Multivariable Semiparametric Regression Model with Combined Estimator of Fourier Series and Kernel

Khaerun Nisa[1]*, I Nyoman Budiantara[1] and Agnes Tuti Rumiati[1]

[1]Department of Statistics, Faculty of Mathematics and Natural Sciences, Institut Teknologi Sepuluh Nopember (ITS). Arif Rahman Hakim Street, Surabaya 60111 Indonesia

Email: khaerunnisaramli@yahoo.co.id

Abstract. We purpose a combined estimator of Fourier series and Kernel on semiparametric regression. This method is used to resolve the problem of regression modeling, when the relationship between the response variable and the predictor variables most follow a certain pattern, partly have a repetitive pattern, and some others not follow a specific pattern. Moreover, this method depends on oscillation, smoothing parameter and bandwidth. The purpose of this research are to obtained the estimator of semiparametric regression model with combined estimator of Fourier series and Kernel using Penalized Least Square method (PLS). The result show that the PLS estimation produces the estimator of parametrik linier regression, the estimator of Fourier series, the estimator of Kernel, and also the combined estimator of Fourier series and Kernel in semiparametric regression model.

1. Introduction

Semiparametric regression is a statistical methods used to estimate the relationship pattern of the predictor variable and response variable, when a case in regression analysis contains two components; parametric component and nonparametric component. Based on several previous research that has been conducted by researchers, mostly they use the same estimator approach for all or some of the predictor variables. Meanwhile, in many cases, the data pattern of each predictor variables are not always identical. Therefore, to solve these problems, we need a more proper estimator that can be used to approximate the data pattern. In semiparametric regression, there are many functions that can be used to approximate the data pattern, especially Fouries series and Kernel.

According to [1], estimation of Fourier series is capable of handling data that is smooth character and follow the pattern repeated at certain interval, previous research have been investigate by researchers as Amato [2], Asrini [3], Bilodeau [1], Pane [4], and Sudiarsa [5]. The kernel is one of the frequenlty used estimators in semiparametric and nonparametric regression. This estimator has more flexible shape, simple mathematics calculation, and achieves convergence more rapidly. Several researches on Kernel estimator have been conducted by some researchers such as Nadaraya [6], Budiantara et al [7], and Speackman [8].

Therefore, this research focus on semiparametric regression model with combined estimator between Fourier series and Kernel obtained through PLS method.

IOP Conf. Series: Earth and Environmental Science **58** (2017) 012028 doi:10.1088/1755-1315/58/1/012028

2. Semiparametric Regression Model

Given the data $\left(x_{1i}, x_{2i}, \ldots x_{pi}, t_i, z_{1i}, z_{2i}, \ldots, z_{ri}, y_i\right)$ with predictor $\left(x_{ji}, t_i, z_{mi}\right)$ and response variables y_i it can be approximated by a semiparametric regression model as follows :

$$y_i = \sum_{j=1}^{p} f_j(x_{ji}) + g(t_i) + \sum_{m=1}^{r} h_m(z_{mi}) + \varepsilon_i,\ i = 1, 2, \ldots, n \tag{1}$$

The $f_j(x_{ji})$ curve is assummed referring to linear pattern. Meanwhile regression curve $g(t_i)$ and $h_m(z_{mi})$ are assummed unknown and smooth that can be approached by Fourier series function and Kernel function, then ε_i is random error that follows normally and independently distributed with mean zero and variance σ^2. The linear parametric components, Fourier series component and kernel component are defined as follows :

Linear function with *p* predictor variable can be written as :

$$f_j\left(x_{1i}, x_{2i}, \ldots x_{pi}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} \tag{2}$$

where :

$\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ are parameters of the linier function

Fourier series function with one predictor variable can be written as :

$$g(t_i) = bz_{ki} + \frac{1}{2}a_0 + \sum_{k=1}^{K} a_k \cos kz_i \tag{3}$$

where :

$b, a_0, a_{k,} k = 1, 2, \ldots, K$ are parameters of the Fourier series function

and then Kernel function with *r* predictor variable can be written as :

$$\hat{h}_m(z_{1i}, z_{2i}, \ldots, z_{ri}) = n^{-1} \sum_{j=1}^{n} W_{\alpha i}(z) y_i \tag{4}$$

where :

$$W_{\alpha i}(z) = \frac{\frac{1}{\alpha} K\left(\frac{z - z_i}{\alpha}\right)}{n^{-1} \sum_{j=1}^{n} \frac{1}{\alpha} K\left(\frac{z - z_i}{\alpha}\right)},$$

