Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

Multivariate Adaptive Regression Spline (MARS) Model On Dengue Hemorrhagic Fever (DHF) Sufferers In Semarang

D.R.S. Saputro[1], D.H Puspitaningrum[2], N.A. Kurdi[3], Respatiwulan[4]
[1,2,3]Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia
[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Sciences, Universitas Sebelas Maret, Indonesia

E-mail : dewiretnoss@staff,uns.ac.id, dianheppy@yahoo.com, arfa@staff.uns.ac.id, respatiwulan@staff.uns.ac.id

Abstract. Regression model examines a functional relationship between response variables and predictor variables. If data are non-patterned and no a priori model information about the retrogression curve is available, a nonparametric regression model called Multivariate Adaptive Regression Spline (MARS) can be used. MARS is the combination between spline and Recursive Partitioning Regression (RPR), and therefore it can yield continuous estimation of regression functions. Three components constructing MARS model include basis function, knot, and interaction. Recursive partitioning approximates an unknown function using a developed basis function. Dengue Hemorrhagic Fever (DHF) is one of health problems of which incidence shows an increase year after year. DHF analysis is carried out on survival period and can be modeled using MARS. Survival period is defined as individual's probability function to survive in certain time; in this case, individual is fully recovered. The present research aims at finding out a model of DHF sufferers' survival period using MARS and its influencing factors. Data of the research include the 2013 medical record data obtained from Semarang Department of Health. The research results in survival period model (MARS) as well as its influencing factors, such as age (X_1), sex (X_2), trombocyte (X_3), hemoglobin (X_4), hematocrit levels (X_5), and immunologic response (X_6).

1. Introduction

Regression analysis denotes a set of statistical techniques underlying statistical inference on the relationship between the measured variables. The construction of a regression model depends on its goals; several goals, however, frequently overlap. The purposes of regression analysis fall into four categories; prediction, variable selection, model specification, and parameter estimation. Regression analysis seeks to examine the functional relationship between response variables and predictor variables, either parametrically or nonparametrically. Parametric regression refers to a

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

method used to find out the pattern of a relationship between independent variables and dependent variables. Its regression curve is assumed by the researcher. In contrast, nonparametric regression is applied with regards to the nonpatterned data and the absence of a priori model information about the regression curve (Eubank [1], Hardle[2]). According to Budiantara [3], this produces highly flexible and objective results.

Methods of nonparametric regression include spline and Recursive Partitioning Regression (RPR). The former presents a drawback in determining manually the number and the location of knots, while the latter has disjoint functions leading to discontinuity at the knot. In 1990, Friedman [4] developed Multivariate Adaptive Regression Spline (MARS), a nonparametric regression method to deal with the aforementioned drawbacks.

MARS is a combination between spline and recursive partitioning, and therefore it can result in continuous estimation of regression functions. Recursive partitioning approximates an unknown function using a developed basis function. The implementation of MARS on time series data was conducted by Lawless [5]. The present research examines MARS model and survival period modeling using MARS on DHF sufferers in Semarang.

2. Research Method

The present study belongs to a theoretical research on MARS and its implementation on survival period of DHF sufferers in Semarang. Data and research procedures are explained as follows:

2.1 Data. The data of the research include the 2013 medical record data of 97 DHF patients obtained from Semarang Department of Health. The dependent variable (Y) used refers to Nisa' and Budiantara's research [6], involving survival period—DHF patients' length of stay in hospitals until full recovery. Meanwhile, the independent variables include such medical record data as age (X_1), sex (X_2), trombocyte (X_3),

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

hemoglobin (X_4), hematocrit levels (X_5), immunologic response (X_6), and the incidence of bleeding (X_7).

2.2 Research Procedures. The research procedures included estimating distribution of survival period, obtaining martingale residuals through the estimation of survival function and the cumulative hazard function, indentifying the pattern of a relationship between martingale residual and independent variables, and modeling survival period using MARS.

3. Survival Function

Survival function is defined as the probability that an individual can survive to time t (Lawless [7]). If T represents random variable of individual's survival period in interval $[0, \infty)$, then the survival function $S(t)$ can be expressed:

$$S(t) = P(T > t) = 1 - P(T \leq t) = 1 - \int_0^t f(t)\, dt = 1 - F(t), \quad (3.1)$$

, where $f(t)$ is the probability density function and $F(t)$ is the cumulative distribution function. In addition, the relationship between $f(t)$ and $S(t)$ is obtained:

$$f(t) = \frac{dF(t)}{dt} = \frac{-dS(t)}{dt} = -S'(t). \quad (3.2)$$

4. Cumulative Hazard Function

Lawless [7] defines hazard function as a probability that an individual during interval t and $t + \Delta t$ (short interval). If the individual survives until time t, then hazard function $\lambda(t)$ is denoted:

$$\lambda(t) = \lim_{\Delta t \to 0} \left[\frac{P(t \leq T < t + \Delta t | t \geq T)}{\Delta t}\right] = \lim_{\Delta t \to 0} \left[\frac{1}{\Delta t} \frac{F(t+\Delta t) - F(t)}{1 - F(t)}\right] = \frac{f(t)}{S(t)}. \quad (4.1)$$

