

NBER COMPUTER RESEARCH CENTER NOTES

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. Notes on research in progress, as well as abstracts of working papers, are a regular feature in the Annals. Following are progress reports on the Mathematical Programming and Data-Analysis Projects, and the abstract of a working paper produced by the Econometrics Project. Documentation of computer programs mentioned in the progress reports, and the full text of the working paper, are available at cost from the NBER Computer Research Center, 575 Technology Square, Cambridge, Massachusetts 02139 (Attention: Support Staff).

THE MATHEMATICAL PROGRAMMING PROJECT, APRIL 1974

The Center's most important accomplishment of the past year in the area of mathematical programming was the completion of SESAME, a system for linear programming and its extensions. SESAME was built by William Orchard-Hays, Michael Harrison, and William Northup. It is capable of solving general linear programming problems with 2,000 rows or more and over 10,000 variables. For linear programming models with structures that occur in transportation and scheduling optimization, SESAME has the capability to solve problems with up to 50,000 constraints and an almost unlimited number of variables. SESAME includes interactive facilities for linear programming sensitivity analysis. DATA-MAT, a language for generating linear programming models from basic data and logical relations is under development and should be completed this year. Preliminary documentation of SESAME—including an overview, primer, and reference manual—have been published. A workshop in mathematical programming centered around the SESAME system was held at the Center on March 28–29; it was attended by about thirty mathematical programming researchers and practitioners from universities, government agencies, and industry throughout the U.S.

Experiments with dual methods of integer programming and other combinatorial optimization models were recently completed by Marshall Fisher, William Northup, and Jeremy Shapiro. They presented a paper on their work at the Mathematical Programming Symposium at Stanford University in August 1973. The dual integer programming methods will be one of the innovative features of the large-scale mixed integer programming system now under development at the Center. This system will also make heavy use of SESAME to solve linear programming approximations of integer programming problems.

Work has also proceeded on the development of programs to compute economic equilibria by fixed-point approximation. The programs are being constructed by Odunayo Olagundoye; they are being applied to equilibrium analyses of urban housing markets by Jeremy Shapiro and Joseph Ferreira of the Urban Studies and Planning Department at M.I.T. Marshall Fisher and F. J. Gould of the University of Chicago are collaborating with the Center in research on the use of these methods to solve nonlinear programming problems. (Herbert Scarf of Yale University, the originator of the fixed-point approximation methods, is a consultant to the Center.)

The Center's staff is involved in a number of applications of mathematical programming using SESAME and other programs. Included are supply, demand, and distribution studies of the U.S. natural gas pipeline system; stochastic programming models for water-resource planning; and refinery-location models for a long-range world petroleum model.

Jeremy Shapiro

THE DATA-ANALYSIS PROJECT, APRIL 1974

Over the past year the Data-Analysis Project at the Center has progressed toward the goal of making serious interactive data analysis possible through the Center's computer facilities—especially the TROLL system and NBERNET, the NBER's digital telecommunications network. The progress includes both research on, and implementation of, several new data analytic methods. (Most of the resulting new programs are now documented in recent installments of the serial publication entitled *TROLL Experimental Programs*.)

One of our major interests is robust regression—i.e., regression methods that are not sensitive to a few wild values or to small changes in the model or the data. In June 1973, David Hoaglin organized a two-day working conference on robust regression, which was attended by about twenty statisticians and economists who have contributed to this growing field. The conference, which was held at the Center, provided us with substantial feedback on our own plans for a Monte Carlo study of robust regression methods.

We have carried out the first stage of the Monte Carlo study. The well known method of minimizing the sum of absolute residuals—"least absolute residuals" (LAR)—has served as a starting point, and methods for improving upon LAR have been examined. We have found that a single weighted least-squares iteration with weights based on the LAR residuals provides a significant improvement on LAR in both Gaussian and non-Gaussian cases. We are currently studying the role of the data matrix in these methods, as well as methods for setting confidence intervals for the regression parameters. Principal researchers in this part of the project are Richard Hill, David Hoaglin, Roy Welsch, and myself.

David Hoaglin has continued his research on asymptotic variances of location-parameter estimates. This work is fundamental to the study of regression problems in that it provides a base line for multivariate problems and suggests new methods in a simple context where it is easier to see what is going on. The Monte Carlo study has also led us to examine random number generators and their properties.

