Journal of the Korean
Statistical Society
Vol. 20, No. 2, 1991

On Sensitivity Analysis in Principal Component Regression

Soon-Kwi Kim* and Sung H. Park**

ABSTRACT

In this paper, we discuss and review various measures which have been presented for studying outliers, high-leverage points, and influential observations when principal component regression is adopted. We suggest several diagnostics measures when principal component regression is used.

A numerical example is illustrated. Some individual data points may be flagged as outliers, high-leverage point, or influential points.

1. Introduction

Consider the ordinary linear regression model

$$\underline{y}=X\underline{\beta}+\underline{\varepsilon}, \tag{1.1}$$

where

$\underline{y}$ is an $n\times 1$ observation vector of dependent variable ;
X is an $n\times p$ $(n>p)$ full rank matrix of independent variables ;
$\underline{\beta}$ is a $p\times 1$ vector of unknown coefficients ; and
$\underline{\varepsilon}$ is an $n\times 1$ vector of error terms.

In addition, we assume that the independent variables are linearly transformed so that $X'X$ is the correlation matrix of the independent variables.

The values of principal components(PCs) for each observation are given by

$$Z=XP$$

where the (i,k)th element of Z is the value of the kth PC for the ith observation, and P is a $p\times p$ matrix whose kth column is the kth eigenvector of $X'X$.

Because P is orthogonal, $X\underline{\beta}$ can be rewritten as $XPP'\underline{\beta}=Z\underline{\alpha}$, where $\underline{\alpha}=P'\underline{\beta}$. Equatin (1.1)

* Dept. of Statistics, Kangnung University.
** Dept. of Computer Science and Statistics, Seoul National University

can therefore be written as

$$\underline{y}=Z\underline{\alpha}+\underline{\varepsilon}. \tag{1.2}$$

Principal component regression(PCR) uses the model(1.2) or the reduced model

$$\underline{y}=Z_g\underline{\alpha}_g+\underline{\varepsilon}_g \tag{1.3}$$

where $\underline{\alpha}_g$ is a $g\times 1$ vector which is a subset of elements of $\underline{\alpha}$, Z_g is an $n\times g$ matrix whose columns are the corresponding subset of columns of $Z=XP$, and $\underline{\varepsilon}_g$ is the appropriate error term. Then the resulting estimators

$$\underline{\hat{\alpha}}_g=(\lambda_1^{-1}\underline{p}_1'X'\underline{y},\ \lambda_2^{-1}\underline{p}_2'X'\underline{y},\ \cdots,\ \lambda_g^{-1}\underline{p}_g'X'\underline{y})' \tag{1.4a}$$

$$\underline{\hat{\beta}}_g=P_g\underline{\hat{\alpha}}_g=\sum_{i=1}^{g}\lambda_i^{-1}\underline{p}_i\underline{p}_i'X'\underline{y} \tag{1.4b}$$

where it is assumed that $\lambda_1\geq\lambda_2\geq\cdots\geq\lambda_p>0$ and $\lambda_{g+1},\ \lambda_{g+2},\ \cdots,\ \lambda_p$ are small eigenvalues of $X'X$.

By defining $P=[P_g : P_s]$ where P_g is $p\times g$, P_s is $p\times s$, and $A=\begin{bmatrix}\Lambda_g & 0\\ 0 & \Lambda_s\end{bmatrix}$ in which Λ_g is the $g\times g$ diagonal and Λ_s is the $s\times s$ diagonal matrix, respectively, (1.4a) and (1.4b) can be written as,

$$\underline{\hat{\alpha}}_g=\Lambda_g^{-1}P_g'X'\underline{y}. \tag{1.5a}$$

$$\underline{\hat{\beta}}_g=P_g\Lambda_g^{-1}P_g'X'\underline{y} \tag{1.5b}$$

2. Leverage and Residuals in Principal Component Regression

Using the estimator (1.5a), the vector of fitted values is

$$\underline{\hat{y}}^*=Z_g\underline{\hat{\alpha}}_g=Z_g(Z_g'Z_g)^{-1}Z_g'\underline{y}=XP_g\Lambda_g^{-1}P_g'X'\underline{y} \tag{2.1}$$

Therefore, the matrix $H^*=Z_g(Z_g'Z_g)^{-1}Z_g'$ plays the same role as the hat matrix H in the least squares method(LSM). The ith fitted value can be written as

$$\hat{y}_i^*=\sum_{j=1}^{n}h_{ij}^*y_j$$

where h_{ij}^* is $i-j$th element of H^* for $i,j=1,\ 2,\ \cdots,\ n$, and consequently, $\partial\hat{y}_i^*/\partial y_i=h_{ii}^*$. The PC hat diagonals h_{ii}^* can be interpreted as leverage in the same sense as the hat diagonal in LSM.

