Korean J. Comp. & Appl. Math. Vol. **4** (1997), No. 2 , pp. 441 - 452

DETECTION OF OUTLIERS IN WEIGHTED LEAST SQUARES REGRESSION

Bang Yong Sohn and Guk Boh Kim

Abstract. In multiple linear regression model, we have presupposed assumptions (independence, normality, variance homogeneity and so on) on error term. When case weights are given because of variance heterogeneity, we can estimate efficiently regression parameter using weighted least squares estimator. Unfortunately, this estimator is sensitive to outliers like ordinary least squares estimator. Thus, in this paper, we proposed some statistics for detection of outliers in weighted least squares regression.

AMS Mathematics Subject Classification : 62J20, 62J05
Key words and phrases : Studentized residuals, high leverage point, outliers, weighted least squares regression.

1. Introduction

Model criticism is called model diagnostics and when we refer to a regression model, the model criticism is called regression diagnostics. Regression diagnostics consists of mainly detection of outliers, assessment of influence and examination of muticollinearity. We consider a mutiple linear regression model:

$$Y = X\beta + \epsilon \tag{1.1}$$

where Y is an $n \times 1$ vector of dependent observations, X is an $n \times p'$ full column rank matrix of known explanatory variables possibly including

Received November 25, 1996. Revised February 4, 1997
This work was supported by the Dae-Jin University Research Grants in 1996.

a constant predictor, β is a $p' \times 1$ vector of unknown parameters to be estimated, and error term ϵ is an $n \times 1$ vector of independent random errors with zero mean and unknown variance σ^2.

In fitting the multiple linear regression model (1.1) by the method of ordinary least squares (OLS), Cook and Weisberg [1982, Chapter 2] explained the method of detection of outliers in the y-direction through analysis of residuals and various plotting methods. Especially, when there are several outliers, the usual identification method does not always find them, because it is based on the sample mean and covariance matrix, which are themselves affected by the outliers. The OLS approach masks outliers in a similar way. To avoid the masking effect, Rousseeuw and Zomeren (1990) proposed that outliers may be unmasked by using a highly robust regression method and classified the data into regular observations, vertical outliers, good leverage points, and bad leverage points.

2. Weighted Least Squares Regression

In this section we briefly review the theory of weighted least square (WLS). The weighted least square model is given by

$$Y = X\beta + \epsilon, \tag{2.1}$$

where all quantities are the same as defined in (1.1), except that $Var(\epsilon) = \sigma^2 D_w^{-1}$ and D_w is a known $n \times n$ diagonal matrix with $w_i > 0$. The w_i are often called case weights.

In fitting the multiple linear regression model (2.1) by the method of WLS, the WLS estimator $\hat{\beta}_w$ is obtained by minimizing $\sum w_i(y_i - x_i^T\beta)^2$, that is,

$$\sum w_i(y_i - c\hat{\beta}_w)^2 = \min_{\beta} \sum w_i(y_i - x_i^T\beta)^2 \tag{2.2}$$

$$= \min_{\beta}(Y - X\beta)^T D_w (Y - X\beta), \tag{2.3}$$

where x_i^T is the row vector of explanatory of ith case. Therefore, we have

$$\hat{\beta}_w = (X^T D_w X)^{-1} X^T D_w Y. \tag{2.4}$$

The WLS theory provides us the following result:

(1) The variance of $\hat{\beta}_w$ is

$$\mathrm{Var}(\hat{\beta}_w) = \sigma^2 (X^T D_w X)^{-1}. \tag{2.5}$$

(2) The $n \times 1$ vector of weighted fitted values is

$$\begin{aligned} \hat{Y}_w = X\hat{\beta}_w &= X(X^T D_w X)^{-1} X^T D_w Y \\ &= H_w D_w Y, \end{aligned} \tag{2.6}$$

where

$$H_w = X(X^T D_w X)^{-1} X^T. \tag{2.7}$$

This matrix is symmetric ($H_w^T = H_w$) and idempotent ($H_w D_w H_w = H_w$) in a weighted Euclidean space. The variance of $\hat{Y}_w$ is

$$\begin{aligned} \mathrm{Var}(\hat{Y}_w) &= \mathrm{Var}(H_w D_w Y) \\ &= H_w D_w \mathrm{Var}(Y) D_w H_w \\ &= \sigma^2 H_w D_w H_w \\ &= \sigma^2 H_w. \end{aligned} \tag{2.8}$$

