Estimation of Quantile Nonparametric Regression Model with Linear Penalized Spline

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

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

Abstract*: Nonparametric regression is one method that is flexible and is used when data plots do not follow parametric patterns. Another problem that is often found in real data is outliers in data. One regression method that has been developed by several researchers to overcome data that contains outliers is quantile regression. The advantage of quantile regression is flexibility in modeling data with the heterogeneous conditional distribution. This is very useful in modeling vulnerable data containing outliers. The use of nonparametric regression has also been developed in the quantile regression model because of the flexibility of the regression. In this article, we chose the penalized spline estimator in estimating the nonparametric quantile regression model. The penalized spline is an estimator in a spline that uses knots and smoothing parameters simultaneously, which can analyze irregularly patterned data with more efficient estimated curve results. In this study using penalized quantile spline with the estimation method used is the least absolute deviation.*

Keywords— least absolute deviation; nonparametric regression; penalized spline; quantile regression

1. INTRODUCTION

Regression analysis is a tool in assessing specifically the pattern of relationships and influences between variables. The analysis method has the ultimate goal to estimate or predict the value of one variable to another [1]. Information about the functional relationship between the predictor variable and the response variable can be estimated by looking at the shape of the relationship pattern in a scatter plot. There are several methods used for parameter estimation in the regression equation, one of which is the standard approach for determining the classical regression model and the parameter estimation is the ordinary least squares (OLS) method. The OLS method is known to be sensitive to the deviation of assumptions in the data. Sometimes to overcome this, researchers will transform the data with the intention that the assumptions are met [2]. However, not all problems with data can be solved by the transformation. For such data, we usually have to use other statistical approaches that are suitable for the conditions of the data.

Nonparametric regression is one method that is flexible and is used when data plots do not follow parametric patterns. Plot data can be obtained from scatter plots or from previous research information [3]. In nonparametric regression, the regression function is unknown in shape so it must be estimated through an estimator, including spline truncated [4], spline smoothing [5], spline penalized [6], kernel [7], and fourier series [8]. Spline estimators have been developed in cross-section data and longitudinal data with various assumptions. Among them, the assumption of multicollinearity [9], heteroscedasticity [10], and autocorrelation [11].

Another problem that is often found in real data is outliers in data. In this case, several researchers have also developed approaches that can overcome or model data containing outliers, for example, robust regression consisting of five estimation methods, including M-estimator, least median square estimator, least trimmed square estimator, S-estimator, and MM-estimator [12] and quantile regression [13]. One regression method that has been developed by several researchers to overcome data that contains outliers is quantile regression. This method has been used in two-way analysis for survey data [14] and body mass index modeling in the Ethiopian population [15]. Also, this method has been developed for several applications that contain outliers [16] and is used in longitudinal data which is an extension of the regression model in conditional quantile where the conditional quantile distribution of response variables is expressed as a covariate function [17]. The advantage of quantile regression is flexibility in modeling data with the heterogeneous conditional distribution. This is very useful in modeling vulnerable data containing outliers. The use of nonparametric regression has also been developed in the quantile regression model because of the flexibility of the regression. Research on nonparametric quantile regression has been carried out, among others, estimating small area quantile through spline regression [18], estimating conditional quantile functions for panel data models with additional individual fixed effects [19] and estimating quantile distribution functions for length-biased and right-censored data [20].

In this article, we chose the penalized spline estimator in estimating the nonparametric quantile regression model. Penalized splines involve knots and smoothing parameters together so that they have an excellent interpretation, are more flexible, can handle data whose behavior changes, and produce a smooth regression curve [21]. Research on penalized spline has been widely used, among others, for bivariate splines on triangulation and energy as a penalty [22], identifying patterns of changes in blood sugar in diabetics [23] and adaptive penalized splines for data smoothing [24]. However, some of these studies have not used quantile regression. Therefore, this study uses a penalized quantile spline with the estimation method used, which is the least absolute deviation (LAD). LAD is a classical least squares alternative for statistical analysis of linear regression models that minimizes the absolute number of errors [25]. In this article, what is presented is a description of the estimation with LAD because so far the development is still limited in the parametric approach.

