Getting the Computational Job Done Electronically

By Hyman Weingarten

This account of computational procedures followed by the Agricultural Economics Division, Agricultural Marketing Service, is presented for readers interested in applications of an electronic computer to statistical and economic problems. A brief description of the machine is given to acquaint the reader with a few of the basic characteristics of this mechanical assistant. AMS machine facilities are under the direction of V. H. Nicholson, Chief of Data Processing in the Statistical Standards Division. His experienced staff of programmers and machine operators readily provide valuable assistance to those seeking machine aid. The author is one of several persons from other divisions who have learned programming techniques under Mr. Nicholson's direction.

THE CENTRAL computing unit of the Agricultural Economics Division has been plagued for years by a twofold problem—first, a lack of trained clerical personnel and, second, an ever-increasing workload. These problems are not unique. They are characteristic of many computing sections and hamper normal research functions everywhere.

Much of the statistical methodology utilized in the research activities of this division is in the area of least-squares regression analysis. Those familiar with its methodology know that a least-squares multiple regression problem is time-consuming and can become an unwelcome chore when a number of problems are presented at once. Because the research analyst is keenly aware of this situation, he often spends considerable time in an attempt to determine a priori the best statistical relationship to fit. He is hesistant to experiment—to test alternatives. In addition, the relatively recent development of computational techniques associated with the limited information method for handling simultaneous equations has added immeasurably to the workload. Problems of the latter type are generally larger and more complex.

The advent of the electronic computer has been a great factor in helping to overcome these handicaps. It has aided research personnel virtually to jump from a comparatively "primitive" era of statistical computations into a modern one of greater efficiency and practically unlimited bounds of application. The Agricultural Economics Division has been utilizing electronic equipment regularly in the recent past. First, under contract with the National Bureau of Standards, statistical computations were performed on its electronic computer, the SEAC. When Standards decided to eliminate SEAC operations, contract work was transferred to the Service Bureau Corporation. During this latter period, our personnel became familiar with computer techniques. As a result, a relatively easy transition was made when in May 1958, the Agricultural Marketing Service set up its own facilities and acquired a magnetic drum data processing type IBM 650. Since that time, the machine has provided results to economic research problems both speedily and efficiently.

The Computer and Its Language

The same battery of questions come from those who are unfamiliar with electronic computers. What is it? What does it do? How does it work? Is it actually a "brain"?

Answering the last question first, "No, it is not a brain"! A computer cannot think for itself. The machine must be told what to do every step of the way. There is nothing mysterious or superhuman about it. If a programmer fails to instruct the machine, it can go no further. It is merely a machine constructed to relieve man of burdensome computations, and this it does in a most spectacular way.

The AMS installation consists basically of three components: (*a*) The read-punch unit, (*b*) the console unit and (*c*) the power unit.

Data, previously punched on cards, are fed into the computer through the read-punch unit. Cards are read, one at a time, at a maximum rate of 200 a minute. After specified computations are performed by the computer, answers are punched by this same unit at a maximum rate of 100 cards a minute. Proper wiring of the unit's control panel permits the variation of card formats among problems, thus allowing for flexibility of input and output.

The second component of the installation, the console, contains the magnetic drum, the control and calculating units. The magnetic drum—the heart of the computer—is a cobalt-nickel-plated cylinder about 4 inches in diameter and 16 inches in length. It provides general storage, or "memory." When the unit is in operation, the drum revolves at a speed of 12,500 revolutions a minute. The control section permits access to machine storage at any time. If necessary, information can be read into general storage or memory manually through certain storage-entry switches. Stored information can be read out and ascertained from the display lights on the control unit. Lastly, the calculating units, as the name implies, perform all the mathematical operations. These occur in conjunction with registers identified as the distributor, the upper accumulator, and the lower accumulator. The accumulators are somewhat similar to those on a desk calculator. Algebraic signs are automatically controlled.

The power unit contains the power supply for all machine units. It contains also the circuitry needed to translate input from decimal form into machine code and machine code back into decimal form for output.

All processing operations are controlled by instructions consisting of numerical words of 10 digits and a plus or minus sign. A meaningful sequence of instructions leading to the solution of a problem is labeled a "program."

