Statistical Diagnosis and Gross Error Test for Semiparametric Linear Model

DING Shijun  ZHANG Songlin  JIANG Weiping  WANG Shouchun

Abstract  This paper systematically studies the statistical diagnosis and hypothesis testing for the semiparametric linear regression model according to the theories and methods of the statistical diagnosis and hypothesis testing for parametric regression model. Several diagnostic measures and the methods for gross error testing are derived. Especially, the global and local influence analysis of the gross error on the parameter $X$ and the nonparameter $s$ are discussed in detail; at the same time, the paper proves that the data point deletion model is equivalent to the mean shift model for the semiparametric regression model. Finally, with one simulative computing example, some helpful conclusions are drawn.

Keywords  parametric regression; semiparametric linear model; influencing analysis; statistical diagnosis; gross error testing

CLC number  P207; P234

Introduction

Over the last 20 years, the diagnosis and analysis of linear regression model has been freely developed [1-3]. Craven and Wahba[4, 5] has gotten some results of diagnosis and analysis for the spline estimation of the nonparametric model regression. However, there is little literature, except the reference [6] that had some discussion on the statistical diagnosis of the semiparametric nonlinear model. There is no report about the statistical diagnosis and gross testing of the semiparametric linear model except for the paper [7] in surveying data processing [4]. Hence, this paper will talk about these aspects.

In the parametric regression analysis, the matrix of the regression vectors that is transformed from observation vectors is called the hat matrix. Hat matrix is very important for the regression diagnosis of the residuals, because its elements will appear in the estimations of the variance and covariance of the residuals. The parametric regression model can be written as $[1, 8]$

$$L = BX + \Delta$$

(1)

where $B$ is a known $n \times t$ nonsingular matrix of rank $t$, $X$ is a $t \times 1$ dimension unknown parameters vector, $L$ is an observation vector of $n \times 1$ dimension, and $\Delta$ is a $n \times 1$ dimension noise vector with mean zero. From Eq. (1), we can get

$$\hat{L} = LH$$

(2)

$$V = -(I - H)L$$

(3)

Where $\hat{L}$ is a vector of fitted value, $V$ is a $n \times 1$ residual vector, $I$ is a $n \times n$ unit matrix, and

$$H = B(B^T PB)^{-1}B^T P$$

and $P$ is a weight matrix of the error $\Delta$. The matrix $H$, which is frequently called the hat matrix, satisfies $H^2 = H$ and $H^T = H$. The elements of $H$ satisfy

Received on June 11, 2009.

Supported by the National Natural Science Foundation of China (No. 406604001), the National High Technology Research and Development Program of China (No. 2007AA12Z312).

DING Shijun, School of Geodesy and Geomatics, Wuhan University, 129 Luoyu Road, Wuhan 430079, China.

E-mail: Shjding@sgg.whu.edu.cn
If \( h_{ij} = 0 \) for all \( i \neq j \), then
\[
h_{ij} = 1 \quad (5)
\]
and
\[
D_{ij} = \sigma_0^2 (I - H) P^{-1} \quad (6)
\]

Where \( D_{ij} \) is the variance and covariance matrix of the residuals, and \( \sigma_0^2 \) is the posteriori variance factor. In the parametric regression, the predictive residual \( e_i \), the studentized residual \( r_i \), the studentized predictive residual \( r_i^* \), and ordinary residual \( v_i \) satisfy the following equations:
\[
e_i = \frac{v_i}{1 - h_{ii}} \quad (7)
\]
\[
r_i = \frac{v_i}{\hat{\sigma}_0 \sqrt{(1 - h_{ii})} P_i^{-1}} \quad (8)
\]
\[
r_i^* = \frac{v_i}{\hat{\sigma}_0^{-i} \sqrt{(1 - h_{ii})} P_i^{-1}} \quad (9)
\]

Where \( \hat{\sigma}_0^2 = V^T P V / (n - t) \) is the estimate of \( \sigma_0^2 \), and \( \hat{\sigma}_0^2 \) is the estimate of \( \sigma_0 \) that is computed using the dataset \( L_i, \cdots, L_n, r_i \), which approximately obeys a standard Normal distribution, and \( r_i^* \) obeys a t-distribution with \( n - t - 1 \) degrees of freedom.

