〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 16, No. 1, 1987

On Simultaneous Considerations of Variable Selection and Detection of Influential Cases+

Byoung Jin Ahn* and Sung Hyun Park**

ABSTRACT

The values of statistics used for variable selection criteria can be reduced remarkably by excluding only a few influential cases. Furthermore, different subsets of regressors change leverage and influence patterns for the same response variable. Based on these motivations, this paper suggests a procedure which coniders variable selection and detection of influential cases simultaneously.

1. Introduction

The problem of selecting a subset of predictor variables is usually described in an idealized setting. That is, it is assumed that the analyst has "good" data at hand from which to draw the eventual conclusion. In practice, the lack of satisfaction of this assumption may render a detailed subset selection analysis a meaningless exercise. Variables that look important or, conversely, appear to be of little value based on standard indicators may, in fact, appear so because of only a few influential observations or outliers. Furthermore, different subsets of predictor variables drastically change leverage and influence patterns for the same response variable. For these reasons, it seems necessary that variable elimination be considered with case influence simultaneously.

Consider the following full rank linear model with n cases, given by

* Department of Applied Statistics, Kon-Kuk University, Seoul 133, Korea.
** Department of Computer Science and Statistics, Seoul National University, Seoul 151, Korea.
+ This work was partially supported by the Ministry of Education, through the Research Institute for Basic Sciences, Seoul National University, 1986~1987.

$$y=X\beta+X_2\beta_2+\varepsilon, \tag{1.1}$$

where y is $n\times 1$, X is $n\times p$, X_2 is $n\times q$, β is $p\times 1$, β_2 is $q\times 1$, $E(\varepsilon)=0$, and $\mathrm{Cov}(\varepsilon)=\sigma^2 I$. Alternatively, we will consider a fixed subset model of the form

$$y=X\beta+\delta. \tag{1.2}$$

In general, (1.2) will provide biased estimates of $E(y)$, but, as is well known, it may smaller mean squared error than (1.1) for estimating expected values of y. See Hocking(1974).

Let $V=(v_{ij})=X(X'X)^{-1}X'$ be the orthogonal projection onto the column space of X, and define U to be the orthogonal projection onto the column space of (X, X_2), *i.e.*,

$$U=(u_{ij})=(X\ X_2)\begin{pmatrix} X'X & X'X_2 \\ X_2'X & X_2'X_2 \end{pmatrix}^{-1}\begin{pmatrix} X' \\ X_2' \end{pmatrix} \tag{1.3}$$

It is convenient here to apply a linear transformation to X_2 so that the variables not in the subset model are orthogonal to those included in the subset. To this end, define

$$Z=(I-V)X_2 \tag{1.4}$$

which is the projection of X_2 onto the orthogonal complement of X. The full model (1.1) can then be rewritten as

$$y=X\beta+Z\gamma+\varepsilon. \tag{1.5}$$

The results in this paper are derived using (1.5) in place of (1.1). In practice, however, the transformation need not be computed. Suppose we let $V^c=Z(Z'Z)^{-1}Z'$, then one can show that $U=V+V^c$. From this fact, we can obtain the same value of $\hat{y}$ through either (1.1) or (1.5).

The fitted value for the i-th case for the subset model is denoted by $\tilde{y}_{p,i}=x_i'\hat{\beta}$ and for the full model by $\hat{y}_i=x_i'\hat{\beta}+z_i'\hat{\gamma}$, where x_i' and z_i' are the i-th rows of X and Z, respectively.

Because of the orthogonalization, the least squares estimator, $\hat{\beta}=(X'X)^{-1}X'y$, will be the same for the models (1.2) and (1.5). The mean squared error of $\tilde{y}_{p,i}$ is given by

$$\begin{aligned} MSE(\tilde{y}_{p,i}) &= E(\tilde{y}_{p,i}-E(y_i))^2 \\ &= v_{ii}\sigma^2+(z_i'\gamma)^2. \end{aligned} \tag{1.6}$$

We will propose and discuss a weighted sum of mean squared errors($WMSE$) as a criterion of variable selection in Section 2. In Section 3, a numerical example is shown to demonstrate the procedure discussed in Section 2. In Section 4, concluding remarks are mentioned briefly.