The singular value decomposition(SVD) (Mandel, 1982 and Jolliffe, 1986, p.37) allows X to be decomposed as $X=U\Lambda^{1/2}P'$,
where

(1) U,P are $n\times p$, $p\times p$ matrices respectively, each of which has orthonormal columns so that $U'U=P'P=I_p$;

(2) $\Lambda^{1/2}$ is a $p \times p$ diagonal matrix;
(3) p is the rank of X.

Using the SVD, the PC leverage of the ith point can be written as

$$h_{ii}^* = \sum_{j=1}^{g} u_{ij}^2$$

since $H^* = U\left[\begin{array}{c|c} I_g & \\ \hline & O_{p-g} \end{array}\right]U'$ where I_g ia an identity matrix of dimension g, O_{p-g} is a zero matrix of dimension $p-g$ and u_{ij} is i-jth element of U for $i=1, 2, \cdots, n$ and $j=1, 2, \cdots, g$.

Several important facts can be deduced from the preceding expression. First, $h_{ii}^* < h_{ii}$ for $i=1, 2, 3, \cdots, n$; that is, for every observation the PC leverage is smaller than the corresponding least squares(LS) leverage. Second, from the fact that $H=UU'$, the leverage increases montonically as g increases since h_{ii} can be written as

$$h_{ii} = \sum_{j=1}^{p} u_{ij}^2.$$

The preceding discussion suggests that the influence can be affected as g increases. Remember, however, that influence is not only a function of leverage but also of the residual. Althouth the leverage of every point monotonically decreases as p decreases, the effect of this increment on the residual is far less clear.

The ith PC residual is defined as

$$e_i^* = y_i - \hat{y}_i^* = y_i - \underline{z}_{g\cdot i}' \underline{\hat{\alpha}}_g, \tag{2.2a}$$

which, using the SVD, can be written as

$$\begin{aligned} e_i^* &= e_i + (\hat{y}_i - \hat{y}_i^*) \\ &= e_i + \sum_{j=1}^{n} y_j \left[\sum_{m=g+1}^{p} u_{im}\ u_{jm} \right] \end{aligned} \tag{2.2b}$$

where $\underline{z}_{g\cdot i}$ is the ith row vector of Z_g

The form of (2.2b) makes it hard to tell the behavior of e_i^*. Notice, however, that the second term on the right hand side of (2.2b) can be either positive or negative; thus the PC residual for any given case can be either larger or smaller than the corresponding LSM residual.

3. Measures Based on the Influence Curve

In this section, we will focus our attention on the detection of a single influential observation. Several measures have been proposed for this purpose, however, they suffer from the problem of masking. That is, there exist some cases that can disguise or mask the prtential influence of other cases.

In case the influence function(IF) is a vector. it must be normalized so that observatins can be orderd in a meaningful way. Thus one may use

$$D_i(M, c) = \frac{(IF_i)'M(IF_i)}{c} \tag{3.1}$$

to assess the influence of the ith observation on the regression coefficients relative to M and c (see Chatterjee and Hadi, 1988). When PCR is used, we want to examine two diagnostic measures which are Welsh-Kuh distance and Cook's distance.

Welsh-Kuh's Distance

In the ordinary linear regression model, the influence of the ith observation on the predicted value $\hat{y}_i$ can be measured by the change in the prediction at $\underline{x}_i$ when the ith observation is omitted, relative to the standard error of $\hat{y}_i$, that is

$$\frac{\hat{y}_i - \hat{y}_{i(i)}}{\sigma\sqrt{h_{ii}}} = \frac{\underline{x}_i'(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})}{\sigma\sqrt{h_{ii}}} \tag{3.2}$$

where $\hat{y}_i$ is the ith row of $H\underline{y}$, $\hat{\underline{\beta}}_{(i)}$ is the estimate of $\underline{\beta}$ when the ith observation is omitted and $\hat{y}_i(\bar{i}) = \underline{x}_i'\hat{\underline{\beta}}_{(i)}$. Belsley et al. (1980) and others suggested using $\hat{\sigma}_{(i)}$ as an estimate of σ in (3.2). Then, a version of Welsh-Kuh's distance can be suggested as

$$WK_i^* = \frac{\underline{x}_i'(\hat{\underline{\beta}}_g - \hat{\underline{\beta}}_{g(i)})}{s^*_{(i)}\sqrt{h_{ii}^*}} \tag{3.3}$$

where $s^*_{(i)}$ is the square root of the residual mean square without the ith case when PCR is used. Note that we have replaced $(s^*)^2$, the residual mean square, by $(s^*_{(i)})^2$.

When the ith observation is omitted, the reduced model in (1.3) can be written as

$$\underline{y}_{(i)} = X_{(i)}P_g^*\underline{\alpha}_{g(i)} + \underline{\varepsilon}_{(i)} \tag{3.4a}$$

$$= X_{(i)}\underline{\beta}_{g(i)} + \underline{\varepsilon}_{(i)} \tag{3.4b}$$

where P_g^* is the $p \times g$ martix whose columns consist of g normalized eigenctors $\underline{p}_1^*$, $\underline{p}_2^*$, $\cdots$, $\underline{p}_g^*$, which correspond to g largest eigenvalues λ_1^*, λ_2^*, $\cdots$, λ_g^* of $X_{(i)}'X_{(i)}$, respectively.