Kernel estimator depends on the Kernel function K and bandwidth parameter α. There are several types of Kernel functions included a Gaussian Kernel of the form, as follows :

$$K(x) = \frac{1}{2} e^{\frac{x^2}{2}}, -\infty < x < \infty$$

Hence, a semiparametric regression model in equation (1) can be written as follows :

$$\tilde{y} = X\vec{\beta} + G\vec{a} + D\vec{y} + \vec{\varepsilon} \tag{5}$$

where :

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ 1 & x_{31} & x_{23} & \cdots & x_{3p} \\ \vdots & & & \ddots & \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \vec{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}$$

$$G = \begin{bmatrix} t_{11} & \frac{1}{2} & \cos t_{11} & \cdots & \cos Kt_{11} \\ t_{12} & \frac{1}{2} & \cos t_{12} & \cdots & \cos Kt_{12} \\ \vdots & & \vdots & \ddots & \vdots \\ t_{1n} & \frac{1}{2} & \cos t_{1n} & \cdots & \cos Kt_{1n} \end{bmatrix}$$

$$\vec{a} = \begin{bmatrix} b & a_0 & a_{11} & \cdots & a_K \end{bmatrix}^T$$

$$D = \begin{bmatrix} n^{-1}W_{\alpha 1}(z_1) & n^{-1}W_{\alpha 2}(z_1) & \cdots & n^{-1}W_{\alpha n}(z_1) \\ n^{-1}W_{\alpha 1}(z_2) & n^{-1}W_{\alpha 2}(z_2) & \cdots & n^{-1}W_{\alpha n}(z_2) \\ \vdots & \vdots & \ddots & \vdots \\ n^{-1}W_{\alpha 1}(z_n) & n^{-1}W_{\alpha 2}(z_n) & \cdots & n^{-1}W_{\alpha n}(z_n) \end{bmatrix}, \vec{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}.$$

3. Multivariable Semiparametric Regression with Combined Estimator between Fourier Series and Kernel

The combination estimator of Fourier series and Kernel in semiparametric regression, $\hat{\mu}\left(x_{1i}, x_{2i}, \ldots x_{pi}, t_i, z_{1i}, z_{2i}, \ldots, z_{ri}\right)$ is obtained by employing PLS optimization, as follows :

$$\underset{f,g,h}{Min}\left\{R(f,g,h) + \lambda J(g)\right\} = \underset{f,g,h}{Min}\left\{n^{-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} f_j(x_{ji}) - g(t) - \sum_{m=1}^{r} h_m(z_{mi})\right)^2 + \lambda\int_0^{\pi}\frac{2}{\pi}\left(g''(t)\right)^2 dt\right\} \quad (6)$$

Several lemmas are needed to complete this PLS optimization.

Lemma 1

If the function of Fourier series is $g(t)$ as in equation (3), then the penalty is

$$J(g) = \int_0^{\pi}\frac{2}{\pi}\left(g^{(2)}(t)\right)^2 dt = \sum_{k=1}^{K} k^4 a_k^2, \quad (7)$$

this lemma has been proven by [3] and [8].

Lemma 2

If $f_j(x_{ji})$ is approach by multiple linier function as in equation (2), $g(t)$ is approach by Fourier series function as in equation (3) and $h_m(z_{mi})$ is approach by Kernel as in equation (4), then the goodness of fit $R(f,g,h)$ is obtained as follows :

$$R(f,g,h) = n^{-1}\left(\vec{y} - X\vec{\beta} - G\vec{a} - D\vec{y}\right)^T\left(\vec{y} - X\vec{\beta} - G\vec{a} - D\vec{y}\right)$$

Proof

In general, goodness of fit $R(f,g,h)$ is defined as follows :

$$R(f,g,h)=n^{-1}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{p}f_j(x_{ji})-g(t)-\sum_{m=1}^{r}h_m(z_{mi})\right)^2$$

It could be shown

$$R(f,g,h)=n^{-1}\sum_{i=1}^{n}\left(\begin{array}{l} y_i-\beta_0-\beta_1x_{i1}-\beta_2x_{i2}-\cdots-\beta_px_{ip}-bt_{1i}-\frac{1}{2}a_0-\sum_{k=1}^{K}a_k\cos kt_{1i} \\ -\cdots-bt_{si}-\frac{1}{2}a_0-\sum_{k=1}^{K}a_k\cos kt_{si}-\sum_{i=1}^{n}\frac{\frac{1}{\alpha}K\left(\frac{z-z_i}{\alpha}\right)}{\sum_{i=1}^{n}\frac{1}{\alpha}K\left(\frac{z-z_i}{\alpha}\right)}Y_i \end{array}\right)^2$$