2. Case Weighting for Variable Selection Criteria

One good criterion to check "closeness" of $\tilde{y}_p$ to y which has been used by many authors is the integrated mean squared error:

$$Q=\int_R MSE(\tilde{y}_p)\,dW \tag{2.1}$$

where R is the region of interest and W is the weight function in the region R.

Since $E(y)=x'\beta+z'\gamma$

and

$$MSE(\tilde{y}_p)=x'(X'X)^{-1}x\sigma^2+(z'\gamma)^2,$$

the integrated mean squared error Q becomes

$$Q=\sigma^2 tr[(X'X)^{-1}M_{pp}]+\gamma' M_{rr}\gamma \tag{2.2}$$

where $M_{pp}=\int_R xx'dW$ and $M_{rr}=\int_R zz'dW$. See Park(1977).

If relative importance is imposed on the data points by letting $W(\cdot)=w_i$ at the data points and $W(\cdot)=0$ elsewhere, then $M_{pp}=X'WX$ and $M_{rr}=Z'WZ$ where $W=\mathrm{diag}(w_1, w_2, \cdots, w_n)$. The w_i's are case weights which reflect relative importance.

For this particular case, we obtain

$$\begin{aligned} WMSE&=\sigma^2 tr[(X'X)^{-1}X'WX]+\gamma'Z'WZ\gamma \\ &=\sigma^2\sum_i w_i v_{ii}+\sum_i w_i(z_i'\gamma)^2 \\ &=\sum_i w_i MSE(\tilde{y}_{p,i}). \end{aligned} \tag{2.3}$$

Hence, if relative importance is imposed on the data points, the integrated mean squared error becomes a weighted sum of mean squared errors ($WMSE$) at the data points.

Many of subset selection criteria are simple functions of the residual sum of squares for the subset model (1.2) (RSS_p). If we assign some weights w_i for each case, we can obtain the following weighted RSS_p as an estimator of $WMSE$,

$$\begin{aligned} WRSS_p&=\sum_i w_i(y_i-\tilde{y}_{p,i})^2 \\ &=\sum_i w_i e_{p,i}^2\,, \end{aligned} \tag{2.4}$$

where $e_{p,i}$ is the i-th residual for the subset model.

In the following, we show that some well known selection criteria are particular cases of $WRSS_p$.

(i) If $w_i=1/(n-p)$ for all i,

$$WRSS_p=\sum_i e_{p,i}^2/(n-p)=RMS_p,$$

which is equal to the residual mean of squares for the subset model.

(ii) If $w_i=1/TSS$ for all i where TSS is total sum of squares, $(1-WRSS_p)$ is equivalent to the squared multiple correlation coefficient R_p^2.

(iii) If $w_i=(n+p)/[n(n-p)]$for all i, $WRSS_p$ is equivalent to the average prediction variance J_p. See Hocking(1972).

(iv) If $w_i=1/[(n-p)(n-p-1)]$ for all i, $WRSS_p$ is equivalent to the average prediction mean squared error S_p. See Tukey(1967).

(v) If $w_i=e_{p,i}(i)/e_{p,i}=1/(1-v_{ii})$ where $e_{p,i}$ (i) is the i-th predicted residual from the subset model fitted to the data with the i-th case excluded, $WRSS_p$ is equivalent to the standardized residual sum of squares RSS_p*. See Schmidt(1973).

