Journal of Mathematics Education and Application (JMEA)
Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

Estimation of Multivariate Adaptive Regression Splines (MARS) Model Parameters by Using Generalized Least Square (GLS) Method

Nurul Azizah Rahmadani Ritonga[1*], Sutarman[2]
[1]Bachelor of Mathematics and natural sciences, Universitas Sumatera Utara, Indonesia
[2]Lecturer at Master of Mathematics Education, Universitas Sumatera Utara, Indonesia
*Corresponding Author. E-mail: nurulazizah0609@gmail.com

Article Info

Article History
Received : 06 Januari 2023
Accepted: 10 Mei 2023
Published: 30 Juni 2023

Keywords:
Parameters Estimation, MARS Model, GLS Method

ABSTRACT

The regression analysis method for estimating the regression curve is divided into 3 (three) categories, namely parametric regression analysis, non-parametric regression analysis, and semi-parametric regression analysis. One form of non-parametric regression model is spline which can be developed into Multivariate Adaptive Regression Splines (MARS). The OLS estimation method will get good estimation results compared to other methods if the classical assumptions are fully met. However, if the classical assumptions cannot be fulfilled, this method is not good enough to use. The GLS method can be used if the classical assumptions required by the OLS method are not met. This study aims to estimate the parameters of the MARS model using the GLS method. The GLS method can be used if the classical assumptions required by the OLS method are not met. An example of a case used in the application of non-parametric estimation of the MARS model is the data on the number of doctors and gross enrollment rates for tertiary institutions in 32 districts/cities in North Sumatra in 2021. The best MARS model obtained in this study was obtained with a knot point of 21.2, 24 .2 and 27.2, with BF=6, MO=3, MI=0 with a GCV value of 6628.965. The best model obtained based on this research is as follows:

$$\hat{Y} = 203.3691 - 31.60352BF_1 - 5.383057BF_2 + 15.04785BF_3 + 15.04785BF_4 - 150.7559BF_5 - 14.72168BF_6$$

To cite this article:

INTRODUCTION

Regression analysis is a statistical analysis technique that is often used in solving statistical problems that can describe the relationship between the response variable and the predictor variable. The regression analysis method for estimating the regression curve is divided into 3 (three) categories, namely parametric regression analysis, non-parametric regression analysis, and semi-parametric regression analysis which is a combination of parametric and non-parametric regression.

One form of non-parametric regression model is spline. Spline is a truncated polynomial in the form of

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

a truncated curve so that the spline can handle data changes at certain intervals. Multiple regression spline developments
nonparametric models for adaptive response and multivariate response are examples of this*Multivariate Adaptive Regression Splines* (MARS) and*Recursive Partitioning Regression* (Breiman, Olshen, & Stone, 1993).

Mars is the model proposed by Friedman (1991). The MARS model focuses on addressing problems of high dimensions and large sample sizes, which require complex and complex value-based calculations *Generalized Cross Validation* (GCV) is the smallest. Parameter estimation is the estimation of population characteristics based on the characteristics of the sample. There are two types of parameter estimation, namely point estimation and interval estimation.

Ordinary Least Square (OLS) method is one of the various methods of regression analysis to see the relationship of the predictor variable to the response variable. The OLS method provides the best estimate compared to other methods when all the classical assumptions are met. However, if the classical assumptions are not met, this method is not good enough to use. Then the Generalized Least Square (GLS) method can be used to overcome this.

Generalized Least Squares (GLS) is a method used to estimate parameters whose values are unknown in a linear regression model when there is a correlation level between the residuals in the regression model. In such cases, the use of the OLS method is statistically inefficient or the results obtained are very poor. Then the GLS is used when the assumptions required by the OLS (*homokedasticity* and*non autocorrelation*) cannot be met.

1. Regression Analysis

Regression analysis is a research technique that tries to explain the nature of the relationship between the variables that influence the pattern of the relationship.

2. Parametric Regression

Parametric regression requires assumptions such as normally distributed residuals and constant variance. Mathematically, the form of parametric regression can be expressed as follows

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon_i$$

3. Non Parametric Regression

Nonparametric regression is a regression approach in which the shape of the curve is unknown. In this model the regression curve only assumes a smooth shape (*smooth*) which means that it is contained in a certain form of function space.