(3) The $n \times 1$ vector of residuals is

$$e_w = Y - \hat{Y}_w = (D_w^{-1} - H_w) D_w Y, \tag{2.9}$$

and we define e_w as the weighted residual. The variance of e_w is

$$\begin{aligned} \mathrm{Var}(e_w) &= \mathrm{Var}\{(D_w^{-1} - H_w) D_w Y\} \\ &= \sigma^2 (D_w^{-1} - H_w) D_w D_w^{-1} D_w (D_w^{-1} - H_w) \\ &= \sigma^2 (D_w^{-1} - 2H_w + H_w D_w H_w) \\ &= \sigma^2 (D_w^{-1} - H_w). \end{aligned} \tag{2.10}$$

(4) An unbiased estimator of σ^2 is

$$\hat{\sigma}^2 = \frac{e_w^T D_w e_w}{n - p'}, \tag{2.11}$$

since

$$\begin{aligned}E(e_w^T D_w e_w) &= E\cdot \mathrm{tr}(e_w^T D_w e_w)\\ &= \mathrm{tr}\{D_w E(e_w e_w^T)\}\\ &= \sigma^2 \cdot \mathrm{tr}\{D_w(D_w^{-1} - H_w)\}\\ &= \sigma^2(n-p').\end{aligned}$$

3. Detection of outliers

In WLS regression, we can detect outlier in the y-direction by using the Studentized weighted residuals and by displaying plots of Studentized weighted residuals versus $\hat{Y}$ or X. In OLS regression, the studentized residual has been studied by Srikantan (1961), and Ellenberg (1973) provides the joint distribution of a set of studentized residuals, assuming that (1.1) holds. Also, like OLS regression, H_w identifies high leverage design points (Chatterjee and Hadi, 1986) in WLS regression. In this section, we discuss these weighted residuals and the role of weighted hat matrix.

Theorem 1. *Let $\hat{\beta}_{w(i)}$ denote the WLS estimator computed without the ith observation. Then*

$$\hat{\beta}_{w(i)} = \hat{\beta}_w - \frac{w_i(X^T D_w X)^{-1} x_i e_{w,i}}{1 - w_i h_{w,ii}}, \tag{3.1}$$

where $e_{w,i} = y_i - x_i^T \hat{\beta}_w$ and $h_{w,ii} = x_i^T (X^T D_w X)^{-1} x_i$.

The proof is omitted to save the space. Readers may refer to author's doctoral thesis [Sohn 1994, pp. 31-32].

Theorem 2. *Let $\hat{\sigma}^2_{(i)}$ denote the weighted mean square when the ith observation is omitted. Then*

$$\hat{\sigma}^2_{(i)} = \frac{(N-p')\hat{\sigma}^2 - w_i e^2_{w,i}/(1 - w_i h_{w,ii})}{n - p' - 1}. \tag{3.5}$$

Readers may refer to Sohn [1994, pp. 32-33] forn the proof.

Studentized weighted residuals can be defined as follows:

(1) Internally Studentized weighted residual is defined as

$$t_{w,i} = \frac{e_{w,i}}{\hat{\sigma}\sqrt{w_i^{-1} - h_{w,ii}}} \tag{3.9}$$

$$= \frac{\sqrt{w_i}e_{w,i}}{\hat{\sigma}\sqrt{1 - w_i h_{w,ii}}} \tag{3.9a}$$

where $e_{w,i}$ is the element of the $n \times 1$ vector, $h_{w,ii}$ is the ith diagonal element of H_w (cf. 2.7) and $\hat{\sigma}^2$ is defined as in (2.11).

(2) Externally Studentized residual is defined as

$$t^*_{w,i} = \frac{e_{w,i}}{\hat{\sigma}_{(i)}\sqrt{w_i^{-1} - h_{w,ii}}} \tag{3.10}$$

$$= \frac{\sqrt{w_i}e_{w,i}}{\hat{\sigma}_{(i)}\sqrt{1 - w_i h_{w,ii}}}, \tag{3.10a}$$

where $\hat{\sigma}^2_{(i)}$ is the weighted mean squares computed without ith observation.

