# Sequential Influence Diagnostics in Multiple Regression
## —축자적 회귀진단 절차의 개발—

김 복 만*
최 성 운**


## ABSTRACT

This paper proposes a new procedures for assessing the influence of individual or groups of cases when any regressors are included. At first, various influence measures pickout influential cases when one regressor is deleted. Next find influential subsets by using heuristic approach and perform group deletion. Retaining or removing any regressors may depend on the presence or absence of one or few cases. Then, we can identify the interrelationships that exist among regressors and cases and examine their impact on the fitted regression equation.

We conclude with an example using fuel consumption data.


## 1. INTRODUCTION

Consider the linear model $Y=X\beta+\varepsilon$, where Y is an n×1 vector of values of response variables, X is an n×p' known and fixed matrix of values of predictors, $\beta$ is a p'×1 unknown but constant vector, and $\varepsilon$ is an n×1 error vector distributed as $N(0, \sigma^2 I)$.

For liner least squares, the vector of ordinary residuals e is given by

$$e = Y - \hat{Y} = (I-H)Y$$

Where $H=(h_{ij})=(X^TX)^{-1}X^T$ and $\hat{Y}\triangleq(y_i)$ is the vector of fitted values. Obsered values of the quantities X and Y are denoted by subscripted lowercase letters : $(x_i, y_i)$ are the observations on X and Y for the ith case in the study. We also use the subscript notion "(i)" or "[j]" to represent the deletion of the ith case or the jth regressor, individually, Thus, for example $X_{(i)}$ is the matrix X with the ith row deleted, and $X_{[j]}$ is the matrix X with the jth column omitted.

We study the interrelationships that exist among regressors and cases examine their impact on the fitted regression equation. To identify this, we propose a new sequential influence diagnastic peocedures. At first, four influence measures pickout influential cases when one regressor is deleted. Next find influential subsets by using heuristic approach. Finally perform group deletion using selected influential subsets.

## 2. Sequential Influence Diagnostics

A new sequential influence diagnostic procedures are represented in Table 1.

### 2.1 Influence Measures When Deleting One Regressor

#### 2.1.1 Externally Studentized Residuals (Cook et al., 1982)

*Dept. of Industrial Engineering, University of Ulsan, Ulsan, Korea.
**Dept. of Industrial Engineering, Kyungwon University, Sungnam, Korea.
접수 : 1992. 4. 25.
확정 : 1992. 5. 2.

$$t_{i[j]}=\frac{e_{i[j]}}{\hat{\sigma}_{(i)[j]}\sqrt{1-h_{ii[j]}}},\ i=1,\ 2,\ \cdots,\ n.\ j=1,\ 2,\ \cdots,\ p. \tag{1}$$

2.1.2 Leverage (Hoaglin et al., 1978)

$$h_{ii[j]}=X_{i[j]}(X_{[j]}^{T}X_{[j]})^{-1}X_{i[j]},\ i=1,\ 2,\ \cdots,\ n.\ j=1,\ 2,\ \cdots,\ p. \tag{2}$$

2.1.3 Cook's Distance (Cook, 1979)

$$D_{i[j]}=\frac{(\hat{\beta}_{(i)[j]}-\hat{\beta}_{[j]})(X_{[j]}^{T}X_{[j]})(\hat{\beta}_{(i)[j]}-\hat{\beta}_{[j]})}{p'_{[j]}\hat{\sigma}^{2}_{[j]}},\ i=1,\ 2,\ \cdots,\ n.\ j=1,\ 2,\ \cdots,\ p. \tag{3}$$

2.1.4 DEFITS(Belsley et al., 1980)

$$DEFITS_{i[j]}=\frac{\hat{Y}_{(i)[j]}-\hat{Y}_{[j]}}{\hat{\sigma}_{(i)[j]}\sqrt{h_{ii[j]}}},\ i=1,\ 2,\ \cdots,\ n.\ j=1,\ 2,\ \cdots,\ p. \tag{4}$$

### 3.1 Group Influence Measures

3.1.1 Externally Studentized Residuals

$$t_I=\frac{e_I}{\hat{\sigma}_{(I)}\sqrt{1-h_{II}}} \tag{5}$$

Where I is an index set corresponding to a subset of cases.