(vi) If $w_i=e_{p,i}^2(i)/e_{p,i}=1(1-v_{ii})^2$, $WRSS_p$ is equivalent to $PRESS_p$. See Allen(1974)

The weighted RSS_p is a biased estimator of $WMSE$ since

$$E(WRSS_p)-WMSE=\sigma^2 tr(W(I-2V)). \tag{2.5}$$

An unbiased estimator of $WMSE$ in (2.3) can be obtained by replacing $MSE(\hat{y}_{p,i})$ with an unbiased estimator $C_{p,i}$:

$$WC_p=\sum_i w_i C_{p,i} \tag{2.6}$$

where $C_{p,i}=(z_i'\hat{\gamma})^2+(2v_{ii}-u_{ii})\hat{\sigma}^2$ and $\hat{\sigma}^2$ is the residual mean of squares for the full model.

If w_i is equal to one for all i, $WC_p/\hat{\sigma}^2$ is just the Mallows's C_p statistic, which is given by

$$C_p=RSS_p/\hat{\sigma}^2+2p-n, \tag{2.7}$$

since $\sum_i (z_i'\hat{\gamma})^2=RSS_p-(n-p-q)\hat{\sigma}^2$, $\sum_i v_{ii}=tr(V)=p$, and $\sum_i u_{ii}=tr(U)=p+q$.

The idea of weighting cases to lessen the case influence of a few influential observations seems natural, but the problem how to determine the case weights which are efficient in downweighting outliers or high leverage points is not simple. Subsequently, we will consider the problem of determining case weights.

Suppose that the i-th case is suspected to be an outlier in a subset model. A useful

framework used to study outliers is the mean shift outlier model (See, for example, Cook and Weisberg(1982).),

$$y=X\beta+d_i\alpha+\varepsilon, \tag{2.8}$$

where d_i is an n-vector with i-th element equal to one, and all other elements equal to zero. Non-zero values of α imply the i-th case is an outlier. The least squares estimator of α is

$$\begin{aligned}\hat{\alpha}&=d'_i(I-V)y/d_i'(I-V)d_i\\&=e_{p,i}/(1-v_{ii}).\end{aligned} \tag{2.9}$$

The extra sum of squares due to fitting α in the model (2.8), is given by

$$R_{p,i}=e_{p,i}^2/(1-v_{ii}). \tag{2.10}$$

This statistic measures the effect of outlier and is the special case of the statistic Q_k discussed by John and Draper(1978).

Note that

$$\begin{aligned}E(R_{p,i})&=E(e_{p,i}^2)/(1-v_{ii})\\&=\sigma^2+(z_i'\gamma)^2/(1-v_{ii}).\end{aligned} \tag{2.11}$$

The above motivations may lead to consider the following case weights given by

$$w_i=\begin{cases}C & \text{if } R_{p,i}/\hat{\sigma}^2\leq h(p)\\ Ch(p)\hat{\sigma}^2/R_{p,i} & \text{otherwise,}\end{cases} \tag{2.12}$$

where C is an appropriate constant and $h(p)$ is a function of p.

Now, we are to decide the function of $h(p)$. Hoaglin and Welsch (1978) determined high-leverage points by looking at the diagonal elements of projection matrix V and paying particular attention to data points which have $v_{ii}>2p/n$. For this reason, we make take the function of $h(p)$ as described in (2.13) by replacing v_{ii} and $(z_i'\gamma)^2$ in (2.11) with $2p/n$ and $\max(y_i-\bar{y}_{p,i})^2$, respectively.

$$h(p)=1+n\mathrm{Max}(y_i-\bar{y}_{p,i})^2/\hat{\sigma}^2(n-2p). \tag{2.13}$$

The deletion of a case corresponding to an outlier in y will tend to result in a marked reduction in the residual sum of squares. The residual sum of squares, therefore, is a diagnostic measure for detecting influential cases arising because of an outlier in y. So, we may determine the case weights which are different from (2.12) as

$$w_i=C\ RSS_p(i)/RSS_p, \tag{2.14}$$

where C is an appropriate constant and $RSS_p(i)$ is the residual sum of squares with i-th case deleted.

These case weights provide the analyst with valuable information about the influence patterns of various subset models.

