Comment

Roy E. Welsch

INTRODUCTION

By the time this paper is in print, it will be about 10 years since the early work on regression diagnostics began to appear. It is appropriate that there be a review paper with discussion to sum up where we stand and take a look at where this area might be headed. In a few years I hope we have another such paper about similar work for generalized linear regression models (McCullagh and Nelder, 1983).

One of my teachers told me it takes about 10 years for new ideas (new approaches) to go from research paper to widespread use. I think this has been the case for regression diagnostics. Many regression texts (a recent example is Myers (1986)) incorporate some of the material and we now have three books specializing in this area: Belsley, Kuh, and Welsch (1980), Cook and Weisberg (1982), and Atkinson (1985). However, I am sorry to see that very few basic statistics texts which cover early regression ideas also mention diagnostics.

Perhaps of even more value for the rapid diffusion of a new idea is the incorporation of computational support in a variety of data analysis systems. This has certainly been the case for regression diagnostics, although much more remains to be done.

Since the field of regression diagnostics now includes the work of many people, there are naturally different viewpoints, different notations, and even heated discussions. This is as it should be, but a review paper should make some attempt to sort out the issues and provide a coherent base. Chatterjee and Hadi have given us a push in this direction, but not without a few complications.

GENERAL COMMENTS

I am hardly one to complain about notation since I will never live down DFFITS, etc. (DFFITS was originally DIFFIT in the computer and it became DFFITS when we scaled it. I have tried in recent work to rename it DFITS, but with mixed success.) Cook chose D which I am told does not stand for Dennis, but could stand for many distances. Chatterjee and Hadi have tried to use last names to denote DFITS and D. DFITS$_i$ becomes WK$_i$, D_i becomes C_i, and Welsch-Kuh-Atkinson becomes C_i^*!? If an asterisk is to denote a diagnostic with s replaced by $s(i)$, then calling this C_i^* is most confusing. Other changes such as using P for H when others use V or H and the further confounding of p_i and the use of the asterisk in p_i^* are not helpful. I think H, D, DFITS, and DBETAS will stand the test of time. Since "studentized" residual can mean either internally or externally studentized, it pays to be specific for these and an asterisk is acceptable as long as it is used consistently.

Throughout this paper various cutoffs (calibration points) are proposed. In fact, formal tests and cutoffs can often be devised when we condition on X or are so bold as to give X a distribution. Without one of these assumptions, theoretical results are difficult. Simulation has real possibilities, but cannot be done casually. Most of the cutoffs suggested are ad hoc and should not be sanctified in any way. Good plots with informal cutoffs seem to work well because the cutoff can be taken in the context of the rest of the points on the plot.

Over the years, I have favored DFITS over D for two reasons. I like to know the sign of the change in fit (we could use (signum)D) and I like a robust scale. I sometimes replace $s(i)$ by a very robust scale called MAD (median absolute deviations from the median). This treats the scale as a nuisance parameter that should be estimated less efficiently but very robustly.

It would also be nice to estimate the metric X^TX robustly (equivalently find a robust distance analogous to h_i) instead of using $X^T(i)X(i)$ as I suggested in Welsch (1982). This is not as easy to do, but the literature on robustness provides some possibilities.

I am sorry to see that Chatterjee and Hadi endorse the term "added variable plot" when X_j is part of the original model. If X_j is a new regressor, then added variable is a good term. When X_j is already part of the model, I would like another term. Originally, I thought it should be "partial regression plot" but everyone confused this with a partial residual plot. Hence we added the word leverage. At the risk of further confusion, we might try "adjusted partial residual plot" since the unadjusted one plots $e + \hat{\beta}_j X_j$ against X_j and the adjusted one plots $e + \hat{\beta}_j X_j$ (adjusted) against X_j (adjusted). These plots are not hard to compute (Velleman and Welsch, 1981).

