A Multivariate Analysis Computer Program

A. J. B. ANDERSON and BRIDGET I. LOWE

Rothamsted Experimental Station

SUMMARY

This paper describes an electronic computer program for the analysis of multivariate data. The program is interactive in the sense that its control structure is designed to facilitate a step-by-step process of analysis extending over several computer runs. Especially comprehensive input routines allow the user flexibility in data presentation. A group structure can be imposed and within- and between-group analyses are possible. The program includes all the usual multivariate techniques, but special attention has been paid to providing a wide repertoire of operations useful in multiple-regression analysis.

1. INTRODUCTION

SINCE the advent of the electronic computer, numerous programs have been written for the examination of multivariate data. Most of these have lacked the flexibility required in a branch of statistical analysis that is essentially interactive. On-line console control languages provide means for the rapid inspection of data with a very simple structure. However, for more complex analyses, the statistician requires a longer time lapse between appeals to the computing system. Even where the originator of the data can be readily consulted, several computer runs may be required to extract the greatest amount of information and to elucidate the underlying structure. This is particularly true where multiple linear regression is concerned, and Healy (1963) has indicated some of the requirements of a program system for such analysis.

The multivariate analysis program (MAP) in use on the Orion computer at Rothamsted has been designed to deal with three main categories of problem:

(a) the routine analysis of small quantities of data, usually involving multiple-regression calculations;

(b) the extensive examination of large quantities of data with many variates and frequently having a group structure imposed on the units—canonical variates, canonical correlations or principal component analysis may be required;

(c) the estimation of parameters in certain types of non-linear model by iterative methods (Nelder, 1968).

For all types of analysis, the data can be presented to the MAP program in various forms and on various input media. The commonly imposed restriction to fixed format unit-by-unit input from cards would be intolerable in a program that, again because of its interactive nature, must be able to communicate with other programs such as the General Survey Program (Anderson, 1966) and the General Factorial Program (Yates and Anderson, 1966).

As with most statistical programs, the output section of MAP is the largest. The items printed by default at any stage of an analysis are those suitable for problems of category (a), which includes most work at Rothamsted. However, these default settings can be readily altered by the user, and the list of quantities that can be output by the *PRINT* instruction (Section 6) should be adequate to meet all needs. So that the program can be used for research into the methodology of statistical analysis, particularly in category (c), a means is provided for switching into the autocode language in which MAP was written (Extended Mercury Autocode). This enables special *ad hoc* segments to be added to the main program.

2. General Form of the Program

MAP can handle observations of up to 48 variates on up to 2,000 units in up to 48 groups. The control language consists of statements that always begin with a directive in capital letters. A directive can be abbreviated provided it is uniquely identifiable, and the first three letters always suffice. Each directive must start on a new line, and some directives are followed by a variable number of arguments; these arguments are separated by spaces or commas, a comma being obligatory as a continuation symbol at the end of a line. Statements are obeyed sequentially except where looping is introduced (Section 10.1) and may appear in any order subject to the restriction that their operands must have been previously defined. These statements can be punched on five- or seven-track paper tape or on cards or any combination of these.

Normally, variates or groups are referenced by number, but the user may also identify such entities by name, so that no hand annotation of output is required. A name consists of up to eight characters, of which the first must be a capital letter, from the set

A to *Z* 0 to 9 + . / * % & ′ ( ) -

The character π may be used in a name to indicate a space and will appear as such in any output.

It is frequently useful to define a set of variates or groups by means of a list. A list is a series of expressions separated by commas or spaces, and must be enclosed in brackets if it extends over more than one line. Each of the expressions in a list may take one of the following forms:

(i) the number of a single variate or group;
(ii) $p(q)r$, meaning numbers $p, p+q, p+2q, \ldots, r$;
(iii) $p-r$, meaning $p, p+1, \ldots, r$, that is, $p(1)r$;
(iv) the name of a previous list. (A list may be named by terminating it with "= name", where "name" is any suitable name as defined above.)

For example,

2 5 8–10 20 31(2)41 = *LISTA*

Variate or group names can be used instead of numbers and the scan counters $\pi 0$ to $\pi 9$ (Section 10.1) may also appear. In this description a list will be denoted by (L).

