

COMPUTER CENTER NOTES

COMPUTER RESEARCH CENTER FOR ECONOMICS AND MANAGEMENT SCIENCE

The NBER Computer Research Center for Economics and Management Science has been engaged, since its formation in 1971, in developing new software systems for quantitative social science research. Prototype systems for exploratory data analysis, mathematical programming, and econometrics are now in various stages of design and implementation. General summaries of research in progress, as well as abstracts of specific Research Reports, are a regular feature in the *Annals*. Following are progress reports on the data-analysis and mathematical programming projects, and abstracts of five recent Research Reports.

THE DATA-ANALYSIS PROJECT (MARCH 1973)

The term "data analysis" has become associated in the past few years with a movement in the statistics profession whose primary aim is to reorient statistical theory to face the realities of empirical data. The acknowledged leaders of this movement are John Tukey of Princeton University and his students. (See Tukey [1970-1971].) The central themes of data analysis are that a body of data should be met on its own terms and that sensitive as well as sensible tools are required to extract the messages from the data. The data analysis project at the NBER Computer Research Center is committed to applying this point of view to the problems of empirical research in economics and management science.

The project is conducting research in two broad areas: (1) resistant techniques for fitting linear models, and (2) the analysis of discrete multivariate data. Our work in these areas is discussed in the following paragraphs.

Resistant Techniques for Fitting Linear Models

Linear models play important explicit and implicit roles in the analysis of multivariate data and also provide a flexible framework for describing substantive theories in such fields as economics and sociology. Examples of linear models include multiple regression, the analysis of variance and covariance, and systems of linear stochastic equations. Rather than concentrating on the multivariate-Normal and least-squares theories of these models, we are exploring areas which we believe to be of more urgent practical importance: the impact of recent developments in robust estimation and multiparameter Bayesian methods.

Generally speaking, robust methods of estimation, unlike more standard estimation procedures, are not adversely affected by moderate departures from the basic model—e.g., the presence of a few wild or contaminated data values, or moderate changes in the shape of the underlying distribution. Recent research (Andrews *et al.* [1972], Hoaglin [1971], Huber [1972]) has concentrated on a comparatively simple situation, estimating location and scale parameters in the symmetric univariate model. These problems are becoming well enough understood that we are beginning to transfer our insights to multivariate situations such as regression and the analysis of variance.

During the fall of 1972, David Hoaglin, Roy Welsch, and I studied previous work on robustness to see what might be usefully generalized to multiple regression problems. We concentrated primarily on understanding M-estimators as described in Huber [1972]. One output of this effort was an experimental program, built from facilities available in the Center's TROLL system, for a form of robust regression. Convergence problems with this program led to several fruitful discussions with the Center's numerical analyst, Virginia Klema, and a visiting analyst, Jim Douglas, from the University of Chicago. During the spring of 1973, Hoaglin is teaching a graduate seminar on robust estimation at Harvard University. This seminar provides us a continuing focus for our work on robust regression.

Bayesian methods have been applied to many problems in statistics, but one of the most successful applications has been to the problem of estimating a large number of parameters. The "simultaneous estimation problem", as it is sometimes called, is now receiving greater attention (Efron and Morris [1972], Lindley [1971], Fienberg and Holland [1972]). It needs even greater attention in the estimation of simultaneous equation systems, where the number of parameters is often gigantic. Theoretical work goes back to James and Stein [1961]. The estimators that are emerging from this line of research are neo-Bayesian in character, and the improvement of estimation over "least squares" can be substantial when the number of estimated parameters is large. Since the original work, much research has been done on related problems for more general models. We intend to build on these results to produce new or modified data-analytic tools that are sensitive to the problems of estimating many parameters.

An important pilot study on the effectiveness of various competitors to ordinary least squares estimation in regression has recently been completed by Dr. Nanny Wermuth [1972] under the direction of Professor Arthur Dempster at the Department of Statistics, Harvard University. This Monte Carlo simulation study compared over fifty regression methods for the simplest multiple regression problem; the study suggests that some simple neo-Bayesian estimators provide a consistent improvement over ordinary least squares and deal effectively with the problem of multicollinearity. We expect to pursue this type of simulation study, extending it to study the effects of outliers and complex autocorrelation structures.

One problem we considered in the fall was whether or not there was a reasonable way we could successfully mix robust and neo-Bayesian regression methods. We came up with the following suggestion, which we will explore in the next few months. In regression we distinguish the "fitted values" (i.e., \hat{f}) from the "parameter estimates" (i.e., $\hat{\beta}$). Robust regression methods are really concerned with providing values of \hat{f} which are not unduly sensitive to the effects of only a few observations. The successful neo-Bayesian methods, on the other hand, concentrate on $\hat{\beta}$ and provide estimates which are more accurate than the corresponding least squares estimators. Traditional econometric practice using least squares intermixes the problems of finding \hat{f} and $\hat{\beta}$ (under the guise of providing a simple solution to both problems), but it seems to us from what we now know that they should be separated. We are continuing to investigate this approach to multiple regression problems to see if it may provide more insightful data-analysis tools than we now have.