ON DETECTING INFLUENTIAL DATA AND SELECTING REGRESSION VARIABLES*

by

Shanti S. Gupta and Deng-Yuan Huang
Purdue University Feng Chia University

Technical Report #89-28C

Department of Statistics
Purdue University

December 1989
Revised February 1992
Revised August 1992
Revised January 1993
Revised March 1993
Revised July 1993
Revised July 1994
Revised August 1994
Revised October 1994
Revised March 1995

* This research was supported in part by NSF Grants DMS–8606964, DMS–8702620, DMS–8923071 at Purdue University.

◇

ON DETECTING INFLUENTIAL DATA AND SELECTING REGRESSION VARIABLES*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
Feng Chia University

Abstract

The analysis of residuals may reveal various functional forms suitable for the regression model. In this r, we investigate some selection criteria for selecting important regression variables. In doing so, we use statistical selection and ranking procedures. Thus, we derive an appropriate criterion to measure the influence and bias for the reduced models. We show that the reduced models are based on some non-centrality parameters which provide a measure of goodness of fit for the fitted models. In this paper, we also discuss the relationships of influence diagnostics and the statistic proposed earlier by Gupta and Huang (1988). We introduce a new measure for detecting influential data as an alternative to Cook's measure.

AMS 1991 Subject Classification: Primary 62F07; Secondary 62J05, 62J20

Key Words: linear model; influential data; selection criteria; inferior models.

* This research was supported in part by NSF Grants DMS–8606964, DMS–8702620, DMS–8923071 at Purdue University.

On Detecting Influential Data and Selecting Regression Variables*

by

Shanti S. Gupta
Purdue University

Deng-Yuan Huang
Feng Chia University

1. Introduction

We consider the following linear model

$$\underline{Y} = X\underline{\beta} + \underline{e}, \tag{1}$$

where $\underline{e} \sim N(\underline{0}, \sigma_0^2 I_n)$, I_n denotes the identity matrix of order n, $\underline{Y}$ is an $n \times 1$ vector of responses, X is an $n \times p$, $(n > p)$, matrix of known constants of rank p, $\underline{\beta}$ is a $p \times 1$ parameter vector. Several authors have studied the influence on the fitted regression line when the data are deleted. Let $\underline{\hat{\beta}}$ be the usual least squares estimator of $\underline{\beta}$ based on the full data and let $\underline{\hat{\beta}}_A$ be an alternative least squares estimator based on a subset of the data. The empirical influence function for $\underline{\hat{\beta}}$, IF_A is defined to be

$$IF_A = \underline{\hat{\beta}}_A - \underline{\hat{\beta}}. \tag{2}$$

For a given positive definite matrix M and a nonzero scale factor c, Cook and Weisberg (1980) defined the distance $D_A(M, c)$ between $\underline{\hat{\beta}}$ and $\underline{\hat{\beta}}_A$ as follows:

$$D_A(M, c) = \frac{(IF_A)' M (IF_A)}{c}. \tag{3}$$

They suggest that the matrix M can be chosen to reflect specific interests.

They pointed out that in some applications, measurement of the influence of cases on the fitted values, $\underline{\hat{Y}} = X\underline{\hat{\beta}}$, may be more appropriate than measuring influence on $\underline{\hat{\beta}}$. They mentioned an example to describe the fact that if prediction is the primary goal it may be convenient to work with a reparameterized model where the regression coefficients are not of interest. They tried to treat their measurement of the influence on the fitted values $X\underline{\hat{\beta}}$ and used the empirical influence function for $\underline{\hat{Y}}$ defined by $X(IF_A)$. In this paper, we attempt to measure the influence on residuals and on $X\underline{\hat{\beta}}$. The large influence on the residual should have much influence on $\underline{\hat{\beta}}$ though the converse may not hold. Furthermore,

* This research was supported in part by NSF Grants DMS–8606964, DMS–8702620, and DMS–8923071 at Purdue University.

Welsch (1982) pointed out that in an earlier paper Cook (1977) chose to measure influence by

$$D = \frac{(\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})' X' X (\hat{\underline{\beta}} - \hat{\underline{\beta}}_{(i)})}{s^2 p} \tag{4}$$

where s^2 is the residual mean square for full data and $\hat{\underline{\beta}}_{(i)}$ is the least squares estimator of $\underline{\beta}$ based on the data set with the ith component in $\underline{Y}$ deleted. Welsch (1982) gave an example to explain that when all of the observations but one lie on a line, (4) can give potentially confusing information since it may indicate that some observations on the line are more influential than the one observation not on the line. This is counterintuitive since the deletion of this one observation leads to a perfect fit. Therefore, finding a more reasonable measurement is very important. We propose a new statistic to measure the influential data and make a comparison with Cook's D. We shall consider the case of one at a time data deletion. Since, for the case of deletion of a subset, computations can be similarly carried out, we refer to Cook and Weisberg (1980), and Gray and Ling (1984).

Next we propose a selection criterion to combine the influence measure and variable selection. We derive a suitable choice of M and c in (3) to measure the influence and bias for the reduced model. Then, inferior reduced models can be determined. An example (Daniel and Wood (1980)) is studied to explain the idea for the proposed criterion.

In this paper, we discuss the analysis of the model structure, and try to obtain reasonable models.

