Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2023, 2023:37

https://doi.org/10.28919/cmbn/7935

ISSN: 2052-2541

THE USE OF THE BINARY SPLINE LOGISTIC REGRESSION MODEL ON THE NUTRITIONAL STATUS DATA OF CHILDREN

ANNA ISLAMIYATI[1,*], ANISA[1], MUHAMMAD ZAKIR[2], UMMI SARI[3], DEWI SARTIKA SALAM[1]

[1]Department of Statistics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar, 90245, Indonesia

[2]Department of Mathematics, Faculty of Mathematical and Natural Sciences, Hasanuddin University, Makassar, 90245, Indonesia

[3]Hasanuddin University Teaching Hospital, Makassar, 90245, Indonesia

Abstract: The nutritional status of children can be grouped into two, namely normal and abnormal nutrition so that the data are analyzed using a binary logistic regression model. In this study, the use of binary logistic regression was developed on the use of the spline estimator as part of the nonparametric regression approach. This model is able to model qualitative response data by showing several trends that may occur in the data. Data on the nutritional status of children were analyzed based on the child's weight through a linear spline estimator and the optimal model was obtained at the use of a one-knot point, which was 4.7 kg. These results indicate that if a child has a body weight of 4.7 kg and above then there is a chance that the child has an abnormal nutritional status of 9.013 times compared to a child with a body weight below 4.7 kg. This could be due to the fact that children are no longer getting breast milk, so they need to get attention regarding their nutritional status.

[*]Corresponding author

E-mail address: annaislamiyati701@gmail.com

Received March 03, 2022

Keywords: binary; logistic; nutritional; spline.

2020 AMS Subject Classification: 92C15.

1. INTRODUCTION

Logistic regression is a regression developed for categorical responses. One form of logistic regression used for two-categorical responses is binary logistic regression. The response variable of the binary logistic regression model is assumed to follow the Bernoulli distribution [1]. Logistic regression developments include binary logistic regression for mixed effects [2], binary lasso logistic regression for those experiencing multicollinearity data [3], and weighted binary logistic regression [4]. Its use in data has also been widely used in various fields of science, including the use of the concept of logistic regression in medical data [5], cable television user survey data [6], and data on the credit risk of consumer loans in banking institutions [7].

In some cases, we often find that a lot of data is not balanced between different classes so the usual use of logistic regression is less accurate. This is because the classification tends to eliminate opportunities from the minority class because the predicted value will tend to be in the majority class. Therefore, in the next development, researchers have made a logistic regression model using a nonparametric regression function estimator. There are several estimators in nonparametric regression, including truncated spline [8], spline smoothing [9], penalized spline [10]–[12], local polynomial [13], kernel [14], and the Fourier series [15]. In this study, we use a truncated spline estimator involving knot points in the estimation criteria. The knot point is the point where the pattern of change occurs so that in a model there can be several segmentations on continuous data that are selected based on the minimum GCV value [16]. This is one of the advantages of the truncated spline so it is widely used by researchers in their research. For example, health data found 4 patterns of changes in blood sugar in diabetic patients based on the time of hospitalization [17] and carbohydrate diet [18]. The study showed that some segmentation of the pattern of changes that could be explained by truncated splines led to more accurate identification of problems in a health case. Therefore, data on the nutritional status of children under five were analyzed based on weight through a nonparametric binary logistic regression model with the estimator used spline truncated.

The nutritional status of children under five is measured by several indicators, and one of them is the weight factor. Nutritional status can be divided into two categories, namely normal and abnormal so that the data can be analyzed using binary logistic regression. There have been many studies on the nutritional status of children under five by considering many factors, including maternal education and income [19], number of family dependents [20], and body weight [21], all of which affect the nutritional status of children under five. However, the study did not show how the probability level of nutritional status could occur in certain weight intervals. Therefore, in this article, we analyze the data on the nutritional status of these toddlers using a spline logistic regression model at several optimal knot points. Furthermore, this article is divided into 4 parts, namely the second part describes the data material and analysis methods. The third section describes the results and discussion related to the nutritional status model based on body weight through binary spline logistic regression. The last section contains the conclusions of our article.