Hence, goodness of fit $R(f,g,h)$ can be written as follows :

$$R(f,g,h)=n^{-1}\left(\vec{y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right)^T\left(\vec{y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right) \tag{8}$$

Optimization using PLS can be solved by combining goodness of fit $R(f,g,h)$ and penalty $J(g)$ as follows :

$$\begin{aligned}\underset{f,g,h}{Min}\left\{R(f,g,h)+\lambda J(g)\right\}&=\underset{\vec{\beta},\vec{a}}{Min}\left\{n^{-1}\left(\vec{y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right)^T\left(\vec{y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right)\right\}\\&=\underset{\vec{\beta},\vec{a}}{Min}\left\{Q(\vec{\beta},\vec{a})\right\}\end{aligned}$$

The estimation of $\hat{\beta}$ and $\hat{a}$ are obtaining by using partial derivative of $Q(\vec{\beta},\vec{a})$ to $\vec{\beta}$ and $\vec{a}$. First, consider $Q(\vec{\beta},\vec{a})$ the function as follows :

$$\begin{aligned}Q(\vec{\beta},\vec{a})&=n^{-1}\left(\vec{Y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right)^T\left(\vec{Y}-X\vec{\beta}-G\vec{a}-D\vec{y}\right)+\lambda\vec{a}^TU\vec{a}\\
&=n^{-1}\left[\left(I-D\right)\vec{y}-X\vec{\beta}-G\vec{a}\right]^T\left[\left(I-D\right)\vec{y}-X\vec{\beta}-G\vec{a}\right]+\lambda\vec{a}^TU\vec{a}\\
&=n^{-1}\left[\vec{y}^T\left(I-D\right)^T-\vec{\beta}^TX^T-\vec{a}^TG^T\right]\left[\left(I-D\right)\vec{y}-X\vec{\beta}-G\vec{a}\right]+\lambda\vec{a}^TU\vec{a}\\
&=n^{-1}\vec{y}^T\left(I-D\right)^T\left(I-D\right)\vec{y}-n^{-1}\vec{y}^T\left(I-D\right)^TX\vec{\beta}-n^{-1}\vec{y}^T\left(I-D\right)^TG\vec{a}\\
&\quad-n^{-1}\vec{\beta}^TX^T\left(I-D\right)\vec{y}+n^{-1}\vec{\beta}^TX^TX\vec{\beta}+n^{-1}\vec{\beta}^TX^TG\vec{a}-n^{-1}\vec{a}^TG^T\left(I-D\right)\vec{y}\\
&\quad+n^{-1}\vec{a}^TG^TX\vec{\beta}+n^{-1}\vec{a}^TG^TG\vec{a}+\lambda\vec{a}^TU\vec{a}\\
&=n^{-1}\vec{y}^T\left(I-D\right)^T\left(I-D\right)\vec{y}-2n^{-1}\vec{\beta}^TX^T\left(I-D\right)\vec{y}-2n^{-1}\vec{a}^TG^T\left(I-D\right)\vec{y}\\
&\quad+n^{-1}\vec{\beta}^TX^TX\vec{\beta}+n^{-1}\vec{\beta}^TX^TG\vec{a}+n^{-1}\vec{a}^TG^TG\vec{a}+\lambda\vec{a}^TU\vec{a}\end{aligned}$$

The derivation of $Q(\vec{\beta},\vec{a})$ to $\vec{\beta}$ and $\vec{a}$ as follows :

$$\frac{\partial Q(\vec{\beta},\vec{a})}{\partial\vec{\beta}}=-2n^{-1}X^T(I-D)\vec{y}+2n^{-1}X^TX\vec{\beta}+2n^{-1}X^TG\vec{a}$$

$$\frac{\partial Q(\vec{\beta},\vec{a})}{\partial\vec{a}}=-2n^{-1}G^T(I-D)\vec{y}+2n^{-1}G^TX\vec{\beta}+2n^{-1}G^TG\vec{a}+\lambda U\vec{a}$$

Equalizing the first derivatif of sum square error with zero, for $\vec{\beta}$ and $\vec{a}$ are obtained :

$$\hat{\beta} = \left(X^T X\right)^{-1} X^T (I-D)\vec{y} - \left(X^T X\right)^{-1} X^T G\vec{a}$$

$$\hat{a} = \left(X^T X\right)^{-1} X^T (I-D)\vec{y} - \left(X^T X\right)^{-1} X^T G\left[\left(G^T G + n\lambda U\right)^{-1} G^T (I-D)\vec{y} - \left(G^T G + n\lambda U\right)^{-1} G^T X\vec{\beta}\right]$$