4. MARS (Multivariate Adaptive Regression Splines)

Model *Recursive Partioning Regression* (RPR) has the disadvantage that the resulting model is not

continuous at knots. The use of the MARS model to overcome the weaknesses of the RPR model by creating a continuous node model that can distinguish between linear and composite functions. The function of the MARS model is as follows (Friedman, 1991).

$$f(x) = \alpha_0 + \sum^{M} \alpha_m \prod^{K_m} [S_{km} \cdot (x_{v(k,m)} - t_{km})]$$

5. Ordinary Least Square (OLS)

OLS is a regression method that minimizes the value of the number of errors (*error*) squared. In the OLS method, in estimating and testing population regression parameters the regression model must meet the BLUE assumption*(Best Linear Unbiased Estimator).*The parameter estimates are as follows:

$$\hat{\beta} = (X^T X)^{-1}(X^T Y)$$

6.Generalized Least Square (GLS)

GLS is the method used when the assumptions of the OLS method are not met. According to Greene (1997), the handling of heteroscedasticity cases can be done by estimating through weighted which can also be said to be the generally accepted least squares or called Generalized Least Squares (GLS). GLS parameter estimates are as follows:

$$\hat{\beta} = (X^T V^{-1} X)^{-1}(X^T V^{-1} Y)$$

RESEARCH METHOD

The solution flow is as follows:

1. Study Literature

   At this stage a collection of literature used in this study was carried out. The literature used is: Parameter Estimation, Model*MARS*, *Generalized Least Square* and value of *Generalized Cross Validation (GCV)*and other supporting theories.
2. Determine appropriate Case Examples

   At this stage, an example case will be determined that can be used in the MARS model. In this study, one predictor variable and one response variable were used.
3. Create*Scatter plot*

   Scatter plots are used to see patterns from the data. The pattern of predictor and response variable data used must follow a nonparametric pattern.
4. Parameter estimation in the model*Multivariate Adaptive Regression Splines* using method*Generalized Least Square*.
5. Calculate the value*Generalized Cross Validation (GCV)*
6. Record results and conclusions

RESULTS AND DISCUSSION

Research data

In this study, the data is used as a tool for the application of the parameter estimation of the MARS model using the method*Generalized Least Square*. The data used in this study are data on the number of doctors and gross enrollment rates for tertiary institutions in 32 districts/cities in North Sumatra in 2021. The

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

data obtained is as follows:

Table 1. Data on Higher Education Gross Participation Rates and Number of Doctors for 32 Regencies/Cities in North Sumatra 2021

Regency	College APK (%)	Number of Doctors (People)
N i a s	13.44	50
Mandailing Natal	16.13	117
Tapanuli Selatan	22.29	61
Tapanuli Tengah	19.63	88
Tapanuli Utara	22.41	85
Toba	9.2	115
Labuhanbatu	9.62	248
A s a h a n	20.09	152
Simalungun	25.21	215
Dairi	13	76
Karo	14.07	171
Deli Serdang	21.84	356
Langkat	16.29	243
Nias Selatan	15.65	45
Humbang Hasundutan	13.57	49
Pakpak Barat	11.61	35
Samosir	13.18	59
Serdang Bedagai	15.07	243
Batu Bara	13.71	99

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

Padang Lawas Utara	12.56	59
Padang Lawas	16.44	74
Labuhanbatu Selatan	13.84	90
Labuhanbatu Utara	14.87	161
Nias Utara	16.63	22
Nias Barat	9.48	15
Sibolga	19.94	92
Tanjung Balai	14.27	71
Pematang Siantar	34.16	286
Tebing Tinggi	16.04	224
Binjai	29.73	400
Padang Sidimpuan	39.76	128
Gunung Sitoli	23.21	60

MARS Model Parameter Estimation Using the Method *Generalized Least Square*

The regression equation using the MARS estimator is as follows:

$$f(x) = \alpha_0 + \sum_{m=1}^{M} \alpha_m \prod_{k=1}^{K_m} [S_{km} \cdot (x_{v(k,m)} - t_{km})] + \varepsilon$$

When in matrix form it can be written as:

$$\boldsymbol{Y} = \boldsymbol{B\alpha} + \boldsymbol{\varepsilon}$$

So that the estimated GLS parameter is obtained as follows:

$$\boldsymbol{\alpha}_{GLS} = (\boldsymbol{B}^T V^{-1} \boldsymbol{B})^{-1} (\boldsymbol{B}^T V^{-1} Y)$$

Parameter Estimation of the MARS Model Using GLS on Data on the Number of Doctors and Gross Enrollment Rates for Higher Education in 32 Regencies/Cities in North Sumatra 2021

1. Parameter estimation of the MARS model with BF=2

The combinations used are BF=2, MI=0, and MO=0. The following is a nonparametric regression model using the MARS approach with BF=2.

