NU Science Journal 2014; 11(1): 1 - 14

Comparison of Estimating Parameter in Parametric Regression, Nonparametric Regression, and Semiparametric Regression models in Case of Two Explanatory Variables

Autcha Araveeporn*

Department of Statistics, Faculty of Science, King Mongkut's Institute of Technology Ladkrabang, Bangkok 10520, Thailand

* Correspondence author. E-mail : kaautcha@kmitl.ac.th, kaautcha@hotmail.com

ABSTRACT

In this paper, we compare the estimating parameter in the parametric regression model, nonparametric regression model, and semiparametric regression model between response variable and two explanatory varaibles. The parametric regression model uses the least square method for estimating parameter. The penalized spline method based on nonparametric regression method is proposed for estimating function of nonparametric regression model, and semiparametric regression model. The minimum of Mean Square Error (MSE) is a criterion for choosing the optimal model. Here, we simulate the response variable and two explanatory variables that correlated a nonlinear data based on uniform distribution. The real data can be applied of these models to illustrate the methodology. The estimated values of nonparametric regression model is a good performance in both of simulated data and real data.

Keywords: nonparametric regression model, parametric regression model, penalized spline method, semiparametric regression model

INTRODUCTION

Regression analysis or parametric regression analysis is a commonly used method to investigate the relationship between variables, and obtains the coefficient of regression function. The simplest of regression function consists of a response variable (y) and a single explanatory variable (x). The multiple regression function is extended from the simple regression function to include additional explanatory variables. To obtain useful the regression analysis, the assumption should investigate before data analysis such as; the variables x and y describes a linear relationship, the variables y are normal distribution with continuous variable, the observed variables are independent, and two explanatory variables (x) occur correlated variation called multicollinearity.

A parametric regression analysis requires an assumption in underlying regression function. A serious drawback of parametric modeling is that a parametric model may be too restrictive in some application. If an inappropriate parametric regression model is used, it is misleading conclusions from the regression analysis. To overcome the difficulty cause by the restrictive assumption of the regression function, on may remove the restriction that the regression function belongs to a

parametric family. This approach to so-called nonparametric regression(Wu and Zhang, 2006).

The gain of nonparametric regression method is to estimate the smoothing function which is produced a smoother directly, rather than to estimate regression coefficients. The smoother is a tool for summarizing the trend of a response variable as a function of one or more explanatory variables. The simple nonparametric regression model is often called scatter plot smoothing which is used one explanatory variable. The nonparametric regression method presents more than one explanatory variables called the nonparametric additive model (David *et al*., 2003). There are many approaches to estimate nonparametric regression model, e.g., a local polynomial regression method (Wand and Jones, 1995 and Fan and Gijbels, 1996), regression splines method (Eubank, 1988 and Eubank, 1999), smoothing splines method (Wahba, 1990 and Green and Silverman, 1994), and penelized splines method (Ruppert *et al*., 2003). Moreover, the nonparametric regression model has been developed for time series data which may have a nonlinear relationship. Robinson (1983) suggested the nonparametric estimation in the context of time series data.

The method of nonparametric regression has a long history in the smoothing method. Wahba (1990) defined the natural polynomial spline model that used to measure the roughness of curve. Green and Silverman (1994) emphasized the simple case of the natural polynomial spline so-called the natural cubic spline. The smoothing spline method is considered a least square criterion to fit the natural cubic spline. Stone (1977) examined the consistency properties of nonparametric regression estimators in local polynomial regression. Fan (1992, 1993) demonstrated the desirable mean square error properties, as well as establishing the local linear regression based on kernel regression. In the local polynomial regression method, the local neighborhoods are specified by a bandwidth but Eubank (1988, 1999) introduced the regression spline that the local neighborhoods are specified by a group of locations. Penalized spline method has developed from regression spline and smoothing spline, which is used of fitting and flexible choice of knots and smoothing parameter in nonparametric regression model. Ruppert, Wand, and Carroll (2003) described penalized spline method based on reduced-knot truncated power function basis with penalties on the untransformed coefficients, fitted as a mixed model, and motivated as a simple low-rank smoothing spline.

Normally, the estimating function of parametric regression method is reliability and accuracy when the data is set as a continuous variable following the assumption. Nonparametric regression method is a choice to relax the assumption of the parametric regression method. The modeling of nonparametric regression method is a performance method when the explanatory variable is analyzed by the data with nonparametric statistics as a discrete variable and fitted the smoothing function. The semiparametric regression consists of continuous variable estimated by parametric regression method and discrete variable fitted by nonparametric regression method.

