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 $DBETAS_{ij}(D^*_{ij})$ 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 Unpublished Ph.D. thesis, Dept.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.

names for statistics that were provided good names by their creators.

I must also quibble with the authors' definition of "influential." The issue (as Chatterjee and Hadi point out) is, "influencing what?" They propose that a point is influential if its deletion would alter the coefficients or predicted values sufficiently. While this definition makes sense, it encourages us to ignore other extreme points. Many statistical analyses begin and end with correlation matrices. Diagnostic statistics are an important tool for demonstrating the danger of such methods. While the "influential" points according to Chatterjee and Hadi are certainly important to correlation analyses, a correlation may be "significant" only because of an outlying data point that is *not* "influential" because it lies on the line described by the remaining data points. I suggest that any point extreme enough to affect the correlation, t, or F statistics should also be dubbed "influential" lest we fail to note it in diagnosing the data.

While inconsistent terminology and notation may have slowed the spread of regression diagnostic methods, they cannot account for the fact that we have a collection of useful, practical, and computationally affordable methods that are not widely used in practical regressions. Statisticians have known of these methods for a decade or more, but the consumers of statistics, who look to statistics as a source of tools for understanding their data, have not adopted diagnostics. Readable survey articles such as this one can help to promote regression diagnostics to this audience, but they are not sufficient. I believe that much of the additional progress must come from statistical computing.

Until a few years ago, it was awkward to compute and work with diagnostics on the major statistics packages. Most modern statistics packages provide some diagnostic statistics. Diagnostic displays such as partial regression leverage plots are somewhat less common (and are unfortunately difficult to construct on many packages. I often use the task of constructing such a plot from basic regression and graphics commands as a simple test of a statistics package's flexibility.)

Diagnostic methods also require flexibility from the analyst. Published guidelines for when a point is influential are vague and in some cases contradictory. I think that an argument can be made for avoiding an arbitrary categorization of points as influential or not, and simply examining any point that is extreme on any of several diagnostic measures without regard to arbitrary trigger points.

Judgment is also needed simply to decide what diagnostics to consider. A sheaf of partial regression leverage plots, leverage stem-and-leaf displays, and Cook's distance-based confidence ellipsoids can easily overwhelm even the most diligent analyst.

Thus, a regression analysis with diagnostics is likely to require extra analysis steps (e.g., to examine histograms or stem-and-leaf displays of leverage or DFFITS values, to construct new displays, or to find trial regressions with a point or two deleted). Such analyses emphasize the data analysis *process* rather than the final analysis itself. By poking around among the possibly influential data values (as well as assessing the need for transformations and checking for collinearities), we learn more about our data and about competing regression models.

Computing environments that support the process of regression analysis are not yet widely available. Because diagnostics require both judgment and the freedom to try out alternative models, the data analyst requires a flexible interactive environment in which he can try alternatives (many of which he will reject) without spending much time or money. As microcomputers grow in power and flexibility, we will see the desktop workstation become a data analysis tool rather than just a small mainframe computer. Prototype data analysis environments that take advantage of the graphics and intense interaction of personal machines are under development at several places.

At first, some researchers may resist diagnostic statistics. It is, after all, much easier to dump the numbers into a stepwise regression and publish the resulting model. The point that a stepwise regression is almost certainly not the best and is very likely to be simply wrong has not been made clearly, loudly, or often enough by statisticians. (One of the best arguments against such automated regression-finding algorithms is a good diagnostic analysis of the data.) Nevertheless, most researchers really do want to discover what is happening in their data, if only because analyses that raise new questions address one of the really important problems of science: "What should my next grant proposal be about?"

Artificial intelligence-based "expert" programs that use diagnostic statistics (and other measures) to guide a regression analysis may offer a painless way to introduce diagnostic methods to the consumers of statistics. However, the need to combine human judgment with diagnostics makes it unlikely that this approach will provide high quality analyses in the near future. I think it may be more productive in the short run to use artificial intelligence methods to sift through the morass of diagnostics, selecting the displays and statistics that seem most informative about the data.

Certainly survey articles such as this one are an important step in making regression diagnostics accessible to the consumers of statistics. If we want the best statistical methods to influence the way researchers analyze data, we will need to devote attention to making them practical, accessible, and easy to use.