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

$$f(x) = \alpha_0 + \alpha_1 [S_1 . (x - t)] + \alpha_2 [S_2 . (x - t)]$$

Or it can be written in the following form:

$$f(x) = \alpha_0 + \alpha_1 BF_1 + \alpha_2 BF_2$$

Calculations are carried out using the help*software R Studio* in the appendix so that the 10 knot point values and the smallest GCV are obtained as follows:

Table 2 10 Knot Point Values and Smallest GCV MARS Model BF=2

No	Titik Knot	GCV	ASR
1	36,2	7166,458	6942,506
2	34,2	7166,458	6942,506
3	37,2	7166,458	6942,506
4	39,2	7166,458	6942,506
5	35,2	7166,458	6942,506
6	38,2	7166,458	6942,506
7	22,3	7196,731	6971,833
8	32,2	7251,862	7025,207
9	31,2	7311,820	2083,325
10	30,2	7367,931	7137,683

Based on Table 4.2 above, the optimal knot is obtained at point 36.2 with BF=2, MO=0, MI=0 with a GCV value of 7166.458. The estimated results of these parameters are as follows:

$$\hat{\alpha} = \begin{bmatrix} 299.2469 \\ -48.10305 \\ -8.765508 \end{bmatrix}$$

In order to obtain the MARS model with BF=2 and use the GLS estimation method as follows:

$$Y = 83.01989 + 6.62164BF_1 \quad + 21.96076BF_1$$

2 Parameter Estimation of the MARS Model with BF=4

The following is a nonparametric regression model with the MARS approach with BF=4.
$$f(x) = \alpha_0 + \alpha_1 [S_1 .(x - t_1)] + \alpha_2 [S_2 .(x - t_1)] + \alpha_3 [S_3 .(x - t_2)] + \alpha_4 [S_4 .(x - t_2)]$$

Or it can be written in the following form:

Calculations are carried out using the help*software R Studio* in the appendix so that the 10 knot point values and the smallest GCV with BF=4 and MO=0,1,2,3 are obtained as follows:

Table 3 10 Knot Point Values and Smallest GCV MARS Model BF=4, MO=0,1,2,3

MO	Titik Knot 1	Titik Knot 2	GCV	ASR
0	39.2	39.2	7981.943	6984.200
	37.2	37.2	8191.112	7167.223
	30.2	30.2	8393.368	7344.197
	35.2	35.2	8519.435	7454.506
	34.2	34.2	8519.435	7454.506
	12.2	12.2	8690.498	7604.185
	10.2	10.2	8721.453	7631.272
	13.2	13.2	8750.264	7656.481
	9.2	9.2	8876.620	7767.042
	18.2	18.2	10070.755	8811.910
1	26.2	27.2	6813.356	5961.686
	39.2	40.2	8096.360	7084.315
	27.2	28.2	8110.888	7097.027
	32.2	33.2	8167.946	7146.953
	33.2	34.2	8188.533	7164.967
	38.2	39.2	8195.287	7170.876
	30.2	31.2	8289.162	7253.016
	35.2	36.2	8574.264	7502.481
	12.2	13.2	8612.110	7535.596
	9.2	10.2	8768.252	7672.220
2	28.2	30.2	7341.805	6424.080
	34.2	36.2	7963.660	6968.202
	31.2	33.2	7998.211	6998.435
	38.2	40.2	8042.113	7036.849
	36.2	38.2	8049.037	7042.907
	37.2	39.2	8113.230	7099.076
	30.2	32.2	8335.027	7293.149
	20.2	22.2	8892.593	7781.019
	21.2	23.2	8924.758	7809.164
	39.2	41.2	8956.773	7837.176
3	24.2	27.2	7459.731	6527.265
	27.2	30.2	7481.034	6545.905
	39.2	42.2	7974.945	6978.077
	12.2	15.2	8378.216	7330.939
	36.2	39.2	8513.292	7449.130
	32.2	35.2	8515.895	7451.408
	30.2	33.2	8525.698	7459.986
	11.2	14.2	8602.522	7527.207

