International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 7 Issue 1, January - 2023, Pages: 150-153

Estimation of Biresponse Semiparametric Regression Model with Weighted Spline Truncated

Ahmad Yani[1], Anna Islamiyati[2], Nurtiti Sunusi[3]

[1,2,3]Department of Statistics, Hasanuddin University, Makassar, 90245, Indonesia
annaislamiyati701@gmail.com

Abstract: Semiparametric regression is a combination of parametric regression and nonparametric regression. Because there is a relationship between predictor variables and response variables that are known patterns and unknown patterns. The purpose of this research is to get the best model of biresponse data. The parametric regression approach uses linear regression, while the nonparametric regression approach uses weighted spline truncated. Determination of the best parametric regression model is based on the largest deterministic coefficient, while the determination of the nonparametric regression model is based on the CGV criterion to get the best model.

Keywords—biresponse; GCV; knot; semiparametric; truncated spline.

1. INTRODUCTION

The regression approach aims to see the effect of predictor variables on response variables. There are three types of regression based on data patterns, which are parametric regression, nonparametric regression and semiparametric regression. The data pattern between the predictor and the response is the basis for using the three regressions. Data patterns can be known from initial data plots, information from previous research, existing theoretical studies, or initial assumptions made by researchers for certain reasons. Parametric regression is used when the data pattern follows a parametric pattern and nonparametric regression is used when the data pattern does not follow a parametric pattern. Semiparametric regression is a combination of parametric and nonparametric regression. This means that semiparametric regression is used when there are parametric patterns and nonparametric patterns so that semiparametric regression estimation is equivalent to estimating parameters on parametric and nonparametric components [1].

Several semiparametric regression studies on one response have been developed by researchers, including [2] developing semiparametric uniresponse with spline smoothing. [3] developed ridge regression in a semiparametric regression approach. [4] has developed a uniresponse semiparametric regression model using spline smoothing. [5] used Cox semiparametric. Another estimator that can be used in the biresponse case is the spline truncated estimator which has an easier visual interpretation. Spline truncated considers knot points that show the point where the pattern of change in the data occurs. This is the advantage of the spline truncated because it is able to explain several patterns of change that can occur in one model [6]. The use of spline truncated is widely used in some real data because of these advantages, including [7] applying to tuberculosis data, [8] developing its use in diabetes data.

Along with the development of data, the problem of the number of responses is not only limited to one, but has developed in the number of responses of two or more. [9] developed biresponse data analysis with local linear estimator, [10] with smoothing spline on longitudinal biresponse data. [11] developed penalized spline on longitudinal data of responses, [12] with mixed estimator between kernel and fourier series. If the response case is more than one and assumed to be correlated, then the data analysis process cannot be modeled with the usual approach as in the case of uniresponse. [13] has developed a weighted partial spline for correlated data. [14] developed a weighted smoothing spline. [15] developed a weighted local polynomial. [16] developed penalized weighted spline with variance covariance matrix in the birresponse case.

Considering the estimation procedures carried out by previous researchers in this study, a spline truncated approach is given. The expectation of this study is to provide mathematical calculations and statistical interpretations that are simple and easy to apply. The spline truncated function is used to approximate nonparametric regression on semiparametric biresponse regression functions. The optimization used in estimating the regression parameters is the *Weighted Least Square* (WLS) method because the response used is more than one. Based on the problem formulation above, the purpose of this study is to model biresponse semiparametric regression with a weighted spline truncated approach.

2. REGRESSION

Regression analysis is a statistical analysis used to see the relationship and functional influence between the independent variable and the dependent variable. The most common and frequently used approach is the parametric approach. This approach assumes that the form of the regression model is known. Suppose a multivariable linear parametric regression model is given as follows

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i \qquad (1)$$

with y is the response variable, x is the predictor, β is the regression coefficient, ε is the error and i is the number of

International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 7 Issue 1, January - 2023, Pages: 150-153

samples running from $1,2,\dots,n$. The next, equation (1) can be written in matrix form, then the parametric regression equation becomes:

$$\mathrm{Y} = \mathrm{X}\beta + \varepsilon \qquad (2)$$

As for data for which no information about the model and regression function is known, nonparametric regression is used. Nonparametric regression provides flexibility in modeling, because there is no assumption of a certain curve shape. [17] gives the equation for the nonparametric regression model as follows

$$.\,\mathrm{y_i} = \mathrm{f(t_i)} + \varepsilon_\mathrm{i} \qquad (3)$$

with y is the response variable, $f(t_i)$ is a function of unknown shape, ε is the error and i is the number of samples running from $1,2,\dots,n$. Next, equation (2) can be written in matrix form, then the parametric regression equation becomes:

