HONAM MATHEMATICAL JOURNAL
Volume 5, Number 1, September, 1983

The Methods for examing Observations in the transformed Regression

By Byung-Yub Cho & Ju-Sun Ahn

1. Introduction

In recent years there has been considerable interest in finding the outlying or the influential observations in regression. The plots of the residuals and the examination of standardized residuals or of studentized residuals have been used to find those observations.

Behnken & Draper (1972) also have illustrated the estimated variances of the residuals contain more relevant informations than that funished by residual plots or studentized residuals.

The quantities used to find such observations have been proposed by Cook (1977) and Andrews & Pregibon (1978). By these methods and quantities we may not detect which explanatory variables can provide the greatest influence for the residuals and which observations can have the same properties clustering them.

It is the purpose of this paper to describe and exemplify the methods which detect such outliers and explanatory variables and which find clusters of the observations.

The residuals and statistics are discussed to compare the methods of this paper with that of other paper in section 2. The transformations of variables and parameters are introduced in section 3. The methods for displaying the transformed regression and for examing observations is discussed in section 4. And we illustrate the plots of the transformed regression as an example whose data have two explanatory variables in section 5.

2. Residuals and statistics

In the following sections a linear model

$$\underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{e}$$

is discussed where $\underset{\sim}{Y}$ is an $n\times 1$ vector of response datas (or dependent variables), X is an $n\times p$ full rank matrix of known constants (or independent variables), $\underset{\sim}{\beta}$ is a $p\times 1$ vector of unknown parameters and $\underset{\sim}{e}$ is an $n\times 1$ vector of randomly distributed errors (or noise) such that $E(\underset{\sim}{e})=\underset{\sim}{O}$ and $V(\underset{\sim}{e})=I\sigma^2$.

The least squares estimate of $\underset{\sim}{\beta}$ is given by

$$\underset{\sim}{b}=(X'X)^{-1}X'\underset{\sim}{Y}$$

The corresponding residual vector is

$$\begin{aligned}\underset{\sim}{R}=(R_k)&=\underset{\sim}{Y}-\underset{\sim}{\hat{Y}}=\underset{\sim}{Y}-X\underset{\sim}{b},\quad k=1,\cdots,n\\&=[I-M]\underset{\sim}{Y}=[I-M]\underset{\sim}{e}\end{aligned}$$

where $M=X(X'X)^{-1}X'$.

The expectations of $\underset{\sim}{Y}$, $\underset{\sim}{b}$ and $\underset{\sim}{R}$ are respectively

$$\begin{aligned}&E(\underset{\sim}{Y})=X\underset{\sim}{\beta}\ (\text{or } E(\underset{\sim}{\hat{Y}})=X\underset{\sim}{b}=\underset{\sim}{\hat{Y}}),\\&E(\underset{\sim}{b})=E[(X'X)^{-1}X'\underset{\sim}{Y}]=\underset{\sim}{\beta},\\&\text{and } E(\underset{\sim}{R})=\underset{\sim}{O}.\end{aligned}$$

The covariance matrices of $\underset{\sim}{\hat{Y}}$, $\underset{\sim}{b}$ and $\underset{\sim}{R}$ are respectively

$$\begin{aligned}&V(\underset{\sim}{\hat{Y}})=X(X'X)^{-1}X'\sigma^2=M\sigma^2,\\&V(\underset{\sim}{b})=(X'X)^{-1}\sigma^2,\\&\text{and } V(\underset{\sim}{R})=E(\underset{\sim}{R}\underset{\sim}{R}')=(I-M)\sigma^2.\end{aligned}$$

For the kth observation, it follows that

$$\begin{aligned}&\hat{y}_k=\underset{\sim}{x_k}'\underset{\sim}{b},\\&V(\hat{y}_k)=\underset{\sim}{x_k}'(X'X)^{-1}\underset{\sim}{x_k}\sigma^2=v_k\sigma^2,\\&\text{and } V(R_k)=(1-v_k)\sigma^2.\end{aligned}$$

where $\underset{\sim}{x_k}$ is the vector of explanatory variables for the kth observation and $v_k=\underset{\sim}{x_k}'(X'X)^{-1}\underset{\sim}{x_k}$.

