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.

how one should determine which observations appear most influential on the precision of $\hat{\beta}$. For example, on the basis of CVR, Table 6 lists observations 1, 17, 20, 15, and 7, whose CVR values are 0.04, 12.51, 0.45, 0.51, and 0.73. Only the first two of these, however, lie outside the suggested cutoffs at $|\,\mathrm{CVR}_i - 1\,| > 3p/N = 0.9$ (i.e., 0.1 and 1.9). Table 5 shows that observations 2, 3, 4, 9, 14, 16, and 18 all have $\mathrm{CVR}_i > 1.9$. By contrast, in Figure 5, which plots CW_i against i, observations 1, 17, 20, 15, and 7 clearly stand out. A stem-and-leaf display or an index plot of CVR_i would call attention to these same five observations. The message from the batch of CVR_i is that values in the range 1.3 to 2.8 are common (15 of 20), whereas values below 1.0 or above 3.0 are not.

The conflict arises from the difference between applying an absolute cutoff ($1 \pm 3p/N$ for CVR_i) and seizing upon apparent structure within the batch of values of the diagnostic measure. In this example the cutoff values for CVR_i treat increasing $\det[\mathrm{cov}(\hat{\beta})]$ by a factor of 1.9 as roughly equal in importance to decreasing it by a factor of 10 (in part because p/N is not small). We prefer to regard change by a factor of 2 as comparable to change by a factor of ½. This attitude leads us to prefer $\log(\mathrm{CVR}_i)$ as a measure of influence on the precision of $\hat{\beta}$. Thus, we focus on observations 3, 9, and 2 (with $\mathrm{CVR}_i = 2.75$, 2.45, and 2.39) before turning to observations 20, 15, and 7. Use of $\log(\mathrm{CVR}_i)$ is equivalent to use of the statistic of Cook and Weisberg (CW_i) in equation (25), but it does not rely on distributional assumptions for its interpretation. In practice we also make some allowance for the fact that simply deleting a noninfluential observation will tend to produce a modest increase in CVR_i (or its logarithm).

Most uses of regression diagnostics are likely to give more weight to WK_i or D^*_{ij}, so we turn briefly to the role of cutoffs for these measures in identifying influential observations. The values of WK_i flagged by asterisks in Table 5 turn out to be precisely those with $|\,\mathrm{WK}_i\,| > 2\sqrt{p}/\sqrt{N} = 1.095$. In Table 8, however, four unflagged entries exceed the proposed cutoff $|\,D^*_{ij}\,| > 2/\sqrt{N} = 0.447$, suggesting that observations 15 and 19 may be influential on $\hat{\beta}_2$ and observations 1 and 2 may be influential on $\hat{\beta}_3$. Three of these four have larger magnitudes than $D^*_{5,1}$, which is flagged. Although we do not believe that good diagnosis can be as cut and dried as simply checking diagnostic measures against cutoffs, it requires clearly stated criteria and guidelines.

We are also puzzled that Chatterjee and Hadi ignore the influence of cases on the intercept estimate. When the goal of the analysis is prediction, all the parameter estimates play key roles. A large value of WK_i might be due to a large D^*_{i0} value.

RESIDUAL VERSUS LEVERAGE PLOTS WITH INFLUENCE CONTOURS

In addition to studying tables or simple index plots of various influence measures, as Chatterjee and Hadi suggest, our influence analysis includes plots of (externally) studentized residuals (t_i^*) against leverage values (h_i). We supplement these plots with contours corresponding to constant values of particular influence measures and labels of "interesting" points. Outlying and high leverage observations are easily identified as those at the extremes of the vertical and horizontal directions, respectively. The contours give the relative influence of observations and indicate how particular measures depend on the leverage and residual values. Our Figures 1 and 2 present these plots for the least squares fit of the full model to Moore's data with contours corresponding to DFITS and COVRATIO, respectively. Welsch (1983), Krasker and Welsch (1983), and Gray (1983, 1985) propose similar plots using different scales for the axes.