An outlier in the response-factor space is a point (x_i^T, y_i) with large $t_{w,i}$ or $t^*_{w,i}$. Outlier are usually detected by plotting $t_{w,i}$ and $t^*_{w,i}$ versus other quantities such as $\hat{Y}$ or each column of X (i.e. X_j).

(3) The scalar form of weighted fitted value $\hat{Y}_w$, seen as in (2.6) is

$$\begin{aligned}\hat{y}_{w,i} &= x_i^T(X^TD_wX)^{-1}X^TD_wY \\ &= \sum_{j=1}^{n} x_i^T(X^TD_wX)^{-1}x_jw_jy_j \\ &= \sum_{j=1}^{n} w_jh_{w,ij}y_j \qquad (3.11) \\ &= w_ih_{w,ii}y_i + \sum_{\substack{j=1 \\ j\neq i}}^{n} w_jh_{w,ij}y_j. \qquad (3.12)\end{aligned}$$

The elements of H_w, especially the diagonal element $h_{w,ii}$, play an important role in the technique of WLS regression diagnostics, which aim at discovering whether individual observations have unusually great influence on the weighted fitted regression model. To illustrate the interpretation of $h_{w,ii}$, we examine how the weighted fitted value $\hat{y}_{w,i}$ changes when y_i varies. If we add an increment Δy_i to y_i, then y_i become, from equation (3.11),

$$\hat{y}_{w,i} + \Delta\hat{y}_{w,i} = \sum_{j=1}^{n} w_j h_{w,ij} y_j + w_i h_{w,ii} \Delta y_i. \tag{3.13}$$

Thus

$$\Delta\hat{y}_{w,i} = w_i h_{w,ii} \Delta y_i. \tag{3.14}$$

We see that the impact on $\hat{y}_{w,i}$ by the change in y_i is that change multiplied by w_i and $h_{w,ii}$. Therefore, we can interpret $w_i h_{w,ii}$ as the amount of leverage of the response value y_i on the corresponding value $\hat{y}_{w,i}$ by y_i. Hoaglin and Welsch (1978) suggested a direct use of the diagnoal elements of the hat matrix as a diagnostic to idetify high leverage points (outliers in the x-direction).

4. A Numerical Example: Stack Loss Data

We will give a numerical example with Brownlee's stack loss data [Atkinson 1986] listed in Table 1, using various diagnostic methods in section 2 and 3 in order to detect outliers in robust regression based on M-estimator [Li 1985]. Because we can obtain the M-estimator $\hat{\beta}_M$ by using internal weights, like WLS estimator $\hat{\beta}_w$. The fully iterative estimator $\hat{\beta}_M$ is written as

$$\hat{\beta}_M = \hat{\beta}_w = (X^T D_w X)^{-1} X^T D_w Y. \tag{4.1}$$

Although the $\hat{\beta}_{w(i)}$ in (3.4) is not equivalent to $\hat{\beta}_{M(i)}$, we hope the difference of these estimator is insignificant. If so, we can substitute $\hat{\beta}_{w(i)}$ for $\hat{\beta}_{M(i)}$ for the practical purposes. Therefore, using the updating formula, the $\hat{\beta}_{w(i)}$ yields

$$\hat{\beta}_{w(i)} = \hat{\beta}_w - \frac{w_i (X^T D_w X) x_i e_{w,i}}{1 - w_i h_{w,ii}}, \tag{4.2}$$

where $D_w = diag(w_1, \dots, w_n)$ and

$$w_i = \frac{\psi'\{(y_i - x_i^T \hat{\beta}_M)/\tilde{\sigma}\}}{\{(y_i - x_i^T \hat{\beta}_M)/\tilde{\sigma}\}}.$$

The stack loss data is obtained from 21 days of operation of a chemical plant that oxidizs ammonia $[NH_3]$ to nitric acid $[HNO_3]$. This data consists of three explanatory variables; X_1=air flow, X_2=temperature of the cooling water in the coils of the absorbing tower, X_3=concentration of nitric acid in the absorbing liquid, and the response observation Y=percent of the ingoing ammonia that is lost by escaping in the absorbed nitric oxides.