Chatterjee and Hadi say that "if estimation of β is of primary concern, then measuring the influence of observations on $\hat{\beta}$ is appropriate. . . ." In reality, we

Roy E. Welsch is Professor of Statistics and Management Science, Sloan School of Management, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139.

cannot talk about an estimate without a measure of precision. Thus we should always be concerned with influence on $\hat{\beta}$ and on the covariance of $\hat{\beta}$. COVRATIO is one idea; any reasonable scalar summary (trace, determinant, condition number, etc.) of the covariance will probably do. Because determinants often scare people, we introduced

$$\text{FVARATIO} = \frac{\text{var}(\hat{y}_i(i))}{\text{var}(\hat{y}_i)} = \frac{s^2(i)}{s^2(1 - h_i)}$$

which has many of the same properties as COVRATIO.

In fact, we might like a measure that combined influence on the center of the confidence region ($\hat{\beta}$ and $\hat{\beta}(i)$) with its volume or "size" (related to $s^2(X^TX)^{-1}$ or $s^2(i)(X^T(i)X(i))^{-1}$). Our original idea was

$$\text{DTSTAT}_j = \frac{b_j}{s\sqrt{X^TX_{jj}^{-1}}} - \frac{b_j(i)}{s(i)\sqrt{(X^T(i)X(i))_{jj}^{-1}}}$$

which got relegated to a footnote in Belsley, Kuh, and Welsch (1980) because it seemed far better to look at center and size separately but simultaneously rather than in one omnibus statistic. This idea can be extended to the fit, a linear combination, or a one-dimensional confidence interval. The likelihood distance also combines (in a different way) center and size. However, all of these combined measures are messy and it is probably better to keep change in center separate from change in size. A plot comparing the two is best. Plots with a function of leverage on one axis and a function of the residuals on the other usually get the whole story across. Since both COVRATIO and DFITS are functions of these quantities, they can both be located on the same plot providing the actual magnitude of each is appropriately coded (color-coded observation numbers to denote points makes this easy.) Related ideas are contained in Krasker and Welsch (1983) and Samarov and Welsch (1982).

SOME HISTORY AND PHILOSOPHY

My own ideas on diagnostics grew out of theoretical work on robust estimation and applied work on econometric models. One model we were working on just did not agree with economic theory. After looking at residuals and some plots without noting anything, I questioned the economic theory. The economists said our estimated model could not be right, so I checked all of the data again. One observation had been entered incorrectly several iterations back and it seemed to me that I should have had a better clue to that problem long before I questioned the economic theory. Leaving each observation out one-at-a-time was easy to talk about and after a few weeks of programming easy to do in practice.

George Box and others might call this "data criticism." Really it is just part of checking assumptions related to data, systematic models, and stochastic models. Words like stability, sensitivity, and perturbation are surely as important as any other words in statistics. Assumptions should be checked and one rewarding outcome of regression diagnostics is that data analysts are pausing to check assumptions.

Naturally we need to provide useful and coherent ways to check assumptions. We also need to provide guidance on how to proceed when assumptions are violated. There is a lot of work left to do in both of these areas.

SUBSETS OF DATA

Chatterjee and Hadi briefly touch upon the problem of influential groups of observations. All of the measures discussed in their paper can be shown to fail for clumps of influential observations, leverage points, and/or outliers.

The naive natural extensions (two, three, etc. at-a-time) to setting aside one point are computationally expensive but no longer infeasible in a computing environment of networked workstations. Background asynchronous concurrent computing can do the job when there are so many machine cycles available due to the advent of advanced personal workstations in virtually every office. Supercomputers can handle the very large problems.

However, it often provides great insight to try to be more elegant (efficient) and less naive. Kempthorne (1985 and 1986) has provided new ideas and improved versions of some simple subset approaches in Belsley, Kuh, and Welsch (1980) among others.

A more radical approach is to work from the bottom up by leaving only the minimum number of points to fit a model with p parameters, namely p. This means that only $\binom{n}{p}$ subsets need be examined instead of $\binom{n}{1} + \binom{n}{2} + \binom{n}{3} + \cdots$. Some things are lost, but perhaps not very much. An interesting discussion is contained in Hawkins, Bradu, and Kass (1984).

Since we assume that at least half of the data is good we see that, in general,

$$\binom{n}{n/2} \gg \binom{n}{p}$$

and the naive set-aside approach will take far more work than the $\binom{n}{p}$ approach (which is not easy either). Some subsets of size p will contain only "good" data but we do not know which ones.