2. Influential Observations in Linear Regression Model

Let $\underset{n\times p}{X} = \binom{X_{(i)}}{\underline{X}_i'}_{1\times p}^{(n-1)\times p}$, $\underset{n\times 1}{\underline{Y}} = \binom{\underline{Y}_{(i)}}{Y_i}_{1\times 1}^{(n-1)\times 1}$, $\hat{\underline{Y}} = X\hat{\underline{\beta}}$, $\underline{e} = \binom{\underline{e}_{(i)}}{e_i}_{1\times 1}^{(n-1)\times 1}$, $e_i = Y_i - \underline{X}_i'\underline{\beta}$, $i = 1, 2, \ldots, n$, $\hat{\underline{\beta}} = (X'X)^{-1}X'\underline{Y}$. Then

$$\begin{aligned}
\underline{e}_{(i)}'\underline{e}_{(i)} &= (\underline{Y}_{(i)} - X_{(i)}\underline{\beta})'(\underline{Y}_{(i)} - X_{(i)}\underline{\beta}) \\
&= (\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)})'(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)}) \\
&\quad + (X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta})'(X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta}),
\end{aligned}$$

Thus

$$\underline{e}_{(i)}'\underline{e}_{(i)} = SSE_{(i)} + R_{(i)},$$

where

$$\begin{aligned}
SSE_{(i)} &= (\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)})'(\underline{Y}_{(i)} - X_{(i)}\hat{\underline{\beta}}_{(i)}) \\
R_{(i)} &= (\hat{\underline{\beta}}_{(i)} - \underline{\beta})' X_{(i)}' X_{(i)} (\hat{\underline{\beta}}_{(i)} - \underline{\beta})
\end{aligned}$$

and

$$SSE_{(i)} = \inf_{\underline{\beta}} \underline{e}'_{(i)}\underline{e}_{(i)}.$$

We have,

$$\frac{\underline{e}'_{(i)}\underline{e}_{(i)}}{SSE_{(i)}} = 1 + \frac{R_{(i)}}{SSE_{(i)}}. \tag{5}$$

Define

$$D_{(i)} = \frac{R_{(i)}}{ps^2_{(i)}}, \text{ where } s^2_{(i)} = \frac{1}{n-p-1} SSE_{(i)}. \tag{6}$$

If $D_{(i)}$ is large, we see from (5) that deleted i-th data will heavily influence the fitted line. Since

$$\begin{aligned} R_{(i)} &= (\hat{\underline{\beta}}_{(i)} - \underline{\beta})' X'_{(i)} X_{(i)} (\hat{\underline{\beta}}_{(i)} - \underline{\beta}) \\ &= (X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta})'(X_{(i)}\hat{\underline{\beta}}_{(i)} - X_{(i)}\underline{\beta}), \end{aligned} \tag{7}$$

it is the Euclidean distance between $X_{(i)}\underline{\beta}$ and its ordinary least square (OLS) estimate $X_{(i)}\hat{\underline{\beta}}_{(i)}$.

We define a statistic $\hat{D}_{(i)}$ to measure the influence in (5) for the fitted line as follows:

$$\hat{D}_{(i)} = \frac{\hat{R}_{(i)}}{ps^2_{(i)}} = \frac{(\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}})' X'_{(i)} X_{(i)} (\hat{\underline{\beta}}_{(i)} - \hat{\underline{\beta}})}{ps^2_{(i)}}. \tag{8}$$

Let the hat matrix

$$H = \underset{n\times p}{X} (X' \underset{p\times p}{X})^{-1} \underset{p\times n}{X}' = (h_{ij})_{n\times n},$$

where

$$h_{ij} = \underline{X}'_i (X'X)^{-1} \underline{X}_j, \quad i,j = 1,\ldots,n.$$

Since $tr(H) = p, H' = H$ and $H^2 = H$, hence $\sum_{j=1}^{n} h_{ij}^2 = h_{ii}, 0 \le h_{ii} \le 1, i = 1,\ldots,n$ and $\sum_{i=1}^{n} h_{ii} = p$. We have

$$\begin{aligned} (X'_{(i)}X_{(i)})^{-1} &= (X'X - \underline{X}_i\underline{X}'_i)^{-1} \\ &= (X'X)^{-1} + \frac{(X'X)^{-1}\underline{X}_i\underline{X}'_i(X'X)^{-1}}{1-h_{ii}}, \\ \underline{X}'_i(X'_{(i)}X_{(i)})^{-1}\underline{X}_i &= \frac{h_{ii}}{1-h_{ii}}. \end{aligned} \tag{9}$$

Thus

$$\begin{aligned}\underline{\hat{\beta}}_{(i)} &= (X'_{(i)}X_{(i)})^{-1}X'_{(i)}\underline{Y}_{(i)} \\ &= (X'X - \underline{X}_i\underline{X}'_i)^{-1}(X'\underline{Y} - \underline{X}_iY_i) \\ &= \underline{\hat{\beta}} - \frac{(X'X)^{-1}\underline{X}_i\hat{e}_i}{1-h_{ii}}, \text{ where } \hat{e}_i = Y_i - \underline{X}'_i\underline{\hat{\beta}},\end{aligned}$$

hence

$$\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}} = -\frac{(X'X)^{-1}\underline{X}_i\hat{e}_i}{1-h_{ii}}. \tag{10}$$

The i-th predicted residual is

$$\hat{e}_{(i)} = Y_i - \underline{X}'_i\underline{\hat{\beta}}_{(i)}, \quad i = 1, 2, \ldots, n. \tag{11}$$