Therefore, we would like to compare the estimating function of parametric regression model, nonparametric regression model, and semiparametric regression model that a response and explanatory variables have to be a nonlinear relationship, and two explanatory variables are multicollinearity based on continuous and discrete

variables. For data analysis, the parametric regression model uses the least square error method to estimate a coefficient of regression function. The nonparametric regression model is concerned the penalized spline method to estimate the parameter of smoothing function. The semiparametric regression is a mixed model between parametric regression model and nonparametric regression model based on penalized spline method. This paper is organized as follows: Section 2 describes the parameter estimation of parametric regression model, nonparametric regression model, and semiparametric regression model. Section 3 shows the process and the results of simulated data and discuss the results. We apply our proposed models to real data in Section 4. Finally, in Section 5, we present the conclusions.

PARAMETER ESTIMATION

Parametric Regression Model

The parametric regression model consists of a response variable, a single explanatory variable, and multiple explanatory variables. In this case, we study two explanatory variables denoted by the $(x_{1t}, x_{2t}),\ t = 1, 2, ..., n$, the regression parameters $\underset{\sim}{\theta} = (\beta_0, \beta_1, \beta_2)^T$, the regression estimators $\hat{\underset{\sim}{\theta}} = (b_0, b_1, b_2)^T$, and the errors $\underset{\sim}{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$. The parametric regression model can be written as

$$y_t = \beta_0 + \beta_1 x_{1t} + \beta_2 x_{2t} + \varepsilon_t, \quad t = 1, 2, ..., n. \tag{1}$$

The assumptions of error variables are contained mean $E(\varepsilon_t) = 0$, and variance $V(\varepsilon_t) = \sigma^2$. By the error term, ε_t and ε_k are not correlated. The sample regression model is obtained as follows

$$\hat{y}_t = b_0 + b_1 x_{1t} + b_2 x_{2t}, \quad t = 1, 2, ..., n.. \tag{2}$$

An estimator $\hat{\underset{\sim}{\theta}} = (b_0, b_1, b_2)^T$ is estimated by the Least Square Error (LSE) that specified to minimize the Sum of Square Error (SSE) as

$$SSE = \sum_{t=1}^{n} (y_t - b_0 - b_1 x_{1t} - b_2 x_{2t})^2 . \tag{3}$$

It is convenient to use matrices to approximate the regression estimators by solving the normal
equation for $\hat{\underset{\sim}{\theta}}$:

$$X^T X \hat{\underset{\sim}{\theta}} = X^T \underset{\sim}{y}, \tag{4}$$

where $\underset{\sim}{y}$ is a $n\times 1$ vector, X is a $n\times 3$ matrix, and $\underset{\sim}{\hat{\theta}}$ is a 3×1 vector, given by

$$\underset{\sim}{y}=\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \qquad X=\begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \dots & \dots & \dots \\ 1 & x_{1n} & x_{2n} \end{bmatrix}, \qquad \underset{\sim}{\hat{\theta}}=\begin{bmatrix} b_0 \\ b_1 \\ b_2 \end{bmatrix}.$$

The coefficient of parameter with parametric regression model is evaluated by

$$\underset{\sim}{\hat{\theta}}=\left(X^T X\right)^{-1} X^T \underset{\sim}{y}. \tag{5}$$

Nonparametric Regression Model

The nonparametric regression method base on a smoothing technique which is produced a smoothing function as a smoother . A smoother is a tool for summarizing the trend of a response variable as a function of one or more explanatory variables.

We mention a penalized spline method for data analysis in the class of nonparametric regression model. Penalized spline method is quite similar to smoothing spline method especially more flexible choice of the spline model, the basis function, and the penalty function.

The simple nonparametric regression model consists of single explanatory variable and response variable, but in this case we focus two explanatory variables called nonparametric additive model which is written as

$$y_t = f(x_{1t}) + f(x_{2t}) + \varepsilon_t, \qquad \text{t} = 1,\dots,\text{n}, \tag{6}$$

where y_t is a response variable, $f(x_{1t})$ is a smooth function of x_1, $f(x_{2t})$ is a smooth function of x_2, and ε_t is an error term.

