〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 19, No. 2, 1990

A Local Influence Approach to Regression Diagnostics with Application to Robust Regression+

Myung-Hoe Huh* and Sung H. Park**

ABSTRACT

Regression diagnostics often involves assesment of the changes that result from deleting multiple cases. Diagnostic methodology based on global influence measure, however, needs prohibitive computing time. As an alternative, Cook(1986) developed ocal influence approach in which it is checked whether a minor modification of specification influences key results of an analysis. In line with Cook's development, we propose and study an influence derivative method that yields both the magnitude and direction of case influences. The utility of our methodology is highlighted when case influence derivatives are plotted in a lower demensional space. Such plots are especially effective in unmasking "masked" observations in least squares regression and in robust regression also. We give several illustrations.

1. Introduction

Consider the following linear regression model :

$$y_i = x_i'\beta + \varepsilon_i,$$

for i=1, ⋯, n, or, in matrix notation,

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

where y is an n×1 vector of responses, X is an n×p matrix of fixed constants, β is a p×1 vector of regression parameters, and ε is an n×1 vector of independent normal random errors with mean 0 and variance $\sigma^2 D_w^{-1} = \mathrm{diag}(\sigma^2/w_1, \cdots, \sigma^2/w_n)$; it is assumed tentatively that $\mathrm{Var}(\varepsilon_i) = \sigma^2/w_i$, for i=1, ⋯, n. Then the weighted least squares estimate of β is given by

$$\hat{\beta}_w = (X'D_wX)^{-1} X'D_wy, \tag{1.2}$$

where $D_w = \mathrm{diag}(w_1, \cdots, w_n)$. Thus $w_1, \cdots, w_n$ are weights for individual observations (cases). Note that

$$\partial\hat{\beta}_w/\partial w_i = -(X'D_wX)^{-1} X'D_iX(X'D_wX)^{-1} X'D_wy + (X'D_wX)^{-1} X'D_iy$$

+ This Research was partially supported by Korea Research Foundation 1989.

$$= -(X'D_wX)^{-1}\, x_i(x_i'\,(X'D_wX)^{-1}\,X'D_wy-y_i)$$
$$= (X'D_wX)^{-1}\, x_i(y_i-x_i'\,\hat{\beta}_w), \qquad (1.3)$$

where $D_i=\mathrm{diag}(0,\cdots,\ 0,\ 1,\ 0,\cdots,\ 0)$; the i-th diagonal of D_i is one and all other diagonals are zero. We will call (1.3) a case influence derivative of $\hat{\beta}_w$. For instance, influence derivative of the least squares estimate $\hat{\beta}$ is given by

$$[\partial\hat{\beta}_w/\partial w_i]_{w=1_n}=(X'X)^{-1}\, x_i e_i, \qquad (1.4)$$

where $e=(e_1,\cdots,\ e_n)'=y-X\hat{\beta}$, $w=(w_1,\cdots,\ w_n)'$, and 1_n equals $n\times 1$ vector whose elements are all equal to one. Expression (1.4) is well known in the regression diagnostics literature (Belsley, Kuh, and Welch, 1980 ; Cook and Weisberg, 1982).

Cook(1986) developed case-weight perturbation scheme with the concept of likelihood displacement $LD(w\mid w_o)$. In general context, denote

$$LD(w\mid w_o)= L(\hat{\beta}_{wo}\mid w_o)-(L(\hat{\beta}_w\mid w_o)$$
$$= -[(y-X\hat{\beta}_{wo})'\,D_{wo}(y-X\hat{\beta}_{wo})-(y-X\hat{\beta}_w)'\,D_{wo}(y-X\hat{\beta}_w)]/2\sigma^2.$$

where $L(\hat{\beta}_w\mid w_o)$ is the log-likelihood function with assumed weights w_o evaluated at $\hat{\beta}_w$. Then one can easily show that

$$[\partial^2 LD(w\mid w_o)/\partial w_i^2]_{w=w_o}=([\partial\hat{\beta}_w/\partial w_i]_{w=w_o})'\,(X'D_{wo}\,X)\,([\partial\hat{\beta}_w/\partial w_i]_{w=w_o}/\sigma^2.$$

Hence we see that case influence derivatives determine likelihood displacement locally (up to the second order), and that, as a by-product, $X'D_{wo}X$ is the normalizing factor in calculating absolute magnitudes of case influence derivatives.