2. PRELIMINARIES

Data on the nutritional status of children under five were obtained from the Community Health Center which in Indonesia is abbreviated as Puskesmas. The secondary data came from the Puskesmas in Barru Regency, Indonesia, with a total sample of 432. The data consisted of the nutritional status of children under five in response to two categories. The abnormal category was coded 0 and the normal was coded 1. The nutritional status of the child was analyzed with the predictor variable being the toddler's weight.

If the response variable y_i is in the form of two categories, then the regression model used is the binary logistic regression model. The model assumes that the data are Bernoulli distributed and independent between observations with the probability distribution function as follows:

$$f(y_i) = \pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}, y_i = 0,1 \tag{1}$$

where $\pi(x_i)$ is the probability of success. If $y_i = 1$, then $f(y_i) = \pi(x_i)$ and if $y_i = 0$ then $f(y_i) = (1-\pi(x_i))$.

Furthermore, the model used in this study is a binary logistic nonparametric regression model with a truncated spline estimator. For example, the order of the spline is expressed as q and the knot point $\boldsymbol{\tau}$ is m, then the spline logistic regression model can be expressed as follows [22]:

$$\pi(x_i) = \frac{\exp\left(\beta_0+\sum_{j=1}^{q}\beta_j x_i^j+\sum_{k=1}^{m}\beta_{q+k}(x_i-\tau_k)_+^q\right)}{1+\exp\left(\beta_0+\sum_{j=1}^{q}\beta_j x_i^j+\sum_{k=1}^{m}\beta_{q+k}(x_i-\tau_k)_+^q\right)} \qquad (2)$$

Through the logit transformation, Equation (2) can be expressed in the form:

$$g(x_i) = \beta_0 + \sum_{j=1}^{q}\beta_j x_i^j + \sum_{k=1}^{m}\beta_{q+k}\,(x_i-\tau_k)_+^q$$

where x_i is the i *-th* predictor variable, β_0 is the intercept, β_j is the coefficient of nonparametric spline truncated logistic regression, $\boldsymbol{\tau}_k$ is the knot point with = 1,2, ..., *m*, and $(x_i-\tau_k)_+^q$ is a truncated polynomial function which can be expressed as following:

$$(x_i-\tau_k)_+^q = \begin{cases}(x_i-\tau_k)_+^q & ; x_i \geq \tau_k \\ 0 & ; x_i < \tau_k\end{cases}$$

Parameter estimation in the spline truncated binary logistic regression model in Equation (2), is done using the maximum likelihood method, which is maximizing the likelihood function. The probability density function is known as in Equation (1) so the likelihood function can be written as follows:

$$\ell(\boldsymbol{\beta}) = \prod_{i=1}^{n}\pi(x_i)^{y_i}\left(1-\pi(x_i)\right)^{1-y_i}, \text{where } \boldsymbol{\beta} = \left(\beta_0,\beta_1,\dots,\beta_q,\beta_{q+1},\dots,\beta_{q+k}\right)^T$$

Furthermore, it is made into the form of ln likelihood so that we get:

$$\ln\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n}\{y_i(\pi(x_i)) - \ln[(1-\pi(x_i))]\}$$

In the next step, the ln likelihood function is derived from the beta parameter until an implicit parameter estimation result is found so that the Newton-Raphson iteration process is carried out. The estimation results of binary logistic regression parameters with a truncated spline estimator can be expressed as follows.

$$\widehat{\boldsymbol{\beta}}_{r+1} = \widehat{\boldsymbol{\beta}}_r + (\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\left(\boldsymbol{X}^T\big(\boldsymbol{y} - \pi(\boldsymbol{X})\big)\right)$$

where $\widehat{\boldsymbol{\beta}}_{r+1}$ is the spline binary logistic regression coefficient at $r+1$ iteration, $\widehat{\boldsymbol{\beta}}_r$ is the spline binary logistic regression coefficient at the r-th iteration, $\boldsymbol{y}$ is the response vector expressed as $\boldsymbol{y} = (y_1, y_2, \dots, y_n)^T$, $\boldsymbol{W}$ is the diagonal matrix i.e. $\boldsymbol{W} = diag\Big(\pi(x_1)\big(1 - \pi(x_1)\big), \pi(x_2)\big(1 - \pi(x_2)\big), \dots, \pi(x_n)\big(1 - \pi(x_n)\big)\Big)$, and the $\boldsymbol{X}$ matrix is expressed as follows:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 & x_1 - \tau_1 & \cdots & x_1 - \tau_m \\ 1 & x_2 & x_2 - \tau_1 & \cdots & x_2 - \tau_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n - \tau_1 & \cdots & x_n - \tau_m \end{bmatrix}.$$