2. PENALIZED SPLINE LINEAR QUANTIL REGRESSION MODEL

Quantile regression was first introduced by Koenker and Bassett in 1978. Quantile nonparametric regression is an approach in regression analysis that is used to estimate the regression function when the assumptions about the shape of the regression curve are unknown and are only assumed to be smooth by involving quantile values [26]. Quantum regression minimizes the absolute number of errors weighted and predicts the model by using conditional quantile functions of a data distribution [27]. The general equation of linear quantile regression for conditional quantile $Q_{y|x}(\theta)$ of the response variable y_i, namely:

$$y_i = \beta_0(\theta) + \beta_1(\theta)x_{i1} + \cdots + \beta_k(\theta)x_{ik} + \varepsilon_i(\theta) \qquad (1)$$

If the relationship between the response variable and the predictor is expressed in a function f whose form is unknown and can be approached with a nonparametric quantile regression model, then the form of the regression equation is as follows:

$$y_i(\theta) = f(x_{i1}, x_{i2}, \dots, x_{ik}\ ) + \varepsilon_i(\theta) \qquad (2)$$

The function $f(x_{i1}, x_{i2}, \dots, x_{ik}\ )$ in equation (2) is assumed to be additive and is approached by a quantile regression function with the penalized spline. Where penalized spline is one of the estimators used in nonparametric regression in estimating the nonparametric regression function. The penalized spline estimator involves the vertex and smoothing of parameters simultaneously in controlling the smoothness of the curve [28]. then obtained a quantile regression model with penalized spline as follows:

$$y_i(\theta) = f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) + \varepsilon_i(\theta) \qquad (3)$$

$$= \sum_{i=1}^{k} f(x_{ik}) + \varepsilon_i(\theta)$$

Based on equation (3), the function f approaches the quantile regression function with penalized spline with the order linear and the number of knots $K_1, K_2, \dots, K_r$ as follows:

$$f(x_{ik}) = \sum_{i=0}^{1} \beta_{ik}(\theta) x_{ik}^{l} + \sum_{h=1}^{r} \beta_{(1+h)k}(\theta)(x_{ik} - K_{hk})_+^1 \qquad (4)$$

where

$$(x_{ik} - K_{hk})_+^1 = \begin{cases} (x_{ik} - K_{hk})^1, jika\ x_{ik} \geq K_{hk} \\ \quad 0\ \ , jika\ x_{ik} < K_{hk} \end{cases}$$

the function described in equation (3) is stated as follows:

$$\sum_{i=1}^{k} f(x_{ik}) = f(x_{i1}) + f(x_{i2}) + \cdots + f(x_{ik}) \qquad (5)$$

by describing the function f and separating between parameters and variables, the quantile penalized spline regression model can be expressed in the form of a matrix as follows:

$$y(\theta) = \boldsymbol{X}[K]\boldsymbol{\beta}(\theta) + \boldsymbol{\varepsilon}(\theta) \qquad (6)$$

Where $y(\theta) = [y_1(\theta), y_2(\theta), \dots, y_n(\theta)]'$ is a $n \times 1$ column vector of the response variable y in the θ quantile, $\boldsymbol{X}[K] = [\mathbf{1} \quad \boldsymbol{X_1} \quad \cdots \quad \boldsymbol{X_k}]$ is a matrix $\boldsymbol{X}$ in the form of a spline of the order q and r knots of size $n \times (k+1)$ with n observations on k variables x, $\boldsymbol{\beta}(\theta)$ is a column vector sized $(k+1) \times 1$ of the β parameter in the θ quantile and $\boldsymbol{\varepsilon}(\theta)$ is a $n \times 1$ column vector of error (ε) in the θ quantile.