Since the average value of the variances of the predicted values at the observation points is given by

$\overline{V(\hat{y})}=\sum_{k=1}^{n}V(\hat{y}_k)/n=\mathrm{tr}\{X(X'X)^{-1}X'\sigma^2\}/n=\dfrac{p\sigma^2}{n}$ where X is an $n\times p$ matrix, the average variance of the residuals is

$$\overline{V(R)}=\sum_{k=1}^{n}V(R_k)/n=\frac{(n-p)\sigma^2}{n}$$

The standardized residuals N_k and the studentized residuals T_k are defined as

$$N_k = \frac{R_k}{\{(n-p)\widehat{V(R_k)}\}^{1/2}} = \frac{R_k}{S(1-v_k)^{1/2}}$$

$$T_k = \frac{R_k}{\{\widehat{V(R_k)}\}^{1/2}} = \sqrt{(n-p)}N_k$$

where $S^2 = R'R$, the residual sum of squares and v_k is the kth diagonal element of $X(X'X)^{-1}X'[b]$.

If the variances of the residuals varied a great deal, it would be worthwhile to examine the $\frac{R_k}{(1-v_k)^{1/2}}$ instead of R_k in the usual residuals plots and to use N_k instead of $\frac{R_k}{S}$ as the "normal deviate form" of the residuals.

Behnken and Draper (1972) have recommended the studentized residuals T_k as more appropriate than the standardized residuals.

A wide variation in the $V(R_k)$ reflects a peculiarity of the X matrix, namely a nonhomogeneous spacing of the observations and will thus often direct attention to data deficiencies. From the standpoint of detection outliers, it is always important to recognize the changing magnitude of $V(R_k)$.

Even if the residual variance is reasonably constant, it will still be useful to consider the magnitude of R_k relative to $s\left\{\frac{n-p}{n}\right\}^{1/2}$ rather than to s (the estmate of σ) in cases where n is not large relative to p.

Andrews-Pregibon statistic is defined as

$$R^k_{ij\cdots}(X^*) = \frac{D^{(k)}_{ij\cdots}|X^{*\prime}X^*|}{|X^{*\prime}X^*|}$$

where $|X^{*\prime}X^*| = R'R \cdot |X'X|$ and $D^{(k)}_{ij\cdots}|X^{*\prime}X^*|$ are the operator which deletes the elements associated with the K observations from $|X^{*\prime}X^*|$.

If the deletion of an observation has a large (small) effect on $|X'X|$, the observation has a large (small) influence on the resulting estimates. Also the more a particular observation deviates from the fitted values, the more the deletion of this observation will reduce the residual sum of squares $R'R$. By combining these two ideas, small values of $R^k_{ij\cdots}(X^*)$ are associated with deviant and/or influential observations.

Cook statistic is defined as

$$D_k = \frac{T_k^2}{p} \cdot \frac{V(\hat{Y}_k)}{V(R_k)}$$

Note that D_k depends on three relevant quantities the number of parameters, p, the studentized residual, T_k which is a measure of the degree to consider kth observation as an outlier from the assumed model, and the ratio $V(\hat{Y}_k)/V(R_k)$ which measure the relative sensitivity of the estimate, $\underset{\sim}{b}$, to potential outlying values at each data point.

3. Transformations

Suppose that the predicted values are given as follows

$$\hat{y}_k = \underset{\sim}{x}_k' \underset{\sim}{b}$$

where $\underset{\sim}{x}_k' = (x_{1k}, x_{2k}, \cdots, x_{pk})$, $k=1, 2, \ldots, n$ and $\underset{\sim}{b}' = (b_1, \ldots, b_p)$.

The corresponding residual is

$$R_k = y_k - \underset{\sim}{x}_k' \underset{\sim}{b}.$$

Let suppose that $b_j > 0$ for all j and $x_{jk} \neq x_{jk'}$ for some $k \neq k'$.