The purpose of this paper is to study the statistical diagnosis and hypothesis testing for the semiparametric model based on the parametric diagnosis and hypothesis testing. In Section 1, it introduces the principle of penalized least squares estimation for semiparametric model and describes methods of statistical diagnosis in Section 2. According to the semiparametric mean shift model, outlier detection will be discussed in Section 3. Finally, simulative computations of some conclusions are given in Sections 4 and 5.

1. Penalized least squares for semiparametric model

The semiparametric linear model can be expressed as
\[
L_i = h_i^T X + s(t_i) + \Delta, \quad i = 1, \cdots, n \quad (10)
\]
\[
E(\Delta) = 0, \quad D(\Delta) = \sigma_0^2 P^{-1} \quad (11)
\]

Eq.(10) can be rewritten in its vector form as follows:
\[
L = BX + s + \Delta \quad (12)
\]

Where \( L \) is an \( n \times 1 \) observation vector, \( B \) is a known column nonsingular design matrix, \( X \) is a \( t \times 1 \) parameter vector, \( \Delta \) is an \( n \times 1 \) noise vector, \( s = (s(t_1), s(t_2), \cdots, s(t_n))^T \) is an \( n \times 1 \) nonparameter vector, \( \sigma_0 \) is the priori variance factor, and \( D(\Delta) \) and \( P \) are variance and weight matrix of the error \( \Delta \), respectively.

The penalized least squares criterion for the semiparametric model is
\[
V^T PV + \alpha \hat{\sigma}_0^2 R \hat{s} = \min \quad (13)
\]

The normal equation of Eq.(13) can be written as
\[
\begin{bmatrix}
B^T P B & B^T P \\
P & P + \alpha R
\end{bmatrix}
\begin{bmatrix}
\hat{X} \\
\hat{\hat{s}}
\end{bmatrix}
= \begin{bmatrix}
B^T P L \\
P L
\end{bmatrix} \quad (14)
\]

The solutions of Eq. (14) are
\[
\hat{X} = (B^T P (I - s) B)^{-1} B^T P (I - s) L \quad (15)
\]
\[
\hat{s} = (P + \alpha R)^{-1} P (L - B \hat{X}) \quad (16)
\]
\[
\hat{L} = H(\alpha) L \quad (17)
\]
\[
V = \hat{\hat{L}} - L = (H(\alpha) - I) L \quad (18)
\]

Where \( H(\alpha) \) is also called the hat matrix, \( H(\alpha) = S + (I - S)B(B^T P (I - s) B)^{-1} BP(I - S) \), and it is a nonidempotent matrix. \( \alpha \) is the smoothing parameter, and the smoothing matrix \( S = (P + \alpha R)^{-1} P \), \( R \) is a design matrix correlative to the nonparameter vector \( s \).

2. Statistical diagnosis for semiparametric model

2.1 Solution of the semiparametric forecasting model

Based on the basic principle of the forecasting model for parametric regression, we delete one group of observations \( (L_i, h_i^T, t_i) \) from the model of Eq. (10). Then, the new model can be expressed as \([4, 8]\)
\[
L^{-i} = B^{-i} X^{-i} + s^{-i} + \Delta^{-i} \quad (19)
\]

The model above is called the data deleted model or the data forecasting model. Eq.(19) denotes that it is obtained by omitting the ith observation.

Suppose that \( \hat{X}^{-i} \) and \( \hat{s}^{-i} \) are the penalized least squares estimates of \( X \) and \( s \) for the data deleted
model of Eq. (19). In order to study the influence of the data point \((L_i, h_i^T, t_i)\) on \(\hat{X}\) and \(\hat{s}\), we simply search the differences between \(\hat{X}\), \(\hat{s}\) and \(\hat{X}^{-i}\), \(\hat{s}^{-i}\). As ground work of our discussion, we provide a theorem [9].

