Behavior Research Methods & Instrumentation
1975, Vol. 7 (5), 485-486

REGM: A multivariate general linear hypothesis program for least squares analysis of multivariate data

LELAND WILKINSON
Yale University, New Haven, Connecticut 06520

Several widely used analysis of variance computer programs are now available for testing general linear hypotheses (Dixon, 1973; Woodward & Overall, 1974; Cramer, Note 1; Finn, Note 2). Each of these programs has limitations. REGM, a multivariate general linear hypothesis program, has been designed for applications unsuited to these other programs.

Applications. REGM has been written to allow maximum control by the user of his analysis. The source code is clear enough for modifications by a programmer and short enough to be compiled inexpensively. Input conventions closely follow the pattern of BMD11V (Dixon, 1973), which requires knowledge of the general linear model and coding of dummy variables if analysis of variance is needed. Applications should therefore be reserved for problems that cannot be solved by other programs. Specifically, REGM will handle designs beyond the size limitations of other programs because of its economy of storage and the stability of numerical routines. This can be especially important for multifactorial and repeated measures designs, which require a large design matrix for a least squares solution. Second, REGM has missing data options for dependent and independent variables. For large multiple regressions on survey data that are missing some values on independent and dependent variables, this can be useful.

Computational Procedures. The model to be analyzed in most cases is $Y = XB + E$, where Y is an $n \times p$ matrix of n observations on p dependent variables, X is an $n \times q$ matrix of q design or independent variables, B is a $q \times p$ matrix of coefficients, and E is an $n \times p$ matrix of errors, with covariance matrix $\sigma^2 I$. The least squares estimate of B can be obtained directly from the system of simultaneous equations $XB = Y$ by orthogonalizing the X matrix and solving the system (Golub & Styan, 1973; Lawson & Hanson, 1974). This method is most accurate computationally and is used in the MANOVA program (Cramer, Note 1). Alternatively, the normal equations $X'XB = X'Y$ may be formed, permitting ordinary solutions to the nonhomogeneous system. While this is less accurate than the orthogonal procedure, accuracy can be maintained by performing all calculations in double precision arithmetic. The normal equations are solved by a Cholesky decomposition of the $X'X$ matrix (Wilkinson & Reinsch, 1971). This is faster and more stable numerically than other pivoting or inversion methods involving the cross-product or covariance matrix of independent variables. In addition, it is faster than the orthogonalization procedure. REGM should solve least squares problems as fast as or faster than any current program.

Tests of hypotheses are made by relevant hypothesis and error matrices. Each takes the form $ABC' = D$, where A is $r \times q$, C is $s \times p$, and D is $r \times p$ (Morrison, 1967). For the usual null hypothesis, D is a null matrix. The error matrix for each hypothesis is symmetric positive definite, and the hypothesis matrix is symmetric positive semidefinite. Each hypothesis is therefore tested by solving the eigenproblem $Hx = \lambda Gx$, where H and G are the hypothesis and error matrices, respectively. Cholesky decomposition reduces the problem to symmetric form, and a Householder transformation with QL shifts is used to derive the latent roots and vectors. While the roots of the symmetric matrix produced via the decomposition are the same as those of $G^{-1}H$, the vectors for the original problem must be transformed from the symmetric solution and renormalized (Wilkinson & Reinsch, 1971). These vectors are standardized by the conditional dependent variable standard deviations for analysis of dispersion problems. Instead of a factoring of $G^{-1}H$, the hypothesis or error matrix may be factored separately in the form of a cross product, covariance, or correlation matrix. For the single group case, in which the dependent variables are regressed on a single unit vector to remove the means, factoring the error matrix is equivalent to a principal components analysis. Normalized varimax rotations may be performed on any factor solution, including that for $G^{-1}H$.

Program Structure. The program is written in ANS FORTRAN IV with two exceptions: IMPLICIT REAL*8 specifications are used at the beginning of main and all subroutines, and the IBM library function DLGAMA is called. There are 16 subroutines:

(1) INPUT handles missing value options and computes cross-product matrices casewise.
(2) CONTR computes hypothesis and error matrices for each contrast $ABC' = D$.
(3) FACT1 factors the linear hypothesis $G^{-1}H$.
(4) FACT2 factors a symmetric matrix (G or H).
(5) SYMDET computes a Cholesky decomposition of a symmetric positive definite matrix.
(6) SYMSOL solves a system of real linear equations given the lower triangular coefficients matrix from SYMDET.
(7) REDUC reduces the eigenproblem $Ax = \lambda Bx$ (A positive semidefinite, B positive definite) to symmetric form via Cholesky decomposition.
(8) TRED2 produces a Householder decomposition for a real symmetric matrix.
(9) IMTQL2 computes latent roots and vectors for a tridiagonalized symmetric matrix.
(10) REBAK transforms the vectors from the symmetric eigenproblem into ones for the original one, $Ax = \lambda Bx$.
(11) PROD performs matrix products.
(12) FCDF is a complement cumulative distribution function for the F statistic.
(13) CHICDF is a cumulative distribution function for the chi-square statistic.
(14) VARMAX performs normalized varimax rotations.
(15) SORT sorts rotated factor loadings to highlight factor structure.
(16) OUTPUT writes rectangular and symmetric matrices.

