Viewpoints

THE USE OF AUTOMATIC INTERACTION DETECTOR AND SIMILAR SEARCH PROCEDURES

IN RECENT years there has been a growing use of stepwise multivariate statistical techniques for examining patterns in data. The Automatic Interaction Detector (A.I.D.), recently illustrated in Heald's paper,[1] is the latest of such techniques to capture the imagination of researchers. It is insufficiently realized, however, that while the output of these procedures are often intuitively appealing, they must be used with extreme care. In particular with A.I.D., there are substantial opportunities for errors due to misinterpretation of the output and spurious results. Most of these errors are illustrated in the paper by Heald.

PROBLEMS WITH A.I.D.

(1) A.I.D. requires very large sample sizes. This was pointed out clearly by Morgan and Sonquist who developed the technique: "a warning to potential users of this program—data sets with a thousand cases or more are necessary; otherwise the power of the search processes must be restricted drastically or those processes will carry one into a never-never land of idiosyncratic results".* With 70 cases, Heald's results must be well within this land. This restriction is easy to understand. Conventional regression analysis is more powerful than A.I.D. because it utilizes the sample observations much more efficiently. In the former, each observation simultaneously measures each relationship so that relatively small sample sizes are needed. With A.I.D., groups are the unit of analysis. In Heald's example, with 6 factors used and if two-way splits are made on each of them, a total sample size of about 2000 is needed to maintain a group sample size of 30. Few statisticians would be prepared to generalize on samples of 3 or 4!

(2) Intercorrelated predictors lead to spurious results. The main advantage of a technique such as regression analysis is of course that, except in extreme cases, it can take care of intercorrelations among the predictors. The regression coefficient represents the net effect of its associated variable. There is no such adjustment with A.I.D. This presents two related problems in interpretation. First, once a given variable is selected, certain other variables become much less likely candidates for inclusion. The more highly correlated a pair of variables, the less likely that both of them will be selected.[2] A common problem that this writer finds is that a second variable may be almost as discriminating as the one

* J. A. SONQUIST, E. L. BAKER and J. N. MORGAN (1971) *Searching for Structure (Alias A.I.D.—III)* p. 1. University of Michigan.

chosen, but if the program is made to split on this, quite a different tree occurs.

The second problem is implied by the above, in that since net effects are not obtained, the analysis can say *nothing* about the "importance" of particular variables in explaining the dependent variable. Heald's conclusion that his tree is "a simple statement . . . that size of store is the most important factor influencing turnover" and "that the number of key-traders in an area is critical to the success of the larger stores" (p. 452), is quite unjustified. This "intuitive" interpretation of A.I.D. results as net effects is a significant source of spurious results.

(3) Bias is created from the model-building process. The search approach to model building, neglecting prior information, is an additional source of spurious results. For example, Heald throws 53 variables into a stepwise regression program and obtains 6 which are significant. However, pure chance would predict this result from random predictors.[3] A similar comment applies to the A.I.D. search. These techniques are quite inappropriate to "reduce the dimension of the problem and to isolate the important factors" (p. 450). The correct approach to the dimensionality problem is to explore the correlations among the predictors and either discard some of them as redundant or use principal components. A further source of bias is caused by sampling error. Even if variables are not selected from a larger population, there are still likely to be some that have some correlation with the dependent variable just by chance. This bias is compensated for in regression analysis by the use of adjusted R^2 which reflects the number of predictors involved. However, no such statistic is available to A.I.D. users.

It should be added that the careful researcher would correct for both sources of bias by splitting the sample, using one part for analysis and the other part for validation.[4]

(4) Bias from noise: this problem has been pointed out by Sonquist.[2] Errors in the data are exaggerated by the problem of multicollinearity with A.I.D. The number of key traders and the catchment population may be almost equally important determinants of turnover—which one is selected will be strongly influenced by the particular sample. Since the initial split strongly affects the whole structure of the tree, this means that the interpretation of the output can be largely determined by noise in the data and measurement errors.

(5) Bias from skewed variables: Morgan and Sonquist have shown that spurious results are generated by either the dependent or independent variables having a skewed distribution. Users, therefore, must determine that their variables are well behaved.

THE EFFECTIVE USE OF SEARCH TECHNIQUES

Search techniques should only be used in the exploratory stage of analysis. If substantial prior information exists, they should not be used at all. It is impor-

tant to understand the motivation for A.I.D. The technique was developed because various researchers were dissatisfied with the statistical assumptions made in using regression analysis, in particular the assumptions of linearity and additivity.[5] It was argued that while the model builder could make hypotheses about the variables to be included in the model, he frequently had little knowledge about the form of the relationships including the types of non-linearities and interactions. Thus A.I.D. was developed to assist model builders to specify the *form* of the model.

Heald's use of stepwise regression *followed* by A.I.D. is consequently quite contradictory. The former requires either the assumption of simple linearity and additivity, or the specification *a priori* of the non-linearities and interactions. A.I.D. was intended to be used first. Once the form of the relationship is suggested, these terms should be built into the regression equation. The latter can then determine the net effects of the predictors.

London Graduate School of Business Studies PETER DOYLE

REFERENCES

[1] G. I. HEALD (1972) The application of the Automatic Interaction Detector (A.I.D.) programme and multiple regression techniques to the assessment of store performance and site selection. *Opl Res. Q.* **23**, 445.
[2] J. A. SONQUIST (1970) *Multivariate Model Building: The Validation of a Search Strategy.* Institute for Social Research, University of Michigan, Ann Arbor.
[3] D. A. AAKER (1971) *Multivariate Analysis in Marketing: Theory and Application.* Wadsworth, Belmont.
[4] D. G. MORRISON (1969) On the interpretation of discriminant analysis. *J. Market Res.* **6**, 156.
[5] J. N. MORGAN and J. A. SONQUIST (1963). Problems in the analysis of survey data and a proposal. *J. Am. statist. Assoc.* **58**, 415.

THE STATE OF RESEARCH IN OR

MR. G. H. MITCHELL, on behalf of the Education and Research Committee, raises interesting and important issues (*Opl Res. Q.* **24**, 3, 1973). I find fault only with the statement that "the peculiar feature of OR as a problem-solving discipline is its use of the scientific method". In practice, closing the scientific "loop" is seldom possible. The OR worker provides a relevant processing of information, "measures" the "value" of various choices in the model world, but rarely, even in that world, does he offer an unequivocal solution to the real-world problem. The future for OR, as problem spaces get larger and more complicated, depends on a fresh look at the "science" of OR and a less physical view of measurement.[1]

The structuring of problems, and of OR itself, is a primary task. It is unsatisfactory that relevant research has been largely carried out in the U.S.A., for it