Letters to the Editor

K-CLUSTERING AND THE DETECTION OF INFLUENTIAL SUBSETS

I read with interest the article by Gray and Ling (1984) and the discussions that followed it. I wish to make three comments.

First, I totally agree with David M. Allen (discussant) that influential cases and variable selection should be dealt with simultaneously, but I am not aware that "leverage and influence have been part of variable selection for a long time" (p. 319). For example, Belsley et al. (1980) stated that "in general, it is best to avoid attempting to discover the model (i.e., explanatory variables) and influential points at the same time" (p. 34).

Second, Equation (3.1) should read

\[ ps^2 D_i = ps^2(1 + h_i^2 e_i h_i)/2 - h_i^2 \]

+ \[ h_j^2 e_j^2 (2 - h_j) + 2e_i e_j h_i h_j (h_i + 1 - h_i h_j) / \sigma_i^2 \]

where \( \sigma_i^2 = (1 - h_i)(1 - h_j) - h_i^2 \geq 0. \)

Third, I agree with the authors that "an effective search for jointly influential cases should make use of the information in the off-diagonal elements of the hat matrix \( H \)..." (p. 307), but I wish to add that such a search also should not ignore the diagonal elements of \( H \). To accommodate for both diagonal and off-diagonal elements of \( H \) or \( H^* \), one may proceed as follows: For \( I = \{i, j\} \), Andrews and Pregibon (1978) measured

\[ R_i = (1 - h_i)(1 - h_j) - h_i^2 = \sigma_i^2, \]

which can be written as

\[ R_i = \sigma_i^2 - [e_i^2 (1 - h_i) + e_j^2 (1 - h_j) + 2e_i e_j h_i h_j] / \text{SSE}. \]

Influential cases are often indicated by large \( D_i \) or small \( R_i \). It can be seen from (1) and (3) that the smaller \( \sigma_i^2 \) and the larger \( e_i e_j h_i h_j \), the larger \( D_i \) and the smaller \( R_i \). The conditions for small \( \sigma_i^2 \) and large \( e_i e_j h_i h_j \) can be seen in the following four configurations.

1. Configuration A: \( e_i e_j > 0 \) and \( h_i h_j > 0 \) (jointly influential)
2. Configuration B: \( e_i e_j < 0 \) and \( h_i h_j < 0 \) (jointly influential)
3. Configuration C: \( e_i e_j > 0 \) and \( h_i h_j < 0 \) (not influential)
4. Configuration D: \( e_i e_j < 0 \) and \( h_i h_j > 0 \) (not influential)

For a simple linear regression case, these configurations are illustrated in Figure 1. (Configurations A and B are adapted from Gray and Ling 1984, figure 3.) It is clear that cases \( i \) and \( j \) are influential in Configurations A and B but not in C and D. It is noteworthy that each of cases \( i \) and \( j \) in Configuration D is an outlier in both \( X \) space and \( Y \) space, yet neither is influential (individually or jointly).

The foregoing argument suggests that the K-clustering methodology should use either of the following "similarity" matrices: (a) the matrix \( S = (s_{ij}) \), whose \( ij \)th element is \( s_{ij} = (e_i e_j h_i h_j) / \sigma_i^2 \), or (b) the matrix \( S^* = (s_{ij}^*) \), whose \( ij \)th element is \( s_{ij}^* = - \sigma_i^2 \).

Using \( S \) or \( S^* \) is advantageous for several reasons.

1. All elements of \( H \) or \( H^* \) are considered.
2. The clustering is carried out only once instead of three times (i.e., once for each of \( |h_i^2| \), \( h_i^2 \), and \( -h_i^2 \)).
3. With respect to \( R_i \), the largest off-diagonal element of \( S^* \) corresponds to the most influential subset of size two, the next largest corresponds to the next most influential subset of size two, and so on. That is, if \( S^* \) is used as a similarity matrix, the method is optimal with respect to \( R_i \) at least for subsets of size two.

Last, but most important, I wish to commend and congratulate Gray and Ling on an excellent article.

Ali S. Hadi
Assistant Professor
Economic and Social Statistics Department
New York State School of Industrial and Labor Relations
Cornell University
Ithaca, NY 14853
REFERENCES

RESPONSE TO ALI S. HADI

In our 1984 article we invited readers to join our efforts in search of answers to many intriguing questions in regression diagnostics, particularly in the area of joint influence. We are very pleased that Ali S. Hadi responded by making three itemized comments (with suggestions) on substantive issues related to our article. Our response follows.

Comment 1. We are in complete accord with his endorsement of David M. Allen’s (discussant) view that influence assessments should be viewed in conjunction with the problem of variable selection. Regarding his question of Allen’s statement that “leverage and influence have been part of variable selection for a long time,” we offer our observation in support of the spirit of Allen’s statement—without intending to respond for him nor to carry any implication that this is how he might have responded: Many statistical ideas have been advanced and practiced by statisticians long before they have been formalized in published sources. Hoaglin and Welsch (1978) stated, “The term ‘hat matrix’ is due to John W. Tukey, who introduced us to the technique about ten years ago” (p. 17). Tukey also introduced the “catcher matrix,” a close kin of the hat matrix, as a diagnostic tool in Beaton and Tukey (1974, p. 160) and Mosteller and Tukey (1977, p. 341).