Using tools developed for the Monte Carlo study, Richard Becker of the Center's programming staff has built a program called ROBUST (for use within the TROLL system) which allows the user to apply various robust regression methods to his data. The user can do LAR regression and various iteratively reweighted least-squares improvements on LAR that have been suggested by Huber, Andrews, and Tukey. In addition, he can obtain a plot, called a "Huber trace", that graphs the values of the regression coefficients as a function of a robustness parameter.

Ridge regression has attracted increasing interest at the Center over the past year. This Bayesian method offers a significant improvement in the estimation of regression coefficients when the data-matrix suffers from near collinearity and when the number of parameters is large. Using sophisticated numerical methods suggested by the Center's numerical analyst, Virginia Klema, Richard Becker wrote a program called RIDGE (also used in TROLL) that gives the user a flexible system for doing ridge regression. The user of RIDGE can obtain various empirical Bayes choices of the ridge parameter, can specify a nonzero prior mean for the regression coefficients, and can obtain a graph called a "ridge trace", in which the estimated regression coefficients are plotted as a function of the ridge parameter. The user may even supply weights computed from robust regression and thus combine both robust and ridge regression methods.

The analysis of discrete data has received attention by Center researchers. By using the equation-solving capability in TROLL, I have found it easy to fit Poisson regression models to data by maximum likelihood. I did this experimentally for some economic data on the generation of new products, and more seriously in collaboration with a sociologist, Samuel Leinhardt, for a model of the distribution of doctors in Pittsburgh. In addition to this, we have implemented three programs in TROLL which will allow the user to fit log-linear models to multidimensional contingency tables, and to solve the inverse problem of adjusting a given data table to prescribed margins. These programs may enable users of cross-sectional data to gain new insights into their problems. Roy Welsch and I will be using these tools in a reanalysis of heart-disease data in association with the Harvard Faculty Seminar on Human Experimentation in Health and Medicine.

One of the most exciting developments for data analysis is computer graphics using inexpensive terminals like the Tektronix 4010. Under the guidance of Roy Welsch, several new graphics subsystems have been implemented by Helge Bjaaland and put into the TROLL system. The basic subsystem, CLOUDS, allows the user to rotate a multidimensional point cloud in any direction and then project it back on any two-dimensional set of axes. Besides making fast scatter plots easy, this technology allows the user to look for multidimensional outliers and important structural features in his data. This program has provided the basis for a number of other plotting facilities including Tukey's "schematic plots" and plots of two-way fits for exploratory data analysis: the ridge-trace plot in the ridge regression program; and the Huber trace in the robust regression program. All these graphical tools can provide the user with extra insight into his data. Residual analysis is made easy using CLOUDS, and a separate program called NORMPLOT can be used to obtain normal probability plots of residuals.

In response to a variety of inputs—robust regression, data analysis, graphics—the Center has begun to develop cluster-analysis methods and to make them available to users. A new member of the data-analysis research group, Donald Olivier, is putting a hierarchical clustering program into TROLL and is active in the development of new clustering methods. Two graphical methods for clustering multivariate data have been implemented. Chernoff's FACES, which represents each multivariate data point as a schematic human face, provides a novel and useful way to cluster data. The STARS plot, which represents each data point as a polygon, is also available and is another potentially useful graphical clustering procedure.

We continue putting useful exploratory data analytic tools into TROLL. Recent additions include LINE, a good version of Tukey's resistant line; and DIPLOT, which gives a diagnostic plot for a two-way table that reveals whether some of the interaction is removable by a power transformation.

I have established a collaborative research project with two sociologists, Samuel Leinhardt at Carnegie-Mellon and James Davis at the University of Chicago and National Opinion Research Center. We are developing new methods for the analysis of small-group social structure. Our work will hopefully provide new methods for analyzing sociometric data and may eventually allow the social scientist to specify and test structural hypotheses and models in a simple way.

Paul W. Holland

WORKING PAPER ABSTRACT

Sarris, Alexander H., and Michael Athans, "**Optimal Adaptive Control Methods for Structurally Varying Systems**", NBER Working Paper No. 24 (December 1973), 77 pp.

The problem of simultaneously identifying and controlling a time-varying, perfectly-observed linear system is posed. The parameters are assumed to obey a Markov structure and are estimated with a Kalman filter. The problem can be solved conceptually by dynamic programming, but even with a quadratic loss function the analytical computations cannot be carried out for more than one step because of the dual nature of the optimal control law. All approximations to the solution that have been proposed in the literature, and two approximations that are presented here for the first time, are analyzed. They are classified into dual and non-dual methods. Analytical comparison is untractable; hence Monte Carlo simulations are used. A set of experiments is presented in which five non-dual methods are compared. The numerical results indicate a possible ordering among these approximations.