In reference to equation (3.2) and (4.1), the following function is obtained:

$$\lambda(t) = \frac{-S'(t)}{S(t)} = \frac{d(-\ln S(t))}{dt}. \quad (4.2)$$

The cumulative hazard function is defined as:

$$\Lambda(t) = \int_0^t \lambda(t)\, dt = \int_0^t \frac{d(-\ln S(t))}{dt}\, dt = -\ln\big(S(t)\big). \quad (4.3)$$

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

5. Martingale Residual

Martingale residual functions as a dependent variable in MARS modeling (Kriner [8]). It is defined as:

$$M_i(t) = N_i(t) - \Lambda_i(t), \quad (5.1)$$

, where $M_i(t)$ is the martingale residual of the i^{th} data, $N_i(t)$ has value of 1 if in the i^{th} data a DHF patient is fully recovered and value of 0 if the patient is not, and $\Lambda_i(t)$ depicts the cumulative hazard function of the i^{th} data.

6. Mutivariate Adaptive Regression Splines (MARS)

As previously stated, MARS refers to a modeling algorithm which combines a nonparametric variable transformation and a recursive partitioning scheme. Such algorithm generates a spline basis function comprising truncated-power splines and selects knot using stepwise regression model. MARS was first introduced by Friedman [4] as a new method constructing prediction models in an accurate way for continuous and binary dependent variables. Several components to consider in constructing MARS model involve:

a. Basis Function (BF). Basis function is a set of functions used to explain a relationship between dependent variables and independent variables. Its maximum value used is twice-four times as many as the independent variables.
b. Knot. Knot is an independent variable value which marks the end of a basis function and the beginning of another. Minimum number of observations between knots (MO) is 0, 1, 2, and 3.
c. Interaction. Interaction is, by definition, the cross product of two interrelated independent variables. Friedman [3] specified a maximum allowable degree of interaction (MI) (1, 2, or 3).

MARS model applied is:

$$\hat{y} = \alpha_0 + \sum_{m=1}^{M} \alpha_m B_m(x); \quad m = 1,2,3,\dots,M \quad (6.1)$$

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

, where $\hat{y}$ is the predicted survival period, α_0 depicts the regression constant of basis function, , α_m is the coefficient of the m^{th} basis function a, M is the maximum basis function, the basis function $B_m(x) = \prod_{k=1}^{km}\left[s_{km}\left(x_{v(k,m)} - t_{km}\right)\right]$, km represents the degree of interaction, s_{km} is the indicator which takes on a value of +1 or -1 and indicates the right or left sense of the associated function, $x_{v(k,m)}$ is independent variable, and t_{km} is the knot value of independent variable $x_{v(k,m)}$.

7. Results and Discussion

7.1 Distribution Estimation. Distribution estimation is conducted on survival period of DHF patients using Anderson-Darling. The results of such test applied in various distributions indicate that the three-parameter log-logistic distribution (LLD3) is the appropriate distribution for survival period since the distribution has p-value of > 0.05 and the smallest value of the Anderson-Darling test. The LLD3 has distribution function:

$$f(t) = \frac{\left(\frac{\beta}{\gamma}\right)\left(\frac{t-\theta}{\gamma}\right)^{\beta-1}}{\left[1+\left(\frac{t-\theta}{\gamma}\right)^{\beta}\right]^2}, \quad 0 < t < \infty \tag{7.1}$$

and its cumulative distribution function is:

$$F(t) = \frac{(t-\theta)^{\beta}}{\gamma^{\beta}+(t-\theta)^{\beta}}. \tag{7.2}$$

Survival function is obtained from equation (7.1):

$$S(t) = 1 - F(t) = \frac{\gamma^{\beta}}{\gamma^{\beta}+(t-\theta)^{\beta}}. \tag{7.3}$$

The cumulative hazard function can be obtained from equation (7.3):

$$\wedge(t) = -\ln\left(S(t)\right) = -\ln\left(\frac{\gamma^{\beta}}{\gamma^{\beta}+(t-\theta)^{\beta}}\right). \tag{7.5}$$

7.2 Application. DHF is one of diseases mostly found in either tropical or subtropical areas in the world, particularly during humid wet season. The Aedes aegypti-borne disease can lead to an outbreak and deaths. The mortality rate for DHF in Central Java has increased year after year. A number of 35 regencies/ cities in the province

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

have been ever included as dengue transmission areas, one of which is Semarang city (Department of Health [9]).