Transform the given y_k and x_k' by $p+1$ real valued functions f_j, $j=0, 1, \ldots, p$ as follows:

$$w_{ok} = f_o(y_k) = \frac{y_k - y_l}{y_u - y_l}\pi \qquad k=1, 2, \cdots, n \qquad \cdots\cdots(1)$$

$$w_{jk} = f_j(x_{jk}) = \frac{x_{jk} - x_{jl}}{x_{ju} - x_{jl}}\pi \qquad \cdots\cdots(2)$$

where $y_u = \max_{1 \leq k \leq n} y_k$, $y_l = \min_{1 \leq k \leq n} y_k$, $x_{ju} = \max_{1 \leq k \leq n} x_{jk}$ and $x_{jl} = \min_{1 \leq k \leq n} x_{jk}$

Consequently, the f_j $(f_j = 0, 1, \cdots, p)$ satisfy the following conditions

(A) $0 \leq f_j(x_{jk}) \leq \pi$ $\cdots\cdots(3)$

(B) f_j is a strictly monotone function in x_{jk}.

Transform the given $\underset{\sim}{b}' = (b_1, \cdots, b_p)$ by d_j, $j=1, 2, \cdots, p$, as follows:

$$d_j = b_j \Big/ \sum_{j=1}^{p} b_j \qquad \cdots\cdots(4)$$

such that $\sum_{j=1}^{p} d_j = 1$ $\cdots\cdots(5)$

From the result of the above transformations (1), (2), and (4), complex numbers C_k' and C_k are corresponded to y_k and $\hat{y}_k = \underset{\sim}{x}_k' \underset{\sim}{b}$, as follows

$$C_k' = \exp(iw_{ok}) \qquad \cdots\cdots(6)$$

$$C_k = \sum_{j=1}^{p} d_j \exp(iw_{jk}) = \sum_{j=1}^{p} d_j (\cos w_{jk} + i \sin w_{jk}) \quad k=1, 2, \cdots, n \qquad \cdots\cdots(7)$$

where $i=\sqrt{-1}$ and d_j is the weight to be assigned to jth variable and satisfies (4). From (3) and (5), we can easily see arguments and absolutes of C'_K and C_K as follows:

(C) $0 \leqq \arg(C'_K) \leqq \pi, \quad 0 \leqq \arg(C_K) \leqq \pi$ ……(8)

(D) $|C'_K| = 1, \qquad |C_K| \leqq 1$ ……(9)

where $|C_K| = 1$ if and only if $w_{jk} = w_{j'k}$ for all j and j'.

4. The method for examing observations

From (6), (7), (8), and (9), we now establish the methods for displaying the transformed regression.

〔Method 1〕 : Display C'_K on the circumference of the upper half of the unit circle in the complex plane according to (6) and C_K in and on it according to (7) where C_K is displayed on it if and only if w_{jk}'s have same angles for all j. In this displayed figure, we can find clusters of the observations and examine the distance between the two points C'_K and C_K instead of R_K, N_K or T_K.

〔Method 2〕 : Display C'_K on the circumference of the upper half of the unit circle, and then from this point display C_K to the opposite direction of C'_K.

In this displayed figure, we can examine the distance of the point C_K from the origin of the complex plane. 〔Method 2〕 is equivalent to display $r_k = C'_K - C_K$ which move C_K of $\overrightarrow{C_K C'_K}$ to the origin.

In case of $b_j < 0$ for some j, since $b_j x_{jk} = (-b_j)\ (-x_{jk})$, $k = 1, \cdots n$, we may change the sign of all values of jth variable and then use the plocedure of section 3 and this section.

In case of $x_{jk} = x_{jk'}$ for all k and k', this cases occur in intercept models, without loss of generality we may change such intercept models to nonintercept models by taking the difference $y_k - b_j x_{jk}$ instead of response data y_k, $k = 1, \cdots n$. Therefore, we can pick out the intercept term because of transfomation (1).

5. Example

This example uses data given by Hald[3]. A fitted first oder model of this data is $Y=52.577+1.468\ X_1+0.662\ X_2$. Table 1 lists x_{1k}, x_{2k}, $y_k-52.577$, R_k, R_k/s, T_k and D_k.

