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., $\hat{y}$) from the "parameter estimates" (i.e., $\hat{\beta}$). Robust regression methods are really concerned with providing values of $\hat{y}$ 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 $\hat{y}$ 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.

Analysis of Discrete Multivariate Data

A rapidly growing statistical literature on the analysis of multidimensional contingency tables is beginning to filter into many disciplines, e.g., sociology, biology, history, and medicine. The central tools in this development are a variety of ways of fitting log-linear models to the cell frequencies of multidimensional tables. I am now collaborating in writing a book aimed at making this technology accessible to a wide audience.

We had originally planned to develop at the Center an interactive program to perform log-linear analyses of multidimensional contingency tables. However, we soon realized that a "statistical language" for manipulating n-dimensional arrays would, if properly done, provide an easy environment in which to write the log-linear analysis program; in addition, it would give the data-analysis researchers an extremely useful tool for solving many other types of problems that arise in our work. Hence, we have been working with the Center's programming staff—primarily, Mark Eisner and Richard Hill—to design such a language and the supporting software system. The system is tentatively named DROSS (for Data-analysis Research-Oriented Statistical System).

When the interactive log-linear analysis program is available, it will allow us to explore the problems of research economists and to ascertain the extent to which this approach can help solve them. This in turn will suggest directions in which the technology needs to be extended. We have found two Center-related projects that may benefit from the log-linear model technology. John Meyer has a six-dimensional table of data on the development of new businesses and their mobility. John Kain and William Apgar are faced with the "inverse problem"; i.e., they know some of the two-dimensional faces of a five-dimensional table and want to fill in the entire table as best they can, based on this margin information. The same technology can be used to attack both of these problems, and this flexibility is one of the appealing features of the log-linear approach to discrete multivariate data.

Paul W. Holland
NBER Computer Research Center

References

Andrews, D. F. *et al.* [1972], *Robust Estimates of Location: Survey and Advances*, Princeton, New Jersey: Princeton University Press.

Efron, B., and C. Morris [1972], "Limiting the Risk of Bayes and Empirical Bayes Estimators—Part II: The Empirical Bayes Case", *Journal of the American Statistical Association*, 67, pp. 130–139.

Fienberg, S., and P. Holland [1972], "Simultaneous Estimation of Multinomial Cell Probabilities" (unpublished manuscript).

Hoaglin, D. C. [1971], "Optimal Invariant Estimation of Location for Three Distributions and the Invariant Efficiencies of Some Other Estimators", unpublished doctoral dissertation, Department of Statistics, Princeton University.

Huber, P. J. [1972], "Robust Statistics: A Review", *Annals of Mathematical Statistics*, 43, pp. 1041–1067.

James, W., and C. Stein [1961], "Estimation with Quadratic Loss", in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1, Berkeley and Los Angeles: University of California Press, pp. 361–379.

Lindley, D. [1971], "The Estimation of Many Parameters", in Godambe, V.P., and Sprott, D. A., eds., *Foundations of Statistical Inference*, Toronto: Holt, Rinehart, and Winston, pp. 435–455.

Tukey, J. W. [1970–71], *Exploratory Data Analysis*, Limited Preliminary Edition, Volumes I–III, Reading, Massachusetts: Addison-Wesley.
Wermuth, Nanny E. [1972], "An Empirical Comparison of Regression Methods", unpublished doctoral dissertation, Department of Statistics, Harvard University.

The Mathematical Programming Project (March 1973)

The Center's mathematical programming project will continue, for the next year, the development of a software package for large-scale linear and mixed integer programming. Complementary activities in which we will also be involved include experimentation with new computational methods of integer programming and discrete optimization, problem formulation and solving in diverse applications areas, and fundamental research in discrete optimization and large-scale optimization.

The linear programming module of our mathematical programming system is progressing well, and the first version of it should be completed by the middle of 1973. This work is being performed by William Orchard-Hays, Michael Harrison and William Northup. Most of the important subroutines have been written and are being debugged. These include the general input and setup procedures, simplex algorithms, and input procedures. A serviceable inversion procedure is in the detailed planning stage. The main task of the coming months will be to encode a more efficient matrix-inversion routine for large models, to be used by the simplex algorithms, and additional service modules to facilitate computational continuity and flexibility. There are many different methods for matrix inversion, and the relation of these methods to special problem structure remains an important research question in large-scale linear programming. Virginia Klema and Tom Magnanti will be involved in studying new methods of matrix inversion. The interesting nature of this system also creates new service concepts. Thus, we envision a continuing evolution of our linear programming module beyond the first completed version, because we will be gaining greater insight into both matrix-inversion theory and a new operational environment.

The linear programming module, once completed, will serve as a building block for our mixed integer programming algorithms. One approach to mixed integer programming is to solve a series of related linear programming subproblems, derived from a given mixed integer programming problem, by fixing the integer variables. This approach has been found to work well for those problems where many or most settings of the integer variables produce feasible linear programming subproblems. For example, consider the problem of where to locate factories at minimal cost in a nationwide production and distribution system, given no capacity restrictions on new factories. (Any solution is feasible except the one in which no factory is built). In this example, the linear programming subproblems are the production and distribution subproblems that result if the selection of factory locations has already been made.