Large values of WK_i^* indicate that the ith observation is influential on the fit of (3.4b).

Cook's Distance

Cook(1977) suggested the meausre

$$C_i = \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})'X'X(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})}{p\hat{\sigma}^2}, \quad i = 1, 2, \cdots, n \tag{3.5}$$

to asses the influence of the ith observation on the center of the confidence ellipsoid or, equivalently, on the estimated coefficients. This measure is called Cook's distance and it can be thought of as the scaled distance between $\hat{\underline{\beta}}$ and $\hat{\underline{\beta}}_{(i)}$.

At least two versions of C_i can be constructed for PCR analysis, namely,

$$C_i^* = \frac{(\hat{\underline{\beta}}_g - \hat{\underline{\beta}}_{g(i)})'(P_g\Lambda_g^{-1}P_g')^-(\hat{\underline{\beta}}_g - \hat{\underline{\beta}}_{g(i)})}{gs^2} \tag{3.6}$$

and

$$C_i^{**}=\frac{(\hat{\underline{\alpha}}_g-\hat{\underline{\alpha}}_{g(i)})'\Lambda_g(\hat{\underline{\alpha}}_g-\hat{\underline{\alpha}}_{g(i)})}{gs^2} \tag{3.7}$$

where the superscript "$-$" denotes the Moore-Penrose inverse matrix. C_i^* and C_i^{**} are based on the fact that $\mathrm{Var}(\hat{\underline{\beta}}_g)=P_g\Lambda_g^{-1}P_g'\sigma^2$ and $\mathrm{Var}(\hat{\underline{\alpha}}_g)=\Lambda_g^{-1}\sigma^2$, respectively. Note that C_i^{**} in (3.7) is not the measure on $\hat{\underline{\beta}}_g$ but the measure on $\hat{\underline{\alpha}}_g$.

WK_i^* gives a measure of the influence of the ith observation on the prediction at $\underline{x}_i$. Similarly, the influence of the ith observation on the prediction at x_r, $r\neq i$, is given by

$$\frac{|x_r'(\hat{\beta}_g-\hat{\beta}_{g(i)})|}{\sigma\sqrt{x_r'P_g\Lambda_g^{-1}P_g'\underline{x}_r}}.$$

However, if $\underline{v}$ is a $k\times 1$ vector, then we note that

$$\sup_{\underline{v}}\frac{|\hat{v}(\hat{\beta}_g-\hat{\beta}_{g(i)})|}{\sqrt{\underline{v}P_g\Lambda_g^{-1}P_g'\underline{v}}}=\sqrt{(\hat{\underline{\beta}}_g-\hat{\underline{\beta}}_{g(i)})'(P_g\Lambda_g^{-1}P_g')^-(\hat{\underline{\beta}}_g-\hat{\underline{\beta}}_{g(i)})}$$

and hence

$$\frac{|\underline{x}_r'(\hat{\underline{\beta}}_g-\hat{\underline{\beta}}_{g(i)})|}{s^*_{(i)}\sqrt{h_{rr}^*}}\leq\sqrt{gs^2C_i^*/(s^*_{(i)})^2},\ \text{for all r.}$$

Thus, if C_i^* does not declare the ith observation to be influential on the prediction at $\underline{x}_i$, then the ith observation does not seem to be influential on the prediction at any other point $\underline{x}_r$. $r\neq i$, when WK_i^* is used as a diagnostic measure. The usual F-distribution can also the used as a rough yardstick for these measures.

4. Measures Based on the Volume of Confidence Ellipsoids

A measure of the influence of the ith observation on the estimated regression coefficients can be based on the change in volume of confidence ellipsoids with and without the ith observation. In this section, we suggest two of these measures, namely,

(1) Andrews-Predgibon statistic, and
(2) Cook-Weisberg statistic.

Andrews-Pregibon Statistic

Using the distribution theory of quadratic forms, we can obtain the following theorem.

Theorem 4.1

Assume that Z_g in the model (1.3) is of rank g and $\underline{\varepsilon}\sim N(0,\ I\sigma^2)$. Then, the quantity below is distributed as noncentral F distribution, with g and n-g degrees of freedom(d.f.) and noncentrality parameters 0, $v=\underline{\beta}'P_s\Lambda_sP_s'\underline{\beta}/\sigma^2$. That is,

$$\frac{(\hat{\underline{\beta}}_g - P_g P_g' \underline{\beta})' X' X (\hat{\underline{\beta}}_g - P_g P_g' \underline{\beta})/g}{\underline{y}'(I-H^*)\underline{y}/(n-g)} \sim F(g, n-g;\ v = \underline{\beta}' P_s \Lambda_s P_s' \underline{\beta}/\sigma^2). \qquad (4.1)$$

where $SSE^* = \underline{y}'(I-H^*)\underline{y}$.