Penalized spline smoother is estimated using truncated power (Ruppert and Carroll, 2000), and the penalized spline regression model is rewritten as

$$y_t = \mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) + \varepsilon_t, \ \text{t} = 1,\dots,\text{n}, \tag{7}$$

where

$$\mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) = \sum_{j=1}^{m-1} \alpha_j^{x_1} x_{1t}^j + \sum_{k=1}^{K_{x_1}} \beta_k \mid x_{1t} - \tau_k \mid^{2m-1} + \sum_{j=1}^{m-1} \delta_j^{x_2} x_{2t}^j + \sum_{k=1}^{K_{x_2}} \gamma_k \mid x_{2t} - \upsilon_k \mid^{2m-1}, t = 1,2,\dots,n, \tag{8}$$

with $\underset{\sim}{\beta} \equiv \left[\beta_1,...,\beta_{K_{x1}}\right]^T \sim N\left(0,\sigma_\beta^2\,\Omega_{x1}^{-1/2}(\Omega_{x1}^{-1/2})^T\right)$, $\Omega_{x1} \equiv \left[\left|x_{1t}-\tau_k\right|_{1\le k\le K_{x1}}^{2m-1}\right]_{1\le t\le n}$, and $\underset{\sim}{\gamma} \equiv \left[\gamma_1,...,\gamma_{K_{x2}}\right]^T \sim N\left(0,\sigma_\gamma^2\,\Omega_{x2}^{-1/2}(\Omega_{x2}^{-1/2})^T\right)$, $\Omega_{x2} \equiv \left[\left|x_{2t}-\upsilon_k\right|_{1\le k\le K_{x2}}^{2m-1}\right]_{1\le t\le n}$.

The penalized spline regression model are specified a group of locations in a range of interval $[a,b]$, where $a<\tau_1<....<\tau_{K_{x1}}<b$ and $a<\upsilon_1<....<\upsilon_{K_{x2}}<b$ introduced by Eubank (1988, 1999) . These locations are known as knots, and $\tau_k, k=1,2,...,K_{x1}$ and $\upsilon_k, k=1,2,...,K_{x2}$ are called interior knots. We will focus on the low-rank thin-plate spline (m=2) which tend to fit with the non-linear data. The low-rank thin-plate spline is presented by

$$\mu(x_{1t},x_{2t},\underset{\sim}{\theta}) = \sum_{j=1}^{2-1}\alpha_j x_1^j + \sum_{k=1}^{K_{x1}}\beta_k \mid x_{1t}-\tau_k\mid^3 + \sum_{j=1}^{2-1}\delta_j x_2^j + \sum_{k=1}^{K_{x2}}\gamma_k \mid x_{2t}-\upsilon_k\mid^3, \tag{9}$$

where $\underset{\sim}{\theta} = (\alpha_1,\beta_1,...,\beta_{K_{x1}},\delta_1,\gamma_1,...,\gamma_{K_{x2}})^T$ is the vector of penalized spline regression model, and $\tau_1<....<\tau_{K_{x1}}$ and $\upsilon_1<....<\upsilon_{K_{x2}}$ are fixed knots. Following Ruppert (2002), we consider a number of knots that is large enough to ensure the desired flexibility, and (τ_k,υ_k) are the sample quartile of x_1 and x_2 corresponding to probability $k/(K_{x1}+1)$ and $k/(K_{x2}+1)$. To avoid overfitting, we minimize

$$\sum_{t=1}^{n}\left\{y_t-\mu(x_{1t},x_{2t},\underset{\sim}{\theta})\right\}^2 + \underset{\sim}{\theta}^T D\,\underset{\sim}{\theta}, \tag{10}$$

$$D = \begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_{x1}+K_{x2})} \\ 0_{(K_{x1}+K_{x2})\times 2} & \Omega_{(K_{x1}+K_{x2})\times(K_{x1}+K_{x2})} \end{bmatrix}, \quad \text{and}$$

$$\Omega = \left[\lambda_{x1}^{2m-1}\left|x_{1t}-\tau_k\right|_{1\le k\le K_{x1}}^{3} \quad \lambda_{x2}^{2m-1}\left|x_{2t}-\upsilon_k\right|_{1\le k\le K_{x2}}^{3}\right]_{1\le t\le n},$$

where $\lambda_{x1} \equiv \dfrac{\sigma_\beta^2}{\sigma_\varepsilon^2}$, $\lambda_{x2} \equiv \dfrac{\sigma_\gamma^2}{\sigma_\varepsilon^2}$ are the smoothing parameter , and D is known positive semi-definite penalty matrix. Smoothing parameter can be approximated by restricted maximum likelihood and approximated best linear unbiased prediction (Robinson, 1991).

