Theory and method of hypothetical test for nonparameters in linear semiparametric model

S. Ding\textsuperscript{1,2}, J. Shi\textsuperscript{1,2} and W. Jiang\textsuperscript{1}

The linear semiparametric regression model is a combination of the linear parametric model and nonparametric model. Based on the penalised least squares theory for the semiparametric model, this paper presents a detailed discussion on the theory and method of the hypothetical test for the non-parameter in the semiparametric model. The hypothetical statistics are derived and the corresponding property is proved. The proposed theory and method of the hypothesis test are confirmed by simulated experiments.

Keywords: Semiparametric model, Penalised least squares, Nonparameter, Hypothetical test

Introduction

For decades, how to eliminate the effect of systematic errors has been a challenging topic in the field of measurement data processing. In the past, the majority of systematic errors are compensated or eliminated by mathematical models, or estimated as unknowns in the parametric model. The remaining systematic errors may be ignored. However, measurements based on modern surveying techniques often contain systematic errors which cannot be well compensated, for example, the satellite orbital errors, the receiver clock and satellite clock errors, the tropospheric and ionospheric errors and multipath errors in Global Positioning System (GPS) applications. Although these systematic errors can be largely mitigated by technologies such as differenciation, linear combinations and so on, some residual still remain. Meanwhile, these residual systematic errors cannot be modelled due to their uncertainties. In the 1980s, statistics brought forward the semiparametric regression model comprising a parametric part and a nonparametric part. Some researchers showed that the semiparametric regression model has many significant merits over parametric models in dealing with complicated relationships between observations and estimated variables, and compensating residual systematic errors.

Up to now, some researches on the algorithms for the linear semiparametric model have been made in various fields (Green and Silverman, 1994; Fischer and Hegland, 1999; Sun and Wu, 2000; Ruppert et al., 2003; Ding and Tao, 2004; Pan, 2006), and the hypothetical test for gross errors have been discussed by Eubank (1984), Ding et al. (2008) and Ding et al. (2009). However, studies on the hypothetical test for non-parameters in the linear semiparametric model are rarely reported. Therefore, this paper will study the theory and method of the hypothetical test for the non-parameter in the linear semiparametric model, and verify the proposed method and theory with simulated datasets.

Penalised least squares estimation for linear semiparametric model

The linear semiparametric model can be expressed as (Green and Silverman 1994; Fischer and Hegland 1999; Sun and Wu 2000)

\[ L = BX + s + \Delta \]  

(1)

\[ E(\Delta) = 0, \quad D(\Delta) = \sigma_{0}^{2}P^{-1} \]  

(2)

where \( L \) is an \( n \times 1 \) observation vector, \( X \) is a \( t \times 1 \) unknown parameter vector, \( s \) is a \( n \times 1 \) non-parameter vector, \( B \) is the coefficient matrix of the unknown parameter \( X \), \( \Delta \) is an independent \( n \times 1 \) noise vector conforming to the normal distribution with the mean value of zero, \( \sigma_{0}^{2} \) is the factor of unit weight variance, \( D(\Delta) \) and \( P \) are the variance and weight matrix of the noise \( \Delta \) respectively.

We can get the following error equation from equation (1)

\[ V = BX + s - L \]  

(3)

where \( V \) is the estimated residual vector of the observation \( L \). \( \hat{X} \) and \( \hat{s} \) are the estimates of the unknown parameters and non-parameters respectively.

Obviously, there are \( n \) observations and \( t+n \) unknowns in equation (1). Therefore, it cannot be resolved under the least squares principle \( V^{T}PV = \text{min} \). In other words, without any constraint on \( s \), equation (3) cannot provide useful estimates for \( X \) and \( s \). In order to place a constraint on \( s \), the minimisation...
The requirement is replaced by the penalised least squares which can be expressed as (Green and Silverman 1994; Fischer and Hegland 1999)

\[ Y^T P V + \alpha s^T R s = \min \]  

(4)

where \( R \) is called the regularisation matrix and \( \alpha \) is the regularisation or smoothing parameter. \( R \) specifies the type of function to be used as an estimate for \( s \) and \( \alpha \) governs the balance between the two quadratic terms in equation (4).