Proof From (1.5b), it follows that

$$X(\hat{\underline{\beta}}_g - \underline{\beta}) \sim N(-XP_g P_g' \underline{\beta},\ H^* \sigma^2).$$

Therefore,

$$Q_0 = (\hat{\underline{\beta}}_g - P_g P_g' \underline{\beta})' X' X (\hat{\underline{\beta}}_g - P_g P_g' \underline{\beta})/\sigma^2 \sim \chi^2(g)$$

where g denotes the d.f. The proof is completed from the fact that

$$\underline{y}'(I-H^*)\underline{y}/\sigma^2 \sim \chi^2(n-g,\ v = \underline{\beta}' P_s \Lambda_s P_s' \underline{\beta}/\sigma^2),$$

where v denotes the noncentrality parameter, and

Q_0 and $\underline{y}'(I-H^*)\underline{y}$ are independent. ■

Let $\Lambda_{g(i)} = \mathrm{diag}(\lambda_1^*, \lambda_2^*, \cdots, \lambda_g^*)$ and $SSE^*_{(i)}$ denote the residual sum of squares in the model (3.4a). Then two versions of the Andrews-Pregibon statistic in PCR based on (3.7) and (3.6) respectively, can be defined as,

$$AP_i^* = 1 - \frac{SSE^*_{(i)}\ |\Lambda_{g(i)}|}{SSE^*\ |\Lambda_g|} \qquad (4.2a)$$

and

$$AP_i^{**} = 1 - \frac{SSE^*_{(i)}\ |X_{(i)}' X_{(i)}|}{SSE^*\ |X'X|} \qquad (4.2b)$$

where the bar denotes determinant. Note that AP_i^* is the measure detecting the sensitivity on $\underline{\hat{\alpha}}_g$. The following theorem shows a property of AP_i^*.

Theorem 4.2

Let $W = (XP_g \vdots \underline{y})$ be an augmented $n \times (g+1)$ martix. Then, AP_i^* in (4.2a) can be written as

$$AP_i^* = 1 - \frac{|W_{(i)}' W_{(i)}|}{|W'W|} \qquad (4.3)$$

Proof Lew $W_{(i)} = (X_{(i)} P_g^* \vdots \underline{y}_{(i)})$ be the augmented $(n-1) \times (g+1)$ matrix. Then, since

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |A|\ |D - CA^{-1}B|,$$

we have

$$
\begin{aligned}
|W'W| &= |P_g'X'X\,P_g| \quad |\underline{y}'(I-H^*)\underline{y}| \\
&= |\Lambda_g|\ SSE^*.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
|W_{(i)}'W_{(i)}| &= |(P_g^*)'X_{(i)}'X_{(i)}P_g^*| \quad |\underline{y}_{(i)}'(I-H^*_{(i)})\underline{y}_{(i)}| \\
&= |\Lambda_{g(i)}|\ SSE^*_{(i)}
\end{aligned}
$$

where $H^*_{(i)}=Z_g^*{}_{(i)}[(Z_g^*{}_{(i)})'Z_g^*{}_{(i)}]^{-1}(Z_g^*{}_{(i)})'$, which completes the proof. ■

Note that the second term in the right hand side in (4.2b) represents the proportion of the volume generated by W that is not due to the ith observation. Hence, large values of (4.2a) and (4.2b) call for special attention.

Cook-Weisberg Statistic

Under normality, the $100(1-\alpha)\%$ joint confidence ellipsoid for $P_gP_g'\underline{\beta}$ can be obtained from (4.1). That is

$$
E=\{P_gP_g'\underline{\beta}:\ \frac{(\hat{\underline{\beta}}_g-P_gP_g'\underline{\beta})'X'X(\hat{\underline{\beta}}_g-P_gP_g'\underline{\beta})}{g(s^*)^2}\leq F(\alpha:g,\ n-g;0,\ v)\}.
$$

Cook and Weisberg(1980) propose the logarithm of the ratio of the volume of the $100(1-\alpha)\%$ confidence ellipsoids with and without the ith observation as a measure of influence. Since the volume of an ellipsoid is proportional to the inverse of the square root of the determinant of the associated matrix of the quadratic forms, the Cook-Weisberg statistic in PCR can be defined as

$$
\begin{aligned}
CW_i^* &= \log\left\{\left|\frac{X_{(i)}'X_{(i)}}{X'X}\right|^{1/2}\left[\frac{s^*}{s^*_{(i)}}\right]^{p}\left[\frac{F(\alpha:g,\ n-g;0,v)}{F(\alpha:g,\ n-g-1;0,v_i)}\right]^{p/2}\right\} \qquad (4.4)\\
&= 1/2\log(1-h_{ii})+p/2\log\left[\frac{(s^*)^2}{(s^*_{(i)})^2}\right]+p/2\log\left[\frac{F(\alpha:g,\ n-g;0,\ v)}{F(\alpha:g,\ n-g-1;0,\ v_i)}\right]\\
&\simeq 1/2\log(1-h_{ii})+p/2\log\left[\frac{(s^*)^2}{(s^*_{(i)})^2}\right],
\end{aligned}
$$

where $v_i=\underline{\beta}P_s^*\Lambda_s^*P_s^*{}'\underline{\beta}/\sigma^2$, P_s^* is the px(p−g) matrix whose columns consist of p−g normalized eigenvectors and Λ_s^* is the (p−g)x(p−g) diagonal matrix, which correspond to p−g smallest eigenvalues of $X_{(i)}'X_{(i)}$, respectively. If this quantity is large and positive, then deleting the ith case will result in a substantial decrease in volume, and if it is large and negative, deleting the ith case will result in a substantial increase in volume.

