Comment

David C. Hoaglin and Peter J. Kempthorne

We thank Chatterjee and Hadi for their review of diagnostics for influence of individual cases in regression. By describing connections and distinctions among a large number of techniques, they have provided a springboard for a timely discussion of specific methods, general approaches, and the contribution of diagnostics to the practice of regression analysis. We offer some criticism of notation, consider cutoffs, rules of thumb, and their role in identifying influential cases, propose simple residual plots which display high leverage, outlying, and influential cases simultaneously, comment on the selection of carrier subsets, discuss approaches to uncovering influential groups of observations, urge more comprehensive presentation of examples, and sketch a step by step diagnostic strategy that should be useful in practice.

NOTATION

It is extremely unfortunate that Chatterjee and Hadi have chosen to introduce new notation for familiar diagnostic quantities such as P for H in equation (5), WK_i for DFITS_i in equations (34) and (35), and D_{ij}^* for DBETAS_{ij} in equations (42) and (44). By the time one gets to the summary in Section 9 and the example in Section 10, only the new names remain; the customary ones have faded from memory, and one must retrieve them (e.g., via the equation numbers in Table 2) to retain contact with other discussions in the literature. Such confusion could easily have been avoided. A consensus on notation for the basic quantities in regression diagnostics would be most welcome.

IDENTIFYING INFLUENTIAL CASES

For labeling cases as having "high leverage," the cutoff $2p/N$ for large h_i is neither the only rule of thumb proposed nor even the most useful rule. Hoaglin and Welsch (1978) proposed $2p/N$ on the basis of limited initial experience, and Velleman and Welsch (1981) suggested that, when $p > 6$ and $N - p > 12$, $3p/N$ is more appropriate. Huber (1981, pages 160–162) prefers to place cutoffs at 0.2 and 0.5, without regard to p and N: "Values $h_i \leq 0.2$ appear to be safe, values between 0.2 and 0.5 are risky, and if we can control the design at all, we had better avoid values above 0.5." In practice we usually examine the h_i in a stem-and-leaf display and identify points of high(er) leverage by considering its appearance in light of the various rules of thumb.

For the example in Section 10, the stem-and-leaf display (for the 6-carrier model) appears in our Table 1. Observation 17 (which Chatterjee and Hadi flag) stands out at .92, exceeding even $3p/N = .9$. The $2p/N$ rule does not flag any additional observations, but the higher of Huber's cutoffs catches observation 2 at .50, and we should probably give further attention to observation 3 at .49. In all this, however, we must recognize that fitting 6 carriers to 20 observations gives only 3⅓ observations per parameter. With the average h_i at $p/N = .3$, it is hardly surprising that 14 of the 20 h_i exceed Huber's lower cutoff. As Weisberg (1981) explained, Moore began with the six carriers in seeking to build a model.

After identifying such cases, it is important to explain the source of their high leverage. By definition, they are outliers in the carrier space and may represent large measurement errors (e.g., miscoded values) or valid observations in the extremes of the ranges of the carriers. Case 17 is an apparent outlier in the empirical distribution of total volatile solids (X_4). Also, we note that the high leverage cases tend to fall near either the start or the end of the 220-day data collection period. A possible explanation is that the carriers vary systematically with the "DAY" variable presented in Weisberg's (1981) description of the experiment.

The discussion in Section 10.2 leads us to question

TABLE 1

Stem-and-leaf display of h_i, the diagonal elements of the hat matrix, for Moore's data with the model whose carriers are 1, X_1, X_2, X_3, X_4, and X_5.

.0	99	
.1	4567	
.2	023568	
.3	4667	
.4	19	20, 3
.5	0	2
.6		
.7		
.8		
.9	2	17

Note: The numbers at the right identify the observations by their rows in Tables 3 and 5.

David C. Hoaglin is Research Associate in Statistics, and Peter J. Kempthorne is Assistant Professor of Statistics, Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, Massachusetts 02138.