So assume we have all p point regression planes chosen from n points. For each plane we can compute residuals, change in fit, or prediction, and a wide variety of other things. (For one suggestion see Rousseeuw, 1985.) Recent work with Alan Zaslavsky at the

Massachusetts Institute of Technology indicates that good subset diagnostics can be constructed in this way. If the computing is overly burdensome in a particular situation, sampling of p from n or simulated annealing can provide useful approximations.

Why do this at all? Some have argued that there are few real examples where clumps occur. Yet, it is easy to construct simulated examples. The reason we have few real examples is because we have no microscope with enough resolution to see the problem. When we do and still find no real examples, then we can go on to other things.

CONCLUDING REMARKS

Perhaps the most difficult task I have undertaken (and by no means completed) in recent years is to develop regression analysis strategies for the guided-computing project at the Massachusetts Institute of Technology (Oldford and Peters, 1985). Even with a vast arsonal of diagnostics, it is very hard to write down rules that can be used to guide a data analysis. So much is really subjective and subtle. Guided computing forces us to consider chance as a possible cause in any diagnostic exploration. It is perhaps a form of controlled magical thinking (Diaconis, 1985). A great deal of what we teach in applied statistics is *not* written down, let alone in a form suitable for formal encoding. It is just simply "lore."

Progress can be made for very restricted problems detected by diagnostics. However, as soon as we try to attack influential data and collinearity, or influential data and model selection, or influential data and transformations, etc. it gets much harder. Multiplicity and simultaneity (of problems and analyses) are additional important words for statisticans to remember. They provide an incredible challenge for the future of diagnostics and statistics.

ACKNOWLEDGMENT

This research was supported in part by National Science Foundation Grant DCR-8116778 and Army Research Office Contract DAAG29-84-K-0207.

ADDITIONAL REFERENCES

ATKINSON, A. C. (1985). *Plots, Transformations, and Regression.* University Press, Oxford.

DIACONIS, P. (1985). Theories of data analysis: From magical thinking through classical statistics. In *Exploring Data Tables, Trends, and Shapes* (D. C. Hoaglin, F. Mosteller, and J. W. Tukey, eds.). Wiley, New York.

HAWKINS, D. M., BRADU, D. and KASS, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics* **26** 197–208.

KEMPTHORNE, P. J. (1985). Identifying rank-influential groups of observations in linear regression modeling. Memorandum NS-539, Dept. Statistics, Harvard Univ.

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. D. (1983). The use of bounded-influence regression in data analysis: Theory, computation and graphics. In *Computer Science and Statistics: Fourteenth Symposium on the Interface* (K. W. Heiner, R. S. Sacher, and J. W. Wilkinson, eds.) 45–51. Springer, New York.

MCCULLAGH, P. and NELDER, J. A. (1983). *Generalized Linear Models.* Chapman and Hall, New York.

MYERS, R. H. (1986). *Classical and Modern Regression with Applications.* Duxbury, Boston.

OLDFORD, R. W. and PETERS, S. C. (1985). DINDE: Towards more statistically sophisticated software. Technical Report 55, Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology, Cambridge.

ROUSSEEUW, P. J. (1985). A regression diagnostic for multiple outliers and leverage points. Abstract 85t-74. *Institute Math. Statist. Bull.* **14** 399.

SAMAROV, A. and WELSCH, R. E. (1982). Computational procedures for bounded-influence regression. In *COMPSTAT 1982: Proceedings in Computational Statistics* (H. Caussinus, P. Ettinger and R. Tomassone, eds.) 412–418. Physica Verlag, Wien.

Comment

Rollin Brant

I strongly agree with the authors' characterization of the number of proposals that have been made regarding outliers and influential points as "bewildering." However, these proposals themselves comprise only a part of a much larger number of methods put forward as useful adjuncts to the criticism of regression models. As a consequence, the conscientious and up to date investigator finds that model validation can be both time-consuming and difficult, requiring the consideration of a myriad of diagnostic quantities and plots. Most difficult of all is the integration of the often fragmentary evidence provided by these procedures into a coherent set of recommendations and/or conclusions. Any attempts at cutting through any portion of this tangled web must be welcomed by all.

Rollin Brant is Assistant Professor, Department of Applied Statistics, University of Minnesota, St. Paul, Minnesota 55108.