3. Input and Transformation of Data

The data input section of MAP is identical with that used in the factorial experiment analysis program. Data can be read variate-by-variate from five- or seven-track paper tape, from cards in free or fixed format, from magnetic tape, or can be read unit-by-unit from paper tape or cards. Any mixture of these forms of input is permitted and input directives may appear as often as required. Missing values are indicated by * and stored as -10^{20} for easy recognition at later stages of analysis. When each variate is read, its maximum, minimum and mean are printed to help detect punching errors and the user is warned if any variate has a particularly skew distribution. A check is made that the correct number of values has been presented. If any fault is detected during input, the remainder of the data is checked but analysis is inhibited.

Associated with data input is a section providing comprehensive variate-transformation facilities. All the usual arithmetic operations are available and also a wide repertoire of functional transformations including those described by Box and Cox (1964). Automatic conversion of scales of measurement is made possible by instructions such as

$$3 \quad CWT/AC = 2 \quad LB/20{\cdot}5FT*13FT$$

This converts variate number 2 which is measured in pounds per 20·5 ft × 13 ft plot to variate number 3 measured in hundredweights per acre.

Original and derived data are always stored variate-by-variate on the magnetic drum; for the Rothamsted Orion the maximum amount of drum store the user can request is 25,000 words. Many MAP directives create new pseudo-variates (e.g. fitted values and their standard errors) and these are similarly stored on the drum and can be operated on as standard variates.

An externally computed matrix of sums of squares and products, mean squares and products, or correlations may be input directly, punched in lower triangular form. The ith row is assumed to relate to variate number i unless it is preceded by a specific variate number. The relevant degrees of freedom must be given and (when correlations are presented) the associated sums of squares.

4. Structure Directives

Directives are provided that permit division of the set of units into groups. Such division may be imposed for the following reasons:

(a) to restrict analysis to specified subsets of units;
(b) to facilitate the construction of certain types of dummy variate;
(c) because a natural stratification is present in the data and a between- and within-groups analysis is required.

The structure information can be supplied to the computer in three ways:

(i) Groups may be defined by the unit values for any specified variate. For example the statement

*GROUPS V*16

sets up a group structure based on variate number 16. If the integer part of the value of unit i is g (where $1 \leqslant g \leqslant 48$), then unit i is placed in group number g. Units for which variate 16 is unknown do not belong to any group. Alternatively the form

*GROUPS C*16

may be used, in which case all units having the same value for variate 16 are placed in the same group, but these values need not bear any relation to the group number.

(ii) The numbers of units in successive groups can be given. For example the instruction
GROUPS 6, 8, 3 * 9
defines five groups consisting of 6, 8, 9, 9 and 9 units respectively.

(iii) The directive *GROUPS* can be followed by lines of the form
group number group name units list

For example, to split 100 units into two groups of odd- and even-numbered units, one might write

GROUPS
1 *ODDS* 1(2)99
2 *EVENS* 2(2)100

The directives

ANALYSE GROUPS (*L*)
and
EXCLUDE GROUPS (*L*)

permit further analysis to be restricted to certain groups. (There is a parallel form for variates.) The restriction is temporary and any succeeding restricting directive cancels the previous one.

5. The Directive Matrix

The directive *MATRIX* causes calculation of

(i) means of variates, and group means if a grouping has been defined;
(ii) a matrix of sums of squares and products.

Only currently included variates and those units in currently included groups are used. If there is no grouping defined, the sums of squares and products are corrected for the mean; if there is grouping, the matrix is "within-groups". (All means are first calculated so that deviations may be used in the formation of the sums of squares and products, thus minimizing rounding errors.) The directive *MATRIX* may be followed by *B*, *U* or *W*.

B gives a between-groups matrix.
U gives an uncorrected matrix.
Wv indicates that the sums of squares and products are to be weighted by variate number *v*. This variate is automatically excluded from the matrix.

When there are missing values, means and group means are based on all available data. In the formation of the elements, s_{ij}, of the matrix of sums of squares and products, only those units known for both the variates being considered are used. If q units are thereby excluded, and there should be d degrees of freedom in the absence of missing values, then s_{ij} is multiplied by $d/(d-q)$. This procedure is unsophisticated and the resulting matrix may not be positive-definite if there are many missing values, but there is not yet a general solution to the problem (Haitovsky, 1968).