According to equations (3) and (4), we can get the following normal equation

\[
\begin{bmatrix}
B^T P B & B^T P \\
PB & P + \alpha R
\end{bmatrix}
\begin{bmatrix}
\hat{X} \\
\hat{s}
\end{bmatrix}
= \begin{bmatrix}
B^T P L \\
PL
\end{bmatrix}
\]  

(5)

The solution of equation (5) is

\[
\begin{aligned}
\hat{X} &= (B^T P (I - S) B)^{-1} B^T P (I - S) L \\
\hat{s} &= S (L - B X) \\
L &= B X + \hat{s} = H(\alpha) L \\
H(\alpha) &= S + (I - S) B B^T P (I - S) B B^T P (I - S) \\
S &= (P + \alpha R)^{-1} P
\end{aligned}
\]  

(6)

where \( S \) is called the smoothing matrix, \( I \) is an \( n \times n \) identity matrix, \( H(\alpha) \) is the non-idempotent hat matrix, \( \hat{X} \) and \( \hat{s} \) are the estimates of the unknown parameters and non-parameters, \( L \) is the adjusted value of the observation.

The smoothing parameter \( \alpha \) can be determined by methods such as the generalised cross-validation (GCV) method (Craven and Wahba 1979), the L-curve method (Hansen and O’Leary 1993; Fischer and Hegland 1999). The methods for determining the regularisation matrix \( R \) can be found in works by Green and Silverman (1994), Fischer and Hegland (1999), Sun and Wu (2000), Pan (2006) and Eubank (1984).

Let \( P_\alpha = \alpha R \), equations (3) and (4) can be rewritten as

\[
\begin{bmatrix}
V_s \\
V_a
\end{bmatrix}
= \begin{bmatrix}
B & I \\
0 & I
\end{bmatrix}
\begin{bmatrix}
\hat{X} \\
\hat{s}
\end{bmatrix}
- \begin{bmatrix}
L \\
0
\end{bmatrix}
\]  

(7)

\[
\begin{bmatrix}
V_s \\
V_a
\end{bmatrix}^T
\begin{bmatrix}
P & 0 \\
0 & P_\alpha
\end{bmatrix}
\begin{bmatrix}
V_s \\
V_a
\end{bmatrix}
= \min
\]  

(8)

With (7) and (8), we can obtain the corresponding normal equation as

\[
\begin{bmatrix}
B^T P B & B^T P \\
PB & P + \alpha R
\end{bmatrix}
\begin{bmatrix}
\hat{X} \\
\hat{s}
\end{bmatrix}
= \begin{bmatrix}
B^T P L \\
PL
\end{bmatrix}
\]  

(9)

Equation (9) is the same with equation (5), so the model using equations (3) and (4) is equivalent with the model using equations (7) and (8).

Let \( \hat{V} = \begin{bmatrix}
V \\
V_s
\end{bmatrix} \), \( \hat{V} = \begin{bmatrix}
B \\
0
\end{bmatrix} \), \( \hat{Y} = \begin{bmatrix}
\hat{X} \\
\hat{s}
\end{bmatrix} \), \( L = \begin{bmatrix}
L \\
0
\end{bmatrix} \), and

\[
\hat{P} = \begin{bmatrix}
P \\
0
\end{bmatrix}
\]  

(10)

Hypothetical test for nonparamater in semiparametric model

Define the null hypothesis as \( H_0: \alpha = 0 \). According to the theory of the linear hypothetic test (Koch 1988), the null hypothesis can be treated as the condition equation for the semiparametric model. Therefore, we can get the following equation

\[
\hat{s} = 0
\]  

(13)

which can also be written as

\[
H \hat{Y} = 0
\]  

(14)

where \( H = [0 \ I] \). Equations (10) and (14) indicate the indirect adjustment with the condition. If \( \hat{Y}_H \) is the solution of equations (10) and (14), it can be expressed as (Yu and Tao 1993; Tao 2007)

\[
\hat{Y}_H = \hat{Y} - N_\alpha^{-1} H^T (H N_\alpha^{-1} H^T)^{-1} H \hat{Y}
\]  