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

	35.2	38.2	8755.618	7661.166
	20.2	23.2	8841.183	7736.035

Based on Table 4. above, optimal knots are obtained at points 26.2 and 27.2, with BF=4, MO=1, MI=0 with a GCV value of 6813.356. The estimated results of these parameters are as follows:

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} 149.7543 \\ 320.892 \\ -15.64629 \\ -348.3157 \\ 12.29957 \end{bmatrix}$$

In order to obtain the MARS model with BF=4, MO=2, MI=0 and using the GLS estimation method as follows:

$$\hat{Y} = 149.7543 + 320.892BF_1 - 15.64629BF_2 - 348.3157BF_3 - 12.29957BF_4$$

3 Parameter Estimation of the MARS Model with BF=6

The following is a nonparametric regression model with the MARS approach with BF=6.

$$f(x) = \alpha_0 + \alpha_1 [S_1 . (x - t_1)] + \alpha_2 [S_2 . (x - t_1)] + \alpha_3 [S_3 . (x - t_2)] + \alpha_4[S_4. (x - t_2)] + \alpha_5 [S_5 . (x - t_3)] + \alpha_6[S_6. (x - t_6)]$$

Or it can be written in the following form:

$$f(x) = \alpha_0 + \alpha_1 BF_1 + \alpha_2 BF_2 + \alpha_3BF_3 + \alpha_4 BF_4 + \alpha_5BF_5 + \alpha_6 BF_6$$

Calculations are carried out using the help*software R Studio* in the attachment so that the 10 knot point values and the smallest GCV with BF=6 and MO=0,1,2,3 are obtained as follows:

Table 4. 10 Knot Point Values and Smallest GCV MARS Model BF=6, MO=0,1,2,3

MO	Titik Knot 1	Titik Knot 2	Titik Knot 3	GCV	ASR
0	35.2	35.2	35.2	7945.761	6952.541
	39.2	39.2	39.2	7948.108	6954.594
	36.2	36.2	36.2	7996.643	6997.062
	37.2	37.2	37.2	8175.380	7153.458
	27.2	27.2	27.2	8655.368	7573.447
	25.2	25.2	25.2	8842.644	7737.313
	9.2	9.2	9.2	8876.620	7767.042
	19.2	19.2	19.2	8897.598	7785.399
	20.2	20.2	20.2	8898.367	7786.071
	24.2	24.2	24.2	9120.273	7980.239
1	29.2	30.2	31.2	7450.377	6519.080
	28.2	29.2	30.2	8154.595	7135.271
	27.2	28.2	29.2	8377.269	7330.110
	39.2	40.2	41.2	8711.512	7622.573

23.2	24.2	25.2	8795.902	7696.415
24.2	25.2	26.2	8844.435	7738.880
16.2	17.2	18.2	9272.180	8113.157
31.2	32.2	33.2	9597.672	8397.963
36.2	37.2	38.2	9948.581	8705.008
18.2	19.2	20.2	10124.838	8859.233
23.2	25.2	27.2	6698.681	5861.346
27.2	29.2	31.2	6970.902	6099.540
26.2	28.2	30.2	6989.801	6116.076
25.2	27.2	29.2	7013.570	6136.874
37.2	39.2	41.2	7964.963	6969.343
38.2	40.2	42.2	8020.809	7018.208
31.2	33.2	35.2	8886.880	7776.020
17.2	19.2	21.2	8930.562	7814.242
10.2	12.2	14.2	8946.553	7828.234
20.2	22.2	24.2	9607.817	8406.840
21.2	24.2	27.2	6628.965	5800.344
29.2	32.2	35.2	6911.002	6047.126
23.2	26.2	29.2	7378.243	6455.963
22.2	25.2	28.2	7507.265	6568.856
34.2	37.2	40.2	8123.022	7107.644
27.2	30.2	33.2	8288.274	7252.240
37.2	40.2	43.2	8531.187	7464.789
25.2	28.2	31.2	9058.177	7925.905
17.2	20.2	23.2	9176.338	8029.296
38.2	41.2	44.2	9228.620	8075.042

Based on Table 4.4 above, optimal knots are obtained at points 21.2, 24.2 and 27.2, with BF=6, MO=3, MI=0 with a GCV value of 6628.965. The estimated results of these parameters are as follows:

$$\hat{\boldsymbol{\alpha}} = \begin{bmatrix} 203.3691 \\ -31.60352 \\ -5.383057 \\ 157.9771 \\ 15.04785 \\ -150.7559 \\ -14.72168 \end{bmatrix}$$