Just as with the linear model, we can generalize penalized spline in general linear mixed model (Brumback *et al*., 1999) as

$$\underset{\sim}{y} = X\underset{\sim}{\alpha} + Z_{Kt}\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}, \tag{11}$$

where $\underset{\sim}{y}$ is a $n\times 1$ vector, X is a $n\times 2$ matrix, $\underset{\sim}{\alpha}$ is a 2×1 vector, $\underset{\sim}{\beta}$ is a $(K_{x1}+K_{x2})\times 1$ vector,

Z_{Kt} is a $n\times(K_{x1}+K_{x2})$ matrix, $\underset{\sim}{\varepsilon}$ is a $n\times 1$ vector, given by

$$\underset{\sim}{y}=\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix}, \qquad X=\begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \dots & \dots \\ x_{n1} & x_{n2} \end{bmatrix}, \qquad \underset{\sim}{\alpha}=\begin{bmatrix} \alpha_1 \\ \delta_1 \end{bmatrix}, \ \underset{\sim}{\beta}=\begin{bmatrix} \beta_1 \\ \vdots \\ \beta_{K_{x1}} \\ \gamma_1 \\ \vdots \\ \gamma_{K_{x2}} \end{bmatrix},$$

$$Z_{Kt}=\begin{bmatrix} |x_{11}-\tau_1|^3 & \dots & |x_{11}-\tau_{K_{x1}}|^3 & |x_{21}-\upsilon_1|^3 & \dots & |x_{21}-\upsilon_{K_{x2}}|^3 \\ |x_{12}-\tau_1|^3 & \dots & |x_{12}-\tau_{K_{x1}}|^3 & |x_{22}-\upsilon_1|^3 & \dots & |x_{22}-\upsilon_{K_{x2}}|^3 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ |x_{1n}-\tau_1|^3 & \dots & |x_{1n}-\tau_{K_{x1}}|^3 & |x_{2n}-\upsilon_1|^3 & \dots & |x_{2n}-\upsilon_{K_{x2}}|^3 \end{bmatrix}, \text{and}$$

$$\underset{\sim}{\varepsilon}=\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}.$$

This class of penalized spline smoothers $\hat{\mu}(x_1,x_2,\underset{\sim}{\hat{\theta}})$ may also be expressed as

$$\hat{\mu}(x_1,x_2,\underset{\sim}{\hat{\theta}})=C(C^TC+D)^{-1}C^T\underset{\sim}{y}, \tag{12}$$

where

$$C=\begin{bmatrix} x_{1t} & x_{2t} & |x_{1t}-\tau_k|^3_{1\le k\le K_{x1}} & |x_{2t}-\upsilon_k|^3_{1\le k\le K_{x2}} \end{bmatrix}_{1\le t\le n},$$

$$D=\begin{bmatrix} 0_{2\times 2} & 0_{2\times(K_{x1}+K_{x2})} \\ 0_{(K_{x1}+K_{x2})\times 2} & \Omega_{(K_{x1}+K_{x2})\times(K_{x1}+K_{x2})} \end{bmatrix},$$

and $\Omega = \left[\lambda_{x1}^3 \left| x_{1t} - \tau_k \right|_{1\le k\le K_{x1}}^3 \quad \lambda_{x2}^3 \left| x_{2t} - \upsilon_k \right|_{1\le k\le K_{x2}}^3 \right]_{1\le t\le n}$.

The penalized spline smoother $\hat{\underset{\sim}{\theta}}$ is computed by

$$\hat{\underset{\sim}{\theta}} = (C^T C + D)^{-1} C^T \underset{\sim}{y} . \tag{13}$$

Semiparametric Regression Model

The semiparametric regression is a mixed model between parametric regression model and nonparametric regression model. The mixed model of penalized spline method allows the two explanatory variables and a response variable on the penalized spline regression model which is written as

$$y_t = \beta_0 + \beta_1 x_{1t} + f(x_{2t}) + \varepsilon_t , \quad \text{t = 1,…,n,} \tag{14}$$

where y_t is a response variable, (β_0, β_1) are coefficient of parametric regression model, $f(x_{2t})$ is a smooth function of x_2, and ε_t is an error term.