3. A Numerical Example

We will apply the procedure discussed in this paper to a specific real data set presented in Weisberg (1981). The data are given in Table 3.1. Suppose that finding a linear model based on a subset of the predictors and at the same time detecting influential cases are of interest. The measured variables are:

Y=log(oxygen demand, mg oxygen per minute)

X_1=biological oxygen demand, mg/liter

X_2=total Kjeldahl nitrogen, mg/liter

X_3=total solids, mg/liter

X_4=total volative solids, a component of X_3, mg/liter

X_5=chemical oxygen demand, mg/liter

Table 3.1 Data from oxygen uptake experiment(Weisberg (1981))

Case	X_1	X_2	X_3	X_4	X_5	Y
1	1125	232	7160	85.9	8905	1.5563
2	920	268	8804	86.5	7388	0.8976
3	835	271	8108	85.2	5348	0.7482
4	1000	237	6370	83.8	8056	0.7160
5	1150	192	6441	82.1	6960	0.3130
6	990	202	5154	79.2	5690	0.3617
7	840	184	5896	81.2	6932	0.1139
8	650	200	5336	80.6	5400	0.1139
9	640	180	5041	78.4	3177	−0.2218
10	583	165	5012	79.3	4461	−0.1549
11	570	151	4825	78.7	3901	0.0000
12	570	171	4391	78.0	5002	0.0000
13	510	243	4320	72.3	4665	−0.0969
14	555	147	3709	74.9	4642	−0.2218
15	460	286	3969	74.4	4840	−0.3979
16	275	198	3558	72.5	4479	−0.1549
17	510	196	4361	57.7	4200	−0.2218
18	165	210	3301	71.8	3410	−0.3979
19	244	327	2964	72.5	3360	−0.5229
20	79	334	2777	71.9	2599	−0.0458

Table 3.2 Cases with values of $w_i \leq 0.8$ in (2.12)

Selected Variables	Cases 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(1)	.56																			
(2)	.48																			
(3)	.22																			
(4)	.57																.57			
(5)																				
(1, 2)	.44																			
(1, 3)	.27																			.63
(1, 4)	.48																			
(1, 5)	.80																			
(2, 3)	.24																			
(2, 4)	.72																.62			
(2, 5)																				
(3, 4)	.26																			
(3, 5)	.21						.67													.31
(4, 5)	.59																			
(1, 2, 3)	.24																			
(1, 2, 4)	.38																			
(1, 2, 5)	.52																			
(1, 3, 4)	.33																			
(1, 3, 5)	.22						.67													.76
(1, 4, 5)	.57																			
(2, 3, 4)	.25																			
(2, 3, 5)	.21						.75									.53				.38
(2, 4, 5)	.53															.68	.59			
(3, 4, 5)	.20						.62													.31
(1, 2, 3, 4)	.29																			
(1, 2, 3, 5)	.22						.79									.56				.41
(1, 2, 4, 5)	.44															.67	.71			
(1, 3, 4, 5)	.20						.59													.32
(2, 3, 4, 5)	.15						.51									.36				.28
(1, 2, 3, 4, 5)	.14						.50									.36				.28

$$**w_i = \begin{cases} 1 & \text{if } R_{p,i}/\hat{\sigma}^2 \leq h(p) \\ h(p)\hat{\sigma}^2/R_{p,i} & \text{otherwise.} \end{cases} \qquad (2.12)$$

Tables 3.2 and 3.3 list the cases with relatively small values of w_i which are derived from the equations (2.12) and (2.14), respectively. These tables exhibit the following facts. First, the influence patterns could be changed as the selected variables vary. Second, the results from Tables 3.2 and 3.3 are similar. Finally, cases 1, 17, and 20 could be influential cases or outliers.