The optimal knot point for the best spline regression model is obtained from the smallest GCV value. The GCV method can be written as follows:

$$GCV(\tau) = \frac{MSE(\tau)}{(n^{-1}trace[\mathbf{I} - \mathbf{A}(\tau)])^2}$$

Where $MSE(\tau) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{y})^2$, $\mathbf{I}$ is the identity matrix, and $\mathbf{A}(\tau) = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

3. Main Results

Data on the nutritional status of toddlers obtained at the Puskesmas Barru Regency, Indonesia, as many as 432 toddlers, namely there are 59 toddlers, or about 13.66% fall into the category of abnormal nutrition, and the remaining 373 toddlers, or about 86.34% who fall into the category of normal nutrition. These percentages indicate that there is a large difference in numbers between the normal and abnormal categories, i.e children under the normal category are 72.68% greater than those in the abnormal category. Therefore, in this study, we modeled data on the nutritional status of children under five based on their weight factor using binary logistic regression with a nonparametric regression estimator, namely linear spline.

For the linear spline binary logistic regression model, we can refer to the spline binary logistic regression model in Equation (3) which is made into a linear form with *m* knot points i.e.:

$$\pi(x_i) = \frac{\exp\left(\beta_0+\beta_1 x_i+\sum_{k=1}^{m}\beta_{1+k}(x_i-\tau_k)_+\right)}{1+\exp\left(\beta_0+\beta_1 x_i+\sum_{k=1}^{m}\beta_{1+k}(x_i-\tau_k)_+\right)} \tag{3}$$

Where $\pi(x_i)$ is the probability of normal nutritional status $(y = 1)$, x_i is the weight factor, β_0 is the intercept, $\beta_1, \beta_{1+1}, \dots, \beta_{1+m}$ is the nonparametric spline truncated logistic regression coefficient, $\boldsymbol{\tau}_k$ is knot points with $k = 1,2,\dots,m$ and $i = 1,2,\dots,432.$

Data analysis begins with the use of 1 knot point so that the spline linear binary logistic regression model in Equation (3) changes to:

$$\pi(x_i) = \frac{\exp(\beta_0 + \beta_1 x_i + \beta_2(x_i - \tau_1)_+)}{1 + \exp(\beta_0 + \beta_1 x_i + \beta_2(x_i - \tau_1)_+)}$$

The knot point is chosen along the predictor variable so it is necessary to choose the optimal knot point for the data. For 1 knot point, the GCV values obtained in the binary spline logistic regression model are shown in Table 1. The optimal knot point is obtained in the model that gives the minimum GCV value as shown in Table 1. For some knot points, it can be seen that the knot point of 4.7 which gives the minimum GCV value is 0.109. Next, we will compare it with the GCV value on the use of 2-knot points. The GCV value in the linear spline binary logistic regression model with 2-knot points is shown in Table 2.

TABLE 1. GCV value of linear truncated spline linear logistic regression one knot point for weight predictor

Knot (τ)	GCV	Knot (τ)	GCV
3.9	0.116	25	0.117
4.7	**0.109**	31	0.117
7	0.116	35.5	0.117
9.5	0.117	42.5	0.117
16	0.116	45	0.117
19	0.116	51.1	0.117

TABLE 2. GCV value of binary spline truncated linear logistic regression two-knot points for the weight factor

Knot (τ_1)	Knot (τ_2)	GCV Value
4.7	**5.9**	**0.110**
4.7	6.5	0.111
5.5	12	0.116
9.5	16	0.117
12.7	27.5	0.117
19	23	0.117
19.5	33.5	0.117

Based on Table 2, for the use of 2-knot points, the minimum GCV value is obtained at 4.7 and 5.9 knots, which is 0.110. However, if the value is compared with the GCV value at the 1-knot point in Table 1, we can decide that the 1-knot point is better to use in the linear spline regression model. Therefore, data on the nutritional status of children under five can be modeled using a linear spline binary logistic regression model with the 1-knot point.