The penalized spline regression model through the mixed model is

$$y_t = \mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) + \varepsilon_t , \text{ t = 1,…,n,} \tag{15}$$

where

$$\mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) = \beta_0 + \beta_1 x_{1t} + \sum_{j=1}^{m-1} \delta_j^{x_2} x_{2t}^j + \sum_{k=1}^{K_{x2}} \gamma_k \mid x_{2t} - \upsilon_k \mid^{2m-1}, t = 1,..., n , \tag{16}$$

with $\underset{\sim}{\gamma} \equiv \left[\gamma_1,..., \gamma_{K_{x2}} \right]^T \square N\left(0, \sigma_\gamma^2 \, \Omega_{x2}^{-1/2} (\Omega_{x2}^{-1/2})^T \right)$, $\Omega_{x2} \equiv \left[\left| x_{2t} - \upsilon_k \right|_{1\le k\le K_{x2}}^{2m-1} \right]_{1\le t\le n}$.

A group of locations in a range of interval $[a,b]$ is $a < \upsilon_1 < .... < \upsilon_{K_{x2}} < b$, and the interior knots are $\upsilon_k, k = 1, 2, ..., K_{x2}$. The low-rank thin-plate spline (m=2) is written by

$$\mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) = \beta_0 + \beta_1 x_{1t} + \sum_{j=1}^{2-1} \delta_j x_2^j + \sum_{k=1}^{K_{x2}} \gamma_k \mid x_{2t} - \upsilon_k \mid^3 , \tag{17}$$

where $\underset{\sim}{\theta} = (\beta_0, \beta_1, \delta_1, \gamma_1, ..., \gamma_{K_{x2}})^T$ is the vector of penalized spline regression model, and $\upsilon_1 < .... < \upsilon_{K_{x2}}$ are fixed knots. We minimize

$$\sum_{t=1}^{n} \left\{ y_t - \mu(x_{1t}, x_{2t}, \underset{\sim}{\theta}) \right\}^2 + \underset{\sim}{\theta}^T D \underset{\sim}{\theta}, \tag{18}$$

$D = \begin{bmatrix} 0_{3\times 3} & 0_{3\times K_{x2}} \\ 0_{K_{x2}\times 3} & \Omega_{K_{x2}\times K_{x2}} \end{bmatrix}$, and $\Omega = \left[\lambda_{x2}^{2m-1} \left| x_{2t} - \upsilon_k \right|_{1\le k\le K_{x2}}^{3} \right]_{1\le t\le n}$,

where λ_{x2} is the smoothing parameter computed by $\lambda_{x2} \equiv \dfrac{\sigma_\gamma^2}{\sigma_\varepsilon^2}$ and D is known positive semi-definite penalty matrix.

We can generalize penalized spline in the matrix form as (19)

$$\underset{\sim}{y} = X \underset{\sim}{\alpha} + Z_{Kt} \underset{\sim}{\beta} + \underset{\sim}{\varepsilon},$$

where $\underset{\sim}{y}$ is a $n\times 1$ vector, X is a $n\times 3$ matrix, $\underset{\sim}{\alpha}$ is a 3×1 vector, $\underset{\sim}{\beta}$ is a $K_{x2}\times 1$ vector,

Z_{Kt} is a $n\times K_{x2}$ matrix, $\underset{\sim}{\varepsilon}$ is a $n\times 1$ vector, given by

$$\underset{\sim}{y} = \begin{bmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{bmatrix}, \qquad X = \begin{bmatrix} 1 & x_{11} & x_{12} \\ 1 & x_{21} & x_{22} \\ ... & ... & ... \\ 1 & x_{n1} & x_{n2} \end{bmatrix}, \qquad \underset{\sim}{\alpha} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \delta_1 \end{bmatrix}, \ \underset{\sim}{\beta} = \begin{bmatrix} \gamma_1 \\ \vdots \\ \gamma_{K_{x2}} \end{bmatrix}.$$

$$Z_{Kt} = \begin{bmatrix} \left|x_{21} - \upsilon_1\right|^3 & ... & \left|x_{21} - \upsilon_{K_{x2}}\right|^3 \\ \left|x_{22} - \upsilon_1\right|^3 & ... & \left|x_{22} - \upsilon_{K_{x2}}\right|^3 \\ ... & ... & ... \\ \left|x_{2n} - \upsilon_1\right|^3 & ... & \left|x_{2n} - \upsilon_{K_{x2}}\right|^3 \end{bmatrix}, \text{ and } \underset{\sim}{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ ... \\ \varepsilon_n \end{bmatrix}.$$