Consider the multiple linear regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon. \tag{4.3}$$

By OLS fitting to the model, the regression equation is given as

$$\hat{Y} = -39.9197 + 0.7156X_1 + 1.2953X_2 - 0.1521X_3, \tag{4.4}$$

and it is very sensitive to outliers and high leverage design points. Therefore, to try robust fitting to the model (4.3), let us use Huber's ψ-function (with tunning constants $k = 2.0$),

$$\psi(t) = \begin{cases} t, & \text{for } |t| \le k \\ k\,\text{sign}(t), & \text{for } |t| > k. \end{cases} \tag{4.5}$$

If we rely on the median of absolute residuals (from the least squares fit) $\tilde{\sigma} = 0.97$ for a resistant estimate of σ, then the robust regression equation can be obtained as

$$\hat{Y} = -39.9898 + 0.8286X_1 + 0.7638X_2 - 0.1089X_3. \tag{4.6}$$

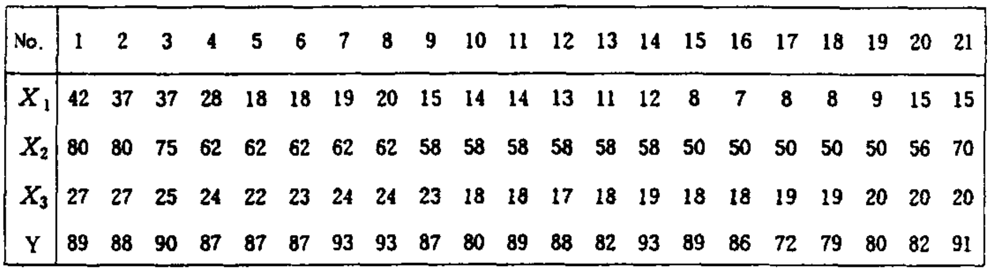
Internal weights and various residuals are listed in Table 2, and we detect observations 1, 3, 4, and 21 as being outliers.

An index plot of OLS residuals versus case number given in Figure 1 (a) shows clearly that these observations are outliers. Figure 1 (b), Figure 1 (c), and Figure 1 (d) are index plots of the robust residuals, internally Studentized weighted residuals, and externally Studentized weighted residuals, respectively, against case number based on Huber's ψ-function. From these plots, observations 1, 3, 4, and 21 can be identfied as outliers. Figure 2 is plots of $w_i h_{ii}$ against case number. Observation 2 is high leverage design point ($w_2 h_{22} = 0.4656$).

5. Conclusion

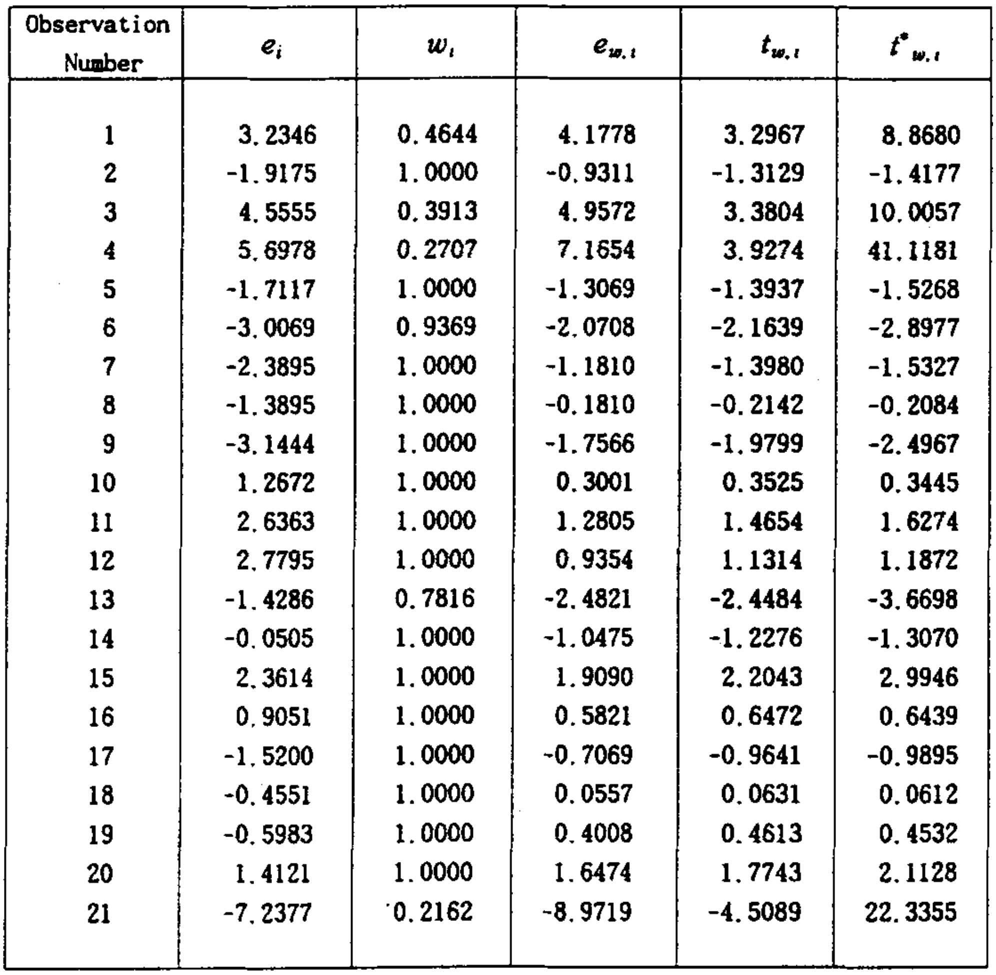
The WLS esmators are sensitive to outliers and high leverage points. To solve these problems, we proposed studentized weighted residuals and weighted hat matrix. Also, we applied WLS regression diagnostics for identfying outliers and leverage points to the robust regression problem. In WLS regression, we belived that some our statistics is practical and effective in identifying outliers and high leverage points.