Next, we try to get more general expression of (1.3) and (1.4) that are suitable in multiple-case perturbation : For the index set $I=\{i_1,\cdots,\ i_m\}$ of m cases, $p\times m$ matrix $\partial\hat{\beta}_w/\partial w_I$ is defined to be

$$\partial\hat{\beta}_w/\partial w_I = (\partial\hat{\beta}_w/\partial w_{i1},\cdots,\ \partial\hat{\beta}_w/\partial w_{im})$$
$$= (X'D_wX)^{-1}X_I'\ diag(y_i-x_{i1}'\,\hat{\beta}_w,\cdots,\ y_{im}-x_{im}'\,\hat{\beta}_w).$$

Thus

$$[\partial\hat{\beta}_w/\partial w_I]_{w=1_n}=(X'X)^{-1}\,X_I'\ diag(e_{i1},\cdots,\ e_{im}),$$

where $X_I=(x_{i1},\cdots,\ x_{im})'$.

These generalizations are particularly useful when we study simultaneous perturbation of m cases. Consider the perturbation scheme which is specified by

$$(w_{i1},\cdots,\ w_{im})=(w'_{i1}(1-u),\cdots,\ w'_{im}(1-u)),\ 0<u<1.$$

This can be used as a local proxy for the scheme of multiple case deletion as will be shown shortly. Then the derivative of $\hat{\beta}_w$ with respect to u is, by the chain rule,

$$\partial\hat{\beta}_w/\partial u=\partial\hat{\beta}_w/\partial w_I\cdot(-w'_i,\cdots,\ -w'_i)'.$$

Thus

$$[\partial\hat{\beta}_w/\partial u]_{u=o}=-(X'D_{wo}X)^{-1}\sum_k w_{ik}\,x_{ik}(y_{ik}-x_{ik}'\,\hat{\beta}_w) \qquad (1.5)$$

and

$$[\partial\hat{\beta}_w/\partial u]_{u=o,\ w=1_n}=-(X'X)^{-1}\sum_k x_{ik}\,e_{ik}. \qquad (1.6)$$

Hence (1.5) and (1.6) are just equal to weighted sums of (1.3) and (1.4), respectively.

It is interesting to note that (1.4) is equivalent to one of influence curves evaluated at the i-th

$$EIC_i = n(X'X)^{-1} x_i e_i.$$

It is also possible to extend the notion of influence curve to handle simultaneous influence of multiple-case perturbation, as will be shown in Section 2.

Influence derivatives (1.3), or (1.4) in particular, are p×1 vectors. Therefore, unless they are represented in lower dimensional space, it is very hard to catch any significant patterns at the stage of real data analysis. Such lower dimensional reduction method will be developed in Section 3.

Robust regression defies usual regression diagnostics methodology based on global measures on multiple cases. Our diagnostics methodology using influence derivatives for weighted regression, however, can be easily adapted for the robust regression, as will be discussed in Section 4. We give several numerical illustrations in Section 5.

2. Influence Curve for Multiple Perturbations

Let (x', y) be the $(p+1)$ random vector following a joint cdf F with $E_F(x\ x') = \Sigma(F)$ and $E_F(x\ y) = \gamma(F)$. Then the functional corresponding to the least squares estimator of β is $T(F) = \Sigma^{-1}(F)\ \gamma(F)$ assuming that $\Sigma^{-1}(F)$ is nonsingular (Cook and Weisberg, 1982, p. 107). The $(p+1)$ dimensional influence curve is

$$\begin{aligned} IC_{T,F}(x_0, y_0) &= \lim_{\varepsilon \to 0} \{T[(1-\varepsilon)F + \varepsilon\ \delta(x_0, y_0)] - T(F)\} / \varepsilon \\ &= \Sigma^{-1}(F)\ x_0(y_0 - x'T(F)), \end{aligned}$$

where $\delta(x_0, y_0)$ is a point mass distribution at (x_0, y_0). Hence $IC_{T,F}(\cdot)$ captures spontaneous change of T(F) when the underlying distribution is perturbed with a point mass distribution. This can be generalized from a point mass to arbitrary distribution function. Now suppose that G is a cdf in $(p+1)$ dimensional Euclidean space, and define

$$IC_{T,F}(G) = \lim_{\varepsilon \to 0} [T[(1-\varepsilon)F + \varepsilon G] - T(F)] / \varepsilon.$$