Table 3.3 Cases with values of $w_i \leq 0.8$ in (2.14)

Selected Variables	Cases 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
(1)	.61																			
(2)	.56																			
(3)	.41																			
(4)	.57																.56			
(5)	.77	.77																		
(1, 2)	.61																			
(1, 3)	.43																			.76
(1, 4)	.58																			
(1, 5)	.72	.79																		.75
(2, 3)	.40																			
(2, 4)	.57																.50			
(2, 5)	.79																			
(3, 4)	.41																			
(3, 5)	.52																			.67
(4, 5)	.71																			
(1, 2, 3)	.44																			
(1, 2, 4)	.58																			
(1, 2, 5)	.69																			
(1, 3, 4)	.43																			.77
(1, 3, 5)	.52																			.68
(1, 4, 5)	.68																.79			.76
(2, 3, 4)	.40																			
(2, 3, 5)	.52																			.73
(2, 4, 5)	.72																			
(3, 4, 5)	.52																			.68
(1, 2, 3, 4)	.44																			
(1, 2, 3, 5)	.52																			.73
(1, 2, 4, 5)	.66																.79			
(1, 3, 4, 5)	.52																			.69
(2, 3, 4, 5)	.51																			.75
(1, 2, 3, 4, 5)	.51																			.75

**$w_i = RSS_p(i)/RSS_p$. (2.14)

Chatterijee and Hadi (1986) applied some diagnostic measures to the same data set given in Table 3.1 and they came to the similar cncolusion that observation numbers 1, 7, 15, 17, and 20 are influential either individually or in groups.

In Table 3.4, the values of $WMSE$ criteria, $RMSP_p$, and C_p are listed. The criteria C_1, C_2 denote the $WRSS_p$ in (2.4) with the case weights given in (2.12) and (2.14), respectively. The criteria C_3, C_4 denote WC_p in (2.6) with the case weights given in

Table 3.4 The values of criteria in the subset models

Selected Variables	C_1	C_2	C_3	C_4	RMS_p	C_p
(1)	0.089	0.086	12.63	11.863	0.112	12.999
(2)	0.205	0.208	51.274	49.261	0.280	56.764
(3)	0.048	0.054	5.162	5.242	0.086	6.321
(4)	0.098	0.098	16.483	16.071	0.139	20.242
(5)	0.078	0.070	6.471	5.554	0.086	6.471
(1, 2)	0.067	0.067	8.443	7.993	0.092	8.637
(1, 3)	0.045	0.051	5.905	5.992	0.085	6.743
(1, 4)	0.072	0.074	10.567	10.102	0.102	10.955
(1, 5)	0.071	0.065	7.372	6.642	0.087	7.486
(2, 3)	0.047	0.052	6.350	6.290	0.087	7.391
(2, 4)	0.102	0.091	18.043	16.459	0.144	21.311
(2, 5)	0.072	0.065	6.807	5.994	0.084	6.808
(3, 4)	0.048	0.052	6.264	6.181	0.086	7.173
(3, 5)	0.031	0.042	1.192*	1.410*	0.064	1.709*
(4, 5)	0.061	0.060	4.988	4.608	0.078	5.178
(1, 2, 3)	0.043	0.047	5.582	5.493	0.078	6.067
(1, 2, 4)	0.057	0.059	7.954	7.612	0.088	8.316
(1, 2, 5)	0.056	0.056	6.036	5.746	0.079	6.272
(1, 3, 4)	0.048	0.050	7.386	7.226	0.087	8.022
(1, 3, 5)	0.032	0.042	2.992	3.126	0.067	3.490
(1, 4, 5)	0.059	0.058	6.669	6.323	0.082	7.008
(2, 3, 4)	0.045	0.050	7.376	7.142	0.088	8.200
(2, 3, 5)	0.030	0.040	1.719	1.983	0.062*	2.326
(2, 4, 5)	0.053	0.054	4.860	4.879	0.076	5.609
(3, 4, 5)	0.030	0.041	2.894	3.084	0.067	3.389
(1, 2, 3, 4)	0.043	0.046	7.155	6.896	0.081	7.520
(1, 2, 3, 5)	0.031	0.040	3.722	3.888	0.066	4.317
(1, 2, 4, 5)	0.049	0.050	6.056	5.972	0.077	6.599
(1, 3, 4, 5)	0.030	0.041	4.617	4.733	0.069	5.121
(2, 3, 4, 5)	0.025*	0.039*	3.234	3.589	0.065	4.002
(1, 2, 3, 4, 5)	0.025*	0.039*	5.150	5.499	0.069	6.000

(2.12) and (2.14), respectively. The symbol '*' in Table 3.4 denotes the minimum value of each criterion.