$$\mathrm{Y} = \mathrm{f(T)} + \varepsilon \qquad (4)$$

Semiparametric regression is one part of inferential statistics that is used to model the relationship between response and predictor variables, some of which have known patterns and others have unknown patterns. Suppose there is paired data $\left(x_i, y_i, t_i\right)$ and the relationship between x_i, y_i and t_i is assumed to follow the following regression model [18]

$$y_i = \beta_0 + \beta_1 \mathrm{x_{1i}} + \beta_2 \mathrm{x_{2i}} + \cdots + \beta_\mathrm{p} \mathrm{x_{pi}} + f\left(t_i\right) + \varepsilon_i \qquad (5)$$

with y is the response variable, x is the predictor,β is the regression coefficient, $f(t_i)$ is a function of unknown shape, ε is the random error which is distributed $N(0,\sigma)$ and i is the number of samples running from *1,2, ..., n*. Equation (3) can be written in matrix form as follows

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{f}(\boldsymbol{T}) + \boldsymbol{\varepsilon} \qquad (6)$$

Spline regression is a nonparametric regression consisting of power cuts of a certain order and connected at tie points. The abscissa value of the tie points is commonly called knots. Knot is defined as a focal point in the spline function, so that the curve formed is segmented at that point [19] An optimum knot value is later selected from various knots that have been tried. This optimum knot is used as a parameter to estimate the regression model.

The p-order spline function can be expressed as follows [20]

$$f\left(x_i\right) = \beta_0 \mathrm{x_i}^0 + \beta_1 \mathrm{x_i}^1 + \beta_2 \mathrm{x_i}^2 + \cdots + \beta_\mathrm{p} \mathrm{x_i}^\mathrm{p} + \sum_{k=1}^{q} \beta_{p+k} \left(x_i - K_k\right)^{+p} \qquad (7)$$

with x is the predictor,β is the regression coefficient, q is the number of knots and K_k, $k = 1,2,\dots,q$.

$$\left(x_i - K_k\right)^{+\boldsymbol{p}} = \begin{cases} \left(x_i - K_k\right)^{\boldsymbol{p}} & , x \geq K_k \\ 0 & , x < K_k \end{cases}, a < K_1 < K_2 < \cdots < K_k < b \qquad (8)$$

where a and b are the smallest and largest values of the data, respectively. The "+" sign in the equation (2.9) indicates that only the result $\left(x_i - K_k\right)^{\boldsymbol{p}} \geq 0$ will be taken while those with values < 0 are assumed to be 0.

If there are a number of n observations, the function matrix is written as follows

$$Y = f(x) = \mathbf{X}\boldsymbol{\beta} \qquad (9)$$

In spline regression, segments are separated by knots. Therefore, the location and number of knots will determine the goodness of spline regression on the data [17]. The optimum knot is selected based on the smallest Generalized Cross Validation (GCV) value [21], which can be calculated by the following equation [22].

$$GCV\left(K_1,\dots,K_r\right) = \frac{MSE\left(K_1,\dots,K_r\right)}{\left(n^{-1} tr\left(\boldsymbol{I} - \boldsymbol{C}\left(K_1,\dots,K_r\right)\right)\right)^2} \qquad (10)$$

With MSE obtained from equation

$$MSE\left(K_1,\dots,K_r\right) = n^{-1}\boldsymbol{Y}'\left(\boldsymbol{I} - \boldsymbol{C}\left(K_1,\dots,K_r\right)\right)'\left(\boldsymbol{I} - \boldsymbol{C}\left(K_1,\dots,K_r\right)\right)\boldsymbol{Y} \qquad (11)$$

$\boldsymbol{C}\left(K_1,\dots,K_r\right)$ is a heat matrix that contains the values of K of the insulation parameters, r is the number of nonparametric component variables [23].

3. ESTIMATION PARAMETERS OF THE BIRESPON SEMIPARAMETRIC REGRESSION MODEL

Suppose given paired birresponse data with two response variables and $p+q$ predictor variables $\left(y_1, y_2, x_1, x_2, \dots, x_p, t_1, t_2, \dots, t_q\right)$. The predictor variables $x_1, x_2, \dots, x_p$ follow a certain pattern while the predictor variable $t_1, t_2, \dots, t_q$ does not have a certain pattern. The biresponse semiparametric regression model containing these variables can be expressed as follows.

$$\begin{aligned} y_{1i} &= \beta_{01} + \beta_{11} x_{11i} + \cdots + \beta_{p1} x_{p1i} + f_1\left(t_{ri}\right) + \varepsilon_{1i} \\ y_{2i} &= \beta_{02} + \beta_{12} x_{12i} + \cdots + \beta_{p2} x_{p2i} + f_2\left(t_{ri}\right) + \varepsilon_{2i} \end{aligned} \qquad (12)$$

for $i = 1,2,\dots,n$ and $r = 1,2,\dots,q$.