Then it is easy to show that

$$IC_{T,F}(G) = \Sigma^{-1}(F)(\gamma(G) - \Sigma(G)T(F)).$$

In particular, when G puts equal mass $1/m$ at $(x_1', y_1), (x_2', y_2), \cdots, (x_m', y_m)$, we have

$$IC_{T,F}(G) = IC_{T,F}(X_0, y_0) = \Sigma^{-1}(F)\ X_0'(y_0 - X_0T(F))/m,$$

where $X_0 = (x_1, \cdots, x_m)'$ and $y_0 = (y_1, \cdots, y_m)'$.

Empirical influence curve, one of sample versions of $IC_{T,F}(X_0, y_0)$ evaluated at empirical distribution $\hat{F}$ or $n \times (p+1)$ (X, y), is

$$\begin{aligned} EIC_{T,\hat{F}}(X_0, y_0) &= EIC_{T,(X,y)}(X_0, y_0) \\ &= n/m \cdot (X'X)^{-1} X_0'(y_0 - X_0\beta), \end{aligned}$$

which is identical to (1.6) when $X_0 = X_1$, except proportional constant $-(n/m)$. Therefore we conclude that the influence derivative approach for multiple perturbation anchors its conceptual basis on influence curve.

3. Principal Components Reduction of Influence Measures

In Section 1, we have seen that p×1 case influence derivatives

influence of multiple case perturbation. Hence, if n such points are visualized in the p-dimensional space, we could recognize which observation or subgroups of observations are influential and could see directions of influence. The problem is that it is hardly possible to represent points in three or higher dimensional space in an efficient way.

We propose principal components reduction of dimensionality of case influence derivatives : locate high-dimensional points in a lower (two or three) dimensional space. Let

$$F=(f_1,\cdots,\ f_n)$$
$$=(X'X)^{-1}\ X'\ diag(y_1-x_1\hat{\beta}_w,\cdots,\ y_n-x_n\hat{\beta}_w). \tag{3.1}$$

Consider the projection of $p\times 1$ f_i's along u satisfying the unit length condition $u'(X'D_wX)u=1$. The main reason why we norm R^p with the semi-positive definite matrix $X'D_wX$ is that statistical distance in the space of parameter estimates of β can be defined with this matrix ;

$$d(\hat{\beta}_w,\ \hat{\beta})=(\hat{\beta}_w-\hat{\beta})'\ (X'D_wX)(\hat{\beta}_w-\hat{\beta}),$$

since

$$Var(\hat{\beta}_w)\ \propto\ (X'D_wX)^{-1}.$$

Second reason was explained in Section 1 after discussing the infinitesimal version of likelihood displacement. Thus the sum of squares of projected f_i's on u is written as

$$\sum_i\ \{f_i'(X'D_wX)u\}^2,$$

Hence, in order to get the best approximation, we need to maximize

$$\{u'(X'D_wX)F\}\{F'(X'D_wX)u\}$$

subject to

$$u'(X'D_wX)u=1.$$

It leads to solve

$$(X'D_wX)(FF')\ (X'D_wX)u=\lambda(X'D_wX)u.$$

Pre-multiplying both sides by $(X'D_wX)^{-1/2}$ and letting $u^*=(X'D_wX)^{1/2}u$, we have

$$(X'D_wX)^{1/2}(FF')\ (X'D_wX)^{1/2}u^*=\lambda\ u^*. \tag{3.2}$$

Hence λ and u^* are eigenvalues and eigenvectors of $(X'D_wX)^{1/2}\ FF'(X'D_wX)^{1/2}$, or,