5. A Numerical Example

The data set which is used for a numerical example is related to the performance of a computerized system for processing military personnel action forms. There are 15 observations on six regressors and one dependent variable(see Table 1). First, we apply principal component analysis (PCA) based on the correlation matrix to the predictors. The correlation matrix and the results of PCA are shown in Table 2 and 3, respectively. Then we select the first four PCs, because the remaining eigenvalues are very small ($\lambda_4=0.4266>>\lambda_5=0.0629$) and the coefficient of determination R^2 is not small compared to the model with all PCs.

Table 4 shows e_i^*, r_{ia}, r_{ia}^*, h_{ii}^* and h_{wii},
where

$$r_{ia}=e_i^*/(s^*\sqrt{1-h_{ii}^{**}}),$$

$$r_{ia}^*=e_i^*/(s^*_{(i)}\sqrt{1-h_{ii}^*}),$$

and h_{wii} is the i-th diagonal element of the hat matrix of $W=(XP_g : Y)$. The scatter plot of r_{ia} versus $\hat{y}_i^*$(Fig. 1) and the normal probability plot(Fig. 2) do not show any gross violation of the usual assumptions. Observations #8 and #15, however, have moderate large residuals. Only one case (#1) has $h_{ii}^*>2(4)/15=.533$, and hence it can be declared to be a high-leverage point.

Fig. 3 shows the boxplot for r_{ia}, h_{ii}^* and h_{wii}. The boxplots for h_{ii}^* and h_{wii}^* show that observations #1, #2, and #8 are separated from the bulk of other observations. Typically h_{wii} picks out observations with large h_{ii}^*(e.g., observation #1) and $|e_i^*|$ (e.g., observation #8) as being different from other observations. In this example, however, h_{ii}^* does not pick out observation #8 ; the reason being that observation #8 lies near the center of the predictor variables and hence has somewhat small h_{ii}^* value($h_{88}^*=.4$).

The L-R plot, defined as the scatter plot of leverage value h_{ii}^* versus $a_i^2=(e_i^*)^2/SSE^*$, for the Hill's data is shown in Fig. 4. Two observations are separated from the bulk of other points. We find the high-leverage point(#1) in the upper-left corner and the outlier (#8 or #15) in the lower-right corner.

Next, we examine the influence measures based on the IF. Thses are also shown in Table 5. The corresponding boxplots (Fig. 5) show that observation #8 is the most influential on $\hat{\beta}_g$. Examination of residuals have not pointed out any peculiarities regarding observation #8. This observation, however, has the second largest standardized residual($r_{ia}=1.81$).

The influential measures based on the volume of confidence ellipsoids are shown in Table 6 and the corresponding boxplots are displayed in Figure 5. According to these measures, observation #2 is the most influential on the volume of confidence ellipsoids. This is, because the points that are remote in the space are the ones that affect the volume of the confidence ellipsoids the most.

With regard to examination of the data for the presence of outliers, high-leverage points, or influential observations, each of which has different characteristics. The L-R plot (Fig. 4) explains the difference among these three observations. Observation #15 is an example of an outlier that is neither a high-leverage point nor influential. #1 is an example of a high-leverage point that is neither an outlier nor influential. Measures based on the IF have pointed #8 as the most influential on $\hat{\beta}_g$ and $\hat{\sigma}$. According to the influential measures based on the volume of confidence ellipsoids #2 is an example of an influential observation that is not an outlier. Note that examination of residuals alone is not sufficient for the detection of influential observations, and C_i^{**} in (3.14) and AP_i^{**} in (4.2b) are not the influential measures of postulated models, but of reduced models in (1.3).

Table 1. Hill's Data

Case	X_1	X_2	X_3	X_4	X_5	X_6	y
1	57.0	6.40	12	293.2	41.1	45.0	61.2
2	53.0	5.00	12	354.3	51.0	29.4	62.3
3	50.3	5.75	14	293.5	24.9	29.4	59.4
4	41.2	4.50	13	299.0	19.4	20.3	66.2
5	36.7	5.15	13	286.0	18.6	17.4	66.0
6	35.5	4.25	10	254.8	17.1	14.9	71.4
7	26.4	3.35	10	270.4	17.6	15.5	75.4
8	25.0	2.50	9	239.2	13.6	13.2	83.2
9	23.5	3.45	11	270.5	14.3	11.7	73.2
10	26.7	6.00	11	298.0	12.9	10.4	71.1
11	25.8	5.70	11	247.0	11.9	15.2	72.8
12	25.7	6.75	12	260.1	12.5	19.5	75.6
13	27.0	4.95	12	228.8	10.5	18.6	76.0
14	24.5	3.65	12	179.4	8.3	19.1	70.2
15	23.1	4.05	11	176.8	8.5	15.9	68.6