Based on equation (6), the penalized spline criteria in the nonparametric quantile regression model can be stated as follows:

$$\text{PLS} = \big(y(\theta) - \boldsymbol{X}[K]\beta(\theta)\big)^T \big(y(\theta) - \boldsymbol{X}[K]\boldsymbol{\beta}(\theta)\big) + \lambda P \qquad (7)$$

where P=$\boldsymbol{\beta}(\theta)^{\boldsymbol{T}}\boldsymbol{D}\beta(\theta)$

3. ESTIMATION OF LINEAR PENALIZED SPLINE QUANTIL REGRESSION MODEL

Regression with the ordinary least square method is estimated by minimizing the number of squares of errors, while quantile regression will minimize the absolute number of errors better known as the least absolute deviation (LAD). In quantile regression, errors are given different weights. For error values greater than or equal to zero, the weight used is θ, and for errors, less than zero $1 - \theta$ is used. Multiplication of weights with errors is called a loss function (ρ_θ) namely:

$$\rho_\theta(\varepsilon) = \sum_{i=1,\varepsilon_{i\geq 0}}^{n} \theta|\varepsilon_i + \lambda P| + \sum_{i=1,\varepsilon_i<0}^{n} (1-\theta)|\varepsilon_i + \lambda P| \qquad (8)$$

In the quantile regression, there is θ quantile function of the variable y with the condition x which takes into account the $\beta(\theta)$estimator, so the solution to the problem can be stated as follows:

$$\min_{\beta\epsilon R^p} \sum_{i=1}^{n} \rho_\theta(\varepsilon) = \min_{\beta\epsilon R^p} \sum_{i=1}^{n} \rho_\theta\big(y_i - Q_{y|x}(\theta) + \lambda P\big) \qquad (9)$$

Where $\rho_\theta(\varepsilon)$ is loss function, θ is quantile index with $\theta \in (0,1)$, $Q_{y|x}(\theta)$ is the quantile function of the variable with the condition x. For $\rho_\theta(\varepsilon)$ in the equation is defined:

$$\rho_\theta(\varepsilon) = \begin{cases} \theta\varepsilon + \theta\lambda P, & \varepsilon \geq 0 \\ (1-\theta)\varepsilon + (1-\theta)\lambda P, & \varepsilon < 0 \end{cases}$$

the function of conditional quantile $Q_{y|x}(\theta)$ is defined as follows:

$$Q_{y|x}(\theta) = \boldsymbol{X}[K]^{\boldsymbol{T}}\boldsymbol{\beta}(\theta) \qquad (10)$$

if y is a known function x and has a probability function $F_{y|x}(y)$, then the θ quantile of the function can be written as in the following equation:

$$\min_{\beta} \theta = \int_{i=1;\varepsilon_i\geq 0}^{n} |\varepsilon_i + \lambda P| dF_y(y) + (1-\theta)\int_{i=1;\varepsilon_i<0}^{n} |\varepsilon_i + \lambda P| dF_y(y) \qquad (11)$$

by considering $\hat{\beta}(\theta)$ to obtain a solution to the problem stated:

$$\hat{\beta}(\theta) = \min_{\beta\epsilon R^p} \left\{ \theta \sum_{i=1;\varepsilon_i\geq 0}^{n} |y_i - \boldsymbol{X}[K]^{\boldsymbol{T}}\boldsymbol{\beta}(\theta) + \lambda P| + (1-\theta) \sum_{i=1;\varepsilon_i<0}^{n} |y_i - \boldsymbol{X}[K]^{T}\boldsymbol{\beta}(\theta) + \lambda P| \right\} \qquad (12)$$

4. REFERENCES

[1] Jain, S., Chourse, S., Dubey, S., Jain, S., Kamakoty, J., and Jain, D. (2016). Regression Analysis – Its Formulation and Execution In Dentistry. Journal of Applied Dental and Medical Sciences, NLM ID: 101671413, 2 (1).

[2] Budiantara, I, N. (2006). Spline Model with Optimal Knot. Journal of Basic Science FMIPA Jember University, 7 (1), 77-85.

[3] Budiantara, I, N., Ratna, M., Zain, I., and Wibowo, W. (2012). Modeling the percentage of poor people in Indonesia using spline nonparametric regression approach. International Journal of Basic & Applied Sciences, 12, 119-124.

[4] 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.

[5] Hidayat, R., Budiantara, I, N., Otok, B, W., and Ratnasari, V. (2020). The regression curve estimation by using mixed smoothing spline and kernel (MsS-K) model. Communications in Statistics-Theory and Methods, 47, 1-12.

