The Korean Communications
in Statistics Vol. 4, No. 2, 1997
pp. 359 - 365

Graphical Methods for Influence Diagnostics

Dae-Heung Jang[1)]

Abstract

Unusual observations can greatly influence the results of least squares estimation. I propose graphical methods which can detect the influential observations.

1. Introduction

We occasionally find that small data points exert a great influence on the fitted regression model. These observations are regarded as influential if their omission from the data result in substantial changes to important features of an analysis. Thus, many numerical measures of influence have been proposed(See Belsely, Kuh and Welsh(1980) and Cook and Weisberg(1982, 1994).). There are hat diagonals, Mahalanobis distance, DFBETAS, Cook's distance, DFFITS, COVRATIO, FVARATIO among numerical measures of influence. These measures almost consider both the location of the data points and the response variable in measuring influence. Cook and Weisberg(1989) and Hocking(1996) show graphical methods for identifying unusual observations, namely, added-variable plot and principal component plot, respectively. Using singular value decomposition and biplot, I propose graphical methods for detecting the influential observations.

2. Singular value decomposition and biplot

The regression model can be written as

$$\underline{y} = X\underline{\beta} + \underline{\varepsilon}$$

where $\underline{y} = (y_1, y_2, \cdots, y_n)'$ is the vector of observed responses, X is the $n \times p$ model matrix, $\underline{\beta}$ is the the $p \times 1$ vector of parameters which appear in the chosen model, p is the number of parameters in the model, and $\underline{\varepsilon} = (e_1, e_2, \cdots, e_n)'$ is the vector of random errors associated with $\underline{y}$.

By singular value decomposition, $n \times p$ model matrix X is presented as a product of three

1) Professor, Department of Applied Mathematics, Pukyong National University, Pusan, 608-737, Korea.

matrices as follows;

$$X = U\Sigma V'$$

where U is a $n \times r$ matrix which columns are orthonormal eigenvectors of XX', V is a $p \times r$ matrix which columns are orthonormal eigenvectors of $X'X$, and Σ is a diagonal matrix of ordered singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$. Here, r is the rank of X. Let X_{-i} denote the $(n-1) \times p$ reduced model matrix with the ith observation eliminated. Then, by singular value decomposition, $(n-1) \times p$ reduced model matrix X_{-i} is presented as

$$X_{-i} = U_{-i}\Sigma_{-i}V_{-i}'$$

where U_{-i} is a $(n-1) \times r$ matrix which columns are orthonormal eigenvectors of $X_{-i}X_{-i}'$, V_{-i} is a $p \times r$ matrix which columns are orthonormal eigenvectors of $X_{-i}'X_{-i}$, and Σ_{-i} is a diagonal matrix of ordered singular values $\sigma_{1,-i} \geq \sigma_{2,-i} \geq \cdots \geq \sigma_{r,-i} > 0$. I propose the singular values plot as a graphical method for detecting the influential observations. The singular values plot is the plot of singular values σ_j and $\sigma_{j,-i} (j = 1,2,\cdots, r; i = 1,2,\cdots,n)$ against eliminated observation. With this plot, we can find the influential observations. If there is the large change of singular values in singular value decomposition of X_{-i} compared with singular values of X, we can consider this observation as the influential observation. But, the defect of this plot is that we can not detect the influential observations when there is the masking effect in dataset.

The biplot is a graphical display of a data matrix by means of two sets of vectors, the rows and columns of any matrix whose rank is 2. The biplot which is devised by Gabriel(1971), has been studied by many researchers(See Gower and Hand(1996).).

The rank-two approximation X^* to X is

$$X^* = \sigma_1 \underline{u}_1 \underline{v}_1' + \sigma_2 \underline{u}_2 \underline{v}_2',$$

where $\underline{u}_1$ and $\underline{u}_2$ are the first two columns of U, and $\underline{v}_1'$ and $\underline{v}_2'$ are the first two rows of V'. To obtain a biplot, it is necessary to write X^* as the product of two matrices J, K as follows ;

$$X^* = JK'$$

where $J = [\sigma_1 \underline{u}_1, \sigma_2 \underline{u}_2]$ and $K = [\underline{v}_1, \underline{v}_2]$. Then, for the biplot the row markers $\underline{j}_1, \underline{j}_2, \cdots, \underline{j}_n$ are the n rows of J and the column markers $\underline{k}_1, \underline{k}_2, \cdots, \underline{k}_p$ are the p rows of K. If we construct a biplot in which the only row markers are plotted, we can detect the influential observations in X through the pattern of the row markers of this biplot. If any observation is isolated from data points group, we can consider this observation as the influential observation. With this biplot, we can detect the influential observations though there is the masking effect in dataset.

Many numerical measures of influence almost consider both the location of the data points and the response variable in measuring influence. But, the singular values plot and the biplot consider only the location of the data points.