Other classes of integer programming problems cannot be solved in this way. In particular, pure integer programming problems are combinatorial or number-theoretic in nature, and it is not possible to fix some of the decision variables to produce more easily solved linear programming subproblems. Such problems

arise, for example, in capital budgeting, airline-crew scheduling, and optimal design of communication networks. We have been experimenting with new methods of pure integer programming which exploit the number-theoretic structure of these problems. These methods have proven effective for pure integer problems, such as the ones just mentioned; and we anticipate greater success in their use when they are incorporated into our production mathematical programming system.

Our experimental pure integer programming algorithm has recently been installed at the Center, and we are in the process of making it available to researchers in universities and industry. We are also concerned with developing a taxonomy for mixed integer programming and discrete optimization problems in order to integrate diverse methods and apply the most effective methods to a given problem. Marshall Fisher, visiting from the University of Chicago, is studying methods for doing this.

Finally, we have started to consider nonlinear programming algorithms and applications, the latter particularly in the areas of robust estimation and constrained nonlinear regression. We are currently installing at the Center the SUMT nonlinear programming code, developed by Fiacco and McCormick, to use for these purposes.

Jeremy F. Shapiro
Massachusetts Institute of Technology
NBER Computer Research Center

ABSTRACTS OF RESEARCH REPORTS

Working papers of researchers at the NBER Computer Research Center are published under the general title NBER-CRC Research Reports and are abstracted here. Of the five reports abstracted below, those by Belsley and Sarris are tentatively scheduled for publication in Volume 2, Number 4 (October 1973) of the *Annals*. The full text of every report is available in limited quantity, at $1.00 per copy, from the NBER Computer Research Center, 575 Technology Square, Cambridge, Massachusetts 02139 (Attention: Support Staff).

Belsley, David A. (Boston College and NBER Computer Research Center), **On the Determination of Systematic Parameter Variation in the Linear Regression Model**, NBER-CRC Research Report W0005 (January 1973), 12 pp.

This paper examines the general problem of time varying parameters in the linear regression model. Systematic, non-stochastic variation of β, linearly dependent upon "outside" variates, is highlighted. A moving-window regression technique is examined as a means of determining relevant outside variates. Computationally efficient algorithms are given for the technique. The procedure is seen to be biased, but not badly so for variates that move slowly over time.

——— **A Test for Systematic Variation in Regression Coefficients**, NBER-CRC Research Report W0006 (January 1973), 7 pp.

This paper offers a statistical test of the constancy of the parameters of a linear regression. The F test is based on transformed residuals which result from OLS applied to the given equation under the null hypothesis of constancy.

Hoaglin, David C. (Harvard University and NBER Computer Research Center), and Edwin Kuh (Massachusetts Institute of Technology and NBER Computer Research Center), **Exploration in Economic Data I: Average Annual Unemployment**, NBER-CRC Research Report W0010 (February 1973), 39 pp.

This paper reports on the first stages in applying techniques of exploratory data analysis to a facet of the structural unemployment problem in the United States. A resistant analysis of average annual unemployment rates in 150 Standard Metropolitan Statistical Areas for the years 1961 to 1971 yields estimates of structural unemployment and reveals some evidence of convergence toward equilibrium for local unemployment rates.

Sarris, Alexander H. (NBER Computer Research Center), **A Bayesian Approach to Estimation of Non-Constant Regression Parameters**, NBER-CRC Research Report W0007 (December 1972), 18 pp.

The problem of estimating non-constant regression parameters is formulated, and its statistical difficulties are examined. A structure is imposed on some of the parameters that allows for a wide class of variations. This structure fixes the number of unknown parameters that must be estimated, and renders the problem amenable to a Bayesian analysis. The solution, with some differences, looks like the Kalman type of estimator. The analysis is also valid for autoregressive processes and random coefficient models. Maximum likelihood is suggested as a way of obtaining estimates of the remaining parameters, and an iterative estimation scheme is presented without numerical tests.

Sutherland, Michael (Hampshire College), Paul W. Holland (NBER Computer Research Center), and Stephen E. Fienberg, **Combining Bayes and Frequency Approaches to Estimating a Multinomial Parameter** (University of Minnesota), NBER-CRC Research Report W0009 (January 1973), 42 pp.

We consider the problem of estimating the vector of cell probabilities for a multinomial random variable. Using a combination of Bayesian models and frequency calculations (i.e., risk functions), we develop a class of estimators that resembles the James–Stein estimator for the multivariate Normal mean. We approximate the risk functions of these estimators to second order, when the number of observations per parameter is moderate and the number of parameters (cells) is large. Then we use these approximations to prove a first-order optimality result for one of the estimators in the proposed class. We briefly consider extensions to the case of cross-classified multinomial data and we apply 10 estimators discussed in the paper to a set of occupational mobility data.