The fitted values of penalized spline smoothers $\hat{\mu}(x_1, x_2, \hat{\theta})$ may also be expressed as

$$\hat{\mu}(x_{1t}, x_{2t}, \underset{\sim}{\hat{\theta}}) = C(C^T C + D)^{-1} C^T \underset{\sim}{y} \quad , \tag{20}$$

where

$$C=\left[1 \quad x_{1t} \quad x_{2t} \quad \left|x_{2t}-\upsilon_k\right|^3_{1\le k\le K_{x2}}\right]_{1\le t\le n}, \quad D=\begin{bmatrix} 0_{3\times 3} & 0_{3\times K_{x2}} \\ 0_{K_{x2}\times 3} & \Omega_{K_{x2}\times K_{x2}} \end{bmatrix},$$

and $\Omega=\left[\lambda_{x2}^3\left|x_{2t}-\upsilon_k\right|^3_{1\le k\le K_{x2}}\right]_{1\le t\le n}$.

The coefficient of parameter with semiparametric regression model is

$$\underset{\sim}{\hat{\theta}}=(C^TC+D)^{-1}C^T\underset{\sim}{y}. \tag{21}$$

SIMULATION STUDY

In this section, we display the process and the results of a simulation experiment that we conducted in order to compare the performance of parametric regression model, nonparametric regression model, and semiparametric regression model. To simulate data, we generated data in the class of a response variable (y) and two explanatory variables (x). The response variable is obtained from

$$y_t = f(x_{1t})+x_{2t}+\varepsilon_t, \text{t=1,…,n} \tag{22}$$

where $f(x_{1t})=x_{2t}^3-(x_{2t}^2\cos(x_{2t}))-\exp\left\{\frac{x_{2t}}{1+\left|x_{2t}\right|}\right\}$, $x_{2t} \sim Uniform(-r,r)$, and $\varepsilon_t \sim Normal(0,1)$.

The uniform distribution of x_{2t} is generated with several values for r = 1, 3, 5, and 7. The data is simulated by R program with sample sizes 25, 50, 100, and 300 , and repeated for fitting model 500 times in each cases. Figure 1 illustrates an example of simulated data with 50 sample sizes (n=50) in scatter plots that shown the relationship between response variable and two explanatory variables.

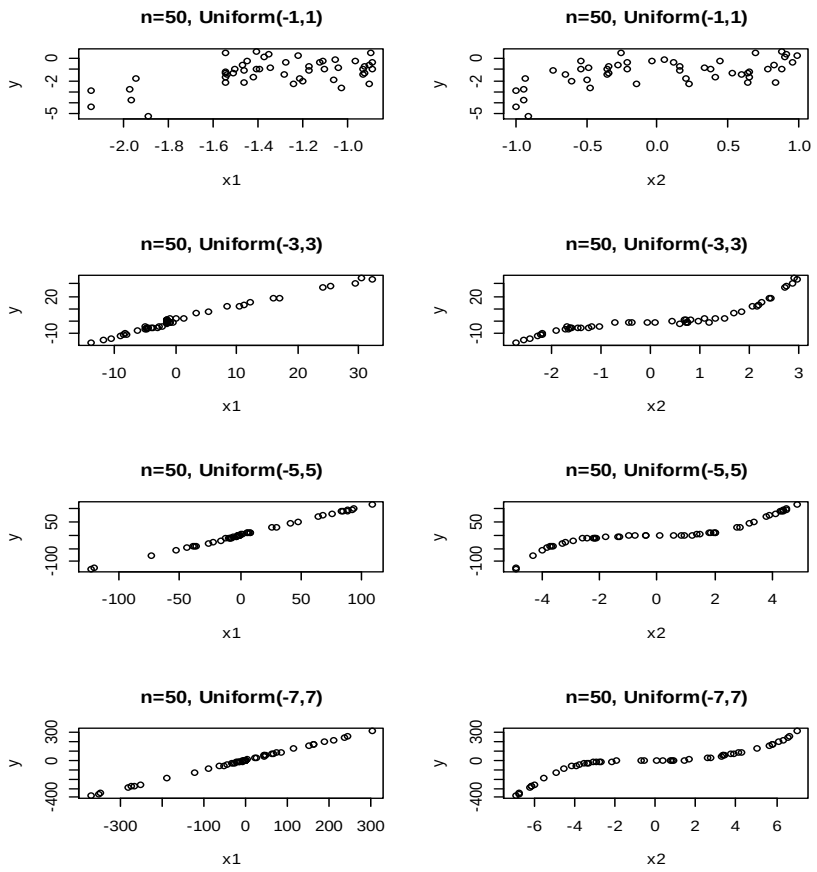

Figure 1 The scatter plot between response variable and two explanatory variables (n=50).