Subroutines SYMDET, SYMSOL, REDUC, TRED2, IMTQL2, and REBAK are FORTRAN translations of ALGOL programs by the same names in Wilkinson and Reinsch (1971). Dummy dimensions are used in all subroutines, so that program dimensions can be modified in a single paragraph at the beginning of the program.

Input. Two cards contain program parameters. The first card defines the problem: (1) Number of independent variables. (2) Number of dependent variables. (3) Number of observations. (4) Number of hypotheses. (5) Input file number (blank if data are in stream). (6) Number of format cards for data. (7) Input option code (order for reading in variables, including or excluding column for constant). (8) Input label code (for including variable labels). (9) Missing data option (no missing values, listwise deletion, pairwise deletion.) All combinations of missing data processing options are possible for independent and dependent variables.

Variable format cards and variable label cards, if used, follow the problem card. The second parameter card specifies a hypothesis. Several options for specifying the A, C, and D matrices are available, ranging from default values created by the program to input as additional data. Options for selecting the

matrix to be factored and number of rotations to be performed are included on this card. Any number of hypotheses may be tested for a single problem, and problems may be stacked in a single job step.

Output. Output from REGM depends on problem parameters. When missing data options are selected, number of observations after deletions are printed. For regressions and analyses of variance, the determinant of the $X'X$ matrix, dependent variable means, regression coefficients, residual cross-product matrix, conditional correlations among the dependent variables, and multiple correlations are printed. For hypotheses involving single dependent variables, the following are printed: contrasts, hypotheses, and error sums of squares and F tests. For multivariate hypotheses, the following are printed: contrasts, hypothesis, and error matrices; Wilks' Lambda; Rao's approximate F statistic, largest root (theta) statistic; canonical correlations; chi-square tests of residual roots; standardized canonical coefficients for the dependent variables, and correlations (loadings) between the dependent variables and the dependent canonical factors. If rotations are requested, rotated canonical coefficients, loadings, and canonical correlations are printed.

Limitations. Because the program is designed to be custom dimensioned for particular applications, limits on problem size are determined only by available memory and error bounds. Data are processed casewise, so array storage depends on the number of dependent and independent variables. If NX represents the number of independent variables, and NY the number of dependent variables, then array storage in bytes for a 32-bit word-length machine can be computed with the following expression: $8(NX[NX + 3NY + 3] + NY[5NY + 4])$. Program size for code from the IBM FORTRAN G1 compiler is less than 80K bytes. The program contains approximately 1,600 statements, so compilation for specific applications is reasonable. On an IBM 370/158, computing time for 435 observations on three dependent variables and 125 independent variables (a 5 x 5 x 5 fixed-factor multivariate analysis of variance) with tests of eight hypotheses is under 50 sec. Smaller designs involving under 10 dependent and independent variables run from ½ to 3 sec.

Finally, computing accuracy is comparable to single-precision orthogonalization programs such as MANOVA (Cramer, Note 1). Double-precision computations are required for machines with less than 60-bit word length.

Availability. The source program may be obtained by sending a tape to Leland Wilkinson, Department of Psychology, Yale University, New Haven, Connecticut 06520. The tape will be returned standard labeled EBCDIC, nine-track, 1,600 bpi density, block size 800, logical record length 80. Special values may be requested for any of these parameters. Please enclose $10 for handling.

REFERENCE NOTES

1. Cramer, E. M. *MANOVA, a computer program for univariate and multivariate analysis of variance.* Unpublished manuscript. L. L. Thurstone, Psychometric Laboratory, University of North Carolina, 1973.

2. Finn, J. D. MULTIVARIANCE: Univariate and multivariate analysis of variance, covariance and regression. Ann Arbor, Michigan: National Education Resources, Inc., 1972.

REFERENCES

Dixon, W. J. (Ed.). *BMD: Biomedical computer programs.* Berkeley: University of California Press, 1973.

Golub, G. H., & Styan, G. P. H. Numerical computations for univariate linear models. *Journal of Statistical Computation and Simulation*, 1973, **2**, 253-274.

Lawson, C. L., & Hanson, R. T. *Solving least squares problems.* New York: Prentic-Hall, 1974.

Morrison, D. F. *Multivariate statistical methods.* New York: McGraw-Hill, 1967.

Wilkinson, J. H., & Reinsch, C. Linear algebra. In F. L. Bauer (Ed.), *Handbook for automatic computation.* New York: Springer-Verlag, 1971.

Woodward, J. A., & Overall, J. E. A general multivariate analysis of variance computer program. *Educational and Psychological Measurement*, 1974, **34**, 653-662.