3.1.2 Leverage (Draper et al., 1981)

$$h_{II}=X_I^{T}(X^{T}X)X_I \tag{6}$$

3.1.3 Cook's Distance (Cook 1979, Cook et al., 1980)

$$D_I=\frac{(\hat{\beta}_{(I)}-\hat{\beta})(X^{T}X)(\hat{\beta}_{(I)}-\hat{\beta})}{p'\hat{\sigma}^{2}} \tag{7}$$

**Table 1. Sequential Influence Diagnostic Procedures**

| |
|---|
| Step 1. Pickout influential cases when one regressor is deleted. |
| ↓ |
| Step 2. Find influential subsets by using heuristic approach. |
| ↓ |
| Step 3. Perform group deletion using selected influential subsets. |
| ↓ |
| Step 4. Identify the interrelationships that exist among regress- and cases and examine their impact on the fitted regression equation. |

3.1.4 DEFITS

$$DEFITS_I=\frac{\hat{Y}_{(I)}-\hat{Y}}{\hat{\sigma}_{(I)}\sqrt{h_{II}}} \tag{8}$$

## 4. Example

As an illustrative example, we adopt the data collected by Christoper Bingham from the American Almanac for 1974. The data as presented by Weisberg(1985) are reproduced in Table 2.

The measured variables are : Y=motor fuel consumption (gallons per person), $X_1$=tax(cents per gallon), $X_2$=percent of population with driver's licenses, $X_3$=average income (thousands of dollars), $X_4$=road (thousands of miles).

A liner model is

$$Y=\beta_0+\beta_1X_1+\beta_2X_2+\beta_3X_3+\beta_4X_4+\varepsilon \tag{9}$$

The steps of sequential influence diagnostics are as follows:

Step 1. $t_{i[j]}$, $h_{ii[j]}$, $D_{i[j]}$, and $DEFITS_{i[j]}$ pickout influential cases when one regressor is deleted (table 3)

Step 2. Cases that appear to be most influential are arranged within each measure in descending order of influence (table 3).
The most common sets arranged within each measure are formed influential subset (table 4). For example, |50, 49, 40| is one of numerous influntial subsets.

Step 3. Perform group deletion using selected influential subsets formed by step 2 (table 4).

Step 4. Identify the interrelationships that exist among regressors and cases and examine their impact on the fitted regression equation (table 4). For example, the models with maximum $R^2$ are $X_1$, $X_2$, and $X_3$ with case number 40, 49, and 50 deleted, and $X_2$ and $X_3$ with case number 40, 45, 49, and 50 deleted.

## 5. CONCLUSION

We proposed a new sequential diagnostic procedures for assessing the influence of individual or groups of cases when any regressors are included. Retaining or removing any regressors might depend on the presence or absence of one or few cases. So we studied the interrelationships that exist among regressors and cases. We recommened using this new procedure as a reference materials for variable selection.