In order to obtain the MARS model with BF=6, MO=3, MI=0 and using the GLS estimation method as follows:

$$\hat{Y} = 203.3691 - 31.60352BF_1 - 5.383057BF_2 + 15.04785BF_3 + 15.04785BF_4 - 150.7559BF_5 - 14.72168BF_6$$

4 Discussion

Estimation was carried out using 3 types of BF, namely: BF=2, BF=4, and BF=6, MO=0, 1, 2, and 3 and MI=0. MI was chosen to be zero because there is only one predictor variable, so there is no interaction between

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

predictor variables. The results of the comparison of the estimation of the MARS model regression parameters using the GLS method are as follows

Table 5 Comparison of Non-Parametric Regression Estimation Results of the MARS Model Using the GLS Method

BF	MO	Point Knot 1	Point Knot 2	Point Knot 3	GCV
2	0	36,2	-	-	7166,458
4	1	26,2	27,2	-	6813.356
6	3	21,2	24,2	27,2	6628,965

CONCLUSION

Based on the discussion carried out in the previous chapter, the results obtained from the estimation of the parameters of the MARS model with the estimation of the GLS parameters are as follows: $\hat{\boldsymbol{\alpha}}_{\boldsymbol{GLS}} =$

$(\boldsymbol{B}^T V^{-1}\boldsymbol{B})^{-\boldsymbol{1}}(\boldsymbol{B}^T V^{-1}Y)$

The application of non-parametric regression estimation of the MARS model to case data on the number of doctors and gross enrollment rates in tertiary institutions in 32 districts/cities in North Sumatra in 2021 using the GLS method obtained the best MARS model with a combination of BF=6, MO=3, MI=0. This can be seen from the GCV value of the MARS model with BF=6, MO=3, MI=0 which are the smallest compared to the others. So that the best MARS model obtained in this study was obtained with knot points of 21.2, 24.2 and 27.2, with BF=6, MO=3, MI=0 with a GCV value of 6628.965. The best model obtained based on this research is as follows:

$$\hat{Y} = 203.3691 - 31.60352BF_1 - 5.383057BF_2 + 15.04785BF_3 + 15.04785BF_4 - 150.7559BF_5 - 14.72168BF_6$$

REFERENCES

Astuti, Desak Ayu Wiri, 2016.*Multivariate Spline Nonparametric Regression Analysis for Modeling Poverty Indicators in Indonesia*. *Mathematics E-Journal,*Vol.5 (3), pp.111-116. ISSN:2303-1751

Breiman, L., Friedman, J., Olshen, R., & Stone, J. (1993). *Classification and Regression Trees.* New York:

Vol. 2, No 2, Juni 2023, pp. 62–72 DOI: https://doi.org/10.30596/jmea.v2i2.13106

Chapman and Hall.

Budiantara, I. N. 2009. "U, GML, CV, and GCV Methods in Spline Nonparametric Regression",*Scientific Magazine of the Indonesian Mathematical Association (MIHMI*), vol. 6, hal.285-290.

Friedman, J.H,.1991*. Multivariate Adaptive Regression Splines*, *The Annals of Statistics.* 1991;19(1). Matter. 1-141.

Greene, W. H. 1997. *Econometric analysis*. Pearson Education, New York.

Gujarati, D. N. and Porter, D. C. 2009. *Basic Econometrics*. Fifth Edition. McGraw Hill/Irwin. New York,

Harini, S. and Turmudi. 2008.*Method Statistics*.Malang: UIN-Malang Press.

Hasan, M. I. 2002. *The Main Materials of Statistics I (Descriptive Statistics).*Jaka PT. Script Earth.

Tehupuring, B. C., & Hadi, S. (2015). Application of the Multivariate Adaptive Regression Spline as a Tool for Modeling the Growth of Broiler Chickens.*VETERINARY Act Indonesiana*, *3*(1), 23-28.

Wasilane, T. 2014.*Ridge Regression Model To Overcome Multiple Linear Regression Models Containing Multicollinearity*. Barekeng Journal, 31-37.

Zhang, H., & Singer, B. H. 2010. *Recursive partitioning and applications*. Springer Science & Business Media.

Zurimi, S. 2019. Analysis of the Multivariate Adaptive Regression Spline (MARS) applicative model on the classification of factors that influence the study period of FKIP students Ambon Darussalam University.*Symmetric Journal*, *9*(2), 250-255.