Table 6.1 Stack Loss Data



No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
X_1	42	37	37	28	18	18	19	20	15	14	14	13	11	12	8	7	8	8	9	15	15
X_2	80	80	75	62	62	62	62	62	58	58	58	58	58	58	50	50	50	50	50	56	70
X_3	27	27	25	24	22	23	24	24	23	18	18	17	18	19	18	18	19	19	20	20	20
Y	89	88	90	87	87	87	93	93	87	80	89	88	82	93	89	86	72	79	80	82	91

Table 2. Internal weights, OLS residuauals, robust residuals, two weighted studentized Residuals



Observation Number	e_i	w_i	$e_{w,i}$	$t_{w,i}$	$t^*_{w,i}$
1	3.2346	0.4644	4.1778	3.2967	8.8680
2	-1.9175	1.0000	-0.9311	-1.3129	-1.4177
3	4.5555	0.3913	4.9572	3.3804	10.0057
4	5.6978	0.2707	7.1654	3.9274	41.1181
5	-1.7117	1.0000	-1.3069	-1.3937	-1.5268
6	-3.0069	0.9369	-2.0708	-2.1639	-2.8977
7	-2.3895	1.0000	-1.1810	-1.3980	-1.5327
8	-1.3895	1.0000	-0.1810	-0.2142	-0.2084
9	-3.1444	1.0000	-1.7566	-1.9799	-2.4967
10	1.2672	1.0000	0.3001	0.3525	0.3445
11	2.6363	1.0000	1.2805	1.4654	1.6274
12	2.7795	1.0000	0.9354	1.1314	1.1872
13	-1.4286	0.7816	-2.4821	-2.4484	-3.6698
14	-0.0505	1.0000	-1.0475	-1.2276	-1.3070
15	2.3614	1.0000	1.9090	2.2043	2.9946
16	0.9051	1.0000	0.5821	0.6472	0.6439
17	-1.5200	1.0000	-0.7069	-0.9641	-0.9895
18	-0.4551	1.0000	0.0557	0.0631	0.0612
19	-0.5983	1.0000	0.4008	0.4613	0.4532
20	1.4121	1.0000	1.6474	1.7743	2.1128
21	-7.2377	0.2162	-8.9719	-4.5089	22.3355

Figure 1. Index Plot of various Residuals: (a) OLS Residuals, (b) Robust Residuals, (c) Internally Studentized Residuals (d) Externally Studentized Residuals