$$(X'D_wX)^{-1/2}X\ diag^2(y_1-x'_1\hat{\beta}_w,\cdots,\ y_n-x'_n\hat{\beta}_w)\ X'(X'D_wX)^{-1/2} \tag{3.3}$$

which is obtained by substituting (3.1) for F. As a consequence of solving the eigenvalue-eigenvector problem (3.2), projections of $f_1,\cdots,\ f_n$ on u are columns of

$$u\cdot u^{*\prime}(X'D_wX)^{1/2}F$$

In other words, principal component scores are elements of $F'(X'D_wX)^{1/2}u^*$. For two or higher dimensional approximation, we just need to compute next larger eigenvalues and corresponding eigenvectors, and similarly we can obtain other principal component scores.

It is interesting to note that, for the case $D_w=I_n$, the largest eigenvalue λ times a contant $2/\sigma^2$ is turned out to be equal to maximal normal curvature that is obtained by the differential geometric approach in Cook(1986).

4. Applications to the Robust Regression

M-estimation for the model (1.1) is obtained by minimizing

$$\sum_i \rho((y_i - x_i'\beta)/\sigma), \tag{4.1}$$

where $\rho(\cdot)$ is a differentiable convex function. Differentiating (4.1) with respect to β, we have

$$\sum_i \varphi((y_i - x_i'\beta)/\sigma)\ x_i = 0, \tag{4.2}$$

where $\varphi(t) = \rho'(t)$. By setting $w(t) = \varphi(t)/t$, (4.2) is equivalent to

$$\sum_i w((y_i - x_i'\beta)/\sigma)\ x_i(y_i - x_i'\beta) = 0. \tag{4.3}$$

Thus, if we set $w((y_i - x_i'\beta))/\sigma) = w_i$, (4.3) can be written as

$$(\sum_i w_i x x_i')\beta = \sum_i w x_i y_i.$$

Thus M-estimate of β can be obtained just like a weighted least squares estimate :

$$\hat{\beta}_w = (X'D_w X)^{-1}\ (X'D_w y).$$

As a consequence, we can think of applying influence diagnositc procedures for the weighted least squares regression developed in Section 3. In the next section, we will give illustrations which show how it works. And in the last section, we will add brief remarks on this heuristic procedure for the robust regression.

5. Numerical Illustrations

Example 1 : The "adaptive score" data (Mickey, Dunn and Clark, 1967) consists of two variables, age of a child in months at first word (X) and Gessel adaptive score (Y), for 21 children. See Fig. 1 for the scatterplot. Consider the linear regression model

$$Y = \beta_0 + \beta_1 X + \varepsilon. \tag{5.1}$$

And try ordinary least squares fitting to the model (5.1), assuming $w = 1_n$ tentatively. Now let us follow the procedures for detecting influential subset explained in Section 3. First compute eigenvalue-eigenvector decomposotion of (3.3) with $w = 1_n$. Its eigenvalues are 109.98 and 53.02. Of course, the two-dimensional principal components plot of influence derivatives does not lose any amount of information, since there is no reduction of dimensionality. In Fig. 2 which shows principal components coordinates of case influence derivatives for the first two axes, we note following two interesting facts. First, observation 19 is the most influential single case. Second, observations 18 and 2 are also influential but in an orthogonal direction to that of observation 19. Therefore, the subset of observations 18 and 2 induces "masking effect" that frequently nullifies single case diagnostic procedures such as Cook's statistic D_i. All these points can be foreseen only with the scatterplot Fig. 1.

Next we try robust fitting to the model (5.1), with Andrew's $\varphi(\cdot)$ function given by,

$$\varphi(t) = \begin{cases} \sin(t/1.5), & \text{for } |t| < 1.5\pi, \\ 0, & \text{for } |t| > 1.5\pi. \end{cases}$$

For the estimate of σ which is needed in fitting, we used the median of absolute residuals (from the least squares fit) 6.67. Then we have $\hat{\beta}_w = (110.31,\ -1.23)'$, whereas the least squares estimate $\hat{\beta}$ is $(109.87,\ -1.13)'$. The internal weights used in computing the robust estimate of β is given by