3. Examples

Our first example is taken from Myers(1986). This dataset 1 consists of 25 observations with 7 explanatory variables and one response variable. We can find the influential observations with the singular values plot. Figure 1 shows the singular values plot for dataset 1. In Figure 1, 'o' on the x-axis of singular values plot, means no elimination of observations in the model matrix X. From Figure 1, We can find that the 23th and the 24th observations make the large change of singular values. There are the large changes of first singular value in the 24th observation and 2nd, 3rd, and 4th singular values in the 23th observation. Figure 2 shows the biplot for dataset 2. From Figure 2, we can find that the 23th and the 24th observations are the isolated observations from data points group, namely, the influential observations.

Table 1 shows several numerical measures of influence for the comparison with singular values plot and the biplot. We can find that the 23th and the 24th observations are the influential observations.

Table 1. Several numerical measures of influence in dataset 1

observation	h_{ii}	DFFITS	Cook's D	COVRATIO
1	0.257	-0.043	0.000	2.181
2	0.161	-0.032	0.000	1.931
3	0.161	-0.202	0.005	1.944
4	0.163	-0.078	0.001	1.911
5	0.148	-0.175	0.004	1.745
6	0.159	-0.248	0.008	1.644
7	0.183	0.356	0.016	1.504
8	0.359	-0.353	0.016	2.269
9	0.281	0.203	0.005	2.143
10	0.130	0.035	0.000	1.858
11	0.124	0.585	0.039	0.603
12	0.202	0.223	0.007	1.848
13	0.080	-0.054	0.000	1.737
14	0.097	-0.007	0.000	1.798
15	0.558	-2.828	0.761	0.254
16	0.402	-0.044	0.000	2.714
17	0.368	-1.016	0.123	1.110
18	0.447	1.129	0.154	1.382
19	0.087	0.311	0.012	1.087
20	0.366	-2.179	0.417	0.093
21	0.070	0.565	0.034	0.269
22	0.785	-3.072	1.079	2.290
23	0.989	-48.518	115.041	0.047
24	0.876	8.537	5.889	0.246
25	0.547	0.472	0.029	3.269

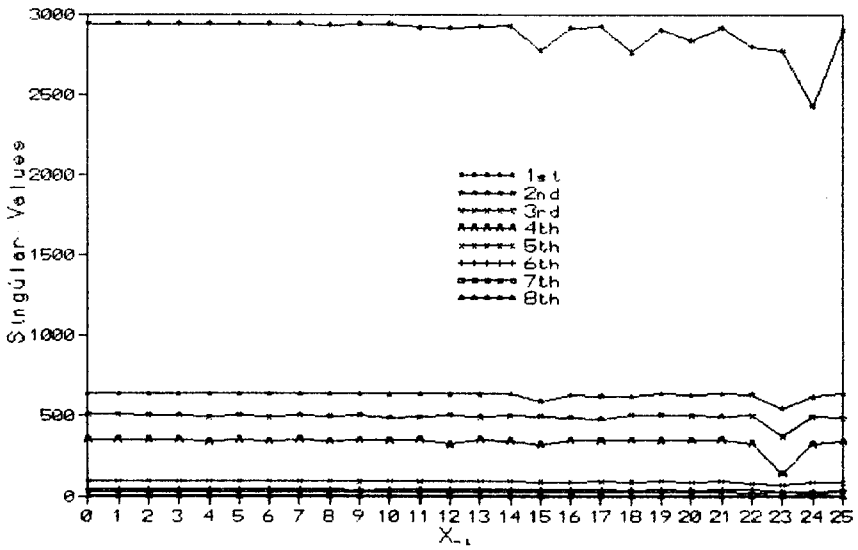

FIGURE 1. SINGULAR VALUES PLOT FOR THE DATASET 1

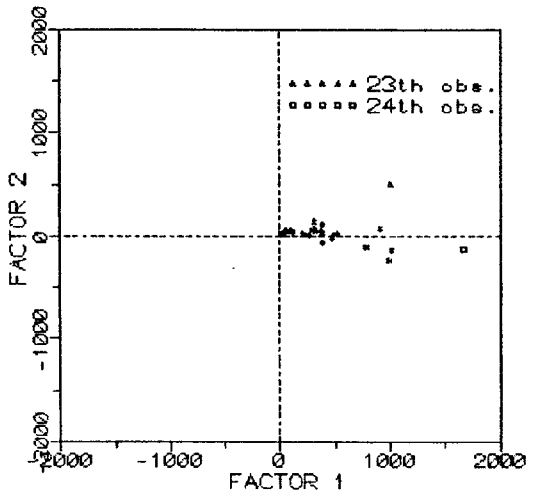

FIGURE 2. BIPLOT FOR THE DATASET 1

Our second example is taken from Hocking and Pendleton(1983). This dataset 2 consists of 27 observations with 3 explanatory variables and one response variable. Table 2 shows several numerical measures of influence. We doubt that the 24th and the 27th observations are the influential observations. But, when the 24th and 27th observations are deleted in diagnostics, respectively, i.e. we consider only the 26 observations for diagnostics, we can find that the 24th and the 27th observations are the influential observations, respectively. Thus, we conclude that the 24th and the 27th observations are influential if the other observation is not included in dataset, but this effect is masked when both are included in dataset. Table 3 shows this fact. The masking is especially dramatic in Cook's distance.

