International Journal of Academic and Applied Research (IJAAR)
ISSN: 2643-9603
Vol. 5 Issue 11, November - 2021, Pages: 166-168

Estimation of Penalized Spline Linear Regression Models through Robust M Estimator

Musafirah[1], Anna Islamiyati[2], Nurtiti Sunusi[3]

[1,2,3]Department of Statistics, Hasanuddin University, Makassar, 90245, Indonesia
annaislamiyati70@gmail.com

***Abstract**: Parametric regression approaches commonly used include simple linear, quadratic, and cubic linear regression. However, its use cannot be used for all data in the real case. Many data have data plots that do not follow parametric patterns, so we must use other approaches, including nonparametric regression. The spline is essentially a generalization of polynomial functions, where the optimization still adopts the concept in the parametric regression approach. The finalized spline regression curve is formed by minimizing the total residual subjects to the size of the spline coefficients. It seems that the least squared spline regression is not resistant to outliers.*

Keywords— penalized spline; outlier; robust M estimator

1. INTRODUCTION

Regression analysis has become one of the statistical methods that is very instrumental in the development of statistical science, especially in looking at the pattern of data pair relationships between predictor variables with response variables. In regression analysis, there are two approaches commonly used to estimate the regression curve, namely the parametric regression approach and nonparametric regression. The parametric regression approach is used if the shape of the regression curve is known. While the nonparametric regression approach is used if information about the form and pattern of the relationship between the predictor variables and the response variable is unknown [1]. When a data does not have information about a data plot it uses nonparametric regression [2].

Parametric regression approaches commonly used include simple linear, quadratic, and cubic linear regression. However, its use cannot be used for all data in the real case. Many data have data plots that do not follow parametric patterns, so we must use other approaches, including nonparametric regression. Several estimators have been developed by researchers, including spline [3], kernel [4], MARS [5], and several others. Spline estimator is one estimator that is widely used because of its flexible nature, which is the data itself which looks for suitable data patterns. Some estimator of spline includes spline truncated [6], spline smoothing [7], penalized spline [8] and b spline [9].

Some previous researchers in the field of financial mathematics have an interest in the penalized spline method, for example in the fields of theoretical and application [10] and nonparametric bayesian analysis [11]. The spline is essentially a generalization of polynomial functions, where the optimization still adopts the concept in the parametric regression approach. One of the advantages of spline is that it can overcome data patterns that show changes in behavior in certain sub-intervals with the help of knots, and the resulting curve is relatively smooth [12]. The calculated spline regression curve is formed by minimizing the total residual subjects to the size of the spline coefficients. So that the formula is limited to the problem of minimal minimization. Apparently, the least squared spline regression is not resistant to outliers [13]. The simple idea is to replace squared residuals with a number of loss functions that are slowly increasing the same as those used in the M-regression estimator [14,15] to reduce the outlier effects of an observation type.

Using estimator M calculations, Lee and Oh [16] propose an iterative algorithm by introducing empirical pseudo data. Furthermore, Finger [17] suggested that the M estimator was obtained by combining a very strong estimator with the least efficient quadratic type estimator for penalized spline regression. One of the criteria for determining the best estimator in nonparametric regression is to use Mean Absolute Deviation (MAD). Mean Absolute deviation (MAD) measures the accuracy of the predictions of the estimated error estimates (the absolute value of each error). MAD is useful to measure error predictions in the same section as the original [18].

2. PENALIZED SPLINE LINEAR REGRESSION MODEL

The spline is a function of the order polynomial chunks q with the joint points of the pieces called knots. The knot point is a combination of two curves that show the pattern of changes in curve behavior at different intervals [19]. If inside $m(x_i)$ in equation (1) below:

$$y_i = m(x_i) + \varepsilon_i, i = 1,2,\cdots,n \quad (1)$$

expressed as a function of the order q spline with knots on K_1, K_2, K, K_d that is:

$$m(x_i) = \sum_{u=0}^{q} \beta_u x_i^u + \sum_{v=1}^{d} \beta_{q+v}(x_i - K_v)_+^q, i = 1,2,\cdots,n \quad (2)$$

Function $m(x_i)$ referred to as a spline nonparametric regression function, with $\beta_0, \beta_1, K, \beta_q, \beta_{q+1}, \beta_{q+2}, K, \beta_{q+d}$ is the spline regression parameter, and $(x_i - K_v)_+^q$ is a truncated element that satisfies equation (3) below:

$$(x_i - K_v)_+^q = \begin{cases} (x_i - K_v)^q, (x_i - K_v \geq) 0 \\ 0 \quad , (x_i - K_v) < 0 \end{cases} \quad (3)$$

