International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 5 Issue 10, October - 2021, Pages: 50-53

Binary Logistic Model in Nonparametric Regression Through Spline Estimator

Dewi S Salam[1], Anna Islamiyati[2], Nirwan Ilyas[2]

[1,2,3]Departement of Statistics, Hasanuddin University, Makassar, 90245, Indonesia
annaislamiyati@unhas.ac.id

Abstract*: Spline Logistic Regression is a modeling solution of binary categorical response data which cannot be modeled by linear regression due to violations of normality assumption. The flexibility of the spline in estimating regression curve creates a modelling approach of the regression equation that is more fitted to data than ordinary logistic regression. Spline logistic regression model parameters are estimated by using Maximum Likelihood Method with Newton Raphson's iteration. The results show that the spline logistic regression function is a non-linear estimation that depends on the number of knots of predictor and the order used in the model.*

Keywords—binary; knot; logistic; nonparametric; truncated spline

1. INTRODUCTION

Regression is a statistical method used to explain the relationship between predictor and response variables. In the case of linear regression, it is assumed that responses and errors are normally distributed. Violation of the normality assumption makes the linear regression model unsuitable for binary response analysis and hence logistic regression is used to model data with binary response. Other methods that can be used to model data with categorical responses include Probit, and log complement. However, only logistic regression can be used to estimate odds ratios for the model predictor. Binary logistic regression has a nominal-scale response variable with two categories, success and failure, which the modeling is derived from The Bernoulli probability distribution function [1]. Various studies of logistic regression include predicting the relationship between multiple predictors and a categorical response using multivariable logistic regression method [2], modelling a binary response data with categorical predictors using binary logistic regression employing closed-form solution [3], testing the reliability of logistic regression in terms of the impact of sample size, nonlinear predictors, and multicollinearity in data [4], and using Group Lasso estimator for logistic regression for high dimensional data [5].

Logistic regression was developed through a nonparametric approach to get a model that was more suitable to the data. Although nonparametric approach looks more complex, it is the more appropriate option because its flexibility characteristic, where data will find its own actual regression curve shape [6]. Some estimators of nonparametric regression include Partitioning, Kernel, k-NN, Least Square, Spline, Neural Network, Orthogonal Series, Penalized Spline, Local Averaging, Semirecursive and Recursive [7]. The Spline regression curves estimator is the most often used because it can adjust effectively to changes in data behavior, so that the model is fitted [8]. Changes in the behavior of the data in spline are shown by knots that occur at the predictor. Spline has been developed in polynomial regression [9], weighting regression [10], and identified regression [11]. Spline regression curves have a very good ability to overcome data behavior changes at certain sub-intervals [12]. The prominent of polynomial truncated spline method in estimating changes in data behavior at certain intervals lies in its truncated function with the best model criteria determined by the number and location of the knots in data [13]. The polynomial truncated spline is easy to interpret [14], and provide simple approximation models of complicated data pattern which are difficult and impracticable to model accurately [15]. Therefore, we use a spline polynomial truncated estimator in logistic regression model with the parameter estimation method employing Maximum Likelihood.

The maximum likelihood method is used as an estimation and inference tool that has optimal properties for large sample sizes [16]. The main purpose of using maximum likelihood is to estimate parameters by maximizing the likelihood function that has been transformed into a log-likelihood. In cases where the log-likelihood function cannot be solved explicitly, Newton-Raphson's numerical method is one of the fastest and most applicable methods for maximizing the log-likelihood function [17].

This article is divided into five parts. In the second part, the development of spline truncated logistic regression models is described for binary response data. The third part explains the estimation of binary spline logistic regression parameters using Maximum Likelihood with the help of Newton Raphson's numerical method to solve the implicit function. Next in the fourth part, we show the application of the method to a simulation data and the fifth part provides conclusions related to the results of this study. Furthermore, the development of logistic regression theory with the spline nonparametric approach is expected to be a reference for the use of data analysis methods, particularly to the data with binary responses.