We fit the data set for estimating parameter of parametric regression which is evaluated by the R program. However, the SemiPar Package in the R Program is used for data analysis with the nonparametric regression and semiparametric models. The performance of parameter estimation is compared by Mean Square Error (MSE) as follows:

$$MSE = \frac{\sum_{t=1}^{n}(y_t - \hat{y}_t)^2}{n},$$

where y_t are the response variable, and $\hat{y}_t$ are the fitted values.

The first and second columns of Table 1 show various sample sizes and its minimum and maximum values from uniform distribution. The third to fifth columns present the average MSE of parametric regression model (Reg), nonparametric regression model (Non), and semiparametric regression model (Sem).

Table 1 The average MSE of parametric regression model (Reg), nonparametric regression model (Non), and semiparametric regression model (Sem).

Sample sizes	Uniform	Reg	Non	Sem
n=25	-1,1	0.9348	**0.8050**	0.8482
	-3,3	0.7481	**0.5795**	0.6067
	-5,5	0.9139	**0.8859**	0.9139
	-7,7	0.4402	**0.2796**	0.3512
n=50	-1,1	0.9434	**0.8946**	0.9114
	-3,3	0.9476	**0.8844**	0.9033
	-5,5	0.8854	**0.8003**	0.8609
	-7,7	0.6926	**0.6718**	0.6855
n=100	-1,1	0.9902	**0.9531**	0.9680
	-3,3	0.9662	**0.9407**	0.9512
	-5,5	0.9693	**0.9375**	0.9502
	-7,7	0.9703	**0.9513**	0.9606
n=300	-1,1	0.9986	**0.9882**	0.9924
	-3,3	0.9911	**0.9810**	0.9847
	-5,5	0.9887	**0.9785**	0.9832
	-7,7	0.9862	**0.9759**	0.9796

By observing the average MSE, the results appear that the nonparametric regression model provides the minimum values in all cases. The average MSE is decreasing when the range of uniform distribution is increasing especially the sample sizes n = 25, 50, and 100. When the sample size is increasing, the average MSE is increasing depended on the asymptotic relative efficiency. However, the average MSE of nonparametric regression model is shown the smallest values since the penalized spline method can be conducted using the truncated power function based on the number of knots which is controlled to trade-off the goodness of fit.

THE APPLICATION IN REAL DATA

In this section, we apply the model described in Section 2 to analyze with the real data. The gold price (US Dollars per Troy Ounce) is denoted the response variable and two explanatory variables are defined by the crude oil price (US Dollars per Barrel) and the number of month. These data consisted of 312 records of monthly volume from January 1988 to December 2013 that can be found at *www.cmegroup.com* and *www.eia.gov.* The estimated parameters are obtain from the

data analysis of parametric regression model, nonparametric regression model, and Semiparametric regression model given in Table 2.

Table 2 The parameter estimation of parametric regression (Reg) model, nonparametric regression (Non) model, and semiparametric regression (Sem) model.

	Reg	Non	Sem
Parameter Estimation	$\hat{\underset{\sim}{\theta}} = \begin{bmatrix} b_0 = 92.4411 \\ b_1 = 11.8667 \\ b_2 = -0.8904 \end{bmatrix}$	$f(x_1), \quad \lambda_{x1} = 23.25, K_{x1} = 35$ $f(x_2), \quad \lambda_{x2} = 352.6, K_{x2} = 2$	$\hat{\underset{\sim}{\theta}} = \begin{bmatrix} b_0 = 145 \\ b_1 = 11.89 \end{bmatrix}$ $f(x_2), \quad \lambda_{x2} = 375.5, K_{x2} = 2$
MSE	45393.51	22462.54	45047.07

From Table 2, it is apparent that MSE by the nonparametric regression model is the smallest values. Therefore, it should be noted that the nonparametric regression model performs better than the parametric regression model and semiparametric regression model since the nonparametric regression model contains the two smoothing function which can be interpolated more than the other models.

Figure 2 shown the fitted values from 3 models and the gold price is the bottom panel. It follows from the Figure 2 that the nonparametric regression model can be close the real values more than two models.