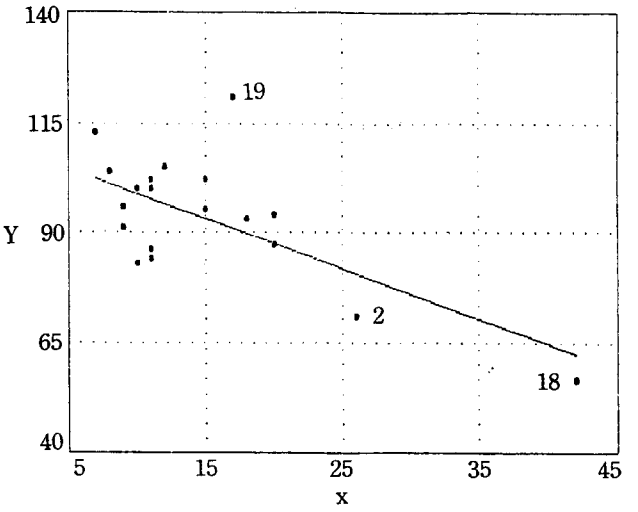

Fig. 1. Scatterplot of "Adaptive Score" Data.

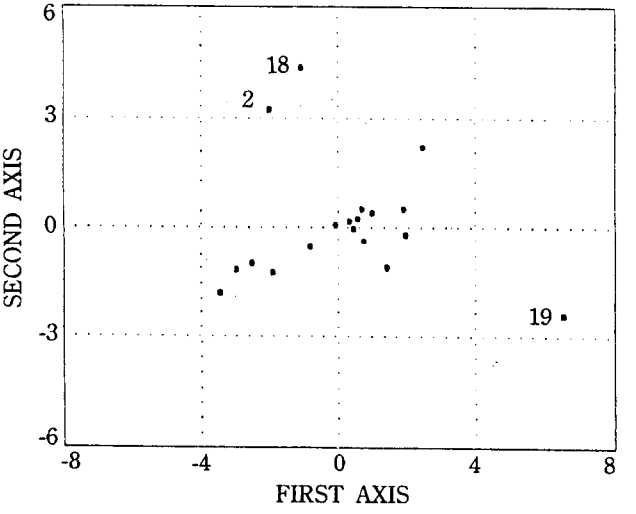

Fig. 2. Case Influence Derivatives for the Least Squares Fit of Adaptive Score Data.

$$w=(0.99,\ 0.91,\ 0.67,\ 0.89,\ 0.84,\ 1.00,\ 0.96,\ 0.99,\ 0.98,\ 0.89,\ 0.80,\ 0.98,\ 0.67,\ 0.75,\ 0.96,\ 1.00,\ 0.86,\ 1.00,\ 0.00,\ 0.82,\ 1.00)',$$

up to proportional constant. With the plot (Fig. 3) of the first two principal components coordinates of case influence derivatives, we note two things. First, observation 19 is identified as *only potentially* influential, since it does not exercise any influence upon the fitted model. Second, observations 18 and 2 are not dominant influential cases any more.

Example 2 : The "stack loss" data(Brownlee, 1965 ; Daniel and Wood, 1980 ; Li, 1985 ; Atkinson, 1986) comes from 21 days of operation of a plant that oxidizes ammonia to nitric acid. It includes

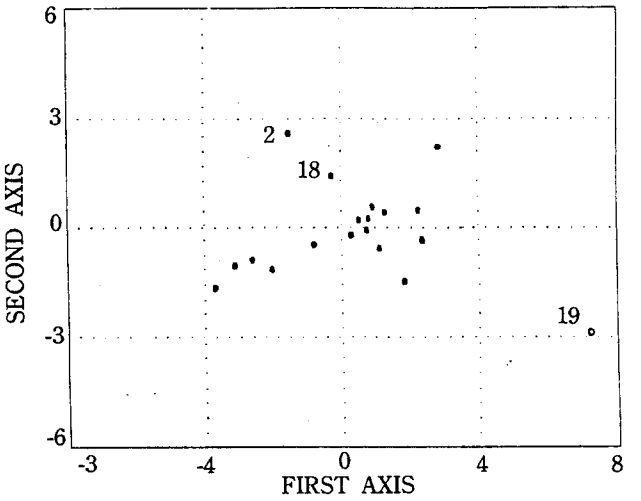

Fig. 3. Case Influence Derivatives for the Robust Fit of Adaptive Score Data.
An unfilled circle represents observation that has zero weight in robust fitting.