The penalized spline estimator is formed from the truncated spline function in the Penalized Least Square (PLS) criteria. The penalized spline estimator uses knots and smoothing parameters together in estimating the nonparametric regression function. Ruppert explains that the function is truncated with the order q which is based on the knot point $a < K_1 < \cdots < K_d < b$, stated by $m(x_i)$ as in the equation (2). Function of $m(x_i)$ can be stated in matrix form, namely:

$$m = X\beta \quad (4)$$

Penalized spline estimators through PLS criteria formed from truncated functions are as follows:

$$\text{PLS} = \sum_{i=1}^{n} (y_i - m(x_i))^2 + \lambda \int_a^b \left[m^{(q)}(x)\right]^2 dx \quad (5)$$

Where λ is a smoothing parameter, *v* is the number of knots (*v = 1,2,. . . ,d)* , *q* is order *spline*, β is a spline regression parameter vector, and **D** is a diagonal matrix containing 0 by $q + 1$ and 1 by knots, or $\mathbf{D} = \text{diag}\left(0_{q+1}, 1_d\right)$

3 ROBUST M ESTIMATION

When a data is contaminated by outliers the consequences can be reduced through the penalized regression estimator can be rearranged by replacing the remaining quadratic functions with the estimator criteria M [11]:

$$\sum_{i=1}^{n} p(y_i - m(x_i,\beta))^2 + \lambda \sum_{j=1}^{k} \beta_{p+j}^2 \quad (6)$$

Where ρ is equivalent, it does not decrease in [0, + ∞) and ρ (0) = 0,

$$p_c(t) = \begin{cases} x^2 & |x| \leq c \\ 2c|x| - c^2 & |x| > c \end{cases} \quad (7)$$

Equation (7) is a huber function with c> 0 and at a certain level $|x| > c$ will increase linearly, with c = 1,345 Mark the derivative of ρ_c as $\psi_c(s)$ = max [-*c*, min(*c*,*s*)]. When residual is fitted $r_i = y_i - m(x_i;\boldsymbol{\beta})$, $i = 1, \ldots, n$ from the known spline regression is known, *estimator* M-scale *robust* can be used to estimate standart derivation $\hat{\sigma}_\epsilon$ of the residuals, as follows:

$$\sum_{i=1}^{n} \varphi_c \left(\frac{yi - m(xi;\boldsymbol{\beta})}{\hat{\sigma}_\epsilon}\right) \quad (8)$$

The value ω determined by

$$P(|r\text{-}r_{0.5}| < \omega) = \frac{1}{2} \quad (9)$$

Where $r_{0.5}$ is the population median. In addition, the standard estimate ω usually uses the median absolute deviation (MAD) statistic, which is defined as:

$$\text{MAD} = \text{Median}\{|r_1\text{-}M_r|,\ldots,|r_n\text{-}M_r|\} \quad (10)$$

Where M_r is the usual sample median of residual fitted and MAD is actually the sample median of values $n\ |r_1 - M_r|, \ldots, |r_n - M_r|$, with a limited sample breakdown point of approximately 0.5 .To place MAD in a more familiar context, it is usually scaled back as to estimate σ_ϵ, especially when residuals sample from the normal distribution. Especially,

$$\text{MADN} = \frac{MAD}{Z_{0.75}} \approx 1.4836 \text{ MAD} \quad (11)$$

By using the spline base function there is computation $\widehat{\boldsymbol{m}}_{\boldsymbol{M}} = \mathbf{X}\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}}$ with the regression coefficient of the M-type identified spline estimator using standard residuals with the regression coefficients as follows:

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{M}} = \text{argmin} \sum_{i=1}^{n} \rho_c \left(\frac{y_i - m(x_i;\beta)}{\sigma_\varepsilon}\right) + \lambda \beta^T \text{D}\, \beta \quad (12)$$

Furthermore, minimization for equation (12) is performed because ρ_c and ψ_c have nonlinear properties to satisfy equation (8). So to calculate the M-type estimator algorithm for penalized spline regression empirical pseudo data is used

$$\text{z} = \widehat{m}_p + \frac{\psi_c(y - \widehat{m}_p)}{2} \quad (13)$$

Next, $\widehat{\boldsymbol{m}}_{\boldsymbol{p}}$ is the least squares estimator according to the pseudo data

$$\widehat{m}_p = \text{X}(\text{X}^\text{T}\text{X} + \Lambda d)^{-1}\text{X}^\text{T}\text{z} \quad (14)$$

