{j}) \) [10]. GAM generalizes the additive model into the form of exponential family distribution. The general form of the GAM model can be formulated as follows:

\[
g(\mu) = \beta_0 + \sum_{j=1}^{p} f_j (X_{j})
\]

where: \( g(\mu) \) = link function \( f_j \) = smoothing function of the predictor variable \( \beta_0 \) = constant coefficient \( X_{j} \) = predictor variable - j

2.3. Penalized Spline (P-spline)
One of the unique aspects of the GAM model is the nonparametric function estimated by the scatter diagram smoothing function. Hastie & Tibshirani (1990) used smoothing spline to produce the best model. Then Eilers and Marx combine spline approach and spline regression into Penalized Spline (P-Spline). P-Spline is the regression obtained based on the least squared with a penalty of roughness. P-spline has many similarities with smoothing spline, but the type of penalty used on P-spline is more general than smoothing spline [11]. The general form of the P-spline can be formulated as follows:

\[
f(x_j) = \beta_0 + \beta_1 x_j + \cdots + \beta_p x_j^p + \sum_{k=1}^{K} (x_j - k_K)_+^p
\]

where: \( \beta_0 \) = constant of P-spline \( \beta_p \) = coefficient of P-spline -p \( x_j \) = predictor variable to-j \( p \) = degree spline \( K \) = number of knots \( k_K \) = value of knots to-K

From the formula the P-spline function indicates that the spline is a broken off polynomial model, but is still continuous on the knots. Knots can be interpreted as a focal point in the spline function so that
the curve formed is segmented at that point. The number of knots is the number of points where functional behavior changes at different intervals.

Knot selection is very important, because it affects the model to be selected. The selection of knots on the P-spline model used Generalized Cross Validation (GCV) criteria. If compared with other methods such as Cross Validation (CV), the GCV method has asymptotically optimal properties [12]. The optimum knot point is marked with a minimum of GCV value. GCV values are defined as follows:

\[
GCV(K) = \frac{MSE(K)}{\left[ n^{-1} tr(I - A(K)) \right]^2}
\]

where: \( n \) = number of data \( I \) = identity matrix \( tr \) = trace \( A = X(X^T X)^{-1}X^T \)

3. Method

This research uses data from the Aceh Health Profile book 2012-2015 period and Aceh Dalam Angka book 2013-2016 period. The observation unit is the district/city in Aceh Province in 2012-2015. The data used consisted of 1 response variable and 9 predictor variables. The amount of data used is 92 observations. The variables are defined as in Table 3.1.

Table 3.1. Variables in this research

| Variable | Explanation | Type of Data |
|----------|-------------|--------------|
| Y        | Number of infant mortality (person) | Numeric |
| X_1      | Childbirth assisted by health personnel (%) | |
| X_2      | ASI Exclusive (%) | |
| X_3      | Low birth weight baby (person) | |
| X_4      | Number of health personnel (person) | |
| X_5      | Healthy house (%) | |
| X_6      | The poor population (%) | |
| X_7      | Percentage of married women under the age of 15 years (%) | |
| X_8      | Year | |
| X_9      | District/City | |

4. Results and discussion

4.1. Descriptive Analysis

The descriptive analysis aims to the representation of the data that has been collected. Here is presented a descriptive analysis like scatter plot matrix from data IMR in Aceh Province 2012 to 2015 in Figure 4.1. Scatter plot matrix is useful for knowing the shape of the histogram, values of Pearson correlation, and regression lines of data. Figure 4.1 show that the histogram shape of response variable Y (IMR) is not symmetrical, and the predictor variable also has not symmetrical histogram shape. The response variable Y (IMR) has a histogram shape extending to the right. The highest Pearson correlation values of the between Y (IMR) and the predictor variables were 0.60, 0.36, 0.14, -0.14, 0.11 and 0.053. The highest linear correlation is Y (IMR) with X_3 (LBWB) which is indicates a strong related of both variables. While the lowest correlation is in Y (IMR) with X_7 (percentage of married women under the age of 15 years) that is equal to 0.053, that indicates is both of variables has weakly related.
4.2. GAM of P-spline bases

GAM model to be formed consists of 9 variables and 2 of them with categorical types of data, that is year and district/city. Both of variables are incorporated into the model as dummy variables and the comparison factors in each variable are year of 2012 and Aceh Barat District. GAM model established using Gamma family and link function log.