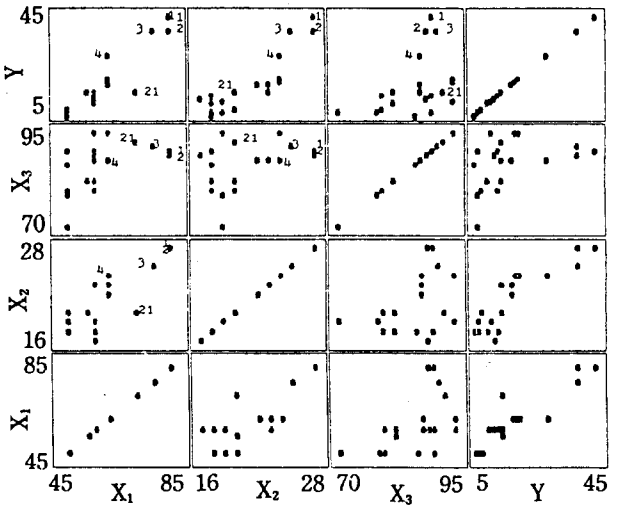

Fig. 4. Scatterplot Matrix of "Stack Loss" Data.

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\varepsilon. \tag{5.2}$$

And try ordinary least squares fitting to the model (5.2). Thus we assume $D_w=I_n$. Now the eigenvalues of (3.3) with $w=1_n$ are 19.72, 10.38, 4.69, and 1.92. Thus the sum of the first two (three) eigenvalues occupies 82% (95%) of the total sum of eigenvalues. At this point, we should decide how many principal axes will be taken into consideration. As in standard principal components analysis, we think there do not exist universally accepted statistical rule for the optimal number of principal axes. Hence it is a matter of analyst's willingness to deal with complexities accompanying higher dimensional treatment. In the remaining analysis, we set the number of principal axis equal to two.

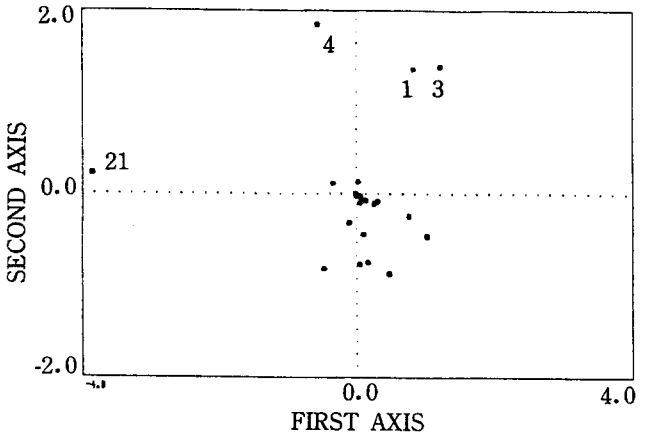

Fig. 5. Case Influence Derivatives for the Least Squares Fit of Stack Loss Data.

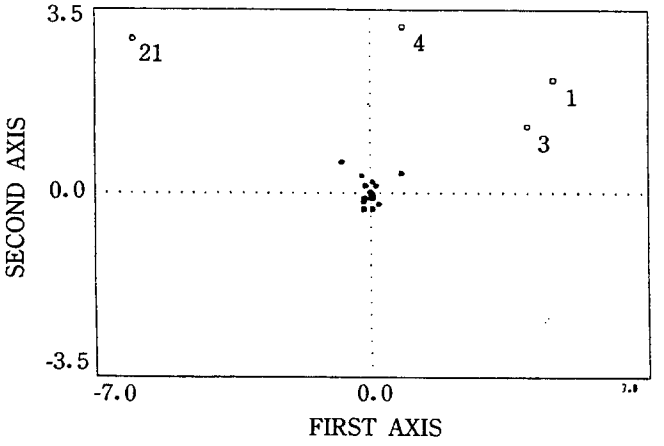

Fig. 6. Case Influence Derivatives for the Robust Fit of Stack Loss Data.
Unfilled circles represent observations that have zero weights in robust fitting.