In the function $f(t_{ri})$ in equation (12) will be approximated with spline truncated function so that the function can be denoted as in equation (7). So that the form of the equation (12) can be written in matrix form

$$[\mathbf{Y}] = [\mathbf{X}\boldsymbol{\beta}] + [\mathbf{T}\boldsymbol{\alpha}] + [\boldsymbol{\varepsilon}] \qquad (13)$$

The semiparametric birresponse regression function shown in equation (13) is estimated using the Weighted Least Square (WLS) method. The WLS estimator combines the goodness of fit function and penalty function by involving a weighting

matrix $\mathbf{V} = \mathbf{diag}\frac{\mathbf{1}}{\mathbf{2}n}$. Suppose the WLS estimator is symbolised by $\boldsymbol{\omega}$, then $\boldsymbol{\omega}$ can be written as follows:

$$\boldsymbol{\omega}(\boldsymbol{\beta}, \boldsymbol{\alpha}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{T}\boldsymbol{\alpha})^{\mathbf{T}}\mathbf{V}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{T}\boldsymbol{\alpha}) \quad (14)$$

To obtain estimators of the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, optimization will be performed on the equation (14) so that the estimator for the parametric component is obtained as follows:

$$\widehat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{K})\mathbf{Y} \quad (15)$$

where $\mathbf{B}(\mathbf{K}) = \mathbf{M}(\mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{X})^{-\mathbf{1}}\{\mathbf{X}^{\mathbf{T}} - \mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{T}(\mathbf{T}^{\mathbf{T}}\mathbf{V}\mathbf{T})^{-\mathbf{1}}\mathbf{T}^{\mathbf{T}}\}$
and the estimator for the nonparametric component:

$$\widehat{\boldsymbol{\alpha}} = \mathbf{A}(\mathbf{K})\mathbf{Y} \quad (16)$$

where $\mathbf{A}(\mathbf{K}) = \mathbf{N}(\mathbf{T}^{\mathbf{T}}\mathbf{V}\mathbf{T})^{-\mathbf{1}}\{\mathbf{T}^{\mathbf{T}} - \mathbf{T}^{\mathbf{T}}\mathbf{V}\mathbf{X}(\mathbf{X}^{\mathbf{T}}\mathbf{V}\mathbf{X})^{-\mathbf{1}}\mathbf{X}^{\mathbf{T}}\}\mathbf{V}$
The estimator of the semiparametric biresponse spline regression model is as follows.

$$\widehat{\mathbf{Y}} = \mathbf{C}(\mathbf{K})\mathbf{Y} \quad (17)$$

where $\mathbf{C}(\mathbf{K}) = \mathbf{X}\mathbf{B}(\mathbf{K}) + \mathbf{T}\mathbf{A}(\mathbf{K})$
The value of *K* The optimal value of the model is obtained using the GCV (Generalized Cross Validation) method whose formula is given as follows:

$$\mathbf{GCV}(\mathbf{K_1}, \dots, \mathbf{K_r}) = \frac{(\mathbf{2n})^{-\mathbf{1}}\mathbf{Y}^{\mathbf{T}}\big(\mathbf{1} - \mathbf{C}(\mathbf{K})\big)^{\mathbf{T}}\big(\mathbf{1} - \mathbf{C}(\mathbf{K})\big)\mathbf{Y}}{\Big((\mathbf{2n})^{-\mathbf{1}}\mathbf{trace}\big(\mathbf{1} - \mathbf{C}(\mathbf{K})\big)\Big)^{\mathbf{2}}} \quad (18)$$

where $\mathbf{K} = (K_1, \dots, K_r)$.

4 CONCLUSION

The birresponse semiparametric regression model with weighted *truncated spline* is a model consisting of two main components, i.e. there is a linear parametric component and a nonparametric component. The equation for this model is $[\mathbf{Y}] = [\mathbf{X}\boldsymbol{\beta}] + [\mathbf{T}\boldsymbol{\alpha}] + [\boldsymbol{\varepsilon}]$ where $\mathbf{Y}$ is the biresponse, $\mathbf{X}\boldsymbol{\beta}$ is the parametric component and $\mathbf{T}\boldsymbol{\alpha}$ is the nonparametric component approximated by a weighted spline truncated. The parameters in the model are approximated using the Weighted Least Square (WLS) method with weights $\mathbf{V} = \mathbf{diag}\frac{\mathbf{1}}{\mathbf{2n}}$. This model is expected to be able to approach the biresponse pattern well because each data is approximated by a curve that fits the data.