Table 2. Correlation Matrix

	X_1	X_2	X_3	X_4	X_5	X_6
X_1	1.000000	.424576	.538386	.688039	.889261	.871502
X_2	.424576	1.000000	.561795	.388890	.295933	.479412
X_3	.538386	.561795	1.000000	.303690	.281019	.503626
X_4	.688039	.388890	.303690	1.000000	.755960	.396400
X_5	.889261	.295933	.281019	.755960	1.000000	.795574
X_6	.871502	.479412	.503626	.396400	.795574	1.000000

Table 3. The Results of PCA Based on Correlation Matrix

	Z_1	Z_2	Z_3	Z_4	Z_5	Z_6
X_1	−.488728	.155276	−.186665	.102896	−.774047	.304062
X_2	−.317532	−.585525	.387861	−.613078	−.024618	.171542
X_3	−.325534	−.607410	−.119596	.677714	.221660	.048460
X_4	−.385216	.294474	.708899	.240910	−.014109	−.451795
X_5	−.452813	.421095	−.049578	.075490	.578309	.524444
X_6	−.448235	.008154	−.543526	.300858	.128340	−.629771
Eigenvalue	3.7999	1.0551	.6235	.4266	.0629	.0317
Proportion	.6333	.1759	.1039	.0711	.0105	.0053
Cumulative Proportion	.6333	.8092	.9131	.9842	.9947	1.0000