In reference to data of Semarang Department of Health, there were 1,844 cases notified with 21 deaths during the period of January-August 2013. The high mortality rate is influenced by patient's survival. Individual patient's survival is detected from the survival period which is influenced by several variables. The pattern of a relationship between the survival period and the influencing can be estimated using both parametric regression and nonparametric regression.

7.3 Survival Function and Cumulative Hazard Function. Survival function and cumulative hazard function are used to find out the relationship between the survival period, and the probability of survival and the death risk. Table 1 denotes the results of the estimation of the survival function and the cumulative hazard function, in which $\hat{\beta} = 3.3047, \hat{\gamma} = 3.8498$ and $\hat{\theta} = 0.4023$

Table 1 The Estimation of the Survival Function and the Cumulative Hazard Function

Survival period (day)	$S(t)$	$\Lambda(t)$	Survival Period (day)	$S(t)$	$\Lambda(t)$
1	0.9978	0.0021	10	0.0466	3.0666
2	0.9482	0.0532	11	0.0340	3.3810
3	0.7858	0.2410	13	0.0195	3.9374
4	0.5557	0.5875	14	0.0152	4.1855
5	0.3574	1.0289	15	0.0121	4.4168
6	0.2250	1.4919	18	0.0066	5.0289
7	0.1443	1.9360			

It is clear from the above table that the longer a DHF sufferer is hospitalized, the lower the probability of survival and the bigger death risk will be.

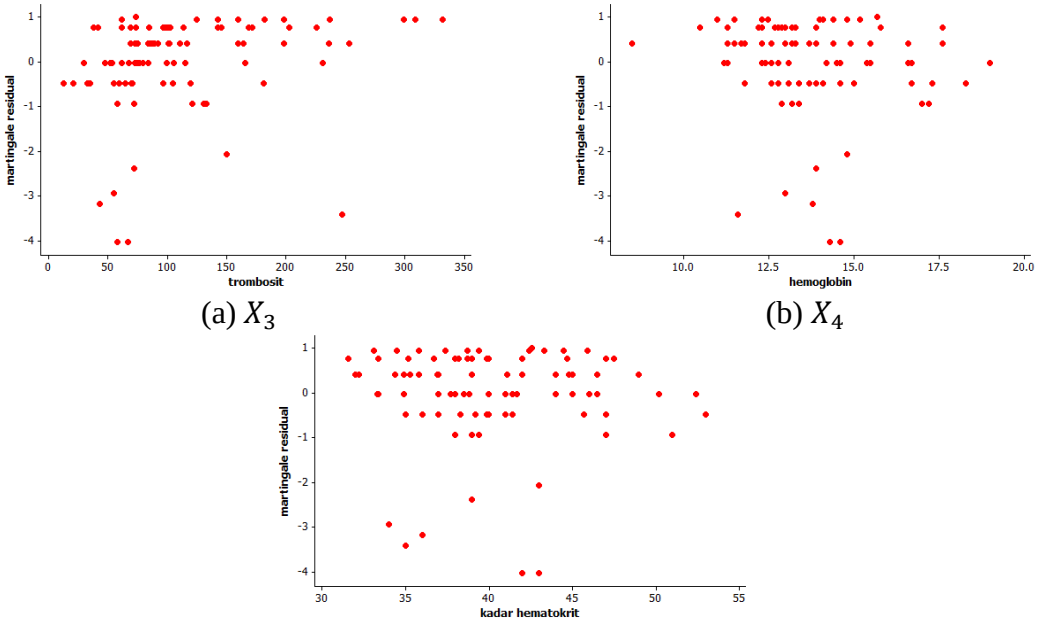
7.4 Martingale Residual. Martingale Residual is utilized as a dependent variable to obtain residual approaching zero in MARS model. Based on equation (5.1), martingale residual $M_i(t)$ is obtained, as shown by Table 2.

Table 2 Martingale residual

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

Survival Period (day)	$M_i(t)$	Survival Period (day)	$M_i(t)$
1	0.9979	10	-2.0666
2	0.9468	11	-2.3810
3	0.7590	13	-2.9374
4	0.4125	14	-3.1855
5	-0.0289	15	-3.4168
6	-0.4919	18	-4.0289
7	-0.9360		

7.5 Identification of the Pattern of the Relationship. Prior to the modeling, the pattern of the relationship between martingale residual and independent variables should be identified. The pattern in scatter plots can be seen in the following Figure 1.



(a) X_3 (b) X_4

(c) X_5

Figure 1 martingale residual scatter plots with independent variables

Figure 1 signifies that the pattern of the relationship between dependent variable and independent variables seems to be random and tends to be nonpatterned. In addition, limited information on form of functions of both dependent variable and independent variables gives a consideration to use a nonparametric regression termed MARS.