The binary logistic regression model with a linear spline at the 1-knot point is as follows:

$$\hat{\pi}(x_i) = \frac{\exp\left(-39.619 + 8.935x_i - 9.013(x_i - 4.7)\right)}{1 + \exp\left(-39.619 + 8.935x_i - 9.013(x_i - 4.7)\right)}$$

It can also be formed in the form of a logit function, namely:

$$g(x_i) = -39.619 + 8.935x_i - 9.013(x_i - 4.7)$$

Based on the estimation results of the spline binary logistic regression model with the 1-knot point, it shows that there are two possible events that can occur in the nutritional status of toddlers based on their weight. First, babies weighing under 4.7 kg tend to have 8.935 times the chance of good nutrition compared to babies weighing over 4.7 kg. Second, babies weighing above 4.7 kg tend to have a very low chance of being able to get normal nutrition, or in other words, we can say, they tend to fall into the category of abnormal nutrition. If it is related to the age of toddlers who weigh around 4.7 kg, it is found that they are infants aged 1-12 months. This means that the chance of normal nutrition is greater when the baby is up to 1 year old. These results are in line with research by Onis and Branca (2016) that normal nutrition is seen in children aged one to two years, after which there is a tendency to experience slowed growth [23].

TABLE 3. Classification results of nutritional status with binary spline logistic regression

Prediksi	Actual		Total
	0 (abnormal)	1 (normal)	
0 (abnormal)	6	1	7
1 (normal)	53	372	425
Total	59	373	432
	Accuracy		87.5%

Furthermore, the results of the classification of the nutritional status of children under five through a spline linear logistic regression model with a 1-knot point are shown in Table 3. The results show an 87.5% level of accuracy in classifying the model, meaning that the linear spline logistic regression model with a 1-knot point is accurate in classifying data. And of course, we can also say that the linear spline logistic regression model is accurate in modeling the nutritional status data of toddlers based on the weight factor.

ACKNOWLEDGMENTS

Many thanks to the Directorate of Research, Technology, and Community Service, Directorate General of Higher Education, Research and Technology of the Ministry of Education, Culture, Research, and Technology for National Competitive Basic Research with research contract No: 090/E5/PG .02.00.PT/2020 dated May 10, 2022.

CONFLICT OF INTERESTS

The authors declare that there is no conflict of interests.

REFERENCES

[1] B.T. Zewude, K.M. Ashine, Binary logistic regression analysis in assessment and identifying factors that influence students' academic achievement : the case of college of natural and computational, J. Educ. Pract. 7 (2016), 1-6.

[2] J.K. Vermunt, Mixed-effects logistic regression models for indirectly observed discrete outcome variables, Multivariate Behav. Res. 40 (2005), 281-301. https://doi.org/10.1207/s15327906mbr4003_1.

[3] A. Rusyana, K.A. Notodiputro, B. Sartono, The lasso binary logistic regression method for selecting variables that affect the recovery of Covid-19 patients in China, J. Phys.: Conf. Ser. 1882 (2021), 012035. https://doi.org/10.1088/1742-6596/1882/1/012035.

[4] D. Eka Apriana Sulasih, S. Wulan Purnami, S. Puteri Rahayu, The theoretical study of rare event weighted logistic regression for classification of imbalanced data, in: International Conference on Science, Technology and Humanity 2015, 159-169.

[5] E.Y. Boateng, D.A. Abaye, A review of the logistic regression model with emphasis on medical research, J. Data Anal. Inform. Process. 07 (2019), 190–207. https://doi.org/10.4236/jdaip.2019.74012.

[6] S.H. Hsieh, S.M. Lee, P.S. Shen, Logistic regression analysis of randomized response data with missing covariates, J. Stat. Plan. Inference. 140 (2010), 927–940. https://doi.org/10.1016/j.jspi.2009.09.020.