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 5 Issue 10, October - 2021, Pages: 50-53

2. SPLINE LOGISTIC REGRESSION

Nonparametric regression model as a relationship form between response and predictor can be stated as follows:

$$y_i = f(x_i) + \varepsilon_i \quad (1)$$

where y is the response, x is the predictor, ε is the error and *i* is the *1,2, ..., n* sample. Suppose the number of predictors considered in the model is *j = 1,2, ..., m*, then the equation (1) becomes:

$$y_i = f(x_{ji}) + \varepsilon_i$$

The function $f(x_{ji})$ is an unknown shape function, assumed to be smooth and contained in a Sobolev Space. In this article, we estimate $f(x_{ji})$ by using the truncated polynomial spline estimator. In nonparametric regression, spline has an ability to estimate the data pattern that tends to be different at different intervals [8]. The ability to estimate the data pattern is shown by the truncated (pieces) attached to the estimator, these pieces are called knots. Knots are joint fusion points that indicate changes in the behavior patterns of functions at different intervals. Knot points are taken at intervals of $a < k_r < b$, where a is the minimum value and b is the maximum value of the data [18].

The truncated spline function which states the relationship between *p* predictors and a single response is expressed with the spline function $f(x_{ji})$ as follows:

$$f(x_{ji}) = \alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} k_{jh})_{+}^{q} \quad (2)$$

where α_{0j} is the j^{th} intercept predictor, α_{jl} is the polynomial parameter at the j^{th} predictor and l^{th} order, $\alpha_{j(q+h)}$ is the truncated parameter at the j^{th} predictor, h^{th} knot point, and q order. $(x_{ji} - k_{jh})_{+}^{q}$ is a truncated polynomial function described as follows:

$$(x_{ji} - k_{jh})_{+}^{q} = \begin{cases} (x_{ji} - k_{jh})^{q}; & x_{ji} \geq k_{jh} \\ 0 & ; x_{ji} < k_{jh} \end{cases}$$

where q is the order of the polynomial spline truncated and k_{jh} is the h^{th} knot point, (*h = 1,2, ..., r*).

For *n* observational data with *k* knot points, Equation (2) can be expressed in the form of a matrix as follows:

$$\boldsymbol{y} = \boldsymbol{X}[k_1, k_2, \ldots k_r]\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where *y* is a vector with $n \times 1$ size, $\boldsymbol{X}$ is a matrix with $n \times (1 + q + r)$ size, α is a vector with $(1 + (q + r) \times 1$ size, and ε is a vector with $n \times 1$ size [19].

Binary logistic regression analysis is used to explain the relationship between a response variable that is dichotomic/ binary and predictor variables that is interval or categorical [20]. The binary logistic regression model is derived from Bernoulli Exponential distribution, which is the distribution of random variables that only has 2 categories, 1 for success and 0 for failure [21]. The Bernoulli distribution probability function is shown as follows:

$$f(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

If $y_i = 0$ then $f(y_i) = 1 - \pi_i$ and if $y_i = 1$ then $f(y_i) = \pi_i$, π_i is the probability of the i^{th} event and i^{th} random variable.

Based on equation (2), the spline logistic regression model that is influenced by *p* predictor variables can be expressed as the expected value of *y* with respect to *x*, so that:

$$E(y_i|x_i) = \frac{\exp(\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})}{1 + \exp(\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})} \quad \ldots (3)$$

where $0 \leq E(y_i|x_i) \leq 1$, y_i has values 0 or 1, the value 1 is the success probability of $E(y_i|x_i)$, so the equation (3) can be stated as follows:

$$\pi(x) = \frac{\exp(\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})}{1 + \exp(\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q})}$$

Logit transformation of $\pi(x)$ yields truncated polynomial spline regression as follows:

$$g(x) = ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}$$

where $\pi(x)$is the probability of occurrence and $g(x)$ is logit estimation value.

Based on the description of the truncated spline regression model and logistic regression, the spline logistic regression model can be developed as follows:

$$\pi(x_{ij}) = \frac{\exp\{\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}\}}{1 + exp\{\alpha_{0j} + \sum_{l=1}^{q} \alpha_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \alpha_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}\}} + \varepsilon_i \quad \ldots (4)$$

where α_{0j}, α_{jl},…, $\alpha_{j(q+h)}$ is the binary spline logistic regression coefficient and k is the knot point.