Then

$$\begin{aligned}\hat{e}_{(i)} &= Y_i - \underline{X}'_i\underline{\hat{\beta}}_{(i)} = Y_i - \underline{X}'_i\left(\underline{\hat{\beta}} - \frac{(X'X)^{-1}\underline{X}_i\hat{e}_i}{1-h_{ii}}\right) \\ &= \hat{e}_i + \frac{h_{ii}\hat{e}_i}{1-h_{ii}} = \frac{\hat{e}_i}{1-h_{ii}},\end{aligned} \tag{12}$$

and

$$\text{Var}\,(\hat{e}_{(i)}) = \frac{\text{Var}\,(\hat{e}_i)}{(1-h_{ii})^2} = \frac{\sigma^2}{(1-h_{ii})}.$$

Thus, we can obtain the following result,

$$\begin{aligned}\hat{R}_{(i)} &= (\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}})'X'_{(i)}X_{(i)}(\underline{\hat{\beta}}_{(i)} - \underline{\hat{\beta}}) \\ &= [-\frac{(X'X)^{-1}\underline{X}_i\hat{e}_i}{1-h_{ii}}]'X'_{(i)}X_{(i)}[-\frac{(X'X)^{-1}\underline{X}_i\hat{e}_i}{1-h_{ii}}] \\ &= (1-h_{ii})h_{ii}\hat{e}^2_{(i)} = (\sum_{j\neq i}^{n} h^2_{ij})\hat{e}^2_{(i)}.\end{aligned} \tag{13}$$

Since $\underline{\hat{Y}} = H\underline{Y}$, hence $\hat{Y}_i = h_{i1}Y_1 + h_{i2}Y_2 + \ldots + h_{ii}Y_i + \ldots + h_{in}Y_n$. Also H is a function of X only, and so $h'_{ij}s$ are fixed. If we now fix Y_i, then $\hat{R}_{(i)}$ is a measure of the influence of Y'_js, $j \neq i$, or $\hat{Y}_i$. Similarly, one can consider conditional influence of Y_i on $\hat{Y}_i$. Now

$$\hat{D}_{(i)} = \left\{\left[\frac{\hat{e}_{(i)}}{s_{(i)}\cdot\frac{1}{(1-h_{ii})^{\frac{1}{2}}}}\right]^2 h_{ii}\right\}\frac{1}{p}. \tag{14}$$

The first factor in (14), within the square brackets, is the studentized residual, that is, the residual divided by its standard error based on a fit to the data with the i-th case excluded, while the second factor is the leverage of the variance of the i-th predicted value.

Since $\hat{\underline{\beta}}_{(i)}$ and $s^2_{(i)}$ are independent (cf. Graybill (1976)), and Y_i and $s^2_{(i)}$ are also independent, it follows that $\hat{e}_{(i)}$ and $s^2_{(i)}$ are independent. Thus

$$t_i = \frac{\hat{e}_{(i)}}{s_{(i)} \cdot \frac{1}{(1-h_{ii})^{\frac{1}{2}}}} = \frac{\hat{e}_{(i)}}{\frac{s_{(i)}}{\sigma} \cdot \frac{\sigma}{(1-h_{ii})^{\frac{1}{2}}}}$$

is t-distribution with $n-p-1$ degrees of freedom. Hence t_i^2 is F-distributed with degrees of freedom 1 and $n-p-1$.

For given α, let C_i satisfy the following equation:

$$P\{\hat{D}_{(i)} \geq C_i\} = \alpha.$$

Since $\hat{D}_{(i)} = t_i^2 h_{ii} \cdot \frac{1}{p}$, we have

$$P\{\hat{D}_{(i)} \geq C_i\} = P\{t_i^2 h_{ii} \cdot \frac{1}{p} \geq C_i\} = P\{t_i^2 \geq \frac{pC_i}{h_{ii}}\} = \alpha,$$

thus

$$\frac{pC_i}{h_{ii}} = F(1,\ n-p-1;\ 1-\alpha).$$

Hence

$$C_i = \frac{1}{p} h_{ii} F(1,\ n-p-1;\ 1-\alpha), \tag{15}$$

where $F(1, n-p-1; 1-\alpha)$ denotes the $100(1-\alpha)$th percentile of the F-distribution. From equation (5):

$$\frac{\widehat{\underline{e}'_{(i)}\underline{e}_{(i)}}}{SSE_{(i)}} = 1 + \frac{\hat{R}_{(i)}}{SSE_{(i)}} \tag{16}$$

The estimated value $\frac{\widehat{\underline{e}'_{(i)}\underline{e}_{(i)}}}{SSE_{(i)}}$ is equal to

$$\begin{aligned} 1 + \frac{p}{n-p-1}\hat{D}_{(i)} &\geq 1 + \frac{p}{n-p-1} \cdot \frac{1}{p} h_{ii} F(1,\ n-p-1;\ 1-\alpha) \\ &= 1 + \frac{1}{n-p-1} h_{ii} F(1,\ n-p-1;\ 1-\alpha). \end{aligned} \tag{17}$$

Thus, if the i-th data is deleted, it will be the influential data, if $\hat{D}_{(i)} \geq C_i$. In the example (Table II) below, we find that the observation number $i = 29$ is an influential data. The amount of the influence for the residual will be at least $\frac{1}{n-p-1} h_{ii} F(1,\ n-p-1;\ 1-\alpha)$, where $F(1, n-p-1; 1-\alpha)$ is the $100(1-\alpha)$th percentile of the central F distribution.