GROUP INFLUENCE

In Section 8, Chatterjee and Hadi allude to the natural generalizations of the singleton measures to groups. Belsley, Kuh, and Welsch (1980) and Cook and Weisberg (1982) discuss these quite extensively, lamenting the large number of groups to consider and the greater difficulty in computing a measure's value for groups of more than one observation. In addition to the greater computational complexity, analyzing group influence raises new conceptual questions. For

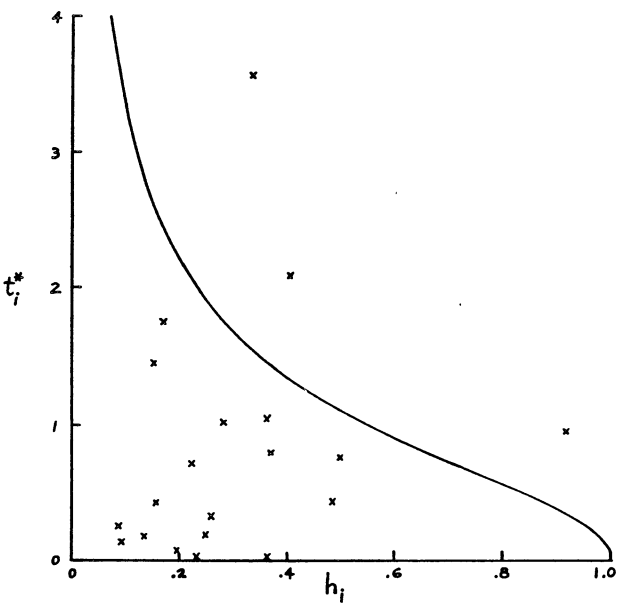

FIG. 1. *Residual versus leverage plot with contour for* $DFITS_i = 2\sqrt{p/N}$.

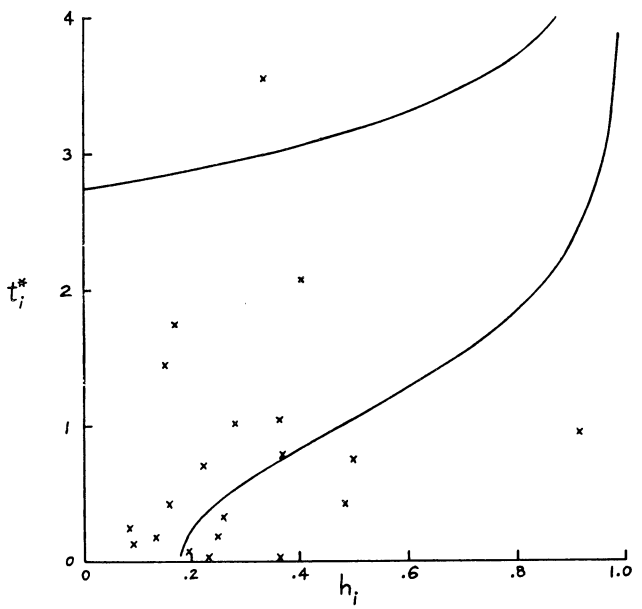

FIG. 2. *Residual versus leverage plot with contours for* $COVRATIO_i = 1 \pm 3p/N$.

a group to be deemed "influential," should its members reinforce each other's impact? If the influence of a group can be explained by a proper subgroup, then should our focus be on that subgroup? Is mutual "masking" of observations a serious problem in practice? If so, how can it be measured, and how can such groups of observations be identified?

In our experience, important influential groups can be overlooked when the group versions of the measures that Chatterjee and Hadi review are applied exhaustively to all groups of a given size. The 10 or 15 "most influential" groups of a given size might all contain the same proper subgroup. In such circumstances, all the observations in any influential group may not reinforce each other's influence, and a subgroup of observations may be responsible for many groups' having significant values of the influence measure. For example, using the influence measure of Cook, C_i, to identify influential pairs and triplets of observations in the simple linear regression fit to Mickey, Dunn, and Clark's (1967) adaptive score data (see Figure 1 of Chatterjee and Hadi), the 19 most influential pairs each contain child 18 (denoted by "I" in Figure 1), and the 18 most influential triplets each contain the pair consisting of children 2 and 18 (the pair lying closest to the lower right corner of Figure 1).