3. ESTIMATION OF SPLINE LOGISTIC REGRESSION PARAMETERS USING MAXIMUM LIKELIHOOD

Based on equation (4), the parameter to be estimated is α by using maximum likelihood estimator. The maximum likelihood method estimates the coefficient α by maximizing the likelihood function and requires that the data must follow a certain distribution. Spline logistic regression is derived from Bernoulli distribution and assumed to be independent, so that the likelihood function can be said as combination of each distribution function. It is known that y_i has a Bernoulli distribution, with the probability density function as follows:

$$f(y_i) = \pi(x_{ij})^{y_i}(1 - \pi(x_{ij}))^{1-y_i}$$

with likelihood function can be stated as follows:

$$L(\boldsymbol{\alpha}) = \prod_{i=1}^{n} \pi(x_{ij})^{yi}(1 - \pi(x_{ij}))^{1-yi} \quad (5)$$

The likelihood function in equation (5) is changed in the form of log natural function so its easier to maximize:

$$ln\, L(\boldsymbol{\alpha}) = \ln\left(\prod_{i=1}^{n} \left(\pi(x_{ij})\right)^{y_i} (1 - \pi(x_{ij}))^{1-yi}\right)$$

$$\ln L(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \left\{ y_i \left(\pi(x_{ij}) \right) - \ln\left[1 + \exp(\pi(x_{ij})\right] \right\} \quad (6)$$

Equation (6) is derived to the parameters α or $\pi(x_{ij})$, so we have:

$$\frac{\partial \ln L(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}} = \sum_{i=1}^{n} \left\{ y_i - \frac{\exp\left(\hat{\pi}(x_{ij})\right)}{1 + \exp\left(\hat{\pi}(x_{ij})\right)} \right\} \quad (7)$$

Estimator α is obtained by solving the equation (7), however the equation is implicit and difficult to solve explicitly, so in order to get α estimator from $L(\alpha)$ nonlinear function, the Newton Raphson method is used as an iteration method to solve nonlinear equations [22]. Based on Newton Raphson's iteration, a second derivative of the likelihood function is obtained for each parameter. Newton Raphson's iteration used to estimate the α parameter can be written as follows:

$$\boldsymbol{\alpha}_{(t+1)} = \boldsymbol{\alpha}_{(t)} - \boldsymbol{H}_{(t)}^{-1} \boldsymbol{D}_{(t)}$$

where $\boldsymbol{D}_{(t)}$ is the first derivative matrix for each parameter α, $\boldsymbol{H}_{(t)}$ is the second derivative matrix for each parameter α. Iteration ends if $\hat{\boldsymbol{\alpha}}_{(t+1)} \cong \hat{\boldsymbol{\alpha}}_{(t)}$ is obtained.

At the end of the iteration, the parameter estimation of logistic spline regression becomes convergent and we obtain the estimated binary spline logistic regression parameters as follows:

$$\hat{\pi}(x_{ij}) = \frac{\exp\{\hat{\alpha}_{0j} + \sum_{l=1}^{q} \hat{\alpha}_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \hat{\alpha}_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}\}}{1 + \exp\{\hat{\alpha}_{0j} + \sum_{l=1}^{q} \hat{\alpha}_{jl} x_{ji}^{l} + \sum_{h=1}^{r} \hat{\alpha}_{j(q+h)} (x_{ji} - k_{jh})_{+}^{q}\}}$$

where $\hat{\pi}(x_{ij})$ is the estimated success probability, $\hat{\alpha}_{0j}$ is the estimated j^{th} intercept predictor, $\hat{\alpha}_{jl}$ is the estimated polynomial parameter at the j^{th} predictor and l^{th} order, $\alpha_{j(q+h)}$ is the estimated truncated parameter at the j^{th} predictor, h^{th} knot point, and q^{th} order.

4 CONCLUSION

Binary categorical response data cannot be modeled using linear regression due to a violation of the normality assumption in errors and responses. One possible approach that can be used to overcome this problem is using binary spline logistic regression where the use of spline plays a role in the flexibility of establishing a regression curve based on data. The method used to estimate binary spline logistic regression parameters is a maximum likelihood estimator with the Newton Raphson iteration $\boldsymbol{\alpha}_{(t+1)} = \boldsymbol{\alpha}_{(t)} - \boldsymbol{H}_{(t)}^{-1} \boldsymbol{D}_{(t)}$. The focus of this article is on the evolution of more flexible estimation methods and in-depth analysis of binary response data modeling, with the expectation that this flexibility can be widely used for the development of data analysis methods and software. This article uses binary categorical responses with the order q spline degree, so further we recommend the development of research on responses that have more than two categories by considering violations of the linear regression assumptions.

The results of the estimation of binary spline logistic regression through simulation data show a smaller AIC value than the classical logistic approach. This becomes a reference that for the analysis of categorical response data, it can be analyzed with a nonparametric regression approach in particular the use of a truncated spline estimator with its optimal knot points. We hope that the application of this method can be used in real data to get more accurate estimation results.