5 REFERENCES

[1] Budiantara, I. N. (2006), Model spline dengan knots optimal. *Jurnal Ilmu Dasar, FMIPA Universitas Jember* 7 (6), 77-85.

[2] Budiantara, I., N. (2005), Model Keluarga Spline Polinomial Truncated dalam Regresi Semiparametrik. *BIMIPA* 15 (3), 55-61.

[3] Hu, H. (2005). Ridge estimation of a semiparametric regression model. *Journal of Computational and Applied Mathematics*, *176*(1), 215–222.

[4] Sugiantari, A. P. & Budiantara, I. N. (2013), Analisis Faktor-faktor yang Mempengaruhi Angka Harapan Hidup di Jawa Timur Menggunakan Regresi Semiparametrik Spline. *Jurnal Sains dan Seni POMITS* 2 (1), D37-D41.

[5] Chandra, N. E. & Rohmaniah, S. A. (2019), Analisis Survival Model Regresi Semiparametrik Pada Lama Studi Mahasiswa. *Jurnal Ilmiah Teknosains*, 5(2), 94-98.

[6] Islamiyati, A. (2017), Spline Polynomial Truncated dalam Regresi Nonparametrik. *Jurnal Matematika, Statistika dan Komputasi* 14 (1), 54-60.

[7] Anggreni, N. P. R., Suciptawati, N. L. P. & Srinadi, I. G. A. M. (2018), Model Regresi Nonparametrik Spline Truncated Pada Jumlah Kasus Tuberkulosis Di Provinsi Bali Tahun 2016. *E-Jurnal Matematika* Vol. 7(3), 211-218.

[8] Ramdhani, Z. A., Islamiyati, A. & Raupong, R. (2020), Hubungan Faktor Kolestrol Terhadap Gula Darah Diabetes dengan Spline Kubik Terbobot. *ESTIMASI: Journal of Statistics and Its Application*, 1(1), 32-39.

[9] Chamidah, N. & Rifada, M. (2016), Local linear estimator in bi-response semiparametric regression model for estimating median growth charts of children. *Far East Journal of Mathematical Sciences* 99 (8), 1233-1244.

[10] Fernandes, A. A. R., Budiantara, I. N., Otok, B. W., & Suhartono. (2015). Spline Estimator for Bi-Responses and Multi-Predictors Nonparametric Regression Model in Case of Longitudinal Data. *Journal of Mathematics and Statistics*, *11*(2), 61–69.

[11] Islamiyati, A., Fatmawati & Chamidah, N. (2020), Penalized spline estimator with multi smoothing parameters in bi-response multi-predictor nonparametric regression model for longitudinal data. *Songklanakarin Journal of Science & Technology* 42 (4), 897-909.

[12] Hidayat, R., Budiantara, I. N., Otok, B. W. & Ratnasari, V. (2020), The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model. *Communications in Statistics-Theory and Methods*, *50*(17), 3942–3953.

[13] Budiantara, I. N., Subanar & Zoyuti. (1997), Weighted Spline Estimator. *Bulletin of the International Statistical Institute* 51 (1), 333-334.

[14] Davies, P. L., & Meise, M. (2008). Approximating Data with Weighted Smoothing Splines. *Journal of Nonparametric Statistics*, *20*(3), 207–228.

[15] Chamidah, N., Budiantara, I. N., Sunaryo, S., & Zain, I. (2012). Designing of Child Growth Chart Based on Multi-Response Local Polynomial Modeling. *Journal of Mathematics and Statistics*, *8*(3), 342–347.

[16] Islamiyati, A. (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.

[17] Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). CRC Press.

[18] Ruppert, D., Wand, M. P., & Carroll, R. J. (2003). *Semiparametric Regression*. Cambridge University Press.

[19] Tripena, A. (2011). Analisis regresi spline kuadratik. *Seminar Nasional Matematika Dan Pendidikan Matematika*, MS 8-MS 18.

[20] Biedermann, S., Dette, H., & Woods, D. C. (2009). *Optimal Designs for Multivariable Spline Models* (No. 823; SFB).

[21] Spiriti, S., Smith, P., & Lecuyer, P. (2018). *Freeknotsplines: Algorithms for Implementing Free-Knot Splines* (1.0.1). R Package.

[22] Takezawa, K. (2006). *Introduction to Nonparametric Regression*. John Wiley & Sons, Inc.

[23] Tripena, A. (2013). Estimator Deret Fourier untuk Estimasi Kurva Regresi Nonparametrik Birespon. *Magistra*, *84*, 6–15.