We believe that a promising approach is to study group influence in terms of the derivatives of weighted least squares estimates of the linear regression parameter. Extending the measures presented in Belsley, Kuh, and Welsch (1980), Kempthorne (1986) proposes the "derivative influence" measure, for which singleton effects are additive in characterizing the influence of groups. The computational complexity of the group analysis is reduced substantially from that required by group measures based on group deletion. In fact, derivative-influential groups can be identified with the aid of a cluster analysis. However, unlike Gray and Ling's (1984) clustering, the positions of observations in the underlying multidimensional configuration have a theoretical foundation. They characterize the magnitudes and the directions of observations' influence on the estimate of the linear regression parameter. In addition, this approach includes a simple measure of the "cohesion" of a group of observations that indicates the degree to which the observations influence the regression parameter estimate in the same direction.

Analysis of the derivative influence of groups of cases in the least squares linear regression fit to Moore's manure data reveals that there are no cohesive derivative-influential groups of multiple cases. The four cases with highest derivative influence among the singletons are, in decreasing order, 1, 17, 20, and 2. For any group of these observations, the cohesion is negative, indicating that the singletons have distinct impacts on the regression parameter estimate (the angle between any pair of direction vectors characterizing observations' influence is obtuse, with negative cosine). Grouping any of the singletons together would seem inappropriate. As an aside, we note that, although Chatterjee and Hadi do not highlight observation 2 (see Table 6), a careful examination of Table 8 indicates that its impact on $\hat{\beta}_3$ is the largest among those with negative D_{ij}^* values. In fact, it exceeds the size-adjusted cutoff $2/\sqrt{N} = .4472$ suggested by Belsley, Kuh, and Welsch.

The discussion in Section 8 of observations "masking" each other's influence is misleading. Contrary to what Chatterjee and Hadi state, in the hypothetical example of points "1" and "2" in Figure 3, the points *are* individually influential, and their impact as a group is well approximated by the sum of their individual impacts. The characterization and assessment of masking effects among observations is an important problem in influence analysis which deserves further study. Such efforts would be better focused if researchers had more real data sets exhibiting this behavior to study.

MODEL (SUBSET) SELECTION

Chatterjee and Hadi comment on the impact of deleting some groups of observations on the choice of "best" subset of explanatory variables based on the minimum RMS criterion. Use of one selection criterion without comment might suggest to readers that experts in regression commonly prefer it for model

selection. To the contrary, no theoretical foundation justifies this criterion. It is a generalization of stepwise regression for all subsets which corresponds to using tests of approximate level .50 (not .05) to compare all pairs of nested subsets. Because the implicit significance levels of the tests do not converge to zero as the sample size N grows large, the procedure is not consistent for identifying a true subset.

Variable selection and outlier exclusion can be considered simultaneously in the context of subset selection by including indicator variables for the observations corresponding to potential outliers. If the purpose of the analysis is to identify the "true" subset and "real" outliers, then a consistent subset selection criterion would seem desirable. A logical choice is the Bayesian Information Criterion, originally proposed by Schwarz (1978), which is asymptotically equivalent to the log of the posterior probability for the model corresponding to a subset, assuming any prior distribution for unknown parameters that gives positive probability to different subsets (models) and positive support to the all regression parameters corresponding to any subset; see also Atkinson (1978, 1981a).

When model selection or hypothesis testing is the purpose of the regression analysis, it is important to assess the influence of observations on the selection or test. None of the measures that Chatterjee and Hadi review are appropriate in this situation. Using a Bayesian decision-theoretic framework for general regression-fitting problems, Kempthorne (1985) develops several influence measures and applies them to the model-selection problem. For special cases of the prior distribution for unknown regression parameters, the measures for a particular observation are simple functions of its residuals and leverage values in the least squares fits to the models under consideration. Because Moore's data has little interesting structure, we do not illustrate these measures here.