From (17), we define

$$IF_{(i)} = \frac{\hat{D}_{(i)}}{C_i}$$

as the measure of the strength of the influence on the residual when the i-th data is deleted. Note that the ith data for which $IF_{(i)}$ greater than one is the influential data.

Now, we rewrite Cook's distance D_i with ith data deleted as follows:

$$D_i = \{[\frac{\hat{e}_i}{s\sqrt{1-h_{ii}}}]^2 \cdot \frac{h_{ii}}{1-h_{ii}}\}\frac{1}{p}.$$

Both D_i and $\hat{D}_{(i)}$ in (14) are influenced by the standardized residual and h_{ii}; however, D_i is influenced by h_{ii} through $h_{ii}/(1-h_{ii})$ and thus will put much more weight on variance more than the residual as h_{ii} gets close to 1.

In Table I (based on simulated data), we see that COOK'S D = 536.901 and DHAT = 6381.64 for the 30th data. This is deleted by D and also by DHAT as influential data and it represents departure from the fitted line:

$$\begin{array}{lcll} X1 & = & 64.506 \; + \; 0.356\, X2 & \text{RSQUARE=92\% C.V.=0.595\%} \\ & & (0.356) \quad (0.02) & \end{array}$$

Note that the values in the parenthesis denote that standard errors of the corresponding coefficients. We find that DHAT is more sensitive than D to detect the data which shows departure from the fitted line.

3. Selecting Important Independent Variables

For the selection of important independent variables in (1), it is necessary to consider the measurement of the influence.

In model (1) let $\underline{Y}' = [Y_1, \ldots, Y_n]$, $X = [\underline{1}, \underline{X}_1, \ldots, \underline{X}_{p-1}]$, $\underline{\beta}' = [\beta_0, \beta_1, \ldots, \beta_{p-1}]$ and $\underline{e} \sim N(\underline{0}, \sigma_0^2 I_n)$; here I_n denotes the identity matrix of order $n \times n$. The model (1) having $p-1$ independent variables is considered as the true model. Any reduced model whose 'X matrix' has r columns is obtained by retaining any $r-1$ of the $p-1$ independent variables $X_1, \ldots, X_{p-1}$, where $2 \leq r \leq p-1$. For each r, $2 \leq r \leq p-1$, there are $k_r = \binom{p-1}{r-1}$ such models. These k_r reduced models of 'size' r are indexed arbitrarily with the indexing variable ℓ going from 1 to k_r. We will refer to a typical model as Model $M_{r\ell}$. If the i-th data is deleted, then the reduced Model $M_{r\ell}$ is denoted by $M_{r\ell(i)}$. A reduced model of size r can be written as

$$E(\underline{Y}|X_{r\ell}) = X_{r\ell}\underline{\beta}_{r\ell}, \quad \ell = 1, 2, \ldots, k_r. \tag{18}$$

The reduced model for deleted i-th data is

$$E(\underline{Y}_{(i)}|X_{r\ell(i)}) = X_{r\ell(i)}\underline{\beta}_{r\ell(i)}, \quad \ell = 1, 2, \ldots, k_r. \tag{19}$$

It should be pointed out that all expectations and probabilities are calculated under model (1).

Usually, we use the residual sum of squares to measure goodness of the fitted model for a random sample. Hence, the expected residual sum of squares is naturally considered as the measurement for the goodness of fit. Large values of this expectation are not desirable. But, the estimate of the expectation is heavily influenced by the influential data. It is important to detect them, and consider them seriously. It should be first noted that our comparisons of models are made under the true model assumptions.

For any r, $2 \leq r \leq p-1$, the residual sum of squares $SS_{r\ell}$ and $SS_{r\ell(i)}$ for the reduced models $M_{r\ell}$ and $M_{r\ell(i)}$, $1 \leq \ell \leq k_r$, $i = 1, 2, \ldots, n$, are respectively as follows:

$$SS_{r\ell} = \underline{Y}' Q_{r\ell} \underline{Y}, \text{ and } S_{r\ell(i)} = \underline{Y}'_{(i)} Q_{r\ell(i)} \underline{Y}_{(i)} \tag{20}$$

where

$$Q_{r\ell} = [I_n - X_{r\ell}(X'_{r\ell}X_{r\ell})^{-1}X'_{r\ell}],$$

and

$$Q_{r\ell(i)} = [I_{n-1} - X_{r\ell(i)}(X'_{r\ell(i)}X_{r\ell(i)})^{-1}X'_{r\ell(i)}].$$

Also

$$\frac{SS_{r\ell}}{\sigma_0^2} \sim \chi^2\{n-r,\ \lambda_{r\ell}\},$$

and (21)

$$\frac{SS_{r\ell(i)}}{\sigma_0^2} \sim \chi^2\{n-r-1, \lambda_{r\ell(i)}\}$$

where

$$\lambda_{r\ell} = (X\underline{\beta})' Q_{r\ell}(X\underline{\beta})/2\sigma_0^2,$$

and

$$\lambda_{r\ell(i)} = (X\underline{\beta})' Q_{r\ell(i)}(X\underline{\beta})/2\sigma_0^2.$$

We note that $Q_{r\ell}$ and $Q_{r\ell(i)}$ are idempotent and symmetric; thus they are positive semi-definite. Hence $\lambda_{r\ell}$ and $\lambda_{r\ell(i)}$ are nonnegative, but not zero, in general.