5 REFERENCES

[1] Hilbe, J. M. (2009). *Logistic Regression Models* (New York: Chapman and Hall/CRC Taylor & Francis Group).

[2] Park, H.A. (2013). An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, 43 (2): 154.

[3] Lipovetsky, S. (2015). Analytical closed-form solution for binary logit regression by categorical predictors *Journal of Applied Statistics*. 42 (1): 37-49.

[4] Bergtold, J.S., Yeager, E.A and Featherstone, A.M. (2018). Inferences from logistic regression models in the presence of small samples, rare events, nonlinearity, and multicollinearity with observational data. *Journal of Applied Statistics*. 45 (3): 528-546.

[5] Meier, L, Van, D.G.S. and Bühlmann. P. (2008). The group Lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*. 70(1): 53–71 .

[6] Islamiyati, A, Fatmawati and Chamidah. N. (2020). Penalized spline estimator with multi smoothing parameters in biresponse multipredictor regression model for longitudinal data. *Songklanakarin Journal of Science and Technology*. 42 (4): 897-909.

[7] Gyorfi, L, Krzyzak, A, Kohler, M and Walk, H. (2002) *A Distribution-Free Theory of Nonparametric Regression* (New York: Springer Inc).

[8] Eubank, R.L. (1999). *Nonparametrik Regression and Spline Smoothing Second Edition* (New York: Marcel Dekker Inc).,

[9] Islamiyati, A, Sunusi, N, Kalondeng, A, Fatmawati F. and Chamidah, N. (2020). Use of two smoothing parameters in penalized spline estimator for bi-variate predictor non-parametric regression model. *Journal of Sciences, Islamic Republic of Iran*. 31 (2): 175-183.

[10] Islamiyati, A, Fatmawati and Chamidah, N. (2018) Estimation of covariance matrix on bi-response longitudinal data analysis with penalized spline regression. *Journal of Physics: Conference Series*. 97**9** (1): 012093.

[11] Kasahara, H and Shimotsu, K. (2019). Identification of Regression Models with a Misclassified and Endogeneus Binary Regressor. *Econometrics Arxiv* 1904.11143.

[12] Cox, D.D. and O'Sullivan, F. (1996). Penalized Likelihood-Type Estimators for Generalized Nonparametric Regression. *Journal of Multivariate Analysis*. 56: 185-206.

[13] Islamiyati, A, Fatmawati and Chamidah, N. (2020). Changes in blood glucose 2 hours after meals in Type 2

diabetes patients based on length of treatment at Hasanuddin University Hospital. *Indonesia Rawal Medical Journal*. 45 (1): 31-34.

[14] Perperoglou, A., Sauerbrei, W., Abrahamowicz, M. and Schmid, M. (2019) A review of spline function procedures in R. *BMC Med Res Methodol*. 19 (46) .

[15] Bessaoud, F., Daures, J.P. and Molinari, N. (2005). Free knot splines for logistic models and threshold selection. *Computer methods and programs in biomedicine*. **77:** 1-9.

[16] Millar, R.B. (2011). *Maximum Likelihood Estimation and Inference with examples in R, SAS, and ADMB* (Chichester, West Sussex: John Wiley & Sons, Ltd)

[17] Harrell, F.E. (2015). *Regression Modeling Strategies with Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis 2nd Edition* (Switzerland: Springer Series in Statistics).

[18] Hardle, W. (1990). *Applied Nonparametric Regression* (New York: Cambridge University).,

[19] Ramli, M., Ratnasari, V. and Budiantara, I.N. (2020). Estimation of Matrix Variance-Covariance on Nonparametric Regression Spline Truncated for Longitudinal Data. *J. Phys.: Conf. Ser.* 1562 012014.

[20] Hosmer, D.W. and Lemeshow, S. (2000). *Applied Logistic Regression 2nd Edition* (New York: John Wiley and Sons Inc).

[21] Islamiyati, A, Fatmawati and Chamidah, N. (2019). Ability of covariance matrix in bi-response multi-prredictor penalized spline model through longitudinal data simulation International. *Journal of Academic and Applied Research*. 3 (3): 8-11.

[22] Agresti, A. (1996). *An Introduction to Categorical Data Analysis* (New York: John Wiley and Sons Inc).