We are puzzled by Chatterjee and Hadi's mention of a set of "most influential variables." A discussion of influential *observations* is complicated enough, and consideration of influential *variables* is particularly difficult when no formal definition is given.

DESCRIPTION OF DATA

The description of the example in Section 10.1 gives far too little information for readers who are not already familiar with the subject to gain even a basic idea of the scientific issues involved and the substantive goals of the analysis. Although Weisberg (1981) does not say a great deal more, he does explain at the start that these data are observations on the same sample of dairy waste over time, and his Table 1 includes the day of each observation (counting from 0 to 220). This alerts the analyst to the possible role of day as a lurking variable (Joiner, 1981). In contrast, the statement by Chatterjee and Hadi that "the data were collected on samples kept in suspension in water for 220 days" seems likely to mislead. We offer these comments on adequacy of description not so much to criticize the present authors (who, after all, did not introduce this data set into the literature) as to urge that examples be made accessible to a much wider audience. Even though journal space is often scarce, almost everyone (including the present discussants) should be able to do better in this area.

A DIAGNOSTIC STRATEGY

Although the two alternative sets of diagnostic measures, $\{WK_i, CW_i, D_{ij}\}$ and $\{C_i^*, CVR_i, D_{ij}^*\}$, tell much about influential observations, we do not feel that they provide an adequately comprehensive picture of the key aspects of a multiple regression data set. In practice we follow a diagnostic strategy much like the following. At almost any step, we may decide that we must modify the model (e.g., by dropping or adding a carrier) or the data (e.g., by setting aside an observation) or both (e.g., by applying a transformation) and then resume diagnosis at some earlier step.

0. Examine the Variables One at a Time. We often use stem-and-leaf displays to get a look at skewness, possible outliers, and other features.

1. Plot the Data. We try to look at scatterplots of Y against each X_j, as well as a scatterplot for each pair of carriers. The scatterplot matrix (Cleveland, 1985) offers a very effective way of organizing this information. (It is surprising that neither Weisberg (1981) nor the present paper includes any of these plots. They aided our understanding of the data when we made a set and studied them.)

2. Check on Leverage. It is often helpful to know whether any design points have high leverage, as measured by the diagonal elements of the hat matrix. We usually study these in a stem-and-leaf display.

3. Examine Residuals. As a main ingredient of the influence measures, the residuals deserve some scrutiny for their own sake. We often use the (externally) studentized residuals (t_i^* in the present paper) in such displays as a stem-and-leaf display and a normal probability plot, and we may plot the ordinary residuals against a variety of carriers. In a related step we also consider the values of $s_{(i)}^2$, for example, by plotting $s_{(i)}^2$ against i.

4. Make Partial Regression Leverage Plots (or added variable plots), one for each carrier.

The order of Steps 5, 6, and 7 depends on the aims of the regression analysis. Others may prefer to substitute some alternative measure of the same type of influence at any of these steps.

5. Study Influence of Individual Observations on Fit. We customarily plot DFITS_i (WK_i) against i.

6. Study Influence of Individual Observations on Estimates of Coefficients. For each j we plot $\text{DBETAS}_{ij}(D_{ij}^*)$ against i, and we look at these plots in parallel.

7. Study Influence of Individual Observations on the Estimated Covariance Matrix of $\hat{\beta}$. Here we plot COVRATIO_i (CVR_i) against i. In Steps 5 and 7 we also examine the residual versus leverage plots with iso-influence contours.

8. Probe for Subsets of Observations That Are Jointly Influential. Although more research is needed in this area, we feel it forms an important part of the diagnostic strategy. The k-clustering approach of Gray and Ling (1984) and the derivative influence techniques of Kempthorne (1986) seem promising. Another, more ad hoc, approach is to drop the observations (say, three or four) that have the most individual influence and then see how much the results change.