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

7.6 Modeling Survival Period of DHF Sufferers Using MARS. The best model on MARS is that with the smallest value of generalized cross-validation (GCV). Based on the combination of $BF = 28$, $MI = 3$, and $MO = 2$, the smallest value of *GCV* is obtained—that is 0.5679 with R^2 of $66.45\,\%$. MARS is built with several input criteria: minspan = 2, trace = 1 that is overview, $nk = 28$, and degree = 3. Therefore, the obtained MARS model is

$$\begin{aligned}\hat{y} = {} & -0.3549 + 0.3692\, x_1(\text{age} \le 14 \text{ years old}) + 0.4019\, x_2(\text{male}) - 0.017\, BF_9 \\ & - 0.0425\, BF_2 + 0.4631\, BF_{19} + 0.0639\, BF_2BF_4 + 0.0622 BF_2BF_5 \\ & - 0.0644\, BF_2BF_6 - 0.0123\, BF_2BF_7 - 0.0109 BF_2BF_8 + 0.0022 BF_9BF_{10} \\ & + 0.0142 BF_2BF_{11} + 0.0156 BF_{12}BF_{19} + 0.0647 BF_{16}BF_{19} \\ & - 0.0177\, X_2(\text{male}) BF_2BF_5 - 0.0019\, X_2(\text{male}) BF_9BF_{10} \\ & - 0.0313\, BF_{16}BF_{17}BF_{19} - 0.0499 BF_{16}BF_{18}BF_{19},\end{aligned}$$

, where

$BF_2 = \max\{0, 101 - X_3\}$,
$BF_4 = \max\{0, X_4 - 13.7\}$,
$BF_5 = \max\{0, 13.7 - X_4\}$,
$BF_6 = \max\{0, X_4 - 14.9\}$,
$BF_7 = \max\{0, X_5 - 38\}$,
$BF_8 = \max\{0, 38 - X_5\}$,
$BF_9 = \max\{0, X_3 - 101\}$,
$BF_{10} = \max\{0, X_5 - 35\}$,
$BF_{11} = \max\{0, X_5 - 42.4\}$,
$BF_{12} = \max\{0, X_3 - 97\}$,
$BF_{16} = \max\{0, 97 - X_3\}$,
$BF_{17} = \max\{0, X_4 - 14.1\}$,
$BF_{18} = \max\{0, 14.1 - X_4\}$,
$BF_{19} = x_6(\text{tidak ada respon imunologis})$.

In reference to the independent variables in the model, it is found out that the variables influencing survival period include age (X_1), sex (X_2), trombocyte (X_3), hemoglobin (X_4), hematocrit levels (X_5), and immunologic response (X_6). Table 3 demonstrates scores of the significant variables.

Table 3 Scores of the Significant Variables in the Model

Independent Variable	Score
X_6	100.0
X_3	61.6
X_4	56.0
X_2	49.4
X_5	49.4
X_1	30.6

Proceeding ICMETA: Volume 1/2017, June 27th 2017 ISBN 978-602-397058-2

8. Conclusion

It is concluded based on research results and discussion that the influencing variables in MARS model can explain diversity of dependent variable (survival period of DHF sufferers in Semarang city), as indicated by value of 66.45%. There are six influencing variables, involving: age (X_1), sex (X_2), trombocyte (X_3), hemoglobin (X_4), hematocrit levels (X_5), and immunologic response (X_6).

References

[1] Eubank, R., (1999), *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, Inc., New York.
[2] Hardle, W., (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge.
[3] Budiantara, I.N., *Penelitian Bidang Regresi Spline Menuju Terwujudnya Penelitian Statistika yang Mandiri dan Berkarakter*, Seminar Nasional FMIPA UNDISKHA. pp. 9-28.
[4] Friedman, J.H., *Multivariate Adaptive Regression Splines*, The Annals of Statistics (1990), Vol. 19, pp. 1-14.
[5] Lewis, P.A.W., & Stevens, J.G. (1991). Nonlinear Modelling of Times Series Using Multivariate Adaptive Regression Splines (MARS). *Journal of the American Statistical Association*. Vol. 86. No. 416. pp. 864-877.
[6] Nisa', S.F. dan Budiantara, *Analisis Survival dengan Pendekatan Multivariate Adaptive Regression Splines pada Kasus Demam Berdarah Dengue (DBD)*, Jurnal Sains dan Seni ITS (2012), Vol. 1, no. 1, D318-D323.
[7] Lawless, J.F., *Statistical Model and Methods for Lifetime Data*, New York:John Wiley and Sons, Inc (1982).
[8] Kriner, M., *Survival Analysis with Multivariate Adaptive Regression Spline*, Munchen University, Jerman, (2007).
[9] Department of Health of Central Java Province, *Profil Kesehatan Provinsi Jawa Tengah*, Semarang, 2011.