6. Output Directives

The directive *PRINT* is followed by a list of items to be output. Variates can be printed unit-by-unit or variate-by-variate and in either case the user can control the number of columns to appear across the paper. General means have their standard errors automatically appended, but standard errors for a table of group means must be specifically requested. Sums of squares, mean squares, standard deviations, sums of squares and products, mean squares and products and correlations may be printed. Matrices are normally printed in lower triangular form, but sub-matrices involving possibly disjoint sets of variates may also be printed and these appear in rectangular form. Any matrix may be printed with the effect of a specified list of variates "partialled out".

Each item in a *PRINT* statement may optionally have a decimal place specification. The possible forms (where p and q denote integers) are:

(i) $p.q$, indicating p places before the decimal point and q after; $p = 0$ gives floating point printing,

(ii) pS, indicating that each element is to be printed with p significant figures,

(iii) pM, indicating that the element of maximum modulus is to be printed with p significant figures, and the decimal points of the remaining elements are to be aligned,

(iv) pA, indicating that each element is to be printed with p figures after the point, aligned as in (iii).

For example, to print means, sums of squares and products to a maximum of six significant figures and correlations with three figures after the point, the user writes

PRINT MEANS SSP/6*M COR*/3*A*

Standard errors are printed to one more significant figure than the corresponding means.

All segmenting of tables and triangular matrices is automatic and unknown entities (such as correlations involving a constant variate) are printed as *. Output is annotated by variate and group numbers or names when these have been given. Textual headings can be printed at any point and, in addition, short headings from a previously stored list of captions can be used to clarify output.

7. Graphical Display

It is often found useful in a preliminary inspection of data to examine univariate and bivariate distributions by the display of histograms and two-way scatter diagrams. These are produced by the directives *HISTOGRAM* and *PLOT* respectively. These directives can also be applied to derived items (such as regression residuals and variates referred to principal axes) which have been stored on the drum by previous directives. For histograms, either the number of classes or the class width must be specified. For scatter diagrams, two variate lists named "*X*" and "*Y*" must previously have been defined; every variate in "*Y*" is plotted against each variate in "*X*", the "*Y*" variates on the vertical axis and the "*X*" variates on the horizontal axis. The range of the abscissa variate is divided into one hundred steps; the same increment is used for the ordinate axis, but the scale is multiplied or divided by a suitable power of 10 to make the vertical dimension of the diagram reasonable.

8. Regression Directives

The directives described in this section relate to the formation of multiple-regression equations and specify the dependent variates (*Y*-variates) for which regressions are to be produced, the independent variates (*X*-variates) to be used, the items to be printed for each regression equation and the quantities to be stored as new variates.

8.1. *Dependent Variates*

The list of dependent variates for which regressions are to be produced at each specification of an X-set is given by a statement of the form

Y (*L*)

Such a Y directive holds until cancelled by another Y directive. However, the form

*Y** (*L*)

may be used to produce output for the listed Y-variates regressed on the current X-set; the current Y-set is unaltered.

8.2. *Independent Variates*

The inadequacies of automatic stepwise regression are well known (Yates, 1968) and MAP does not provide this form of analysis. Instead, it has a wide repertoire of directives for controlling the entry and exit of independent variables from the regression equation. At each alteration of the X-set, appropriate output is produced and derived quantities are stored (Section 8.3). The directives are given in Table 1.