**Theorem.** Set vector \(L' = (L'_1, L'_2, \cdots, L'_{n-1}, L'_n)^T\), which satisfies
\[
L'_j = L_j, \quad j \neq i \quad L'_i = b_i^T \hat{X}^{-i} + \hat{s}^{-i}(t_i), \quad j = i
\]
then \(L'^{-i} = H(\alpha)L'\).

According to the theorem above, it follows that
\[
\hat{L}^{-i} = \sum_{j=1, j \neq i}^n h_j L_j - L_i = \sum_{j=1, j \neq i}^n h_j L_j + h_i (b_i^T \hat{X}^{-i} + \hat{s}^{-i}(t_i)) - L_i
\]
\[
= \sum_{j=1}^n h_j L_j - L_i + h_i (b_i^T \hat{X}^{-i} + \hat{s}^{-i}(t_i) - L_i)
\]
\[
= \hat{L} - L_i + h_i e_i
\]

We can get
\[
e_i = \frac{v_i}{1 - h_i}
\] (20)

Where \(v_i = \hat{L} - L_i\), and \(e_i = b_i^T \hat{X}^{-i} + \hat{s}^{-i}(t_i) - L_i\).

\(\hat{X}^{-i}\) and \(\hat{s}^{-i}\) are the solutions of the following minimum function:
\[
\varphi_1(X, s) = \sum_{j=1, j \neq i}^n P(L_j - b_j^T X - s(t_j))^2 + \alpha s^T R s
\] (21)

At the same time, \(\hat{X}^{-i}\) and \(\hat{s}^{-i}\) are also solutions of the following minimum function:
\[
\varphi_1(X, s) = \sum_{j=1, j \neq i}^n P(L_j - b_j^T X - s(t_j))^2 + P(L'_i - b_i^T X - s(t_i))^2 + \alpha s^T R s
\]

namely
\[
\varphi_1(X, s) = \sum_{j=1}^n P(L'_j - b_j^T X - s(t_j))^2 + \alpha s^T R s = \min
\] (22)

By comparing Eqs. (13) and (22), setting \(d_i=(0, 0, \cdots, 0, 1, 0, \cdots, 0)^T\), in which the \(i^{th}\) element of the vector \(d_i\) is 1, we can get the following from the theorem above and Eq.(15):
\[
\hat{X}^{-i} = (B^T P(I - S) B)^{-1} B^T P(I - S) L'
\]
\[
= (B^T P(I - S) B)^{-1} B^T P(I - S) (L - (L_i - b_i^T \hat{X}^{-i} - \hat{s}^{-i}(t_i)) d_i)
\]
\[
= (B^T P(I - S) B)^{-1} B^T P(I - S) L + (B^T P(I - S) B)^{-1} B^T P(I - S) d_i e_i
\]

Substituting Eq. (23) into the formula above, we get
\[
\hat{X}^{-i} = \hat{X} + \frac{(B^T P(I - S) B)^{-1} B^T P(I - S) d_i e_i}{1 - h_i}
\] (23)

Similarly, following Eq. (16), we get
\[
\hat{s}^{-i} = (P + \alpha R)^{-1} P(L - BX^{-i})
\]
\[
= (P + \alpha R)^{-1} P(L - (L_i - b_i^T \hat{X}^{-i} - \hat{s}^{-i}(t_i)) d_i - BX^{-i})
\]

Substituting Eq.(23) into the formula above then gives
\[
= (P + \alpha R)^{-1} P\{L + d_i e_i - BX
\]
\[
- B[B^T P(I - S) B]^{-1} B^T P(I - S) d_i e_i\}
\]
\[
= \hat{s} + \frac{(P + \alpha R)^{-1} P(I - B[B^T P(I - S) B]^{-1} B^T P(I - S) d_i e_i)}{1 - h_i}
\] (24)

where \(v_i = b_i^T \hat{X} + \hat{s}(t_i) - L_i\).