Second, observation 4 is also influential but in an orthogonal direction to that of observation 21. Third, observations 1 and 3 are jointly influential in a similar direction with that of observation 4. Therefore, the subset of observations 1 and 3 (possibly with observation 4) induces "masking effect".

Next we try robust fitting to the model (5.2), with Andrew's $\varphi(\cdot)$ function as defined in Example 1. For the estimate of σ, we used the median of absolute residuals (from the least squares fit) 0.97. Then we have $\hat{\beta}_w = (-37.14,\ 0.8180,\ 0.5203,\ -0.07250)'$, whereas the least squares estimate $\hat{\beta}$ is $(-39.92,\ 0.7156,\ 1.295,\ -0.1521)'$. As observed in Li(1985), the robust fit is unaffected by rather unusual observations 1, 3, 4, and 21 of which weights are 0. For the influence diagnostics for the fitted model, we compute eigenvalue-eigenvector decomposition of (3.3) with

$$w = (0.00,\ 0.92,\ 0.00,\ 0.00,\ 0.96,\ 0.88,\ 0.99,$$
$$0.96,\ 0.93,\ 1.00,\ 0.95,\ 1.00,\ 0.51,\ 0.84,$$

sum of first two eigenvalues corresponds to 95% of total sum. With the plot of the first two principal components coordinates of case influence derivatives (Fig. 6), we can identify observations 1, 3, 4, and 21 as only *potentially influential*. This reflects one of reliable aspects of statistically robust methodology.

6. Concluding Remarks

We believe our local and graphical approach is effective in identifying possibly multiple influential observations in ordinary and weighted linear regression analysis. Computation is relatively easy compared to solving the combinatorial problem which is inevitable for global influence analysis of regression models.

When our approach is applied heuristically to robust regression, it correctly identifies (potentially) influential observations as was demonstrated from two examples in Section 5. Therefore we think that it is an useful complementary tool following robust fitting of linear regression model. Thus it is an alternative in opposite direction to several variations of robust technology such as bounded influence regression suggested by Mallows(Li, 1985) or least median of squares(LMS) regression(Rousseeuw and Leroy, 1987).

Supplementary Note

After final revision of this article, Professor Yutaka Tanaka of Okayama University pointed out that there exists an independent work of Mr. Sung Ho Moon under the guidence of him. Some results of Section 3 of this paper are very much the same as part of Moon's master course thesis "A Numerical Investigation on Multiple-Case Diagnostics in Regression Analysis"(Okayama University, March 1990). Authors are indepted to Professor Tanaka for the information and comments given during Joint Japanese-Korean Statistics Conference at Fukuoka, Japan in July 1990.

References

1. Atkinson, A.C.(1986). "Masking Unmasked," *Biometrika*, 73, 533-541.
2. Belsley, D.A., Kuh, E., and Welsch, R.E.(1980). *Regression Diagnostics*. Wiley, New York.
3. Brownlee, K.A.(1965). *Statistical Theory and Methodology in Science and Engineering*, 2nd Edition. Wiley, New York.
4. Cook, R.D.(1986). "Assessment of Local Influence," *Journal of Royal Statistical Society*, *B* 48, 133-169.
5. Cook, R.D., and Weisberg, S.(1982). *Residuals and Influence in Regression*. Champman and Hall, New York.
6. Daniel, C., and Wood, F.S.(1980). *Fitting Equations to Data*, 2nd Edition. Wiley, New York.
7. Li, G.(1985). "Robust Regression," in *Exploring Data Tables, Trends, and Shapes*(edited by D.C. Hoaglin, F. Mosteller, and J.W. Tukey). Wiley, New York. 281-343.
8. Mickey, M.R., Dunn, O.J., and Clark, V.(1967). "Note on the Use of Stepwise Regression in Detecting Outliers," *Computers and Biometrical Research*, 1, 105-109.
9. Rousseeuw, P.J., and Leroy, A.M.(1987). *Robust Regression and Outlier Detection*. Wiley, New York.