So it is proven, the least squares estimator is finalized $\widehat{\boldsymbol{m}}_{\boldsymbol{p}}$ converge with $\widehat{\boldsymbol{m}}_{\boldsymbol{M}}$, Initial curve estimation $\widehat{m}_p^{(0)}$ in the estimation algorithm of the regression model with the robust M estimator below, to choose the least square spline regression of the nonrobust penalty as in equation (15)

$$\widehat{m}_{\text{LS}} = \text{X}\,\hat{\beta}_{\text{LS}} = \text{X}(\text{X}^\text{T}\text{X} + \lambda \text{d})^{-1}\text{X}^\text{T}\text{z} \quad (15)$$

The algorithm for estimating the regression model with the Robust M estimator is as follows:

a. Enter the initial curve estimate $\widehat{m}_p^{(0)}$, termination tolerance ε and maximum iteration number Iter_{max}. Meanwhile arrange *k*=0 and do a few loop iterations
b. Estimation $\hat{\sigma}_\varepsilon$ with using $\text{MADN} = \frac{MAD}{Z_{0.75}} \approx 1.4836$ MAD
c. Generate empirical pseudo data according to $\text{z} = \widehat{m}_p + \frac{\psi_c(y - \widehat{m}_p)}{2}$
d. Calculate the least squares identified with the estimator $\widehat{m}_p^{(k+1)}$ defined in $\widehat{m}_p = \text{X}(\text{X}^\text{T}\text{X} + \Lambda d)^{-1}\text{X}^\text{T}\text{z}$ for pseudo data where the parameter penalty λ chosen according to the formula GCV
e. If $\| \widehat{m}_p^{(k+1)} - \widehat{m}_p^{(k)} \| < \varepsilon \| \widehat{m}_p^{(k)} \|$ or *k*= Iter_{max} , terminates and results from penalized spline estimation by type M $\widehat{m}_M = \widehat{m}_p^{(k+1)}$, with k= k+1 . And continue with step b

4 REFERENCES

[1] 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 **979** (1) 012093
[2] Hidayat R Budiantara I N Otok B W Ratnasari V 2019 The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model Communication in Statistics Theory and Methods 1-12

[3] 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
[4] Budiantara I N Ratnasari V Ratna M Zain I 2015 The combination of spline and Kernel estimator for nonparametric regression and its properties Applied Mathematical Sciences 9 (122) 6083-6094
[5] Nisa K Budiantara I N Rumiati A T 2017 Multivariable semiparametrics regression model with combined estimator of Fourier series and kernel IOP Conferences series : Earth and environmental Science 58 01202
[6] Aprilia B Islamiyati A and Anisa 2019 Platelet Modeling Based On Hematocrit in DHF Patients with Spline Quantile Regression International Journal of Academic and Applied Research 3 (12) 51-54
[7] 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
[8] Islamiyati A Raupong and Anisa 2019 Use of penalized spline linear to identify change in pattern of blood sugar based on the weight of diabetes patients International Journal of Academic and Applied Research 3 (12) 75-78
[9] El-Sayed S M Abonazel M R Seliem M M 2019 B Spline Speckman estimator of partially linear model International Journal of Systems Science and Applied Mathematics 4(4) 53-59
[10] Krivobokova T 2006 Theoretical and Practical Aspects of Penalized Spline Smoothing Bielefeld University
[11] Crainiceanu C M Ruppert D Wand M P 2005 Bayesian analysis for penalized spline regression using WinBUGS Journal of Statistical Software 14 (14) 1-23
[12] Islamiyati A Fatmawati 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
[13] Wang B Wenzhong S and Zelang M 2014 Comparative Analysis for Robust Penalized Spline Smoothing Methods Hindawi Publishing Corporation Mathematical Problem in engineering Arc ID 642475
[14] Huber P J 1964 Robust estimation of a location parameter Annals of Mathematical Statistics 35 (1) 73–101
[15] Wilcox R R 2012 Introduction to Robust Estimation and Hypothesis Testing Academic Press New York NY USA
[16] Lee T C M Oh H S 2007 Robust penalized regression spline fitting with application to additive mixed modeling Computational Statistics 22 (1) 159–171
[17] Finger R 2013 Investigating the performance of different estimation techniques for crop yield data analysis in crop insurance applications Agricultural Economics 44 (2) 217–230
[18] Khair U hasanul F sarudin AH and Robbi R 2017 Forecasting Error Calculation with Mean Absolute Deviation and Mean Absolute Percentage Error International Conference on Information and Communication Technology 930 2-7
[19] 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