2.2 **Measures of statistical diagnosis**

Obviously, if the data sufficiently fits the model of Eq. (10), the estimation of \(\hat{X}\) and \(\hat{s}\) will not change too much when a data point is deleted. On the other hand, if the change is not tiny, it means that the deleted data point does not belong to the dataset from which the other point data come. In order to study the influence of the \(i^{th}\) observation \((L_i, b_i^T, t_i)\) on the estimation of \(\hat{X}\) and \(\hat{s}\), let \(\tilde{\beta} = (X^T s^T)^T\), we defined an influencing function as
\[
IF_i = \tilde{\beta}^{-1} - \hat{\beta}
\] (25)

This denotes the influence of the \(i^{th}\) observation on the estimation of \(\tilde{\beta}\). Since the influencing function is a vector, which is not convenient to use in practice, it is necessary to select appropriate data or distance to construct a scalar quantity to measure the difference between the estimates of the two kinds of model. Therefore, we can generalize some kinds of distances that are frequently used in parametric regression analysis as a diagnosis for the semiparametric model.

2.2.1 **Generalized cook’s distance**

Cook’s distance is a commonly used estimate of the influence of a data point when doing least squares regression. It measures the effect of deleting a given observation. Cook’s distance is put forward from the standpoint of the parametric confidence region; here, we will generalize it to the semiparametric regression model. We take a nonnegative definite matrix \(M\) and
a positive real number $c$ as scale factor and define the distance of influencing function $IF$ as $D_i(M,c) = (\hat{\beta}_i - \beta_i) M (\hat{\beta}_i - \beta_i) / c$ as 

$$D_i(M,c) = (\hat{\beta}_i - \beta_i) M (\hat{\beta}_i - \beta_i) / c \quad (26a)$$

$D_i(M,c)$ is called the global influence of the gross error on the parameter $\hat{X}$ and the nonparameter $\hat{s}$, and it depends on $M$ and $c$. One possible choice is:

$$H(\alpha) = \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} I \end{bmatrix} M^{-1} \begin{bmatrix} B \end{bmatrix}^T P L \begin{bmatrix} I \end{bmatrix} P$$

We can obtain the following equation from Eq. (14):

$$\begin{bmatrix} \hat{X} - \hat{\bar{X}} \\ \hat{s} - \bar{s} \end{bmatrix} = \begin{bmatrix} M^{-1} \begin{bmatrix} B \end{bmatrix}^T P (L - L) \\ M^{-1} \begin{bmatrix} B \end{bmatrix}^T P d_i e_i \end{bmatrix}$$

From Eqs. (20), (26), (27), and (28), we get the following formula:

$$D_i(M,\hat{\alpha}) = e_i^T P \begin{bmatrix} B \end{bmatrix}^T \begin{bmatrix} B \end{bmatrix} \begin{bmatrix} I \end{bmatrix} M^{-1} \begin{bmatrix} B \end{bmatrix}^T P d_i e_i / \hat{\sigma}_0^2$$

$$= \frac{h_i}{1 - h_i} r_i^2 \quad (29)$$

where $r_i = \frac{v_i}{\hat{\sigma}_0 \sqrt{(1 - h_i)p_i^{-1}}}$. 

Another possible choice is:

$$M = \begin{bmatrix} B^T P B & B^T P \\ PB & P + \alpha R \end{bmatrix}, \quad c = \hat{\sigma}_0^{-2} \quad (30a)$$

Similarly, we get

$$D_i(M,\hat{\sigma}_0) = \frac{h_i}{1 - h_i} r_i^2 \quad (30b)$$

where $r_i^* = \frac{v_i}{\hat{\sigma}_0 \sqrt{h_i}}$.

Eqs. (29) and (30) show that the generalized Cook's distance depends on the residual $v_i$ and the diagonal element $h_i$ of the hat matrix $H(\alpha)$. This is similar to the Cook’s distance in the parametric regression model.

For the parametric part $D(X)$ or the nonparametric part $D(s)$, it is called the local influence of the gross error on the parameter $X$ or the nonparameter $s$. We define those as follows:

$$D_i(X) = (\hat{X} - X)^T (I_1 [0] M^{-1} [I_0] 0^T) (\hat{X} - X) / \hat{\sigma}_0^2$$

$$D_i(s) = (\hat{s} - \bar{s})^T (0 [I_1] M^{-1} [0] [I_1]) (\hat{s} - \bar{s}) / \hat{\sigma}_0^2 \quad (31a)$$

Using Eqs. (23) and (24), we get

$$D_i(X) = \frac{d_i^T (H(\alpha) - S)d_i}{1 - h_i} r_i^2 \quad (32a)$$

$$D_i(s) = \frac{d_i^T (H(\alpha) - S)d_i}{1 - h_i} r_i^2 \quad (32b)$$

Where $H_i = H + (I - H)[P(I - H) + \alpha R]^{-1} P(I - H)$.