We have

$$E[SS_{r\ell}] = (n-r)\sigma_0^2 + 2\sigma_0^2\lambda_{r\ell},$$

and (22)

$$E[SS_{r\ell(i)}] = (n-r-1)\sigma_0^2 + 2\sigma_0^2\lambda_{r\ell(i)}.$$

Since σ_0^2 is fixed, it is clear from (22) that $\lambda_{r\ell}$ and $\lambda_{r\ell(i)}$, for all i, should not be large for $M_{r\ell}$ to be a good model.

Consider the coefficient of partial determination between the dependent variable Y and X_i, given $X_1, \ldots, X_{p-1}$ except X_i in the model, denoted by $\gamma^2_{Yi\cdot\rightarrow i}$.

It is known that

$$\gamma^2_{Yi\cdot\rightarrow i} = \frac{SSE(X_i|\rightarrow X_i)}{SSE(X_1,\ldots,X_{p-1})} = \frac{SSE(\rightarrow X_i) - SSE(X_1,\ldots,X_{p-1})}{SSE(X_1,\ldots,X_{p-1})}$$

where $SSE(\rightarrow X_i)$ is the residual sum of squares for the model which includes $X_1, \ldots, X_{p-1}$ except X_i. We can write

$$\gamma^2_{Yj\cdot\rightarrow j} = \frac{SSE(\rightarrow X_j)}{SSE(X_1,\ldots,X_{p-1})} - 1 = \frac{SS_{p-1,j}}{SS_{p1}} - 1$$

and

$$\hat{\lambda}_{p-1,j} = \frac{n-p}{2} \cdot \frac{SS_{p-1,j}}{SS_{p1}} - \frac{n-(p-1)}{2}, j = 1, \ldots, p-1.$$

Hence

$$\gamma^2_{Yj\cdot\rightarrow j} = \frac{2}{n-p}\hat{\lambda}_{p-1,j} + \frac{1}{n-p} = \frac{1}{n-p}(2\hat{\lambda}_{p-1,j} + 1).$$

Thus, we can use $\hat{\lambda}_{p-1,j}, 1 \leq j \leq p-1$, for ranking the importance of the independent variables $X_j, 1 \leq j \leq p-1$.

From the assumption $\underline{e} \sim N(\underline{0}, \sigma_0^2 I_n)$, it follows that the statistic

$$V_{ri} = \frac{(SS_{ri} - SS_{p1})/(p-r)}{SS_{p1}/(n-p)}$$

has the noncentral F distribution denoted by $F'(p-r, n-p, \lambda_{ri})$.

Since $\hat{\lambda}_{rj} = \frac{n-p}{2}\frac{SS_{rj}}{SS_{p1}} - \frac{n-r}{2}$, it follows that

$$V_{rj} = \frac{2}{p-r}(\hat{\lambda}_{rj} + \frac{n-r}{2}) - \frac{n-p}{p-r}.$$

And

$$E(V_{rj}) = \frac{(p-r+2\lambda_{rj})(n-p)}{(n-p-2)(p-r)}.$$

We obtain an unbiased estimator $\wedge_{rj}$ of λ_{rj} as

$$\wedge_{rj} = \frac{n-p-2}{2(n-p)}[\hat{\lambda}_{rj} - \frac{(p-r)}{(n-p-2)}].$$

The quantity $\wedge_{rj}$ is the bias of the reduced model M_{rj}.

Gupta and Huang (1988) have proposed some selection procedures for selecting good models based on $\lambda_{r\ell}$'s. Now, we are interested in studying how large is the influence for $\lambda_{r\ell(i)}$ when the i-th observation is deleted.

We have an estimate of $\lambda_{r\ell}$, σ_0^2, $\sigma_{0(i)}^2$ and $\lambda_{r\ell(i)}$ as follows:

$$\hat{\sigma}_0^2 = \frac{SS_{p1}}{n-p} = \frac{SSE}{n-p},\ \hat{\sigma}_{0(i)}^2 = \frac{SS_{p1(i)}}{n-p-1} = \frac{SSE_{(i)}}{n-p-1}$$

$$\hat{\lambda}_{r\ell} = \frac{n-p}{2}\frac{SS_{r\ell}}{SS_{p1}} - \frac{n-r}{2} = \frac{n-r}{2}[\frac{SS_{r\ell}/(n-r)}{\hat{\sigma}_o^2} - 1]. \tag{23}$$

If $SS_{r\ell}/(n-r) \approx \hat{\sigma}_o^2$, then $M_{r\ell}$ is near the true model; and

$$\hat{\lambda}_{r\ell(i)} = \frac{n-p-1}{2} \cdot \frac{SS_{r\ell(i)}}{SS_{p1(i)}} - \frac{n-r-1}{2}.$$

Since

$$(X\underline{\beta} - X_{r\ell(i)}\underline{\beta}_{r\ell(i)})'Q_{r\ell(i)}(X\underline{\beta} - X_{r\ell(i)}\underline{\beta}_{r\ell(i)}) = (X\underline{\beta})'Q_{r\ell(i)}(X\underline{\beta}) = 2\sigma_0^2\lambda_{r\ell(i)}.$$

Hence, $\lambda_{r\ell(i)}$ also measures the influence of the i-th data on fitted values. We define the measurement of the influence for the i-th data as follows:

$$D_{r\ell(i)} = \lambda_{r\ell(i)} = \frac{(X\underline{\beta})'Q_{r\ell(i)}(X\underline{\beta})}{2\sigma_0^2}. \tag{24}$$