Table 3.4 exhibits some interesting facts that C_3, C_4, and $C_p$result in the same conclusion and that C_1, C_2, and RMS_p lead to different conclusions. It may be so because the $WMSE$ criteria C_3 and C_4 are derived from the weighted sum of estimators of $MSE(\hat{y}_{p,i})$, the $WMSE$ criteria C_1 and C_2 are derived from that of e_{pi}^2, and the case

weights in (2.12) and (2.14) are designed to downweight the cases with large residuals. It seems that the *WMSE* criteria C_3 and C_4 are more suitable than the *WMSE* criteria C_1 and C_2 as variable selection criteria. Table 3.4 also shows that the subset (X_3, X_5) or (X_2, X_3, X_5) seems desirable for regressors.

4. Concluding Remarks

A question may be raised whether the procedure presented in this paper is more favorable compared to the following two-step approaches which are used in practice.

Step 1. First, clean the data set by applying some detection rule for infuential cases or outliers.

Step 2. Then, use some variable selection procedure on the remaining cases.

The procedure proposed in this paper which considers the variable selection and detection of influential cases simultaneously may have some following advantages:

(1) The majority of the users who are far from being expert statisticians often run package programs for variable selection directly without Step 1. In the presence of outliers or influential cases, they may face some peculiar results that are difficult to understand. However, the proposed procedure could provide better results for the users to understand.

(2) In multiparameter regression problems outliers are not easily recognized. Even if we apply some detection rule for full data set, it is difficult to adopt case influence to variable selection procedure because different subsets of regressors change the influence patterns for the same response variable. However, the proposed method overcome such problems.

(3) The *WMSE* estimators are resistant to some influential cases. However, they lead to similar conclusion with usual variable selection criteria if the data set does not contain influential cases or outliers. Furthermore, the *WMSE* estimators may give more information than the two-step approaches, because they can make a smooth transition between full acceptance and full rejection of an observation by giving weights between 0 and 1.

References

(1) Allen, D.M. (1974). The relationship between variable selection and data argumentation and a method for prediction, *Technometrics*, Vol. 16, 125~127.

(2) Chatterijee, S., and Hadi, A.S. (1986). Influential observations, high leverage points, and outliers in linear regression, *Statistical Science*, Vol. 1, 379~416.

(3) Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York, Chapman-Hall.

(4) Hoaglin, D.C., and Welsch, R.E. (1978). The hat matrix in regression and ANOVA, *The American Statistician*, Vol. 32, 17~22.

(5) Hocking, R.R. (1972), Criteria for selection of a subset regression: Which one should be used, *Technometrics*, Vol. 14, 967~970.

(6) Hocking, R.R. (1974), Misspecification in regression, *The American Statistician*, Vol. 28, 39~40.

(7) John, J.A., and Draper, N.R. (1978). On testing for two outliers or one outlier in two-way tables, *Technomertrics*, Vol. 20, 69~78.

(8) Park, S.H. (1977). Selection of polynomial terms for response surface experiments, *Biometrics*, Vol. 33, 225~229.

(9) Schmidt, P. (1973). *Methods of choosing among alternative linear regression models*. Univ. of North Carolina, Chapel Hill, North Carolina.

(10) Tukey, J.W. (1967). Discussion (of Anscombe (1967)), *Journal of the Royal Statistical Society*, Vol. 29, 47~48.

(11) Weisberg, S. (1981). A statistic for allocating C_p to individual cases, *Technometrics*, Vol. 23, 27~31.