TABLE 1

Directives specifying independent variates

Directive	*Effect*
$X\ (L)$	Specifies a set of independent variables.
$\left.\begin{array}{l}\mathit{ADD}\\ \mathit{DELETE}\end{array}\right\}\ (L)$	The listed variates are added to/deleted from the current X-set. Redundant alterations are ignored.
$\mathit{SWITCH}\ (L)$	Those variates which are already in the X-set are deleted, and those which are not are added.
$\mathit{TRY}\ (L)$	Each of the listed variates is singly added to or deleted from the X-set. The associated printing is produced but the X-set remains unaltered by the operation.
$\left.\begin{array}{l}\mathit{COMB}2\\ \mathit{COMB}3\\ \mathit{COMB}4\end{array}\right\}\mathit{TRY}\ (L)$	These are similar to the previous directive, but all combinations of up to 2, 3 or 4 variates from the list are examined. To discourage the user from producing an excessive amount of output, the number of possible combinations is limited to 15.
$\left.\begin{array}{l}\mathit{BEST}\\ \mathit{WORST}\\ \mathit{MIN}\end{array}\right\}\ (L)$	The result of these directives depends on an examination of the residual mean squares of the lowest-numbered Y-variate. Each of the listed variates is singly added to (*BEST*), deleted from (*WORST*), added to/deleted from as appropriate (*MIN*) the current X-set. The alteration which gives rise to the smallest mean square is then made or if no improvement is produced, the X-set is left unchanged. Any of the three directives may be preceded by *COMB*2, *COMB*3, ..., *COMB*6 in which case all combinations of up to 2, 3, ..., 6 variates from the list are examined. The mean squares computed by the searching procedure are printed in ascending order.

If any directive is terminated by *, the appropriate output is produced, but the X-set is unchanged after the operation.

When the set of independent variates is changed, adjustments to the regression coefficients are made by the method of Woolf (1951) and no explicit matrix inversion routine is required. Should a linear dependence be detected among the X-variates the offending variate is ignored and a comment printed.

8.3. *Output and Storage*

Various items can be automatically printed at each alteration to the regression equation. By default, only regression coefficients and the analysis of variance appear. Additional output can include the inverse of the X-variates matrix, the variance–covariance matrix of the regression coefficient estimates, the fitted values and residuals with their standard errors, and, where a second-order surface is being fitted, the optimum response and corresponding factor levels. For each regression coefficient its standard error is appended and the probability of the truth of the null hypothesis that the regression coefficient is zero. The multiple correlation of each of the X-variates with the other X-variates is also given as an indication of the internal structure of the X-set. (This quantity is calculated as

$$\sqrt{\{1-(v_{ii}\,c_{ii})^{-1}\}},$$

where v_{ii} is the sum of squares for variate i and c_{ii} is the corresponding element of the inverse matrix.) When the sums of squares and products are uncorrected for the mean, the regression line passes through the origin; otherwise the y-intercept and its standard error are given. The commonly used multiple correlation of the dependent variate with the independent variates is unsatisfactory as an indicator of goodness of fit because no account is taken of the degrees of freedom involved. In place of this, MAP prints the quantity 100 {1 – (residual mean square/original mean square)} as a measure of the percentage variance accounted for.

The fitted values, residuals and standard errors or variances of fitted values may be transferred to the variate store and are then available for use as standard variates. This feature makes possible iterative regression calculations using any of the above sets of values as a weighting variate.

9. Other Multivariate Computations

Three directives are available for initiating Principal Components, Canonical Variates and Canonical Correlation analyses:

(i) *PCOMP* The latent roots and vectors ("rotation loadings") of the sums of squares and products matrix are printed. Any constant variates are automatically excluded. The directive may optionally be followed by

(a) a list of variates, in which case the rotation will be based only on these listed variates,

(b) a positive integer, r, in which case only the r largest latent roots and corresponding vectors will be found,

(c) the letter C, in which case the computation will be based on the correlation matrix.

The rotation loadings may be applied to the variates used in the calculation and the transformed values stored on the drum as new variates.

(ii) *CVAR* Canonical variates for two or more groups are derived using an algorithm described by Gower (1966). The latent roots of the sums of squares and products matrix (each root being the sum of squares accounted for by a canonical variate) are printed in descending order of magnitude together with the χ^2 values for Bartlett's test of significance. A number of other items may be printed, including Mahalanobis's D^2 matrix, the rotation loadings and the rotated group centroids. Transformed data values may be stored as new variates.

(iii) *CCOR* Canonical correlations relative to two variate lists called "*X*" and "*Y*" are printed together with the associated rotation loading matrices. These transformations may be applied to the two sets of variates and the resulting quantities stored.