We estimate $D_{r\ell(i)}$ in (24) by

$$\hat{D}_{r\ell(i)} = \hat{\lambda}_{r\ell(i)} = \frac{n-p-1}{2}\frac{SS_{r\ell(i)}}{SS_{p1(i)}} - \frac{n-r-1}{2} \tag{25}$$

and use $\hat{D}_{r\ell(i)}$ as a statistic to measure the influence. We want to find a constant d such that

$$\inf_{\lambda_{r\ell(i)} \geq \Delta} P\{\hat{D}_{r\ell(i)} \geq d\} = \alpha. \tag{26}$$

where $\Delta > 0$ and α are given. In order to do this, note that

$$V_{r\ell(i)} = \frac{[SS_{r\ell(i)} - SS_{p1(i)}]/(p-r)}{SS_{p1(i)}/(n-p-1)},$$

follows the noncentral F denoted as $F'(p-r,\ n-p-1;\ \lambda_{r\ell(i)})$ (cf. Graybill (1976)).

Using this and the fact that the noncentral F is stochastically increasing in $\lambda_{r\ell(i)}$, it can be seen that (26) is satisfied by

$$P\{\hat{D}_{r\ell(i)} \geq d | \lambda_{r\ell(i)} = \Delta\} = \alpha \tag{27}$$

and that

$$[(d + \frac{n-r-1}{2})\frac{2}{n-p-1} - 1]\frac{n-p-1}{p-r} = F'(p-r,\ n-p-1;\ \Delta) \tag{28}$$

From (28), we have

$$d = \frac{(n-p-1)}{2}\left\{\frac{(p-r)}{(n-p-1)}F'(p-r,\ n-p-1;\ \Delta) + 1\right\} - \frac{n-r-1}{2}. \tag{29}$$

Patnaik (1949) provided an approximation to the noncentral F distribution (cf. Guenther (1979)) by the relation

$$\begin{aligned} F'(p-r, n-p-1; \Delta) &\approx \{[(p-r) + 2\Delta]/p-r\}\, F(p^*,\ n-p-1), \\ &\text{where } p^* = [(p-r) + 2\Delta]^2/[(p-r) + 4\Delta]. \end{aligned} \tag{30}$$

Hence the constant d can be computed as follows:

$$d \approx \frac{(n-p-1)}{2}\left\{\frac{(p-r)+2\Delta}{n-p-1}F(p^*,\ n-p-1) + 1\right\} - \frac{n-r-1}{2}. \tag{31}$$

We summarize the results as follows:

If the i-th data is deleted, and

$$\hat{D}_{r\ell(i)} \geq d, \tag{32}$$

then we conclude that there exists an influential data in the reduced model $M_{r\ell}$.

A reduced model $M_{r\ell}$ is called an inferior model, if there is some i-th data which satisfies the condition (32), where i-th data is not an influential data in model (1). A

method to select important independent regression variables is given in Gupta and Huang (1988).

4. Selection Criteria for Regression Variables

Consider the true model

$$Y_j = \beta_0 + \beta_1 X_{1j} + \cdots + \beta_{p-1} X_{p-1} + \varepsilon_j, \quad j = 1, \ldots, n$$
$$E(\varepsilon_j) = 0, \quad \mathrm{Var}(\varepsilon_j) = \sigma_0^2, \quad \mathrm{cov}(\varepsilon_i, \varepsilon_j) = 0, \ i \neq j.$$

Let

$\mu_j = E(Y_j | X_{1j}, \ldots, X_{p-1,j})$ and $\eta_{ij} = E(Y_j | X_{1j}, \ldots, X_{r-1,j})$ for a reduced model M_{ri}.

The standardized total error is defined by

$$\Gamma_{ri} = \frac{1}{\sigma_0^2} \sum_{j=1}^{n} (\mu_j - \eta_{ij})^2 + \frac{1}{\sigma_0^2} \sum_{j=1}^{n} \mathrm{Var}(\hat{Y}_{ij}),$$

where $\hat{Y}_{ij}$ is the predicted value of η_{ij}. We rewrite the model in matrix form:

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

where $\underline{Y} = [Y_1, \ldots, Y_n]'_{n \times 1}$, $\underline{\beta} = [\beta_0, \beta_1, \ldots, \beta_{p-1}]'_{p \times 1}$, and $\underline{\varepsilon} = [\varepsilon_1, \ldots, \varepsilon_n]'_{n \times 1}$.

It can be shown that

$$\sum_{j=1}^{n} (\mu_j - \eta_{ij})^2 = \sum_{j=1}^{n} [E(Y_j | X_{1j}, \ldots, X_{p-1,j}) - E(\hat{Y}_{ij} | X_{ij}, \ldots, X_{r-1,j})]^2 = 2\sigma_0^2 \lambda_{ri},$$

where $\lambda_{ri} = (X\underline{\beta})' Q_{ri} (X\underline{\beta}) / 2\sigma_0^2$, $Q_{ri} = I - X_{ri}(X'_{ri}X_{ri})^{-1} X'_{ri}$, and $\underline{\hat{Y}}_i = [\hat{Y}_{i1}, \ldots, \hat{Y}_{in}]' = X_{ri}(X'_{ri}X_{ri})^{-1} X'_{ri}\underline{Y}$. Also

$$\sum_{j=1}^{n} \mathrm{Var}(\hat{Y}_{ij}) = \sigma_0^2 \ \mathrm{tr} \ X_{ri}(X'_{ri}X_{ri})^{-1} X'_{ri} = r\sigma_0^2.$$