The 10 digits of an instruction are divided into three basic groups: The operation code, the data address, and the instruction address. As the numerical words representing the original data and instructions are read into the computer by the read-punch unit, they are stored in the console unit in locations on the surface of the drum in the form of magnetic spots. Each word is identified by a particular pattern of spots. The AMS computer can store 2,000 numerical words in storage locations, numbered 0000–1999. Each word is assigned a 4-digit address when it is first stored on the magnetic drum in order to hasten its location at a later time. Thus, in performing an operation, the machine refers to a storage location on the magnetic drum to obtain an instruction. After performing the specified operation, the machine goes to another storage address. There it obtains another instruction and acts upon it. These steps are repeated over and over.

As an illustration, suppose storage location 0101 contains the following instruction: 15 1849 0102 +, where the operation code designated is 15, the data address is 1849, and the instruction address is 0102. This means that the word in location 0101 instructs the machine through operation code 15 to add the contents of location 1849 to the lower accumulator and go to location 0102 to pick up the next instruction. Location 0102, in turn, may have the following word: 20 0147 0103 +. Now the machine is told to store the contents of the lower accumulator in location 0147 and go to location 0103 for its next instruction, and so on.

Using the Electronic Computer in a Multiple Regression Analysis

Because regression analysis is a major concern, our first objective was to secure a complete least-squares regression procedure of operation. A study of many different programs compiled by the manufacturer in his program library was made prior to the delivery of the AMS computer.[1] Five programs were selected. For any least-squares analysis, their operation provides all the data necessary with a minimum of manual computation.

In presenting a regression problem for processing, the originator submits only the observations for each variable in the problem, designating which variable is dependent. When computation is complete, the series of programs calculates

(1) Means;
(2) Standard deviations;
(3) Simple correlations coefficients;
(4) Inverse of the correlation matrix;
(5) Partial correlation coefficients;
(6) Partial regression coefficients;
(7) Standard errors of these regression coefficients;
(8) Residuals;
(9) Predicted values.

The programs do not provide for the computation of the constant term of the regression equation

[1] A "program library" is maintained as a service to those using this equipment. It is an extremely important service, as users benefit from the experiences of many programmers throughout the country. Other manufacturers have similar services.

and the multiple correlation coefficient. However, such values can be obtained easily by a short desk-calculator operation.

The five IBM library programs that make up the regression analysis series are:

1. *Correlation Analysis with Annotated Output, No. 290–1, File No. 6.0.014.*

This program computes the means, standard deviations, and simple correlations for problems with as many as 25 variables and 9,999 observations. Results from this program are obtained both in fixed and in floating decimal point form. Floating decimals are based on a code or characteristic of 50, representing an integer followed by its decimals; for example, 501478200 is the number 1.478200. Cross product terms in fixed and floating decimal form are other options that may be punched out.

2. *Double Precision Matrix Inversion, No. 283, File No. 5.2.009.*

This program can invert a matrix of any order up to a maximum of 25. Input data are the simple correlation coefficients derived from the preceding program. Elements of the inverse matrix have eight-place accuracy, that is, the symmetric elements of the matrix are identical to eight places. Input and output data are in floating decimal point form.

3. *Multiple Regression Analysis, Phase III–2, No. 299, File No. 6.0.001.*

Using the inverse elements and standard deviations computed by the two preceding programs, partial correlation and regression coefficients involving the dependent variable are derived. Both input and output data are in floating decimal point form.

4. *Unbiased Standard Errors of the Regression Coefficients, No. 297, File No. 6.0.008.*

This computation is based on the partial regression and partial correlation coefficients from the preceding program. Input and output data are in floating decimal point form.

5. *Multiple Regression Analysis, Phase IV, No. 299, File No. 6.0.001.*

This program computes either the estimated values of the dependent variable or the residual values between the computed and observed quantities based on the regression coefficients, means and the original observations of preceding programs.

Initally, this five-program series proved to be as efficient a procedure as we could get. Needless to say, efficiency could have been improved with a greater reduction in card handling. It is better, for example, to process a number of problems at one time rather than individually. Programs listed above as items 1, 2, and 4 permitted computation of any number of problems in succession without the necessity of reloading the program deck for each problem. However, this was not true in the cases of Phases III–2 and IV of Program No. 299, items 3 and 5 above. The latter require that the program deck be reloaded along with data for each new problem.