Table 2. Several numerical measures of influence in dataset 2

observation	h_{ii}	DFFITS	Cook's D	COVRATIO
1	0.2226	-0.4202	0.045	1.3759
2	0.0921	0.1784	0.008	1.2434
3	0.0401	0.0156	0.000	1.2432
4	0.0402	-0.0788	0.002	1.2116
5	0.0504	-0.1022	0.003	1.2140
6	0.1139	0.3865	0.037	1.0974
7	0.0662	-0.0728	0.001	1.2621
8	0.2576	0.1736	0.008	1.5839
9	0.1077	0.0624	0.001	1.3309
10	0.1529	-0.1287	0.004	1.3869
11	0.1668	-0.4974	0.061	1.1522
12	0.0549	-0.2704	0.018	1.0117
13	0.1709	-0.3020	0.023	1.3305
14	0.1651	-0.0659	0.001	1.4252
15	0.1077	0.0022	0.000	1.3387
16	0.1772	-0.1386	0.005	1.4285
17	0.0387	1.3886	0.159	0.0122
18	0.2309	-0.0620	0.001	1.5496
19	0.1540	-0.5389	0.071	1.0672
20	0.0775	0.2384	0.014	1.1471
21	0.1067	0.0931	0.002	1.3198
22	0.0538	0.0732	0.001	1.2410
23	0.0473	-0.0796	0.002	1.2252
24	0.4674	0.0387	0.000	2.2424
25	0.1410	0.0927	0.002	1.3776
26	0.0940	-0.2332	0.014	1.1999
27	0.6024	-0.1297	0.004	2.9988

Table 3. Diagnostics for the deleted dataset 2

observations (deleted observation)	h_{ii}	Cook's D
24 (27)	0.97	3.17
27 (24)	0.98	4.38

With the biplot, we can overcome the masking effect in influential observations. Figure 3 shows the biplot for dataset 2. From Figure 3, we can find that the 24th and the 27th observations are the isolated observations from data points group, namely, the influential observations.

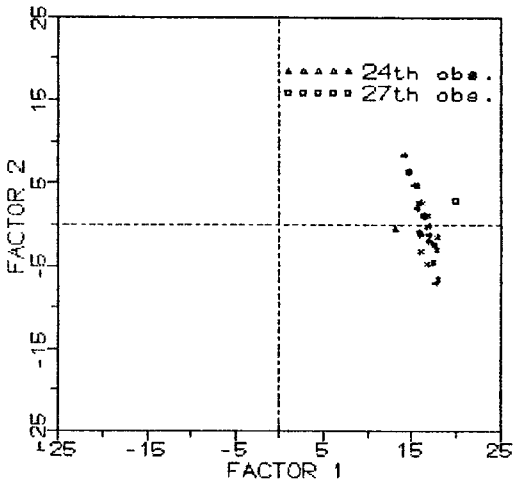

FIGURE 3. BIPLOT FOR THE DATASET 2

4. Conclusion

I propose graphical methods for detecting the influential observations. We can detect the influential observations through the singular values plot and the biplot.

References

[1] Belsley, D. A., Kuh, E., and Welsch, R. E.(1980). *Diagnostics : Identifying Influential Data and Sources of Collinearity*, New York : John Wiley.

[2] Cook, R. D., and Weisberg, S.(1982). *Residuals and Influence in Regression*, London : Chapman and Hall.

[3] Cook, R. D., and Weisberg, S.(1989). Regression Diagnostics with dynamic graphics, *Technometrics*, 31, 277-311.

[4] Cook, R. D., and Weisberg, S.(1994). *An Introduction to Regression Graphics*, New York : John Wiley.

[5] Gabriel, K. R.(1971). The biplot-graphic display of matrices with application to principal component analysis, *Biometrika*, 58, 453-467.

[6] Gower, J. C. and Hand, D. J.(1996). *Biplots*, London : Chapman and Hall.

[7] Hocking, R. R.(1996). *Methods and Applications of Linear models*, New York : John Wiley.

[8] Hocking, R. R. and Pendleton, O. J.(1983). The regression dilemma, *Communications in Statistics - Theory and Methods*, 12, 497-527.

[9] Myers, R. H.(1986). *Classical and Modern Regression with Applications*, Boston : Duxbury Press.