Available online at http://scik.org

Commun. Math. Biol. Neurosci. 2023, 2023:83

https://doi.org/10.28919/cmbn/8072

ISSN: 2052-2541

ABILITY OF ORDINAL SPLINE LOGISTIC REGRESSION MODEL IN THE CLASSIFICATION OF NUTRITIONAL STATUS DATA

SAMSUL ARIFIN, ANNA ISLAMIYATI*, ERNA TRI HERDIANI

Department of Statistics, Hasanuddin University, Makassar 90245, Indonesia

Abstract: In this study, an ordinal spline logistic regression model was developed and used to classify data on the nutritional status of children under five in the Gowa district, Indonesia. The nutritional status of toddlers consists of 3 categories: malnutrition, good nutrition, and excess nutrition. So nutritional status data for toddlers can be modeled by ordinal spline logistic regression. The results of this study indicate that the data on the nutritional status of children is optimal in the ordinal spline logistic regression model using 2-knot points with a GCV value of 0.2158. The estimation results of the ordinal spline logistic regression model show that toddlers aged 18 months and 24 months tend to have a good chance of getting good nutrition. In comparison, toddlers aged 18 to 24 months tend to have a minimal chance of getting good nutrition, and the accuracy of the classification model of the nutritional status of toddlers uses the ordinal spline logistic regression of 92.25%.

Keywords: logistic; nutrition; ordinal; spline; toddlers.

2020 AMS Subject Classification: 92B10.

*Corresponding author

E-mail address: annaislamiyati701@gmail.com

Received June 27, 2023

1. INTRODUCTION

Ordinal logistic regression is a form of logistic regression used for response variables of more than two categories [1], has an order [2], and is assumed to have a multinomial distribution [3]. Logistic regression has been developed using robust [4], principal components [5], and mixed models [6]. In addition, its use in data has also been widely used in various fields of science, including the use of logistic regression on health data [7], education [8], and socioeconomic [9]. In some cases, we often find unbalanced data, so the regular use of logistic regression is less accurate [10]. This is because classification tends to eliminate opportunities from minority classes. For this reason, several logistic nonparametric regressions have been developed, including spline binary logistic regression [11] and local polynomial logistic regression [12]. This development is in line with several well-known estimators in regression, including truncated spline [13], smoothing spline [14], penalized spline [15], and local polynomials [16].

In this study, we used a truncated spline estimator that involved knots in the estimation. Truncated splines are used because they can handle data that has changed behavior at certain intervals and tend to look for data estimates wherever the data pattern moves with the help of knot points [13]. The knot point is where the pattern of changes in functional behavior occurs at different intervals based on the minimum GCV value [17]. Research on health data found two patterns of changes in children's weight [11]. The study used spline binary logistic regression and obtained an accuracy of 87.5%. However, this research has not considered the three categories of response variables with more than two categories and has an order. Therefore, researchers will develop an ordinal spline logistic regression model and then apply the method to the nutritional status data of toddlers in the Gowa district, Indonesia.

The nutritional status of toddlers in Indonesia can be measured by indicators of age, weight, and body length [18]. However, in this study, we used age as a predictor variable. There have been many studies on the nutritional status of toddlers by considering several factors, including breastfeeding [19], nutritional intake [20], formula feeding [21], and mother's knowledge [22]. However, the study did not show the probability level that it could occur in the age interval.

Therefore, we will analyze the nutritional status data, which consists of three categories: undernutrition, good nutrition, and overnutrition. We will use ordinal spline logistic regression with several knot points to see patterns of changes that might occur.

2. PRELIMINARIES

In this study, we used secondary data from the Gowa District Health Office, which consisted of 17600 toddlers who had weighed each Posyandu in the Gowa District, Indonesia. The data consists of 3 categories of response variables: malnutrition, good nutrition, and excess nutrition. Toddler nutritional status was analyzed with the age predictor variable.