The best fitting GAM model with P-spline bases selected based on the optimal knot point. The location of different knot points will result in different models. The optimal knot point is obtained based on minimum GCV criteria. The following table shows the optimal knot points with minimum GCV values.

| Knot | GCV   |
|------|-------|
| 5    | 0.0767|
| 6    | 0.0726|
| 7    | **0.0723**|
| 8    | 0.0740|
| 9    | 0.0743|
| 10   | 0.0746|

Table 1 show that the optimal knot is to use point knot 7 with minimum of GCV value is 0.0723. So the best model fitting of GAM based on P-spline using knot 7. The complete model of GAM based P-spline from Infant Mortality Rate in Aceh Province is as follows:

![Scatterplot matrix of IMR in Aceh Province in 2012-2015 period](image)
\[ g(\mu) = \beta_0 + \sum_{j=1}^7 f(x_j) + \text{factor}(x_j) + \text{factor}(x_j) \]

\[ = 3.9188 + f(x_1) + f(x_2) + f(x_3) + f(x_4) + f(x_5) + f(x_6) + f(x_7) + 0.2304x_8(2013) + \\
0.367x_9(2014) + 0.2131x_10(2015) - 0.3304x_11(2) + 0.4253x_12(3) - 0.7750x_13(4) - \\
0.2488x_14(5) - 0.8382x_15(6) + 0.6851x_16(7) - 0.3099x_17(8) - 0.8441x_18(9) + 0.6157x_19(10) + \\
0.3214x_20(11) - 1.0858x_21(12) - 0.3075x_22(13) + 0.9105x_23(14) - 0.91x_24(15) - \\
0.4449x_25(16) - 0.6540x_26(17) - 0.4713x_27(18) + 0.8232x_28(19) - 0.8005x_29(20) - \\
2.0693x_30(21) - 0.4679x_31(22) - 1.5131x_32(23) \]

### 4.3. Visualization

Based on the model that has been established, the visualization of GAM based on P-spline model is as follows.

![Visualization of GAM based on P-spline](image)

**Figure 4.2.** Visualization of GAM based on P-spline

Figure 4.2 is a plot of the best GAM model based on P-spline bases using knots 7. The shaded area of gray is a 95% confidence interval. Then the y-axis of each plot is a smooth component which is useful for view residual spread and the general shape of the curve. From the seven plots smoothing, the curve of smoothing are different, where the smoother plots are X₂ (ASI exclusive), X₅ (healthy house), X₆ (the poor population), and X₇ (percentage of married women under 15 ). In the X₄ plot (number of health workers) there are a number of health workers far different from others, that indicates there is one district/city with a large number of health workers. While other districts/cities have the number of health workers with an adjacent number. Similarly, plot X₆ there are some poor people with a small amount. While plots X₁, X₂, X₃, X₅, and X₇ have the same criteria, which has a value that is not much different from the others. While the visualization of variables with categorical data types are as follows.
Figure 4.3. The curve of IMR based on (a) year; (b) district/city

Figure 4.3 (a) shows the Infant Mortality Rate continues to increase from year of 2012 to 2014, but in 2015 Infant Mortality Rate is declines. It can be seen also that the highest Infant Mortality Rate in 2014 with the year used as a comparison is 2012. And year of 2013 and 2015 have the value of Infant Mortality is almost the same. Figure 4.3 (b) shows that the estimated infant mortality rate in each district varies. By using the Aceh Barat as comparison factor, the highest infant mortality rate was in the district-14 (Bireuen City) and the lowest was the district-21 (Sabang). While the infant mortality rate approaching to the comparison factor of Aceh Barat District is the 2nd, 5th, 8th, 13th, 16th, 17th, 18th and 22th districts. Each of these districts are Aceh Barat Daya, Aceh Selatan, Aceh Tengah, Bener Meriah, Langsa, Lhokseumawe, Nagan Raya, and Simeulue.

5. Conclusions
1. The best-fitting model of GAM P-spline for Infant Mortality Rate data in Aceh Province period of 2012-2015 using a knot point 7 with minimum GCV is 0.0723.
2. GAM model on P-spline bases with comparative factor year of 2012 and Aceh Barat District is:

\[ g(\mu) = \beta_0 + \sum_{j=1}^{7} f(x_j) + \text{factor}(x_0) + \text{factor}(x_j) \]

6. References
[1] Reidpath, D. D., and Allotey, P. 2003. *Journal of Epidemiology and Community Health*, 57(5), 344-346.
[2] Dinas Kesehatan Aceh. 2015. *Profil Kesehatan Provinsi Aceh Tahun 2015*.
[3] Harahap, A.R. 2014. Pemodelan Angka Kematian Bayi di Provinsi Aceh Dengan Menggunakan Generalized Poisson Regression. *Skripsi*. Universitas Syiah Kuala, Banda Aceh.
[4] Jackson, dkk. 2008. *Journal of the Georgia Public Health Association*, Vol. 1 No. 1.
[5] Miftahuddin. 2017. *Assessment of Sea Surface Temperature in the Indian Ocean using Generalized Additive Models*. Prosiding SEMIRATA 2017.
[6] Hastie, T. and Tibshirani, R. 1990. *Generalized Additive Models 4th ed*. Chapman and Hall.
[7] Hall, P. and Opsomer, J. D. 2005. *92*:1-105.
[8] WHO. 2012. World Health Organization, Amerika.
[9] FEUI. 2010. *Dasar-Dasar Demografi*. Salemba Empat, Jakarta.
[10] Hastie, T. and Tibshirani, R. 1986. *Statistical Science*, 1(3):297-318.
[11] Ruppert, D., Wand, M.P., and Carrol, R.J. 2003. *Semiparametric Regression*. Cambridge
[12] Wahba, G. 1990. *Spline Models for Observational Data*. SIAM Pensylvania.