(15)

The corresponding sum of squares of the residuals can be calculated as

\[
\Omega_H = \hat{V}_H^T \hat{P} \hat{V}_H = (\hat{B} \hat{Y}_H - \hat{L}) \hat{P} (\hat{B} \hat{Y}_H - \hat{L})
\]

\[
= (\hat{B} \hat{Y} - \hat{L}) \hat{P} (\hat{B} \hat{Y} - \hat{L}) + \hat{Y} - \hat{Y}_H)^T N_\alpha (\hat{Y} - \hat{Y}_H)
\]

\[
= \hat{V}^T \hat{P} \hat{V} + (\hat{Y} - \hat{Y}_H)^T N_\alpha (\hat{Y} - \hat{Y}_H)
\]  

(16)

Let \( \hat{\Omega} = \hat{V}^T \hat{P} \hat{V}, \hat{\Omega}_2 = (\hat{Y} - \hat{Y}_H)^T N_\alpha (\hat{Y} - \hat{Y}_H) \), equation (16) can be simplified as

\[
\hat{\Omega}_1 = \hat{\Omega}_1 + \hat{\Omega}_2
\]  

(17)

Substituting equation (15) into the equation to calculate \( \hat{\Omega}_2 \), the following equation can be obtained

\[
\hat{\Omega}_2 = N_\alpha^{-1} H^T (H N_\alpha^{-1} H^T)^{-1} H \hat{Y}^T
\]

\[
N_\alpha [N_\alpha^{-1} H^T (H N_\alpha^{-1} H^T)^{-1} H \hat{Y}]
\]

\[
(\hat{H} \hat{Y})^T (H N_\alpha^{-1} H^T)^{-1} (\hat{H} \hat{Y})
\]  

(18)

From equation (18), we can find that the adjusted results of equations (10) and (11) is

\[
\frac{\hat{\Omega}_1}{\sigma_0^2} \sim \chi^2(f_1)
\]  

(19)

where \( f_1 = n - \text{tr}[H(\alpha)] \) is the degree of freedom of the \( \chi^2(f_1) \)-distribution with \( \text{tr}[H(\alpha)] \) as the trace of the hat matrix \( H(\alpha) \), the variance \( \sigma_0^2 \) is the a-prior variance factor. The estimate \( \sigma_0^2 \) can be obtained by
\( \sigma^2 = \frac{\tilde{V}^T \tilde{P} \tilde{V}}{n - \text{tr}[H(z)]} \) 

(20)

In the next section, the statistic \( \Omega_2/\sigma_0^2 \) will be proved to obey the \( \chi^2(f_2) \) distribution with the degree of freedom \( f_2 = R(H) \), and \( R(H) \) is the rank of the matrix \( H \). If the null hypothesis \( H_0 \) is accepted, \( \Omega_1 \) and \( \Omega_2 \) will be proved as independent statistics. We can further form the \( F \) statistic term as

\[
F = \frac{\Omega_2/f_2}{\Omega_1/f_1} \sim F(f_2, f_1)
\]

(21)

If null hypothesis \( H_0 \) holds, the statistic \( F \) obeys the \( F \) distribution with the degrees of freedom \( f_2 \) and \( f_1 \), and the statistic \( F \) must be less than the critical value \( F_{\alpha} \) with the definite confidence level \( \alpha \). If \( F > F_{\alpha} \), the null hypothesis \( H_0 \) is rejected. In other word, we can obtain \( s \neq 0 \). The semiparametric model (3) is right and its optimal solution is determined by (6) or (12).

**Proof of statistic \( f \)**

\( \chi^2(f_2) \) - distribution of statistic \( \Omega_2/\sigma_0^2 \)

**Theorem 1:** Assume that \( X \) obeys the normal distribution with the mean value \( \xi \) and the finite variance \( D(X) \). \( M \) is the symmetric matrix, and \( MD(X) \) is an idempotent matrix. The quadratic form \( X^T MX \) obeys the decentralised \( \chi^2 \) distribution

\[
X^T MX \sim \chi^2[R(M), \lambda]
\]

(22)

where \( R(M) \) is the rank of the matrix \( M \), and the decentralise parameter is \( \lambda = \xi^T M \xi \).