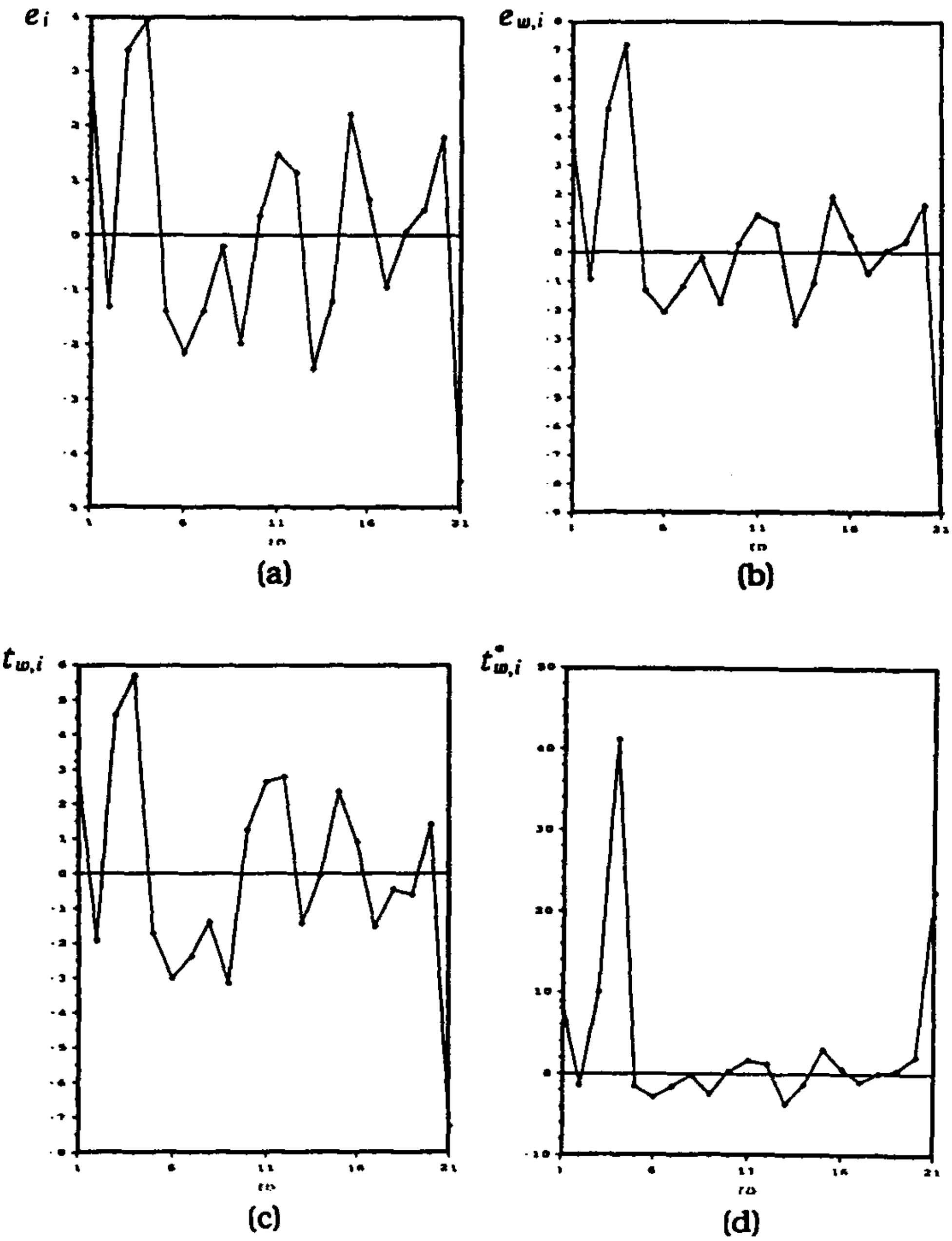
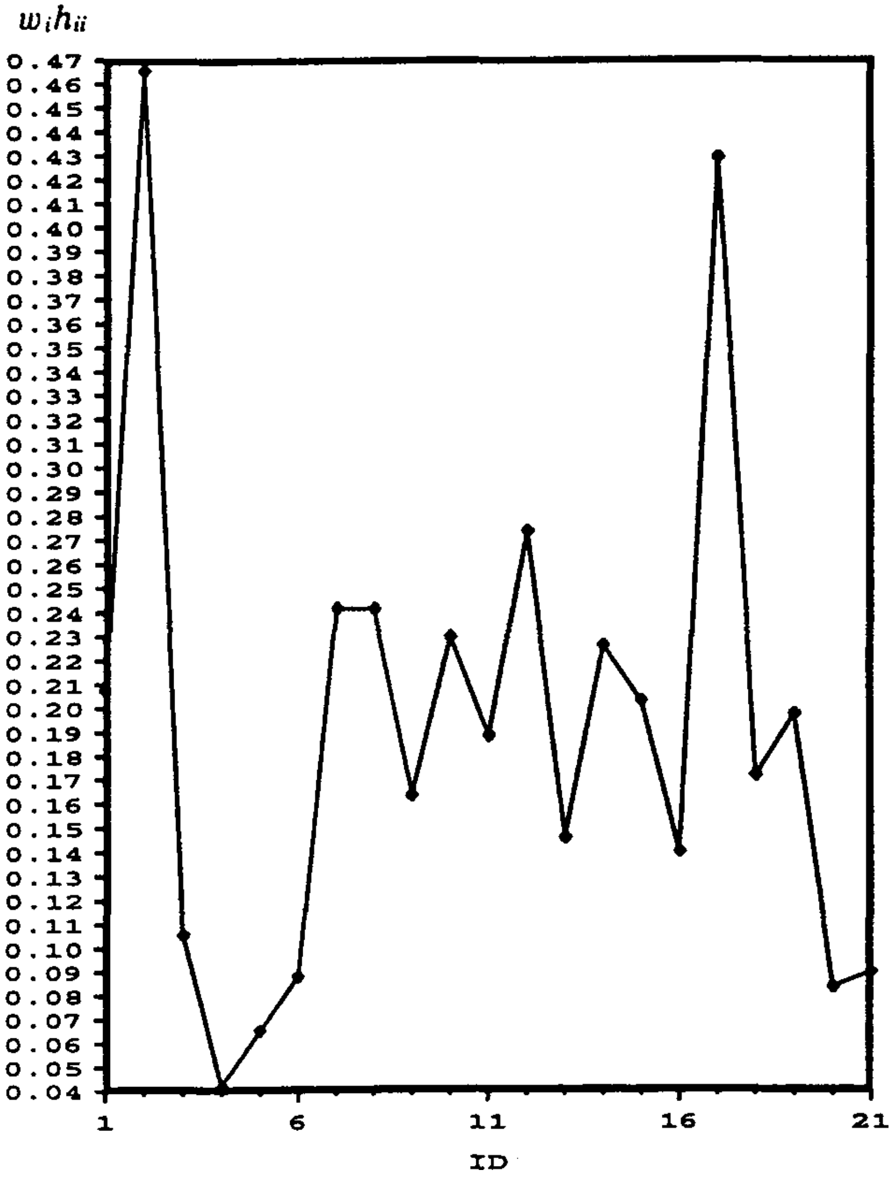