If the response variable yi is in the form of three-level categories, then the regression model used is an ordinal logistic regression model. The model is assumed to have a multinomial distribution and is independent between observations with the probability density function as follows:

$$f(y_i) = \pi_1(x_{ji})^{y_{i1}} \pi_2(x_{ji})^{y_{i2}} \dots \pi_{s-1}(x_{ji})^{y_{i(s-1)}}$$

Furthermore, the model used in this study is an ordinal logistic nonparametric regression model with a truncated spline estimator, which can be stated as follows:

$$\pi(x_{ji}) = \frac{\exp(\alpha_s + \sum_{l=1}^{q} \beta_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \beta_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})}{1 + \exp(\alpha_s + \sum_{l=1}^{q} \beta_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \beta_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})}$$

Parameter estimation is done by decomposing it using a logit transformation as follows:

$$logit[P(y_i \le s | x_{ji})] = \alpha_s + \sum_{l=1}^{q} \beta_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \beta_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}$$

The function $(x_i - k_h)_{+}^{q}$ It is a truncated polynomial that is described as follows:

$$(x_i - k_h)_{+}^{q} = \begin{cases} (x_i - k_h)^q, jika\ x_i \ge k_h \\ 0, \quad jika\ x_i < k_h \end{cases}$$

If, in the above equation, we substitute the value q = 1,2,3, we get a spline function called a linear truncated spline, a quadratic truncated spline, and a cubic truncated spline. Parameter estimation in the ordinal spline logistic regression model uses the maximum likelihood estimation (MLE) method by maximizing the likelihood function. It is known that the model has a multinomial distribution with the probability density function as follows:

$$h(\gamma) = \prod_{i=1}^{n} \pi_1(x_{ji})^{y_{i1}} \pi_2(x_{ji})^{y_{i2}} \dots \pi_{s-1}(x_{ji})^{y_{i(s-1)}}$$

Next, it is necessary to transform the likelihood function into the natural logarithmic form so that the ln-likelihood function is obtained as follows:

$$\ln h(\gamma) = \prod_{i=1}^{n} \ln \pi_1(x_{ji})^{y_{i1}} \ln \pi_2(x_{ji})^{y_{i2}} \dots \ln \pi_{s-1}(x_{ji})^{y_{i(s-1)}}$$

The value of the cumulative probability function for each response category is as follows:

For the first category:

$$\begin{aligned}\pi_1(x_{ji}) &= P(Y = 1|x_{ji}) \\ &= P(Y \leq 1|x_{ji}) \\ &= \frac{\exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})}\end{aligned}$$

For the second category:

$$\begin{aligned}\pi_2(x_{ji}) &= P(Y = 2|x_{ji}) \\ &= P(Y \leq 2|x_{ji}) - P(Y \leq 1|x_{ji}) \\ &= \frac{\exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})} - \frac{\exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})} \\ &= \frac{\exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta}) - \exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})}{[1 + \exp(\alpha_1 + \boldsymbol{X}[k]\tilde{\beta})][1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})]}\end{aligned}$$

For the third category:

$$\begin{aligned}\pi_3(x_{ji}) &= P(Y = 3|x_{ji}) \\ &= 1 - P(Y \leq 2|x_{ji}) \\ &= 1 - \frac{\exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}\end{aligned}$$

$$= \frac{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})} - \frac{\exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}$$

$$= \frac{1}{1 + \exp(\alpha_2 + \boldsymbol{X}[k]\tilde{\beta})}$$

We derive each parameter to find the maximum value of the ln-likelihood function, and then the derivative equals zero. The result of the derivative is a nonlinear function, so it is necessary to use a numerical method to obtain the parameter estimation, one of which is Newton-Raphson iterations. The estimation results of ordinal spline truncated logistic regression parameters are as follows:

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

A suitable method for selecting optimal knot points is the Generalized Cross Validation (GCV) method. The GCV method can be written in the following equation:

$$\text{GCV}(k_1, k_2, \dots, k_r) = \frac{MSE(k_1, k_2, \dots, k_r)}{[n^{-1} trace(\boldsymbol{I} - \boldsymbol{A}(k_1, k_2, \dots, k_r))]^2}$$

Where $MSE(k_1, k_2, \dots, k_r) = n^{-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, k is the knot point, matrix A is $X(X^tX)^{-1}X^t$, I is the identity matrices. The minimum GCV value gives the optimal knot point value.