Comment 2. We thank Professor Hadi for pointing out the typographical errors and omissions in our Equation (3.1).

Comment 3. There were several reasons why we did not use the information in the diagonal elements of $H$ in our methodology for detecting influential subsets.

First, as stated in Section 1 of Gray and Ling (1984), we focused our attention “particularly [on] those subsets whose individual cases interact to produce a high influence that is not accounted for by the main effects of their single cases…” (p. 305). Large diagonal elements in the hat matrix tend to obscure the real effects of joint influence.

Second, during the course of our research leading to the article, we did consider some modified forms of the hat matrix that included the diagonal elements of $H$ in the off-diagonal elements of the “similarity” matrix $S$ used for clustering. In fact, one such form we considered, with

$$s_{ij} = e_i e_j^T h_{ij}/(\delta_{ij} + h_{ij}^2),$$

(1)

is very similar to the $s_{ij} = e_i e_j^T h_{ij}/\delta_{ij}$ proposed by Hadi. The use of (1) was motivated by one of the joint influence measures considered by Belsley et al. (1980, pp. 35–37):

$$MEWDFIT_i = \sum [(i,j) \in I] s_{ij}$$

$$- \sum [(i \in I) MEWDFIT_i$$

$$+ \sum [(i,j) \in I, i \neq j] s_{ij},$$

where $s_{ij}$ is given by (1). We had much less success using this form as the similarity matrix $S$ than as the $H^*$ in our clustering methodology.

Third, when $h_{ij}$ plays a more dominant role in the $s_{ij}$ elements of $S$, as in Hadi’s suggested use of $-\delta_{ij}$ in $S^*$, the characteristic “block structure” we used to identify potentially influential subsets no longer holds, since one large value of $h_{ij}$ will affect the entire row and column of the $S$ matrix. Thus, in theory (relative to the structures in $S$ for clusters), clustering such a matrix to seek influential subsets no longer makes sense.

In general, the forms of dependence of joint influence measures on the signs and magnitudes of the residuals $e$ and the elements $h_{ij}$ of $H$ are extremely varied and complicated, even for subsets of size 2—much more so than Hadi’s attempted characterizations in terms of $\delta_{ij}$ and $e_i e_j^T h_{ij}$. Moreover, there is not much relation between the rankings of “influence” by Cook’s $D$ and Andrews and Pregibon’s $R$, because the latter measures only the geometric extremeness of the configurations or, more specifically, the change (effected by the deletion of a subset) in the volume of the concentration ellipsoid in the $X$-$Y$ space, rather than the changes in the estimated parameters as measured by $D$. Thus, the cases $(i,j)$ in Hadi’s configurations $C$ and $D$, while not influential according to Cook’s measure $D$, would be influential according to $R$. (See Little 1985 for a discussion of the theoretical relation between $D$ and $R$.)

Of the two “similarity” measures suggested by Hadi (for clustering), $S^*$ is clearly inferior to $S$, on theoretical as well as empirical grounds. On applying them to the stack-loss data, each was far less effective compared to the use of $H^*$ in our article.

Hadi’s $S^*$ does have the very nice property that for subsets of size two $s_{ij}$ is monotonically related to $R$ for subset $(i,j)$, so not only can $S^*$ be used to find the “best” size-2 subsets directly (i.e., without applying any clustering routine), but its elements can also be used to rank all subsets of size 2 according to the measure $R$. However, clustering the stack-loss data...
using $S^*$ produced, for subsets of size 4 or less, only 
(4, 21), (1, 2), (1, 2, 21), and (1, 2, 4, 21), which account 
for only two of the subsets in table 1 of Gray and 
Ling (1984) (also see table 3 for the ranks of these 
subsets according to different measures). Clustering 
Hadi's $S$ produced the following subset candidates: 
(1, 3), (2, 21), (7, 9), (11, 12), (2, 4, 21), (6, 7, 9), (10, 11, 
12), and (6, 7, 8, 9).

In conclusion, we thank Professor Hadi for cor-
recting the typographical errors in our Equation (3.1) 
and for offering some stimulating suggestions.

Robert F. Ling  
Department of Mathematical Sciences  
Clemson University  
Clemson, SC 29631

J. Brian Gray  
M. J. Neeley School of Business  
Texas Christian University  
Fort Worth, TX 76129

ADDITIONAL REFERENCES

Series, Meaning Polynomials, Illustrated on Band-Spectroscopic 
Data,” Technometrics, 16, 147–185.
Regression and ANOVA,” The American Statistician, 32, 17–22.
Little, J. K. (1985), “Influence and a Quadratic Form in the 
Mosteller, F., and Tukey, J. W. (1977), Data Analysis and Regres-
sion, Reading, MA: Addison-Wesley.