〈Table 1〉 The 13 observations, residuals and Cook statistics

observation (k)	x_{1k}	x_{2k}	$y_k-52.577$	R_k	R_k/s	T_k	D_k
1	7	26	26.068	-1.6	-0.66	-0.77	0.014
2	1	29	21.666	1.0	0.42	0.48	0.008
3	11	56	51.72	-1.5	-0.62	-0.66	0.026
4	11	31	34.97	-1.7	-0.71	-0.81	0.05
5	7	52	43.3	-1.4	-0.58	-0.61	0.018
6	11	55	56.558	4.0	1.66	1.77	0.186
7	8	71	57.446	-1.3	-0.54	-0.68	0.053
8	1	31	19.89	-2.1	-0.87	-1.00	0.04
9	2	54	40.484	1.8	0.75	0.83	0.094
10	21	47	63.342	1.4	0.58	0.87	0.536
11	1	40	31.248	3.3	1.37	1.52	0.174
12	11	60	56.768	0.9	0.37	0.42	0.014
13	10	68	56.796	-2.9	-1.20	-1.36	0.154

Figure 1 is to display C_k' and C_k corresponding to these given observations and the representing way of observation # 5 by [Method1]. From figure 1, it is seen that the observation # 6 which has the greatest R_k(or $|T_k|$) has the largest distance of C_k from C_k' and since $m(\angle OQR)$ is greater than $m(\angle QC_kP)$, we can see that the varible X_1 provides more influence than the variable X_2 about the observation #5. Also, we can see that the observation #10 which takes the greatest value in the Cook's statistics D_k's deviates from the other observations.

Figure 2 is to display r_k by [Method 2] and here we may find the magnitude of residuals.

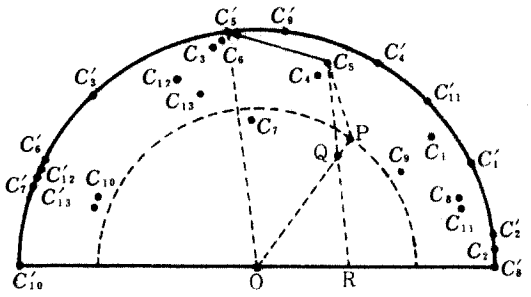

Fig.1 The Graph of C'_K, C_K by Method 1.

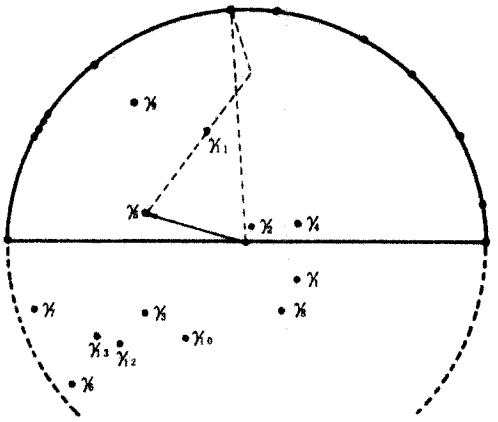

Fig.2 The Graph of γ_K by Method 2.

References

[1] Andrews. D. F. and pregibon. D (1978). *Finding the outliers that matter. J. Roy. Statist. Soc. Ser. B40, 85 – 93*

[2] Atkinson. A. C. (1981). *Two graphical displays for outying and influential observations in regression. Biometrica 68, 1, 13-20*

[3] Behnken. D. W. and Draper. N. R. (1972). *Residuala and their variance patterns. Technometrics 14, 102-111*

[4] Cook. R. D. (1977). *Detection of influential observations in linear regression. Technometrics 19, 15-18*

[5] Doornbos. R. (1981). *Testing for a single outlier in a linear model. Biometrics 37, 705-711*

[6] Draper. N. R. and John. J. A. (1981). *Influetial observations and outliers in regression. Technometrics 23, 21-28*

[7] Searle. S. R. (1971). *Linear models. John Wiley & Sons*

[8] Wakimoto. K. and Taguri. M. (1978). *Constellation graphical method for representing multi-dimensional data. Ann. Inst. Statist. Math. 30, A97-104*