[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] Ratnasari, V., Budiantara, I, N., Ratna, M., and Zain, I. (2016) Estimation of Nonparametric Regression Curve Using Mixed Estimator of Multivariable Truncated Spline and Multivariable Kernel. Global Journal of Pure and Applied Mathematics, 12 (6), 5047-5057.

[8] Mardianto, M, F, F., Tjahjono, E., and Rifada, M. (2019). Statistical modelling for prediction of rice production in indonesia using semiparametric regression based on three forms of fourier series estimator. Journal of Engineering and Applied Sciences, 14 (15), 2763-2770.

[9] Daoud, J, I. (2018). Multicollinearity dan Regression Analysis. Journal of Physics Conf. Series, 949, 012009.

[10] Klein, A, G., Gerhard, C., Buchner, R, D., Diestel, S., Engel, K, S. (2016). The detection of heteroscedasticity in regresion models for psychological data. Psychological Test and Assessment Modeling, 58 (4), 567-592.

[11] Lee, J., and Lund, R. (2004). Revisiting simple linear regression with autocorrelated errors. Biometrika, 91 (1), 240-245.

[12] Chen, C. (2002). Robust Regression and Outlier Detection with the Robustreg procedure. SUGI Proceedings SAS institude Inc Cary NC, 265-271.

[13] 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.

[14] Sauzet, O., Razum, O., Widera, T., and Brzoska, P. (2019). Two Part Models and Quantile Regression for the Analysis of Survey Data with a Spike. The Example of Satisfaction with Health Care Frontiers Publich Health, 7 (146), 1-7.

[15] Yirga, A, A., Ayele, D, W., and Melesse, S, F. (2018). Application of Quantile Regression Modeling Body Mass Index in Ethiopia. The Open Public Health Journal, 11, 221-223.

[16] Davino, C., Furno, M., and Vistocco, D. (2013). Quantile Regression: Theory and Applications. Wiley Series in Probablity and Statistics.

[17] Huang, Q., Zhang, H., Chen, J., and He, M. (2017). Quantile Regression Models and Their Applications: A review Journal of Biometric & Biostatistics, 8 (3), 1-6.

[18] Chen, Z., Chen, J., and Zhang, Q. (2019). Small Area Quantile Estimation Via Spline Regression and Empirical Likelihood. Statistics Canada Catalogue, No. 12-001- X.

[19] Yan, K, X., and Li, Q. (2018). Nonparametric Estimation of a Conditional Quantile Function in a Fixed Effects Panel Data Model. Journal of risk and financial management, 11, 44.

[20] Shi, J., Ma, H., Zhou, Y. (2018). The Nonparametric Quantile Estimation for Length-Biased and Right-Censored Data. Statistics & Probability Letters, 134, 150-158.

[21] Islamiyati, A., Sunusi, N., Kalondeng, A., Wati, 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.

[22] Lai, M, J., and Wang, L. (2013). Bivariate penalized spline for regression. Statistica Sinica, 23, 1399-1417.

[23] 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.

[24] Yang, L., and Hong, Y. (2017). Adaptive Penalized Splines for Data Smoothing. Computational Statistics & Data Analysis, 108, 70-83.

[25] Chen, K., Ying, Z., Zhang, H., and Zhao, L. (2008). Analysis of Least Absolute Deviation Biometrika, 95 (1), 107-122.

[26] Putri, W,N,A., Islamiyati, A., and Anisa. (2020). Penggunaan Regresi Multivariat pada Perubahan Trombosit Pasien Demam Berdarah Dengue. ESTIMASI: Journal of Statistics and Its Application, 1 (1), 1-9.

[27] Buhai, I, S. (2014). Quantile Regression: Overview and Selected Application 2005 Ad Astra 4.

[28] Islamiyati, A., Fatmawati, and Chamidah, N. (2018). Estimation of Covariance Matrix on Bi-Response Longitudinal Data Analysis with Penalized Spline Regression. Journal of Physics Confenrence Series, 979 012093.