Similarly, using Eqs. (23) and (24), we get the following formulas:

$$D_i(X) = \frac{d_i^T (H(\alpha) - S)d_i}{1 - h_i} r_i^2 \quad (33)$$

$$D_i(s) = \frac{d_i^T (H(\alpha) - S)d_i}{1 - h_i} r_i^2 \quad (34)$$

2.2.2 Statistic of Welsch-Kuh

The statistic of Welsch-Kuh is put forward from the standpoint of data fitting. Considering that the influence on the fitted value before or after the $i$th data point is deleted, we get with the above Theorem

$$\hat{L}_j^i - \hat{L}_j = \frac{h_i}{1 - h_i} v_j \quad (35)$$

We consider the variance $\hat{\sigma}_0^2 \sqrt{P_i^{-1}} h_i$ for $\hat{L}_j$, and used it to divide Eq.(35) to get the statistic of Welsch-Kuh ($WK_i$), namely,

$$WK_i = \frac{\hat{L}_j^i - \hat{L}_j}{\hat{\sigma}_0 \sqrt{P_i^{-1} h_i}} = \frac{v_j}{\hat{\sigma}_0 \sqrt{(1 - h_i)P_i^{-1}}} \sqrt{1 - h_i}$$

$$= r^* \sqrt{\frac{h_i}{1 - h_i}} \quad (36)$$

where

$$\hat{\sigma}_0^2 = \sum_{j=1}^n p_j ((L_j - \hat{b}^T \hat{X} - \hat{s}^T (t_j))^2 / (n - tr(H(\alpha)) - 1)$$

2.2.3 Value of lever

The value of lever is very important in linear regression diagnosis. It is the diagonal element $h_i$ of the hat matrix $H(\alpha)$. We can generalize it to the semiparametric model and define the general value of...
lever as
\[
H(\alpha) = S + (I - S)B(B^T P(I - S)B)^{-1}B^T P(I - S)
\] (38)

3 Mean shift model and outlier testing

The data deletion model is quite intuitionistic and also convenient for computation. It is the base for designing an effective diagnosis statistic and is the most important diagnosis model in practice. Another diagnosis model is the mean shift model, and it can be expressed as
\[
\begin{align*}
L_i = b_i^T X + s(t_i) + \Delta, & \quad j \neq i \\
L_i = b_i^T X + s(t_i) + \delta + \Delta, & \quad j = i
\end{align*}
\] (39)

Its vector form is
\[
L = BX + s + d\delta + \Delta
\] (40)

where \(d_i = (0, 0, \ldots, 1, \ldots, 0)^T\) is a unit column vector in which the \(i\)th element’s value is 1. \(\delta\) denotes an outlier. The model shows that the outlier is added to the \(i\)th observation. Setting \(\hat{X}_m, \hat{s}_m\) and \(\delta\) are the penalized least squares solutions of this model. According to Yu and Tao [8], the data deletion model and mean shift model are equivalent for parametrical model. This conclusion holds for the semiparametric model according to the following proof. Therefore, we get the following result:
\[
\hat{X}^{-i} = \hat{X}_m, \quad \hat{s}^{-i} = \hat{s}_m
\] (41)

**Proof.** By definition, we know that \(\hat{X}^{-i}\) and \(\hat{s}^{-i}\) are the resolutions of the following minimal problem:
\[
\varphi(X, s) = \sum_{j, i \neq i}^n p_j (L_j - b_j^T X - s(t_j))^2 + \alpha s^T Rs
\] (42)