[7] E. Costa e Silva, I.C. Lopes, A. Correia, et al. A logistic regression model for consumer default risk, J. Appl. Stat. 47 (2020), 2879-2894. https://doi.org/10.1080/02664763.2020.1759030.

[8] A. Islamiyati, A. Kalondeng, N. Sunusi, et al. Biresponse nonparametric regression model in principal component analysis with truncated spline estimator, J. King Saud Univ. - Sci. 34 (2022) 101892. https://doi.org/10.1016/j.jksus.2022.101892.

[9] B. Lestari, Fatmawati, I.N. Budiantara, et al. Smoothing parameter selection method for multiresponse nonparametric regression model using smoothing spline and Kernel estimators approaches, J. Phys.: Conf. Ser. 1397 (2019), 012064. https://doi.org/10.1088/1742-6596/1397/1/012064.

[10] A. Islamiyati, Fatmawati, N. Chamidah, Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression, J. Phys.: Conf. Ser. 979 (2018), 012093. https://doi.org/10.1088/1742-6596/979/1/012093.

[11] A. Islamiyati, N. Sunusi, A. Kalondeng, et al. Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model, J. Sci. Islam. Repub. Iran 31 (2020), 175-183. https://doi.org/10.22059/jsciences.2020.286949.1007435.

[12] A. Islamiyati, Fatmawati, N. Chamidah, Penalized spline estimator with multi smoothing parameters in bi-response multi-predictor nonparametric regression model for longitudinal data, Songklanakarin J. Sci. Technol. 42 (2020), 897-909.

[13] N. Chamidah, E. Tjahjono, A.R. Fadilah, et al. Standard growth charts for weight of children in East Java using local linear estimator, J. Phys.: Conf. Ser. 1097 (2018), 012092. https://doi.org/10.1088/1742-6596/1097/1/012092.

[14] R. Hidayat, I.N. Budiantara, B.W. Otok, et al. The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model, Commun. Stat. - Theory Methods. 50 (2020), 3942-3953. https://doi.org/10.1080/03610926.2019.1710201.

[15] M.F.F. Mardianto, E. Tjahjono, M. Rifada, Semiparametric regression based on three forms of trigonometric function in fourier series estimator, J. Phys.: Conf. Ser. 1277 (2019), 012052. https://doi.org/10.1088/1742-6596/1277/1/012052.

[16] A. Islamiyati, Raupong, A. Kalondeng, et al. Estimating the confidence interval of the regression coefficient of the blood sugar model through a multivariable linear spline with known variance, Stat. Transit. 23 (2022), 201-212.

[17] A. Islamiyati, Fatmawati, N. Chamidah, Changes in blood glucose 2 hours after meals in type 2 diabetes patients based on length of treatment at Hasanuddin University Hospital, Indonesia, Rawal Med. J. 45 (2020), 31-34.

[18] A. Islamiyati, Spline longitudinal multi-response model for the detection of lifestyle-based changes in blood glucose of diabetic patients, Curr. Diabetes Rev. 18 (2022), 98-104. https://doi.org/10.2174/1573399818666211117113856.

[19] Z.I. Nafia, I.Z. Shodiq, L. Handayani, Nutritional status of children under five years in the work area of Puskesmas Cipadung, Dis. Prev. Public Heal. J. 15 (2021), 125-132.

[20] C. R. Titaley, I. Ariawan, D. Hapsari, et al. Determinants of the stunting of children in Indonesia : a multilevel analysis of the 2013 Indonesia basic health survey, Nutrients 11 (2014), 1-13.

[21] S. Ramadhani H., J.I. Sari, R. Rahmadhani, Study of differences in children nutrition status aged 6-24 months with exclusive and non-exclusive breastfeeding in Mattampa Bulu Village, Green Med. J. 3 (2021), 81-90. https://doi.org/10.33096/gmj.v3i2.85.

[22] D.S. Salam, A. Islamiyati, N. Ilyas, Binary logistic model in nonparametric regression through spline estimator, Int. J. Acad. Appl. Res. 5 (2021), 50-53.

[23] M. de Onis, F. Branca, Childhood stunting: a global perspective, Matern. Child Nutri. 12 (2016), 12-26. https://doi.org/10.1111/mcn.12231.