The expected residual sum of squares is

$$E(SS_{ri}) = E(\underline{Y} - \underline{\hat{Y}}_i)'(\underline{Y} - \underline{\hat{Y}}_i) = (n - r)\sigma_0^2 + 2\sigma_0^2 \lambda_{ri}.$$

Thus

$$\Gamma_{ri} = \frac{E(SS_{ri})}{\sigma_0^2} - (n - 2r) = 2\lambda_{ri} + r$$

and we can estimate λ_{ri} by

$$\hat{\lambda}_{ri} = \frac{n-p}{2} \cdot \frac{SS_{ri}}{SS_{p1}} - \frac{n-r}{2}.$$

Under the assumption that $\underline{\varepsilon}$ is normally distributed, we can obtain an unbiased estimator of Γ_{ri} (as shown in Gupta and Huang (1988)), namely,

$$\hat{\Gamma}_{ri} = \frac{n-p-2}{n-p}[2\hat{\lambda}_{ri} + (p-r)] - (p-2r),$$

which was used as a measure of goodness of fit for the reduced model M_{ri}.

Mallow's C_{ri} is defined as

$$C_{ri} = (p-r)V_{ri} - (p-2r),$$

so that

$$\begin{aligned}\hat{\Gamma}_{ri} &= \frac{n-p-2}{n-p}[C_{ri} + (p-2r)] - (p-2r) \\ &= [1 - \frac{2}{n-p}]C_{ri} - \frac{p-2r}{n-p}.\end{aligned}$$

We know that C_{ri} is a biased estimator of Γ_{ri}, but when $n-p$ is sufficiently large, then $\hat{\Gamma}_{ri}$ is asymptotically equivalent to C_{ri}.

Gupta and Huang (1988) proposed a two-stage procedure to determine the sample size and guarantee the probability of a correct selection.

5. An Example

We take an example for the selection of influential data from Daniel and Wood (1980, p 234). The data were obtained in a laboratory study of the distillation properties of various crude oils with respect to their yield of gasoline. The four independent variables measured were:

X_1: crude oil gravity, $^\circ API$,

X_2: crude oil vapor pressure, psi,

X_3: crude oil ASTM 10% point, $^\circ F$,

X_4: gasoline ASTM end point, $^\circ F$,

Y: gasoline yield, as percentage of crude.

We fit the full model for the data as follows:

$$Y = -6.952 + 0.229X_1 + 0.553X_2 - 0.149X_3 + 0.155X_4.$$

$$R^2 = 0.96, \text{ Root MSE } = 2.235, \text{ C.V. } = \frac{\text{Root MSE}}{\overline{Y}} \times 100\% = 11.37\%.$$

where $\overline{Y}$ is the sample mean of Y_i's. We consider the reduced model as in Daniel and Wood (1980, p. 247):

$$Y = 70.84 - 0.212X_3 + 0.159(X_4 - 332) \tag{33}$$

$$R^2 = 0.95, \text{ Root MSE } = 2.426, C.V. = 12.338\%.$$

In the reduced model (33), there is no influential data.

We have computed some values in TABLE II to illustrate the various statistics in the previous discussion.

We now summarize various notations used in the column heads of TABLE II.

$$\text{Residual} = Y_i - \hat{Y}_i,$$

$$\text{RSTUDENT} = \frac{\hat{e}_{(i)}\sqrt{1-h_{ii}}}{s_{(i)}}$$

$$\text{HAT DIAG } H = h_{ii},$$

$$\text{DHAT} = \hat{D}_{(i)} = (\text{RSTUDENT})^2 \times (\text{HAT DIAG } H)/p,$$

$$C_i = \frac{1}{p} h_{ii} F(1,\ 32-5-1;\ 0.95),$$

$$IF_{(i)} = \hat{D}_{(i)}/C_i,$$

where $F(1,\ 26;\ 0.95) = 4.2252, n = 32$ and $p = 5$.

For the reduced model M_{31} in (33), using (25), we have

$$\begin{aligned}\hat{D}_{31(29)} &= \frac{32-5-1}{2} \times \frac{150.4016}{111.657} - \frac{32-3-1}{2} \\ &= 3.51,\end{aligned}$$

and using (31),

$$d = \frac{32-5-1}{2}\left\{\frac{(5-3)+2\Delta}{32-5-1} F(p^*, 32-5-1) + 1\right\} - \frac{32-3-1}{2}.$$

Letting $\Delta = 1.3$, we get $p^* = [(5-3)+2\Delta]^2/[(5-3)+4\Delta] = 2.94$. For $\alpha = 0.05$, $F(2.94,\ 26;\ 0.95) \approx 3.00$ and we have $d \approx 5.9$. Thus, $\hat{D}_{31(29)} < d$. We have checked that

$\hat{D}_{31(i)} < d$ for all $i = 1, \ldots, 32$. Hence, the reduced model (33) is reasonable to accept as a good model (not an inferior model). Note that the value of Δ can be chosen as in Gupta and Huang (1988). The value of Δ is the amount of bias for a reduced model in (26).

In TABLE II, the value of $\hat{D}_{(29)}$ seems to detect the 29th data as an influential data for the full model.