**Table 2. Fuel Consumption Data : Weisberg (1985)**

| Case | $X_1$ | $X_2$ | $X_3$ | $X_4$ | Y |
|---|---|---|---|---|---|
| 1 | 9.00 | 52.5 | 3.571 | 1.976 | 541 |
| 2 | 9.00 | 57.2 | 4.092 | 1.250 | 524 |
| 3 | 9.00 | 58.0 | 3.865 | 1.586 | 561 |
| 4 | 7.50 | 52.9 | 4.870 | 2.351 | 414 |
| 5 | 8.00 | 54.4 | 4.399 | 0.431 | 410 |
| 6 | 10.00 | 57.1 | 5.342 | 1.333 | 457 |
| 7 | 8.00 | 45.1 | 5.319 | 11.868 | 344 |
| 8 | 8.00 | 55.3 | 5.126 | 2.138 | 467 |
| 9 | 8.00 | 52.9 | 4.447 | 8.557 | 464 |
| 10 | 7.00 | 55.2 | 4.512 | 8.507 | 498 |
| 11 | 8.00 | 53.0 | 4.391 | 5.939 | 580 |
| 12 | 7.50 | 52.5 | 5.126 | 14.816 | 471 |
| 13 | 7.00 | 57.4 | 4.817 | 6.930 | 525 |
| 14 | 7.00 | 54.5 | 4.207 | 6.580 | 508 |
| 15 | 7.00 | 60.8 | 4.332 | 8.159 | 566 |
| 16 | 7.00 | 58.6 | 4.318 | 10.340 | 635 |
| 17 | 7.00 | 57.2 | 4.206 | 8.508 | 603 |
| 18 | 7.00 | 54.0 | 3.718 | 4.725 | 714 |
| 19 | 7.00 | 72.4 | 4.716 | 5.915 | 865 |

| | | | | | |
|---|---|---|---|---|---|
| 20 | 8.50 | 67.7 | 4.341 | 6.010 | 640 |
| 21 | 7.00 | 66.3 | 4.593 | 7.834 | 649 |
| 22 | 8.00 | 60.2 | 4.983 | 0.602 | 540 |
| 23 | 9.00 | 51.1 | 4.897 | 2.449 | 464 |
| 24 | 9.00 | 51.7 | 4.258 | 4.686 | 547 |
| 25 | 8.50 | 55.1 | 4.574 | 2.619 | 460 |
| 26 | 9.00 | 54.4 | 3.721 | 4.746 | 566 |
| 27 | 8.00 | 54.8 | 3.448 | 5.399 | 577 |
| 28 | 7.50 | 57.9 | 3.846 | 9.061 | 631 |
| 29 | 8.00 | 56.3 | 4.188 | 5.975 | 574 |
| 30 | 9.00 | 49.3 | 3.601 | 4.650 | 534 |
| 31 | 7.00 | 51.6 | 3.640 | 6.905 | 571 |
| 32 | 7.00 | 51.3 | 3.333 | 6.594 | 554 |
| 33 | 8.00 | 57.8 | 3.063 | 6.524 | 577 |
| 34 | 7.50 | 54.7 | 3.357 | 4.121 | 628 |
| 35 | 8.00 | 48.7 | 3.528 | 3.495 | 487 |
| 36 | 6.58 | 62.9 | 3.801 | 7.834 | 644 |
| 37 | 5.00 | 56.6 | 4.040 | 17.782 | 640 |
| 38 | 7.00 | 58.6 | 3.897 | 6.385 | 704 |
| 39 | 8.50 | 66.3 | 3.635 | 3.274 | 648 |
| 40 | 7.00 | 67.2 | 4.345 | 3.905 | 968 |
| 41 | 7.00 | 62.6 | 4.449 | 4.639 | 587 |
| 42 | 7.00 | 56.3 | 3.656 | 3.985 | 699 |
| 43 | 7.00 | 60.3 | 4.300 | 3.635 | 632 |
| 44 | 7.00 | 50.8 | 3.745 | 2.611 | 591 |
| 45 | 6.00 | 67.2 | 5.215 | 2.302 | 782 |
| 46 | 9.00 | 57.1 | 4.476 | 3.942 | 510 |
| 47 | 7.00 | 62.3 | 4.296 | 4.083 | 610 |
| 48 | 7.00 | 59.3 | 5.002 | 9.794 | 524 |
| 49 | 8.00 | 45.2 | 5.162 | 3.246 | 551 |
| 50 | 5.00 | 64.8 | 4.995 | 0.602 | 345 |