\(\hat{X}_m\) and \(\hat{s}_m\) are the solutions of the following minimal problem for the model of Eq.(39), namely,
\[
\varphi_m(X, s, \delta) = \sum_{j, i \neq i}^n p_j (L_j - b_j^T X - s(t_j))^2 + p_i (L_i - b_i^T X - s(t_i) - \delta)^2 + \alpha s^T Rs
\] (43)

Therefore, the estimate \(\hat{\delta}\) of the outlier satisfies the following equation:
\[
\frac{\partial \varphi_m(X, s, \delta)}{\partial \delta} = -2p_i (L_i - b_i^T X - s(t_i) - \delta) = 0
\]

\(\delta = L_i - b_i^T X - s(t_i)\) (44)

When substituting Eq.(44) into Eq.(43), we will get \(\varphi(X, s) = \varphi_m(X, s, \delta)\), and find that they are the same minimal problems. Then, we get \(\hat{X}^{-i} = \hat{X}_m\), \(\hat{s}^{-i} = \hat{s}_m\), and the conclusion of Eq.(41) is right.

To construct the procedure by which we can find the outlier in the semiparametric model, we propose a hypothesis testing based on mean shift model:

Null hypothesis:
\[
E(L) = BX + s
\] (45)

Alternative hypothesis:
\[
E(L) = BX + s + d\delta
\] (46)

If the variance of unit weight \(\sigma^2_o\) is known, we can construct the following statistic [1, 6]:
\[
T = \frac{\varphi_m(\hat{X}, \hat{s}) - \varphi_m(\hat{X}, \hat{s}, \delta)}{\sigma^2_o} \sim \chi^2(1)
\] (47)

Under the null hypothesis, the square root of the statistic \(T\) obeys a standard normal distribution
\[
T^{1/2} = \frac{v_i}{\sigma^2_o \sqrt{(1 - h_i)} p_i^{-1}} \sim N(0, 1)
\] (48)

If the variance of unit weight \(\sigma^2_o\) is unknown, then we can construct the following statistics as
\[
r_i = \frac{v_i}{\hat{\sigma} \sqrt{(1 - h_i)} p_i^{-1}} \sim t(n - tr(H(\alpha)) - 1)
\] (49)

\(r_i\) approximately obeys a standard normal distribution, and \(r_i\) obeys a \(t\)-distribution with \(n - tr(H(\alpha))\) degrees of freedom.

4 Simulative computation

According to the simulation example of the paper [7], set a simulated observation \(L = BX + s + \Delta\), \(B = (b_{1i})_{0.0, 2}\), \(b_{1i} = \sin(t_i)\), \(b_{2i} = \sin(3t_i)\sin(2t_i)\), \(i = 1, 2, \cdots, 80\), \(t_i = 2(i - 1)\pi/180\), \(X = [10, 10]^T\), \(s(t_i) = 4(i/35)^2\), \(\Delta \sim N(0, \sigma^2_0 P^{-1})\), \(\sigma^2_0 = 1\), \(P = I\), and the position and numerical value of the gross errors are shown in Table 1.

| Gross error | \(L_{20}\) | \(L_{30}\) | \(L_{40}\) | \(L_{50}\) | \(L_{60}\) |
|-------------|--------|--------|--------|--------|--------|
| \(\sigma_0\) | 3 \(\sigma_0\) | 5 \(\sigma_0\) | 8 \(\sigma_0\) | 5 \(\sigma_0\) | 3 \(\sigma_0\) |
The data processing results with the parametric model: $\hat{X} = [3.246, 11.536]_T$, $\hat{V}^T \hat{P} \hat{V} = 522.3$, and $\hat{\sigma}_0 = \pm 8.813$. The test statistics of the single gross error are shown in Figs.1 and 2. The computing values of $r_i$ and $r'_i$ are less than the critical values 1.96 and 1.98, so the gross error testing is failed in the case of observations including systemic errors.

The data processing results with the semiparametric model: the smoothing parameter $\alpha = 2.76$, $\hat{V}^T \hat{P} \hat{V} = 206.14$, and $\hat{\sigma}_0 = \pm 1.663$. $\hat{X} = [10.05, 10.09]_T$. The singular outlier testing, $r_{30}^*, r_{40}^*$, and $r_{50}^*$ are bigger than the critical value 2. $t_{0.05/2} = 2.0$. Obviously, $r_{30}^*, r_{40}^*$, and $r_{50}^*$ hold outliers. The estimates of residuals are shown in Fig.3, and the estimates of the systematic parameters (the nonparamters) are shown in Fig.4.