TABLE I, ($\alpha = 0.5$)

OBS	DEP VAR $X1$	IND VAR $X2$	COOK'S D	DHAT	C	IF (DHAT/C)	RESIDUAL
1	100.247	100.285	0.000	0.00	0.07301	0.00	0.0356
2	99.347	99.373	0.015	0.01	0.07007	0.14	−0.5396
3	100.709	100.694	0.007	0.01	0.07574	0.13	0.3520
4	99.634	99.611	0.006	0.01	0.07049	0.14	−0.3374
5	99.396	99.338	0.012	0.01	0.07007	0.14	−0.4782
6	98.959	98.962	0.031	0.03	0.07007	0.43	−0.7813
7	100.753	100.710	0.008	0.01	0.07574	0.13	0.3903
8	99.076	99.075	0.025	0.02	0.06986	0.29	−0.7045
9	99.715	99.723	0.004	0.00	0.07070	0.00	−0.2963
10	101.890	101.912	0.078	0.08	0.08812	0.91	1.0994
11	101.008	100.999	0.017	0.02	0.07805	0.26	0.5424
12	98.438	98.432	0.064	0.07	0.07112	0.98	−1.1136
13	100.446	100.396	0.002	0.00	0.07364	0.00	0.1951
14	101.648	101.641	0.057	0.06	0.08476	0.71	0.9539
15	100.412	100.384	0.001	0.00	0.07364	0.00	0.1654
16	99.146	99.132	0.022	0.02	0.06986	0.29	−0.6548
17	99.977	99.960	0.001	0.00	0.07154	0.00	−0.1186
18	100.098	100.100	0.000	0.00	0.07217	0.00	−0.0475
19	99.338	99.283	0.014	0.01	0.07007	0.14	−0.5166
20	101.100	101.144	0.020	0.02	0.07951	0.25	0.5828
21	99.228	99.225	0.019	0.02	0.06986	0.29	−0.6060
22	101.240	101.256	0.027	0.03	0.08056	0.37	0.6829
23	98.677	98.706	0.048	0.05	0.07028	0.71	−0.9722
24	100.407	100.417	0.001	0.00	0.07385	0.00	0.1487
25	100.940	100.936	0.014	0.01	0.07763	0.13	0.4969
26	100.016	100.001	0.000	0.00	0.07175	0.00	−0.0942
27	100.939	100.948	0.014	0.01	0.07763	0.13	0.4916
28	101.185	101.133	0.026	0.03	0.07930	0.38	0.6717
29	100.089	100.076	0.000	0.00	0.07196	0.00	−0.0479
30	90.000	70.200	536.901	6381.64	2.04471	3121.05	0.5000

TABLE II, ($\alpha = 0.5$)

OBS	COOK'S D	DHAT	C	IF = DHAT/C
1	0.024	0.020575	0.11325	0.18
2	0.019	0.019965	0.03051	0.65
3	0.017	0.013772	0.12626	0.11
4	0.016	0.012537	0.15322	0.08
5	0.038	0.026563	0.23942	0.11
6	0.096	0.080673	0.16522	0.49
7	0.002	0.001830	0.09871	0.02
8	0.008	0.005654	0.24339	0.02
9	0.005	0.003888	0.17131	0.02
10	0.044	0.043326	0.06144	0.71
11	0.000	0.000351	0.04183	0.01
12	0.001	0.000642	0.09432	0.01
13	0.083	0.070816	0.14832	0.48
14	0.000	0.000349	0.11764	0.00
15	0.001	0.000537	0.07615	0.01
16	0.025	0.019209	0.18322	0.10
17	0.010	0.008050	0.13547	0.06
18	0.031	0.029679	0.06913	0.43
19	0.040	0.039271	0.06448	0.61
20	0.029	0.027883	0.06144	0.45
21	0.020	0.016869	0.10547	0.16
22	0.023	0.018239	0.15280	0.12
23	0.080	0.059961	0.21314	0.28
24	0.045	0.041968	0.09060	0.46
25	0.035	0.029200	0.14223	0.21
26	0.002	0.001638	0.17714	0.01
27	0.024	0.019000	0.15922	0.12
28	0.033	0.029029	0.10936	0.27
29	0.056	0.060943	0.04927	1.24
30	0.000	0.000037	0.24492	0.00
31	0.000	0.000218	0.12922	0.00
32	0.068	0.047230	0.25742	0.18

Note the value of $\hat{D}_{(29)}$ in the above table indicates the 29th data as influential for the full model.

References

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

[2] Cook, R. D. and Weisberg, S. (1980). Characterizations of an empirical influence function for detecting influential cases in regression. *Technometrics* **22**, 495–508.

[3] Daniel, C. and Wood, F. S. (1980). *Fitting Equations to Data.* 2nd edition. John Wiley & Sons, New York.

[4] Gray, J. B. and Ling, R. F. (1984). K-Clustering as a detection tool for influential subsets in regression. *Technometrics* **26**, 305–318.

[5] Graybill, F. A. (1976). *Theory and Application of the Linear Model.* Duxbury Press, North Scituate, MA.

[6] Guenther, W. C. (1979). The use of noncentral F approximations for calculation of power and sample size. *Amer. Statist.* **33 (4)** , 209–210.

[7] Gupta, S. S. and Huang, D. Y. (1988). Selecting important independent variables in linear regression models. *JSPI* **20**, 155–167. See also *Erratum JSPI* **24** (1990), 269.

[8] Patnaik, P. B. (1949). The noncentral chi-squared and F distributions and their applications. *Biometrika* **36**, 202–232.

[9] Welsch, R. E. (1982). Influence functions and regression diagnostics. Modern Data Analysis (Launer, R. L. and Siegel, A. F.). New York: Academic Press.