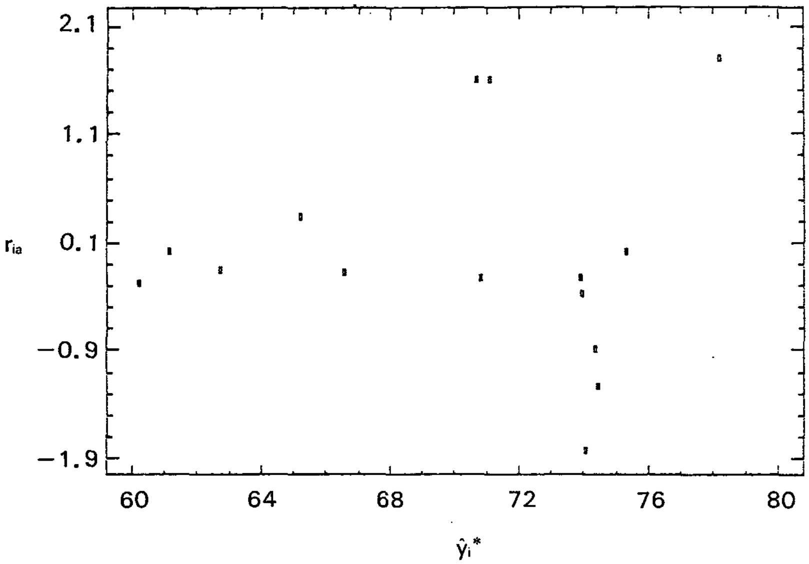

Fig. 1. Scatter Plot of r_{ia} versus $\hat{y}_i^*$

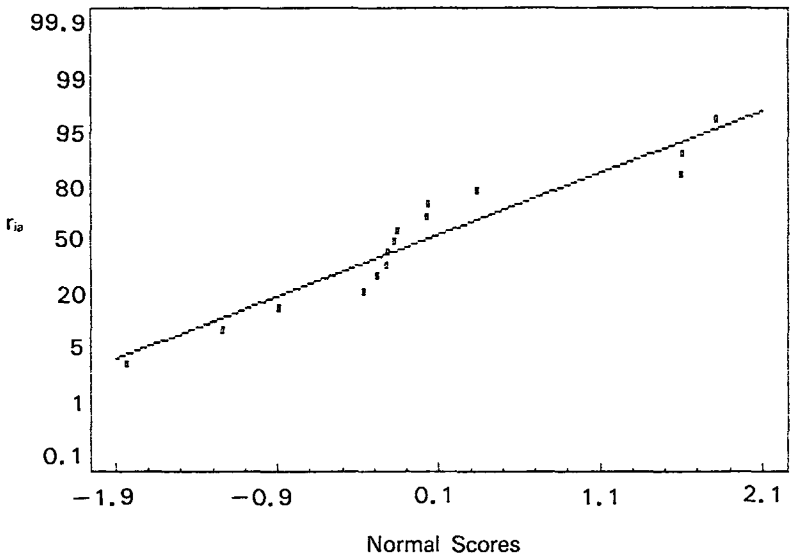

Fig. 2. Normal Probability Plot

Table 4. e_i^*, r_{ia}, r_{ia}^*, h_{ii}^*, and h_{wii}

Row	e_i^*	r_{ia}	r_{ia}^*	h_{ii}^*	h_{wii}
1	.041150	.021710	.020578	.717919	.717932
2	−.406078	−.160996	−.152879	.500482	.501777
3	−.799176	−.279180	−.265281	.356599	.361614
4	1.010373	.339144	.323539	.303119	.311134
5	−.538058	−.172275	−.163371	.234092	.236365
6	−2.936843	−.896670	−.886647	.157711	.225431
7	.082599	.026172	.024785	.217937	.217991
8	5.001830	1.80925	2.043408	.399868	.596302
9	−.696050	−.219866	−.209143	.213084	.216888
10	−3.341525	−1.240794	−1.271169	.430551	.518220
11	−1.139843	−.361975	−.345970	.221432	.231633
12	4.493012	1.598399	1.727911	.379604	.538106
13	5.308437	1.607700	1.750274	.143972	.365226
14	−.611123	−.224886	−.214069	.420175	.423107
15	−5.468735	−1.836095	−2.095293	.303457	.538275

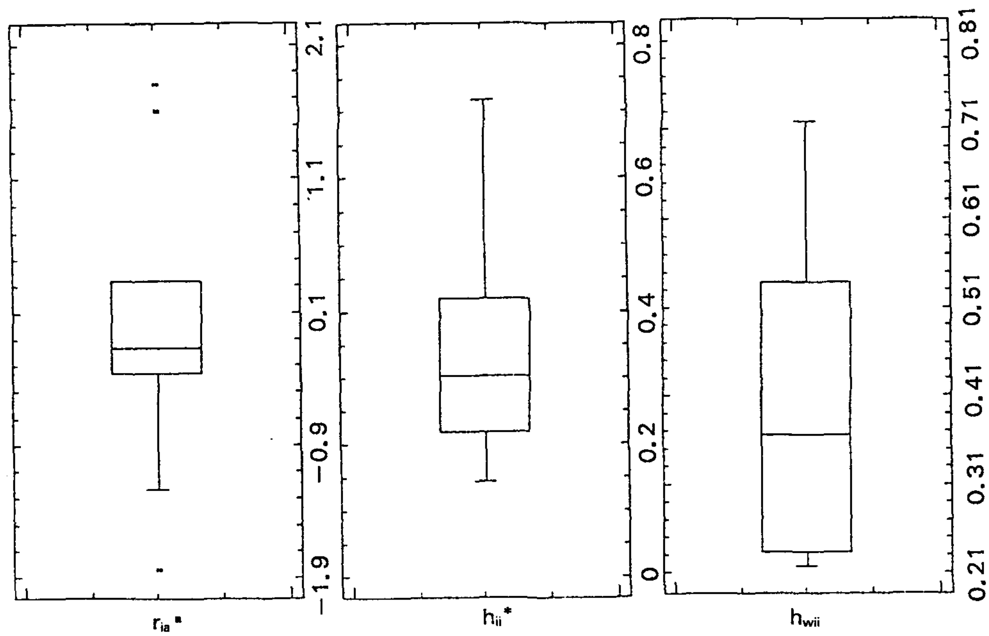

Fig. 3. Boxplots of r_{ia}, h_{ii}^*, and h_{wii}

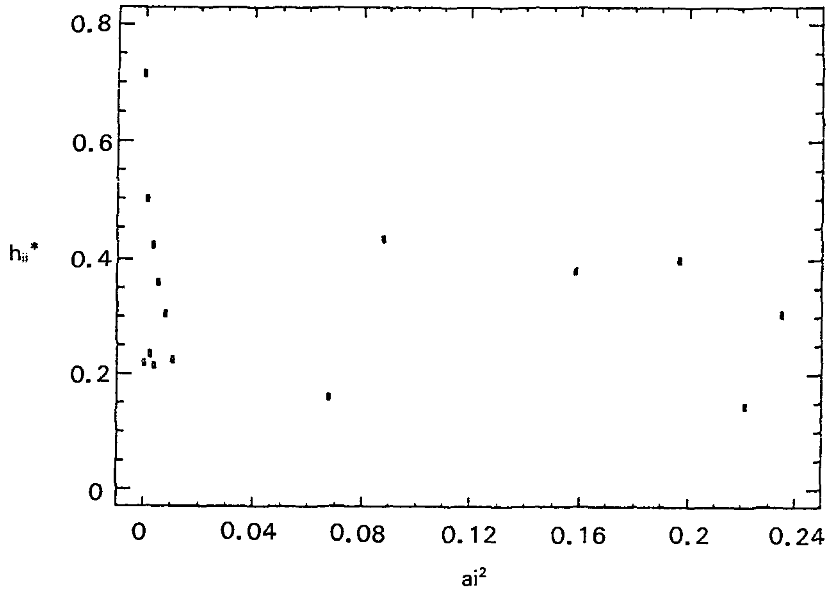

Fig. 4. L-R plot

Table 5. Influence Measures

Row	WK_i^*	C_i^*	C_i^*
1	.173063	.008637	1.223204
2	.118442	.006005	.164332
3	.188799	.011415	.100737
4	.127122	.008788	.011049
5	.094152	.002238	.013899
6	.079869	.020866	.017492
7	.027385	.000161	.006761
8	2.040672	.406936	.193196
9	.059258	.001243	.032152
10	1.044241	.222112	.626418
11	.182800	.006720	.029430
12	1.022652	.301711	.724575
13	.370837	.055851	.030583
14	.168849	.006830	.529440
15	.903245	.266475	.185863