Repeating computation after the gross errors of the observations $r_{30}^*, r_{40}^*$, and $r_{50}^*$ are deleted. The computing results are as follows: $\alpha = 0.98$, $\hat{V}^T \hat{P} \hat{V} = 83.89$, $\hat{\sigma}_0 = \pm 1.067$, $\hat{X} = [9.96, 10.02]_T$, $r_{20}^* = 2.392$, and $r_{60}^* = 3.067$ are bigger than the critical value 2. $L_{20}$ and $L_{60}$ affirmatively hold gross errors and should be eliminated. The estimates of the systematic parameters are shown in Fig.5 after all gross errors are deleted. Comparing Fig.5 with Fig.4, the estimates of the systematic parameters in Fig.5 are near to the simulative values.

5 Conclusion

First, this paper gives the solutions of the penalized least square estimation for semiparametric model discuss.

Second, it derives the relations between the resolutions of the semiparametric forecasting model and original model. Based on this, some measures of statistical diagnosis for the semiparametric regression model can be further derived, and they are the Generalized Cook’s Distance, the Statistic of Welsch-Kuh, and the value of lever. Especially both the global and local influence measures of the gross error on the parameter $X$ and the nonparameter $s$ about the Generalized Cook’s Distance are discussed in detail.

Third, the paper proves that the estimations of the forecasting model and the shift model are equivalent, and their resolutions are the same. Finally, the methods of gross error testing for semiparametric model has put forward based on mean shift model, and two statistics of hypothesis testing are given.

Finally, the experiment shows that the gross error
testing with the parametric model failed and was effective with the semiparametric model when the observations contain systemic errors. It also shows that the estimates of the parameter $X$ and the variance factor $\sigma^2$ are more exact when the systematic errors be rightly separated and also the gross errors from the model.

Acknowledgement

The authors thank Prof. Tao Benzao and Prof. Wang Xingzhou for several helpful suggestions during the preparation of this manuscript.

References

[1] Chen Xiru, Wang Guisong (1987) Classical and modern regression analysis[M]. Hefei:Anhui Education Publishing Hous
[2] Koch K R (1988) Parameter estimation and hypothesis testing in linear models[M]. Berlin, Heidelberg, New York: Springer
[3] Li Deren(1988) Error processing and reliability theory[M]. Beijing :Surveying and Mapping Press
[4] Eubank R L (1984) The Hat matrix for smoothing splines[J]. *Statistics and Probability Letters*, 2: 9-14
[5] Eubank R L (1985) Diagnostics for smoothing splines[J]. *J. R. Statist. Soc. B*, 47(2): 332-341
[6] Zhu Zongyi, Wei Bocheng (2001) Diagnostic and influence analysis for semiparametric nonlinear model[J]. *Acta Mathematicae Application Sinica*, 24 (4): 568-581
[7] Ding Shijun, Tao benzao,Jin Xiangsheng(2006)The research on gross error test for Semparametric models[J]. *Bulletin Surveying and Mapping*, 6:1-5
[8] Yu Zongchou,Tao Benzao, et al. (1993) The papers for theory and application of survey adjustment models and model Errors[M]. Beijing: Surveying and Mapping Press
[9] Green P J, Silverman B W (1994) Nonparametric regression and generalized linear models[M].London: Chapman and Hall
[10] Fischer B, Hegland M (1999) Collocation filtering and nonparametric regression[J], *ZFW*, 1: 17-23
[11] Ding Shijun, Tao Benzao (2003) Semiparametric regression and adjustment models[J]. *Journal Geodesy and Geodynamics*, 23(4): 111-114
[12] Ding Shijun, Tao Benzao (2004) Semiparametric regression model with natural spline and systematic Estimation[J]. *Journal of Wuhan University (Information Science Edition)*, 29(1): 964-967