3. MAIN RESULTS

Data on the nutritional status of children under five were obtained from the Gowa District Health Office, Indonesia; as many as 17600 children under five had malnutrition, good nutrition, and excess nutrition. The results of the nutritional status data plot for toddlers are shown in Figure 1. The percentage for the malnutrition category was 2102 toddlers or 11.94%, for the excess nutrition category were 1766 toddlers or 10.03%, and the rest for good nutrition was 13732 toddlers or 78.03%. This percentage shows a significant difference in numbers between the categories of good nutrition with less and more nutrition. Therefore, in this study, we modeled nutritional status data based on the age factor using ordinal logistic regression with a truncated spline estimator.

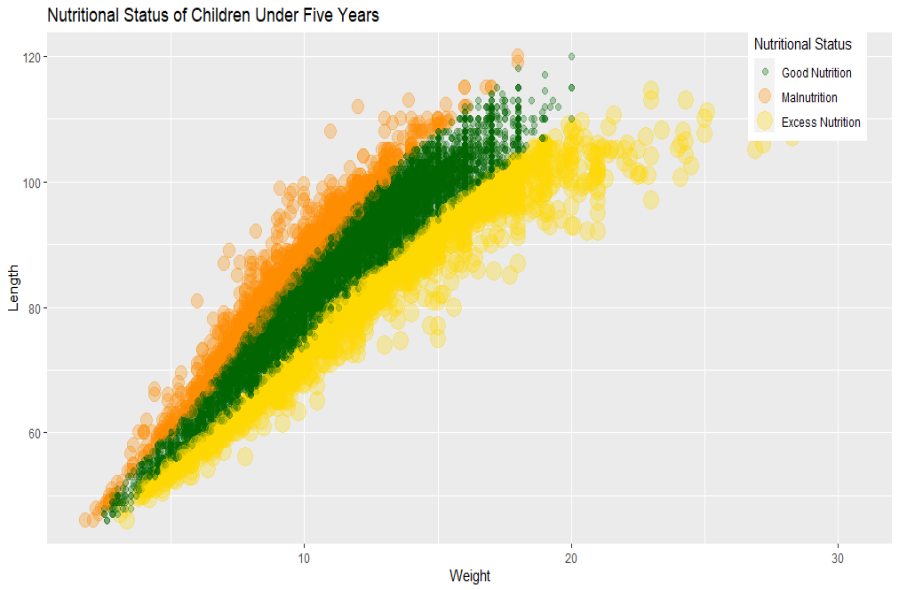

Figure 1. The plot of Nutritional Status Data for Toddlers in Gowa Regency, Indonesia

The knot point is chosen at the predictor variable interval, namely age, so the optimal knot point needs to be chosen to get the optimal model. For 1 knot point, the value obtained using the ordinal spline logistic regression model for each linear, quadratic, and cubic order is as shown in Table 1. For several knot points in Table 1, it can be seen that the 9-knot point in the linear order gives a minimum GCV value of 0.2159. Then it will be compared with the GCV value using 2-knot points in each linear, quadratic, and cubic order. GCV values using the ordinal spline logistic regression model with 2-knot points, as shown in Table 2.

Table 1. The GCV value of the ordinal spline logistic regression model with a 1-knot point

Knot point	The GCV value		
	Linear	Quadratic	Cubic
3	0.2191	0.2179	0.2183
6	0.2170	0.2179	0.2183
9	0.2159	0.2179	0.2184
12	0.2160	0.2180	0.2184
18	0.2165	0.2182	0.2186
24	0.2177	0.2185	0.2187
36	0.2186	0.2187	0.2188
48	0.2188	0.2190	0.2190
60	0.2192	0.2192	0.2192

Table 2. The GCV value of the ordinal spline logistic regression model with a 2-knot point