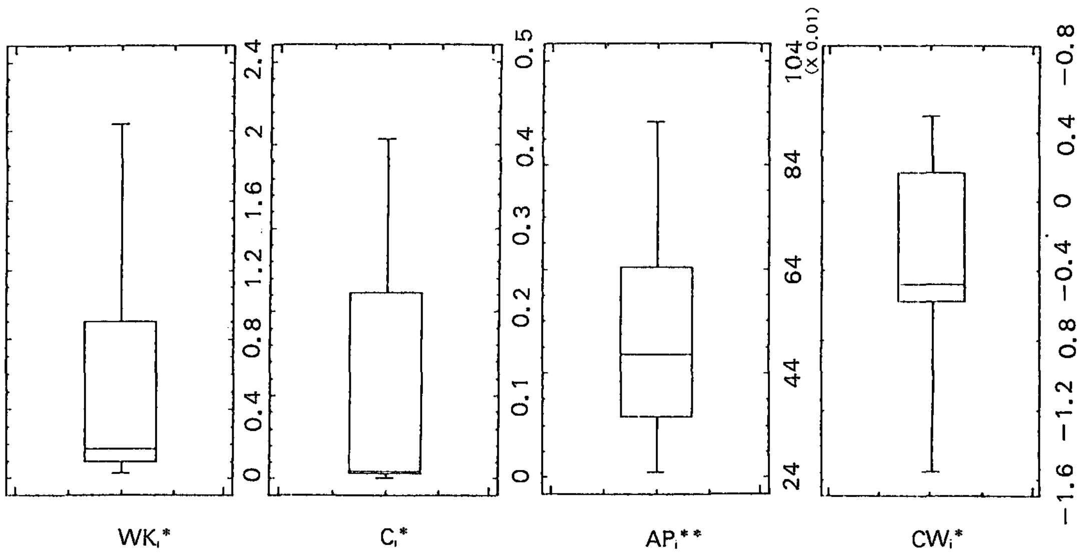

Fig. 5. Boxplots of WK_i^*, C_i^*, AP_i^{**} and CW_i^*

Table 6. Measures Based on the Volume of Confidence Ellipsoids

Row	AP_i^*	AP_i^{**}	CW_i^*
1	.639503	.836638	−1.198021
2	.424886	.921989	−1.554768
3	.288840	.461735	−.584342
4	.240760	.417471	−.517109
5	.166288	.270025	−.446044
6	.160146	.678051	−.562536
7	.147713	.311064	−.484709
8	.528446	.640801	.423012
9	.149045	.357170	−.488106
10	.453835	.522655	−.117628
11	.167085	.249999	−.377584
12	.467466	.605699	−.162883
13	.299292	.365963	.449783
14	.355152	.474885	−.584309
15	.470974	.644044	.490724

6. Concluding Remarks

In Sections 3 and 4, we suggested several diagnostic measures for detection of outliers or influential observations when principal component regression(PCR) was used. We have seen that many of these measures are closely related. Therefore, the analyst should choose some of the diagnostic measures that can assess the influence of each case on the particular features of interest depending on the specific goals of analysis.

To get an idea of the sensitivity of the data, the resulting fit should be examined in detail. To compute and calculate various matrix manipulations, we have used the statistical software, M Matrix Language for Statistics and Matrix Algebra, and Statgraphics has been used to make various statistial figures.

In Section 5, a numerical example was illustrated. Some individual data points may be flagged as outliers, high-leverage points, or influential points. Any point falling into one of these categories should be carefully examined for accuracy(transcription error, etc), relevancy(whether it belongs to the data set or not), or special significance(abnormal conditions. etc).

Acknowledgement

The authors would like to give deep appreciation to the referees for their kind suggestions and comments in improving this work.

References

1. Belsley, D.A., Kuh, E., and Welsh, R.E.(1980). Regression Diagnostics : Indentifying Influential Data and Sources of Collinearity, New York : John Wiley & Sons
2. Chatterjee, S. and Hadi, A.S.(1988). Sensitivity Analysis in Linear Regression, John Wiley & Sons.
3. Cook, R.D.(1977). Detection of Influential Observations in Linear Regession. Technometrics, 19, 15-18.
4. Jolliffe, I.T.(1986). Principal Component Analysis. Springer-Verlag.
5. Mandel, J.(1982). Use of the Singular Value Decomposition in Regression Analysis. The American Statistican, 36, 15-24.