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.