Knot point	Knot point	The GCV value		
		Linear	Quadratic	Cubic
3	9	0.2164	0.2166	0.2168
6	12	0.2160	0.2160	0.2169
9	18	0.2159	0.2159	0.2173
12	24	0.2160	0.2164	0.2164
18	24	0.2158	0.2169	0.2180
24	36	0.2175	0.2182	0.2185
36	48	0.2186	0.2187	0.2187
48	60	0.2188	0.2190	0.2191

Based on Table 2, for 2-knot points, the minimum GCV value is obtained in the linear order at points 18 and 24, equal to 0.2158. If the minimum GCV values obtained at the 1-knot point and 2-knot points are compared, it can be seen that the ordinal spline logistic regression model with 2-knot points in linear order is better to use because it gives the minimum GCV value. Therefore, data on the nutritional status of children under five is modeled using an ordinal spline logistic regression model with 2-knot points. The estimation results of the ordinal logistic regression model are shown in Table 3 below:

Table 3 Estimation of the parameters of the ordinal spline logistic regression model

Parameter	Estimation	Wald	Sig.	Information
α_1	-0.1669	-2.1833	0.000	Significant
α_2	0.7937	10.2892	0.000	Significant
β_{11}	-0.0794	-12.4759	0.000	Significant
β_{12}	0.1376	9.0248	0.000	Significant
β_{13}	-0.0918	10.2892	0.000	Significant

Based on Table 3 shows that age has a significant effect at a level of 5%, and the model is as follows:

$$logit\left[P\left(y_i \leq 1 \middle| x_{ji}\right)\right] = -0.1669 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+$$

$$logit[P(y_i \leq 2|x_{ji})] = 0.7937 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+$$

Furthermore, the probability function of each response category is obtained as follows:

$$\hat{\pi}_1 = \frac{\exp(-0.1669 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+)}{1 + \exp(-0.1669 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+)}$$

$$\hat{\pi}_2 = \frac{\exp(0.7937 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+)}{1 + \exp(0.7937 - 0.0794x_{1i} + 0.1376(x_{1i} - 18)_+ - 0.0918(x_{1i} - 24)_+)}$$

Based on the ordinal spline logistic regression model estimation results with 2-knot points, it shows three possible occurrences of toddler nutritional status based on age. First, toddlers aged 18 months have a 0.7549 times chance of getting good nutrition compared to toddlers over 18 months. Second, toddlers aged 18 to 24 months have a minimal chance of getting good nutrition compared to those under 18. Third, toddlers aged 24 months have a 0.3196 chance of getting good nutrition.

Table 4. The results of the classification of the nutritional status of children under five using the ordinal logistic regression model

Observation	Prediction			Total
	Malnutrition	Good nutrition	Excess nutrition	
Malnutrition	1576	368	0	1944
Good nutrition	526	13189	295	14010
Excess nutrition	0	175	1471	1646
Total	2102	13732	1766	17600
Accuracy				92.25%

Furthermore, the results of classifying the nutritional status of children under five use the ordinal logistic regression model with 2-knot points, as shown in Table 3. These results show a classification accuracy rate of 92.25%, meaning that the ordinal spline logistic regression model with 2-knot points is accurate in classifying data. Of course, the model can be accurate in modeling data on the nutritional status of toddlers in Gowa Regency, Indonesia, with the age factor because it gets a tremendous accuracy value.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interest.

REFERENCES

[1] X. Zhang, B. Li, H. Han, et al. Predicting multi-level drug response with gene expression profile in multiple myeloma using hierarchical ordinal regression, BMC Cancer. 18 (2018), 551. https://doi.org/10.1186/s12885-018-4483-6.

[2] W. Cao, V. Mirjalili, S. Raschka, Rank consistent ordinal regression for neural networks with application to age estimation, Pattern Recogn. Lett. 140 (2020), 325-331. https://doi.org/10.1016/j.patrec.2020.11.008.

[3] M. Rezapour, K. Ksaibati, Application of multinomial and ordinal logistic regression to model injury severity of truck crashes, using violation and crash data, J. Mod. Transport. 26 (2018), 268-277. https://doi.org/10.1007/s40534-018-0166-x.

[4] I.A.I. Ahmed, W. Cheng, The performance of robust methods in logistic regression model, Open J. Stat. 10 (2020), 127-138. https://doi.org/10.4236/ojs.2020.101010.