Substitution $\vec{a}$ to $\vec{\beta}$, so **:**

$$\vec{\beta} = A\hat{y} \tag{9}$$

where :

$$A = P_1 M_1$$

$$P_1 = \left[I - \left(X^T X\right)^{-1} X^T G\left(G^T G + n\lambda U\right)^{-1} G^T X\right]^{-1}$$

$$M_1 = \left[\left(X^T X\right)^{-1} X^T (I-D) - \left(X^T X\right)^{-1} X^T G\left(G^T G + n\lambda U\right)^{-1} G^T (I-D)\right]$$

Then substitution $\hat{\beta}$ into $\vec{a}$, so **:**

$$\vec{a} = B\hat{y} \tag{10}$$

where :

$$B = P_2 M_2$$

$$P_2 = \left[I - \left(G^T G + n\lambda U\right)^{-1} G^T X\left(X^T X\right)^{-1} X^T G\right]^{-1}$$

$$M_2 = \left[\left(G^T G + n\lambda U\right)^{-1} G^T (I-D) - \left(G^T G + n\lambda U\right)^{-1} G^T X\left(X^T X\right)^{-1} X^T (I-D)\right]$$

So the estimator of linear parametric curve is :

$$\vec{f}(x) = X\vec{\beta} \tag{11}$$

$$= C\vec{y},$$

where :

$$C = XP_1 M_1$$

the estimator of Fourier series curve is :

$$\vec{g}(t) = G\vec{a} \tag{12}$$

$$= K\vec{y},$$

where :

$$K = DP_2 M_2$$

then, the estimator of Kernel curve is :

$$\vec{h}(z) = D\vec{y} \tag{13}$$

The estimator of semiparametric with combined Fourier series and Kernel, obtained :

$$\hat{\mu}(x_{1i}, x_{2i}, \ldots, x_{pi}, t_i, z_{1i}, z_{2i}, \ldots, z_{ri}) = \hat{f}(x) + \hat{g}(t) + \hat{h}(z) \tag{14}$$

$$= A\tilde{y} + K\tilde{y} + D\tilde{y}$$

$$= (A + K + D)\tilde{y}$$

$$= N\tilde{y}$$

where :

$$N = A + K + D$$

4. Conclusion

The results shows that the estimator of combination between Fourier series and Kernel function in semiparametric regression was obtined through the PLS optimization, as follows :

$$\underset{f,g,h}{Min}\left\{R(f,g,h) + \lambda J(g)\right\} = \underset{f,g,h}{Min}\left\{n^{-1}\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} f_j(x_{ji}) + g(t) + \sum_{m=1}^{r} h_m(z_{mi})\right)^2 + \lambda\int_0^{\pi}\frac{2}{\pi}\left(g''(t)\right)^2 dt\right\}$$

The estimator of combination between Fourier series and Kernel function is

$$\hat{\mu}(x_{1i}, x_{2i}, \ldots, x_{pi}, t_i, z_{1i}, z_{2i}, \ldots, z_{ri}) = \hat{f}(x) + \hat{g}(t) + \hat{h}(z)$$

where

$$\hat{f}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}$$

$$\hat{g}(t) = \hat{b}z_i + \frac{1}{2}\hat{a}_{k0} + \sum_{k=1}^{K}\hat{a}_k \cos kz_i .$$

References

[1] Bilodeau M 1992 *The Canadian Journal of Statistics* **3** 257-269
[2] Amato U, Antoniadis A, and De-Feis I 2002 *Journal of Computational and Applied Mathematics* **146** 459-479
[3] Asrini L J, Budiantara I N 2014 *ARPN Journal of Engineering and Applied Sciences* **9** 9 1501-1506
[4] Pane R, Budiantara I N, Zain I, and Otok B J 2014 *Apllied Mathematical Sciences* **8** 102 5053-5064
[5] Sudiarsa I W, Budiantara I N, Suhartono, Purnami S W 2015 *Applied Mathematical Science* **9** 100 4997-5010
[6] Nadayara EA 1964 *Theory of Probability and its Applications* 141-142
[7] Budiantara I N, Ratnasari, Ratna and Zain 2015 *Applied Mathematical Sciences* 9 **122** 6038-6094
[8] Speackman P 1988 *Journal Royal Statistical Society* B **50** 413-436