**Table 3. Influential Cases When One Variable is Deleted.**

| Regressor deleted | $t_{i[j]}$ | $h_{ii[j]}$ | $D_{i[j]}$ | $DEFITS_{i[j]}$ |
|---|---|---|---|---|
| None | 50, 40, 49<br>② ③ ① | 50, 37, 7, 6, 12<br>④ ③ ⑤ ② ① | 50 | 50, 49, 45, 19<br>④ ③ ② ① |
| $X_1$ | 40, 50, 49, 45<br>① ④ ③ ② | 37, 7, 12, 49, 19<br>② ③ ⑤ ① ④ | 50 | 50, 49, 45, 19<br>④ ③ ② ① |
| $X_2$ | 40, 50, 19, 45<br>③ ① ④ ② | 50, 37, 12, 6, 7, 45<br>④ ⑤ ③ ② ⑥ ① | 50 | 50, 45, 40, 7, 19<br>④ ⑤ ③ ② ① |
| $X_3$ | 50, 40<br>② ① | 37, 50, 19, 7, 20, 12<br>④ ⑥ ③ ⑤ ① ② | 50 | 50, 40, 19, 7<br>④ ③ ② ① |
| $X_4$ | 50, 40, 49<br>② ③ ① | 6, 7, 50, 19, 20<br>① ② ④ ⑤ ③ | 50 | 50, 49, 19<br>③ ② ① |

**Table 4. The $R^2$ Effects of Group Deletion**

| Case deleted | $P_1$* | $P_2$ | $P_3$ | $P_4$ | $R^2$ | Regressor included |
|---|---|---|---|---|---|---|
| None | 0.579 | 0.000 | 0.002 | 0.473 | 0.477 | 2, 3 |
| 7.27 | 0.609 | 0.000 | 0.006 | 0.460 | 0.428 | 2, 3 |
| 37 | 0.592 | 0.000 | 0.003 | 0.514 | 0.473 | 2, 3 |
| 40 | 0.597 | 0.000 | 0.001 | 0.307 | 0.480 | 2, 3 |
| 19, 40, 45 | 0.978 | 0.002 | 0.000 | 0.083 | 0.513 | 2, 3 |
| 49 | 0.774 | 0.000 | 0.000 | 0.249 | 0.546 | 2, 3 |
| 37, 50 | 0.007 | 0.000 | 0.005 | 0.302 | 0.622 | 1, 2, 3 |
| 50 | 0.006 | 0.000 | 0.004 | 0.255 | 0.625 | 1, 2, 3 |
| 37, 49, 50 | 0.012 | 0.000 | 0.000 | 0.546 | 0.677 | 2, 3 |
| 49, 50 | 0.101 | 0.000 | 0.000 | 0.477 | 0.679 | 2, 3 |
| 37, 40, 49, 50 | 0.007 | 0.000 | 0.000 | 0.701 | 0.701 | 1, 2, 3 |
| 40, 49, 50 | 0.006 | 0.000 | 0.000 | 0.634 | 0.704 | 1, 2, 3 |
| 40, 45, 49, 50 | 0.055 | 0.000 | 0.000 | 0.773 | 0.704 | 2, 3 |

* : p value of regressor $X_1$

## References


Belsely, D. A., Kub, E., and Welsch, R. E.(1980), *Regression Diagnostics : Identifying Influential Data & Sources of Collinearity*, New York : John Wiley & Snons.

Cook, R. D.(1977), "Detection of Influential Observations in Linear Regression," *Technometrics*, 19, 15-18.

_____(1979) "Influential Observations in Linear Regression," *J. Amer. Statist. Assoc.*, 74, 164-74.

_ _ _, and Weisberg, S(1980), "Characterizations of an Empirical Influence Function for Detecting Influential in Regression," *Technometrics*, 22, 495-508.

_____(1982), *Residuals and Influence in Regression*, London : Chapman and Hall.

Draper, N. R., and John, J. A.(1981), "Influential Observations and Outliers in Regression," *Technometrics*, 23, 21-26.

Hoaglin, D. C., and Welsch, R.(1978), "The Hat Matrix in Regression and ANOVA," *Amer. Statistician*, 32, 17-22.

Weisberg, S.(1985), *Applied Linear Regression*, 2nd ed., New York : John Weley & Sons.