IBM programmers have been working toward the objective of developing shortcuts that will further reduce card handling. One modification recently completed, Program No. 290–3, consolidates steps 3 and 4 above. Thus partial correlation coefficients, regression coefficients, and standard errors of the regression coefficients can be obtained now by the same program. The program deck is read into memory once and then solutions for any number of problems can be derived in succession. In addition to coefficients obtained formerly, the new program also provides for the computation of the constant term of the regression equation, the multiple correlation coefficient, the standard error of estimate and computed "t" values.

To give readers some idea of the amount of time needed to obtain the coefficients for a least-squares regression analysis, estimates have been made from timing formulas for the five steps of our original program series. The economic research problems handled by the Division contain an average of approximately 20 observations. The computational time required for deriving coefficients for a typical 20-observation problem having different numbers of variables is summarized below.

Number of variables	Minutes of computation
5	11
6	12
7	13
8	14
9	15
10	17

No allowance is made for card punching, verifying, listing, or handling between programs. If such operations are considered, a rough estimate of total time required for the processing of one problem—punching of data on cards through the listing of results—would be between $1\frac{1}{2}$ and 2 hours. But, as indicated before, any given time estimate will be reduced substantially if a number of problems are processed together. With the substitution of Program No. 290–3 into the AMS series, computation time is reduced even more. Although no formula is yet available, the maximum computation time for the largest problem it can accept, 25 variables, is approximately 5 minutes. Therefore, the computation times given above would be reduced substantially. One criterion developed as a result of these time calculations is that coefficients for any two, three or four-variable problem are derived by desk calculator; problems having five or more variables are processed on the computer.

Another recent addition to the AMS program series is a logarithmic subroutine, File No. 3.0.013. It has been adapted and make an integral part of the series so that regression problems may now be computed based on actual and/or logarithmic data. The desirable feature of this routine is that its output is directly acceptable as input to Program No. 290–1, the first step in the regression analysis program series.

Other Statistical Adaptations

The procedure outlined above represents only one approach used in getting solutions to some of our research problems. Many more facets of our research activity can be adapted to the machine.

For example, in some of our research, we obtain results from both the least-squares and the limited-information methods for comparison purposes. The author is currently working on a program series that will result in an eventual solution by either method and, at the same time, take advantage of computations common to both.[2] So far, a program has been designed to secure the adjusted augmented moments for as many as 28 variables. Once obtained, such factors can then be separated according to specified relationships and used either in a least-squares or limited-information solution of an equation. The program is patterned after methodology described in "Computational Methods for Handling Systems of Simultaneous Equations."[3] The method described in the publication derives coefficients from an inverse of a variance-covariance matrix. On the other hand, the methodology used in the library routines outlined here computes coefficients from an inverse of a correlation matrix. This is the essential difference between the two systems of computation.

The operation of this program produces the following data: Unadjusted augmented moments, means, scale factors, deadjustment factors, and the adjusted augmented moments in fixed and in floating decimal point form. As the program operates in fixed point form exclusively, input data have one limiting aspect. For minimum loss of information owing to rounding error, observations should not exceed four digits—three integers and a decimal. However, additional work is continuing in order to avoid such restrictions and to increase the overall flexibility of the program.

Other library programs have proved useful. For example, many of the computational steps required by the limited information method involve matrix operations. One development in particular requires a triple-product matrix of the form, $C=A\ B^{-1}\ A'$. Its solution is secured by using the inversion program outlined above, a matrix transfer program, and a program for matrix multiplication. Another area of computation deals with the determination of the roots of a characteristic equation. Heretofore, such roots were derived through an iterative method designed primarily for desk calculators. Currently, experimentation is underway to determine whether certain library programs can be adapted to our computational processes.

Because of the computer's limited drum surface and its associated storage-capacity restrictions, a complete program for the solution of simultaneous equations using the limited-information method is not feasible. Certain areas of computation have been adapted to the machine thus far. At present, the best that can be hoped for is to continue to section off various phases of computation and link them into a program series similar to the one established for least squares.

[2] A program for the solutions of equations by the methods of limited information and least squares has been developed by IBM for a larger capacity electronic computer.

[3] Friedman, Joan and Foote, Richard J. *Computational Methods for Handling Systems of Simultaneous Equations.* U.S. Dept. Agr. Agr. Handb. 94, 109 pp., illus. Nov. 1955.