Given equation (18), we can get

\[
\frac{\Omega_2}{\sigma_0^2} = (H \tilde{Y})^T (H N^{-1} \tilde{H}^T)^{-1} (H \tilde{Y})
\]

(23)

where \( H \tilde{Y} \) is a quadratic vector, the corresponding variance is \( \sigma_0^2 (HN^{-1} \tilde{H}^T) \), and \( (HN^{-1} \tilde{H}^T)^{-1} \) is the matrix of quadratic form. Apparently, we have

\[
\frac{(HN^{-1} \tilde{H}^T)^{-1}}{\sigma_0^2} = (HN^{-1} \tilde{H}^T)^{-1} / \sigma_0^2 \quad \text{and} \quad (HN^{-1} \tilde{H}^T)^{-1} = \frac{[H N^{-1} \tilde{H}^T]^{-1}}{\sigma_0^2}
\]

(24)

According to the theorem 1, we can get

\[
\frac{\Omega_2}{\sigma_0^2} \sim \chi^2[R(M), \lambda]
\]

(25)

The degrees of freedom and decentralised parameter are

\[
f_2 = R(M) = R \left[ \frac{(HN^{-1} \tilde{H}^T)^{-1}}{\sigma_0^2} \right] = R(H N^{-1} \tilde{H}^T) = R(H)
\]

(26)

\[
\lambda = \xi^T M \xi = \frac{1}{\sigma_0^2} (H \tilde{Y})^T (H N^{-1} \tilde{H}^T)^{-1} (H \tilde{Y})
\]

(27)

When the null hypothesis \( H_0 \) holds, \( H \tilde{Y} = 0 \). Therefore, \( \lambda = 0 \). According to the above proof, we can get the following statistics with the centralised \( \chi^2 \) distribution

\[
\frac{\Omega_2}{\sigma_0^2} \sim \chi^2[R(H)]
\]

(28)

**Proof of independence between \( \Omega_1 \) and \( \Omega_2 \)**

**Theorem 2:** Random vector \( X \) obeys the normal distribution with the mean value \( \xi \) and the finite variance \( D(X) \), namely \( X \sim N(\xi, D(X)) \). If \( M \) and \( N \) are the symmetrical reversible matrix, and \( MD(X)N = 0 \) or \( ND(X)M = 0 \), then the quadratic forms \( X^T MX \) and \( X^T NX \) are mutually independent.

The sum of squares of residuals is obtained by equations (10) and (12)

\[
\Omega_1 = \tilde{V}^T \tilde{P} \tilde{V} = (\tilde{B} \tilde{N}^{-1} \tilde{B}^T \tilde{P} \tilde{L} - \tilde{L})^T \tilde{P} (\tilde{B} \tilde{N}^{-1} \tilde{B}^T \tilde{P} \tilde{L} - \tilde{L})
\]

(29)

Obviously, \( \Omega_1 \) as the quadratic form of random vector \( \tilde{L} \), obeys the normal distribution with the mean value \( \bar{\xi} \) and the variance \( D(\tilde{L}) \). Substituting equation (12) into equation (18), we can get

\[
\Omega_2 = \tilde{L}^T (\tilde{P} - \tilde{P} \tilde{B} \tilde{N}^{-1} \tilde{B}^T \tilde{P}) \tilde{L}
\]

(30)

So, \( \Omega_2 \) is also the quadratic form of the random vector \( \tilde{L} \) as

\[
\tilde{P} \tilde{B} \tilde{N}^{-1} \tilde{B}^T (\tilde{H}\tilde{N}^{-1} \tilde{H})^{-1} \tilde{H}\tilde{N}^{-1} \tilde{B}^T \tilde{P} \tilde{L} = 0
\]

(31)

we arrive at the conclusion that two quadratic forms \( \Omega_1 \) and \( \Omega_2 \) are mutually independent under the theorem 2.