Figure 2 Index Plot of $w_i h_{ii}$



References

1. Atkinson, A.C., *Regression diagnostics, transformations and constructed variables*, Journal of Royal Statistical Society, B **44** (1982), 1–36.
2. Chatterjee, S. and Hadi, A.S., *Influential observations, high leverage points, and outlier in linear regression*, Statistical Science **1** (1986), 379–416.
3. Cook, R.D. and Weisberg, S., *Residuals and Influence in Regression*, Chapman and Hall, New York, 1982.
4. Ellenberg, J.H., *The joint distribution of the studentized least squares residuals from a general linear regression*, Journal of American Statistical Association **68** (1973), 941–943.
5. Hoaglin, D.C. and Welsch, R., *The hat matrix in regression and ANOVA*, American Statistician **32** (1978), 17–22.
6. Li, G., *Robust regression in Exploring Data Tables, Trends and Shapes (edited by D.C. Hoaglin, F. Mosteller, and J.W. Tukey)*, Wiley, New York, 1985.
7. Rousseeuw, P.J. and Zomeren, B.C., *Unmasking multivariate outliers and leverage points*, Journal of American Statistical Association **85** (1990), 623–651.
8. Sohn, B.Y., *Weighted Least Squares Regression Diagnostics and its Application to Robust Regression*, Doctoral Thesis, Dept. of Statistics Korea University, 1994.
9. Srikantan, K.S., *Testing for a single outlier in a regression model*, Sankhy $\overline{a}$ A **23** (1961), 251–260.

Bang Yong Sohn and Guk Boh Kim
Department of Computer Engineering
Daejin University
Kyunggi-Do, 487-800, Korea
`e-mail : bysohn@road.daejin.ac.kr`