10. Other Features

10.1. *Looping Facilities*

Ten looping counters, indicated by $\pi 0$ to $\pi 9$, are available. These take values in the range 1–48, and may be used to index variates or groups in lists and elsewhere. The form of a loop is

```
SCAN πi = (L)
Body of loop
REPEAT
```

The body of the loop will be repeatedly obeyed with πi taking the values given by the scan range list, (*L*). Loops may be nested to depth 10. More than one counter may be attached to a loop, e.g.

```
SCAN π0 = (5(2)19)   π1 = 21–28
ANALYSE VARIATES 1–4, π0
MATRIX Wπ1
Y  π0
X  1–4
REPEAT
```

10.2. *Routines*

The directive *ROUTINE* followed by a name introduces a series of statements to be regarded as a subroutine, the series being terminated by *RETURN*. The routine is entered by giving its name as a directive. For example, if a routine called "*XYZ*" has been previously defined, the directive *XYZ* causes entry to that routine. Up to eight routines can be defined at any time. A routine cannot call itself recursively.

10.3. *Storage of Analyses on Magnetic Tape*

A feature of MAP which has proved most useful is the ability to store an analysis on magnetic tape for resumption at a later date. The state of the analysis is exactly restored, e.g. data, sums of squares and products matrix, variate and group names and routines are immediately available. Normally, an analysis starts with the directive *BEGIN* and is terminated by *END*. When storage is required, a MAP storage tape is loaded, and the word *BEGIN* is followed by an analysis name, e.g.

```
BEGIN  ABC/1
```

When the directive *END* is encountered, the analysis is stored in the first vacant position on the magnetic tape. To restart an analysis, the first directive is *RESUME* followed by the name and position number of the required analysis, e.g.

```
RESUME  ABC/1  23
```

When *END* is found, the current state of the analysis is again stored in the same position, or, if it is desired to leave intact the previous stage of the analysis, a new position may be specified.

10.4. *Fault Monitoring*

Every effort is made to detect user's errors and to provide adequate diagnostics. When a fault is found, a fault number and the current directive are printed, together with the offending line, underlined as far as it has been decoded, e.g.

```
ERROR  NO.  13  IN  PRINT
PRINT  MEANS  SSP/3,2  COR
```

11. Discussion

During the past two years, more than fifteen hundred special analyses have been completed using MAP, and the program has proved itself well able to deal with the varied requirements of agricultural and biological science. Weaknesses have been removed as soon as they became apparent, and many users' suggestions have been incorporated in the program. The simplicity of the control language has enabled such extensions and amendments to be made with relative ease. This step-by-step method of development ensures that the reasonable requirements of the more sophisticated user can be satisfied whilst at the same time a working version of the basic program is always available for routine analyses.

The exploratory nature of much multivariate work has been emphasized in the introduction to this paper. Although the syntax of MAP was designed with a view to possible operation in on-line mode, we have not yet had practical experience of this. However, extensive use of the magnetic tape dump facilities (Section 10.3) has shown the benefits of the semi-interactive approach, particularly for large amounts of data.

The logical structure of MAP is simple. For example, only one matrix of sums of squares and products is defined at any time. Problems of file handling, therefore, do not arise and the "housekeeping" section of the program is not excessively large. Generalization of the program would be worthwhile only if the syntax of the control language did not thereby become unduly cumbersome for simple analyses.

No Fortran version of MAP is available. However, the GENSTAT program currently being developed at Rothamsted is written in Fortran and will include many of the features of MAP.

References

Anderson, A. J. B. (1966). A note on the construction of a general survey program in Extended Mercury Autocode. *Comput. J.*, **8**, 312–314.

Box, G. E. and Cox, D. R. (1964). An analysis of transformations. *J. R. Statist. Soc.* B, **26**, 211–252.

Gower, J. C. (1966). A *Q*-technique for the calculation of canonical variates. *Biometrika*, **53**, 588–590.

Haitovsky, Y. (1968). Missing data in regression analysis. *J. R. Statist. Soc.* B, **30**, 67–82.

Healy, M. J. R. (1963). Programming multiple regression. *Comput. J.*, **6**, 57–61.

Nelder, J. A. (1968). Weighted regression, quantal response data, and inverse polynomials. *Biometrics*, **24**, 979–985.

Woolf, B. (1951). Computation and interpretation of multiple regressions. *J. R. Statist. Soc.* B, **13**, 100–119.

Yates, F. (1968). Theory and practice in statistics. *J. R. Statist. Soc.* A, **131**, 463–477.

Yates, F. and Anderson, A. J. B. (1966). A general computer programme for the analysis of factorial experiments. *Biometrics*, **22**, 503–524.