**Experiment and analysis**

Example: assume that two observations are \( L_1 = BY + \Delta \) and \( L_2 = BX + s + \Delta \). \( X = [25]^T \) and \( s = [s(t_1), s(t_2), \ldots, s(t_{100})]^T \), where \( s(t_i) = 3 + 5 \sin(2t_i) \) and \( t_i = 2\pi(t_i - 1) / 100 \), \( i = 1, 2, \ldots, 100 \). \( A \sim N(0, \sigma^2)^{-1} \) and \( \sigma^2 = 1 \).

In the above two simulated observations, the observation \( L_1 \) does not contain the non-parameter \( s(t) \), and the observation \( L_2 \) contained the non-parameter \( s(t) \). In this case, the estimated value will be obtained using the parametric and semiparametric models respectively. The statistical test for the non-parameter will be experimented in the semiparametric model. The example is chosen to examine the difference of the estimates between the parametric and semiparametric models, and to test the significance of the non-parameter in the semiparametric model. In order to compare the effect between the standard least squares estimation of the parametric model and the penalised least squares estimation of the semiparametric model, the mean square error (MSE) is introduced as

\[
\text{MSE}(\hat{X}_{LS}) = \frac{1}{n} \text{tr}[J^T J]
\]

(32)
Choosing parametric model $L_1 = \mathbf{X} + \Delta$

The results were shown in Table 1 using the standard least squares estimation. The estimated values are $\hat{X}_1 = 1.918$ and $\hat{X}_2 = 5.058$, the estimate of the unit weight variance is $\hat{\sigma}_0 = 0.962$, and $\text{MSE}(\hat{\mathbf{X}}_{\text{PLS}}) = 0.018$. Figure 1 shows that the residuals of the observations are almost same with the simulated noises.

Choosing semiparametric model $L_1 = B\mathbf{X} + s + \Delta$

The regularised parameter $\alpha$ is determined by the generalised cross-validation (GCV) (Green and Silverman 1994)

$$GCV(\alpha) = \frac{\text{tr}[\mathbf{J}(\alpha \mathbf{R} + \mathbf{M})^{-1}]}{1 - \text{tr}[\mathbf{H}(\alpha)/n]}$$

According to equations (12), (32) and (33), the minimum value of $GCV(\alpha)$ is 91.556 with $\alpha = 94.34$ (see Fig. 2). The estimates of the parameters $\hat{X}_1$ and $\hat{X}_2$ are 1.975 and 5.067, the estimate $\hat{\sigma}_0$ of the unit weight variance is 0.927. For the significance test of the systematic parameter $s$, the statistic $F$ calculated by equation (21) is 0.073. The critical value $F_{0.05}$ is equal to 1.30, which is obtained by the look-up table of the $F$-distribution with the confidence level 0.05 and degrees of freedom of 100 and 94. As $F < F_{0.05}$, then, the null hypothesis $H_0: s = 0$ is accepted. Figure 1 shows that residuals of the observations in the semiparametric model are almost same as the simulated noises. Figure 3 indicates that the estimated systematic error is close to zero. Table 1 shows that the norm $||\mathbf{X} - \hat{\mathbf{X}}||$ based on the standard least squares of the parametric model is greater than the semiparametric model.

Table 1 Results of two models for the observation $L_1$

| Models                | $\hat{X}_1$ | $\hat{X}_2$ | $\hat{\sigma}_0$ | $||\mathbf{X} - \hat{\mathbf{X}}||$ | $||\sigma_0 - \hat{\sigma}_0||$ | $\text{MSE}(\mathbf{X})$ |
|-----------------------|-------------|-------------|-------------------|----------------------------------|---------------------------------|--------------------------|
| Parametric model      | 1.918       | 5.058       | 0.960             | 0.100                            | 0.04                            | 0.018                    |
| Semiparametric model  | 1.975       | 5.067       | 0.927             | 0.072                            | 0.73                            | 0.064                    |
than that obtained by the penalised least squares of the semiparametric model and the penalised least square. On the other hand, MSE(PLS) which indicates that the standard least squares estimation of the parametric model is better than the penalised least squares of the semiparametric model when the observations contain no systematic errors.