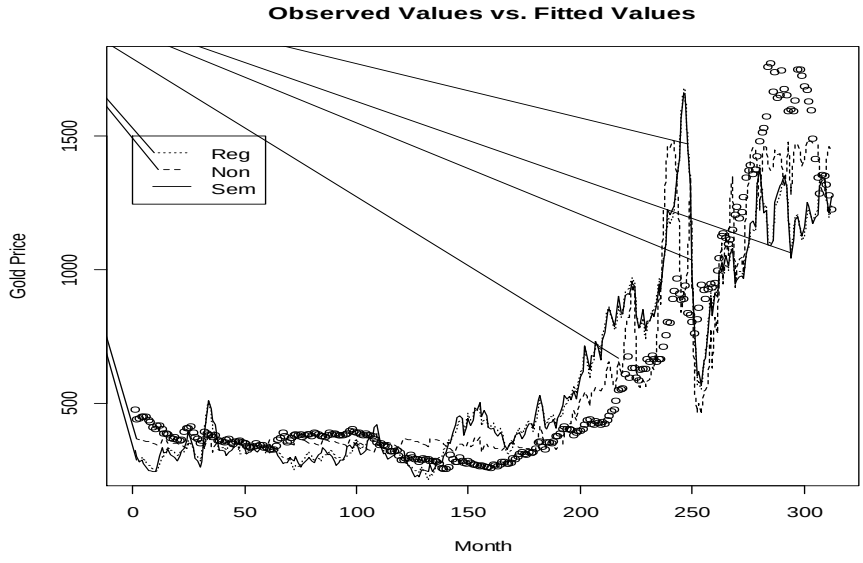

Figure 2 The scatter plot of the gold price and fitted values of Reg model, Non model, and Sem model.

CONCLUSIONS

In this paper, we have estimated parametric regression model, nonparametric regression model, and semiparametric regression model in classes of response variable and two explanatory variables. The parametric regression model is fitted by the least square error method while the penalized spline method is proposed to estimate the nonparametric regression model and semiparametric regression model in terms of smoothing function. Through simulation study, the response variable is generated in terms of correlated function between two explanatory variables. The results show that the nonparametric regression model performs better than other two models based method in terms of minimizing MSE. For real data, we are also interested the financial data to compare the performance of three models by considering the MSE. The results is similar to the simulation data that the nonparametric regression is a superior over parametric regression model and semiparametric regression model. This is expected because the nonparametric regression method consists of smoothing function which controlled the trade-off between fidelity to the data and roughness of estimating function. On future work, we may focus a local polynomial regression method, regression splines method, smoothing splines method based on nonparametric regression method to compare with 3 models.

REFERENCES

Brumback B.A., Ruppert D. and Wand M. P. (1999). Comment on Shively, Kohn, and Wood, *Journal of the American Statistical Association*, 94, 794-797.

David R., Wand M.P., and Caroll R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

Eubank R.L. (1988). *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Eubank R.L. (1999). *Nonparametric Regression and Spline Smoothing*, Marcel Dekker, New York.

Fan J. (1992). Design-adaptive nonparametric regression, *Journal of American Statistics Association*, 87, 998-1004.

Fan J. (1993). Local linear regression smoothers and their minimax efficiency, *The Annals of Statistics*, 21, 196-226.

Fan J. and Gijbels I. (1996). *Local polynomial modelling and its applications*, Chapman and Hall, London.

Green P.J.and Silverman B.W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall, London.

Robinson D. A. (1991). That BLUP is a good thing: the estimation of random effects, *Statistical Science*, 6, 15-51.

Robinson P.M. (1983). Non-parametric estimation for time series models, *Journal of Time Series Analysis*, 4, 185-208.

Ruppert D. (2002). Selecting the number of knots for penalized splines, *Journal of computational and Graphical Statistics*, 11, 735-757.

Ruppert D. and Carroll R.J. (2000). Spatial-adaptive penalties for spline fitting, *Australian and New Zealand Journal of Statistics*, 42, 205-224.

Ruppert D., Wand, M.P. and Carroll R.J. (2003). *Semiparametric Regression*, Cambridge University Press.

Stone C.G. (1977). Consistent nonparametric regression, *The Annals of Statistics*, 5, 595-620.

Wahba G. (1990). *Spline Models for Observational Data*, SIAM, Philadelphia, PA.

Wand M.P. and Jones M.C. (1995). *Kernel Smoothing*, Chapman and Hall, London, UK.

Wu H. and Zhang J.T. (2006) . *Nonparametric Regression Methods for Longitudinal Data Analysis*, A John Wiley & Sons INC., Publication, New Jersey, USA.