For a diagnostic analysis, this strategy constitutes a bare minimum. Often, other areas of diagnosis are critical to the analysis: need for transformation, influence on model choice, or detecting departures from the standard Gauss-Markoff assumptions such as heteroscedastic or correlated errors. Research in these areas among others has been especially active in recent years, including applications of a Bayesian perspective. See, e.g., Atkinson (1982), Cook and Weisberg (1983), Dawson (1985), Johnson and Geisser (1983), and Pettit and Smith (1985).

ACKNOWLEDGMENTS

This work was supported in part by Contract DAAG29-85-K-0262 between the United States Army Research Office and Harvard University and by National Science Foundation Grant SES-8401422.

ADDITIONAL REFERENCES

ATKINSON, A. C. (1978). Posterior probabilities for choosing a regression model. *Biometrika* **65** 39–48.

ATKINSON, A. C. (1981a). Likelihood ratios, posterior odds, and information criteria. *J. Econometrics* **16** 15–20.

CLEVELAND, W. S. (1985). *The Elements of Graphing Data.* Wadsworth, Monterey, Calif.

COOK, R. D. and WEISBERG, S. (1983). Diagnostics for heteroscedasticity in regression. *Biometrika* **70** 1–10.

DAWSON, R. (1985). Diagnosing data and prior influence in a Bayesian analysis. Unpublished Ph.D. thesis, Dept. Statistics, Harvard Univ.

GRAY, J. B. (1983). The L − R plot: a graphical tool for assessing influence. In *1983 Proceedings of the Statistical Computing Section* 159–164. Amer. Statist. Assoc., Washington, D. C.

GRAY, J. B. (1985). Graphics for regression diagnostics. In *1985 Proceedings of the Statistical Computing Section* 102–107. Amer. Statist. Assoc., Washington, D. C.

HUBER, P. J. (1981). *Robust Statistics.* Wiley, New York.

JOHNSON, W. and GEISSER, S. (1983). A predictive view of the detection and characterization of influential observations in regression analysis. *J. Amer. Statist. Assoc.* **78** 137–144.

JOINER, B. L. (1981). Lurking variables: some examples. *Amer. Statist.* **35** 227–233.

KEMPTHORNE, P. J. (1985). Decision-theoretic measures of influence in regression. In *1985 Proceedings of the Business and Economic Statistics Section* 429–434. Amer. Statist. Assoc., Washington, D. C. To appear in *J. Roy. Statist. Soc. Ser. B.*

KEMPTHORNE, P. J. (1986). Identifying derivative-influential groups of observations in regression. Memorandum NS-540, Dept. Statistics, Harvard Univ.

KRASKER, W. S. and WELSCH, R. E. (1983). The use of bounded-influence regression in data analysis: Theory, computation, and graphics. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface* (K. W. Heiner, et al., eds.) 45–51. Springer, New York.

PETTIT, L. I. and SMITH, A. F. M. (1985). Outliers and influential observations in linear models. In *Bayesian Statistics* **2** (J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, eds.) 473–494. North Holland, Amsterdam.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

WELSCH, R. E. (1983). Discussion of Developments in linear regression methodology: 1959–1982, by R. R. Hocking. *Technometrics* **25** 245–246.

Comment

Paul F. Velleman

I congratulate Chatterjee and Hadi on an excellent survey of an area that has developed rapidly in the past decade. One of the disappointments of this area is that these very valuable techniques have been slow to infiltrate the literature of disciplines using regression techniques. We need to turn some of our attention to promoting the use of diagnostic statistics in ordinary practical analyses.

One problem with regression diagnostics has been that terminology has not yet standardized. Unfortunately, Chatterjee and Hadi exacerbate rather than alleviate this problem. I do not believe that we need yet another name and notation for the Hat matrix, nor that we benefit from new and somewhat cryptic

Paul F. Velleman is Associate Professor of Economic and Social Statistics, Cornell University, 358 Ives Hall, Ithaca, New York 14853.