[5] Y. Song, J.A. Westerhuis, A.K. Smilde, Logistic principal component analysis via non-convex singular value thresholding, Chemometrics Intell. Lab. Syst. 204 (2020), 104089. https://doi.org/10.1016/j.chemolab.2020.104089.

[6] A. Moretti, Estimation of small area proportions under a bivariate logistic mixed model, Qual. Quant. 57 (2022), 3663-3684. https://doi.org/10.1007/s11135-022-01530-6.

[7] D. Yang, Y. He, B. Wu, et al. Drinking water and sanitation conditions are associated with the risk of malaria among children under five years old in sub-Saharan Africa: A logistic regression model analysis of national survey data, J. Adv. Res. 21 (2020), 1-13. https://doi.org/10.1016/j.jare.2019.09.001.

[8] L. Niu, A review of the application of logistic regression in educational research: common issues, implications, and suggestions, Educ. Rev. 72 (2018), 41-67. https://doi.org/10.1080/00131911.2018.1483892.

[9] Z. Nie, X. Bai, L. Nie, et al. Optimization of the economic and trade management legal model based on the support vector machine algorithm and logistic regression algorithm, Math. Probl. Eng. 2022 (2022), 4364295. https://doi.org/10.1155/2022/4364295.

[10] F. Thabtah, S. Hammoud, F. Kamalov, et al. Data imbalance in classification: Experimental evaluation, Inform. Sci. 513 (2020), 429-441. https://doi.org/10.1016/j.ins.2019.11.004.

[11] A. Islamiyati, , Anisa, M. Zakir, et al. The use of the binary spline logistic regression model on the nutritional status data of children, Commun. Math. Biol. Neurosci. 2023 (2023), 37. https://doi.org/10.28919/cmbn/7935.

[12] A. Abd Elaziz El-Sayed, S. Boulaaras, N.H. Sweilam, Numerical solution of the fractional-order logistic equation via the first-kind Dickson polynomials and spectral tau method, Math. Methods Appl. Sci. 46 (2021), 8004-8017. https://doi.org/10.1002/mma.7345.

[13] 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.

[14] B. Lestari, N. Chamidah, I. Nyoman Budiantara, et al. Determining confidence interval and asymptotic distribution for parameters of multiresponse semiparametric regression model using smoothing spline estimator, J. King Saud Univ. - Sci. 35 (2023), 102664. https://doi.org/10.1016/j.jksus.2023.102664.

[15] 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.

[16] 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.

[17] 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. Transition New Ser. 23 (2022), 201-212. https://doi.org/10.2478/stattrans-2022-0012.

[18] A.D. Permatasari, F.T. Waluyanti, The correlation between infant and toddler feeding practices by working mothers and the nutritional status, Enfermería Clínica. 29 (2019), 65-69. https://doi.org/10.1016/j.enfcli.2019.04.010.

[19] N.L. Suci, M.A. Azizah, Effectiveness of complementary feeding patterns on nutritional status in toddlers age 6-24 months: A systematic review, Int. J. Res. Publ. 115 (2022), 494-510. https://doi.org/10.47119/ijrp10011511220224363.

[20] Hijrawati, A.N. Usman, S. Syarif, et al. Use of technology for monitoring the development of nutritional status 1000 hpk in stunting prevention in Indonesia, Gaceta Sanitaria. 35 (2021), S231-S234. https://doi.org/10.1016/j.gaceta.2021.10.028.

[21] P.G. Brooker, M.A. Rebuli, G. Williams, et al. Effect of fortified formula on growth and nutritional status in young children: a systematic review and meta-analysis, Nutrients. 14 (2022), 5060. https://doi.org/10.3390/nu14235060.

[22] R.M. Marbun, S.M. Karina, M. Meilinasari, et al. Correlation of characteristics, maternal nutrition knowledge with nutritional status (H/A) in Baduta in Sumbang District, Banyumas Regency, Central Java, Indonesia, Open Access Maced. J. Med. Sci. 10 (2022), 471-474. https://doi.org/10.3889/oamjms.2022.8489.