Setting parametric model \( L_2 = BX + \Delta \)

The estimated parameters \( \hat{X}_1 \) and \( \hat{X}_2 \) are 4.547 and 3.281 respectively, and the estimate \( \hat{\sigma}_0 \) of the unit weight variance is 3.622 based on the standard least squares estimation. Figure 4 shows that the residuals of the observation \( L_2 \) contain significant systematic errors. In this case, the estimates are much greater than the actual values.

Choosing semiparametric model \( L_2 = BX + s + \Delta \)

The computing procedures are the same with step 2. The minimum value of \( GCV(\alpha) \) is 109.47 with the smoothing parameter \( \alpha = 2.118 \) (see Fig. 5). The estimates \( \hat{X}_1, \hat{X}_2 \) and \( \hat{\sigma}_0 \) are 1.382, 5.321 and 0.856 respectively. For the significance test of the systematic parameter \( s \), the calculated value of the statistic \( F \) is 8.465. The critical value is 1.60 with the confidence level of 0.95 and the degrees of freedom of 100 and 67 respectively. So, \( F > F_{0.05} \). Then, the null hypothesis \( H_0 \) is rejected, and \( s \neq 0 \) is accepted. Fig. 4 shows the residuals of the observations by the semiparametric model are approximate to the simulated noises. Figure 6 shows that the systematic errors are significant. Table 2 shows that the norm \( \|X - \hat{X}\| \) based on the parametric model is greater than that obtained by the semiparametric model. The same phenomena can also be detected for the norm \( \|\sigma_0 - \hat{\sigma}_0\| \) and MSE. Therefore, the penalised least squares of the semiparametric model is better than the standard least squares estimation of the parametric model when the observations contain the systematic errors.

Conclusions

For the survey data processing, the choice of the parametric or semiparametric model depends on the significance of the systematic error. For the significance test of the systematic error, the mathematical expectation and the distribution of the observation residuals are often used to test whether the model contains systematic errors in the parametric regression (Ding et al. 2009; Tao 2007). Based on the principle of significance test of parameters in the parametric regression, this paper presented the method of the hypothetical test for the nonparametric parameter in the semiparametric model. The proposed theory and method of the hypothesis test have been confirmed by simulated experiments.

Table 2 Results of two models for observation \( L_2 \)

<table>
<thead>
<tr>
<th>Models</th>
<th>( \hat{X}_1 )</th>
<th>( \hat{X}_2 )</th>
<th>( \hat{\sigma}_0 )</th>
<th>( |X - \hat{X}| )</th>
<th>( |\sigma_0 - \hat{\sigma}_0| )</th>
<th>MSE(X)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parametric model</td>
<td>4.547</td>
<td>3.281</td>
<td>3.622</td>
<td>3.073</td>
<td>2.622</td>
<td>0.265</td>
</tr>
<tr>
<td>Semiparametric model</td>
<td>1.382</td>
<td>5.321</td>
<td>0.856</td>
<td>0.696</td>
<td>0.144</td>
<td>0.192</td>
</tr>
</tbody>
</table>
The simulated experiment and analysis show that the performance of the least squares estimation in the linear parameter model depends on the existence of the systematic error. When the observations contain no systematic errors, the least squares estimation of the parametric model is superior to the penalised least squares estimation of the semiparametric model. In this case, the semiparametric estimation is biased. When systematic errors occur to the observations, the parametric estimation will lead to large deviations. The semiparametric estimations based on the penalised least squares are better than the parametric estimation due to the non-parameter estimate for the consideration of the systematic errors.

Acknowledgements

We would like to thank Professor Benzao Tao for several helpful discussions during the preparation of this manuscript, and are grateful to the support of the National Nature Science Foundation of China (no. 41174009). The editor in charge and anonymous reviewers are also acknowledged for improving the manuscript.

References


