Behavior Research Methods & Instrumentation
1979, Vol. 11 (3), 395-396

A FORTRAN IV program for the estimation of missing data

MARTIJN P. F. BERGER
University of Tilburg, Tilburg, The Netherlands

Missing data are often a problem in using multivariate techniques such as factor analysis, canonical correlation, regression, and discriminant analyses. Although elimination of subjects with one or more missing scores can solve the problem, it often leads to a considerable loss of information.

Several procedures have been proposed to estimate correlation matrices and covariance matrices from incomplete data (see Buck, 1960; Glasser, 1964; Timm, 1970; Wilks, 1932; and Dear, Note 1, for a review of the literature). Frane (1976) and Gleason and Staelin (1975) provide procedures for estimating missing data.

The program described here estimates missing data by regression analysis using all available variables.

Let X be an n by p standardized data matrix with scores of n subjects on p variables. X can be partitioned into two submatrices:

$$X = [X_1 X_2].$$

X_1 is of order n by p_1 and X_2 is of order n by p_2. The partitioned correlation matrix is:

$$R = \frac{1}{n} X'X = \frac{1}{n}\begin{bmatrix} X_1'X_1 & X_1'X_2 \\ X_2'X_1 & X_2'X_2 \end{bmatrix} = \begin{bmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{bmatrix}.$$

If X_1 consists of missing data and X_2 contains all available data, it is possible to estimate X_1 by the regression of X_1 on X_2. When p_1 scores are missing for any subject, then, these scores can be estimated by a transformed regression equation:

$$\hat{x}_1 = -x_2 R^{21} [R^{11}]^{-1}, \qquad (1)$$

where R^{21} and R^{11} are submatrices of the partitioned R^{-1} and $\hat{x}_1$ and $\hat{x}_2$ are row vectors of X_1 and X_2. Computationally this equation should be faster than the well-known regression equation:

$$\hat{x}_1 = x_2 [R_{22}]^{-1} R_{21}, \qquad (2)$$

because more linear equations ordinarily have to be solved in Equation 2 than in Equation 1 (Koopman, 1976). The p_2 by p_1 matrix $-R^{21} [R^{11}]^{-1}$ contains the regression coefficients for standardized variables. Raw data matrices can be handled by transforming these standardized regression coefficients into unstandardized coefficients and by computing the intercepts.

Before estimating missing data by Equation 1, the correlation matrix R must be estimated. Two methods can be used: Wilks (1932) proposes substituting the corresponding column mean for each missing entry; Glasser (1964) estimates R from all available pairs of scores for each pair of variables.

The estimates obtained from Equation 1 can be improved by iteration. The estimates from one iteration can be used to improve the estimate of the correlation matrix for use in a next iteration. The program, however, becomes very expensive to run when the number of iterations is large.

This procedure depends on the assumption that the data are missing at random and the estimates will be quite unsatisfactory when the missing variables do not correlate highly with at least one of the available variables and/or when the proportion of missing entries is large.

Description. The main program uses a set of FORTRAN IV subroutines. These routines implement the estimation procedure and all supplementary operations. The subroutine WILKS computes the correlation matrix R by Wilks' (1932) method; GLASSER computes R by Glasser's (1964) method; TRANSF draws R^{21} and R^{11} from the relevant parts of R^{-1}; REGRES estimates the missing entries according to Equation 1; CORR computes correlations and standard deviations; XMISS generates a data presence-absence matrix; GRINV is a matrix inversion routine that was published by Kaiser and Dickman (1972). The set also includes a matrix printing and punching routine.

Input. The jobdeck consists of the following cards. Card 1 is the alphameric title; Card 2 specifies the number of rows (N) and columns (NP) of the data matrix; Card 3 contains the choice of procedure for computing R (NR) and the number of iterations (NRUN), and it contains an option for punching the data matrix with estimated missing entries (NPU); Card 4 is the data format card (F-type variable format); Card 5 contains the code for missing entries. The next cards contain the raw data matrix. Each row of the data matrix must begin on a new card. The use of more than one card per row is permitted.

Output. The printed output includes the titling information and, for each iteration, the correlation matrix, the data matrix with estimated missing entries, and corresponding means and standard deviations. The proportion of missing entries is also reported; the program gives a warning when this proportion exceeds the limit of .20. The data matrices with estimated missing entries can be punched as desired.

Restrictions. The data matrix may not have more than 200 rows and 10 columns. These limits can easily

0005-7878/79/030395-02$00.45/0

be changed, however, by modifications of the array sizes. The variables IDIM and JDIM specify these array sizes as given by the DIMENSION statements in the main program. The program indicates an error and terminates the execution when the data matrix contains a whole row of missing entries and/or when the number of missing entries in a column is more than N − 2.

Computer and Language. The program is written in standard FORTRAN IV and was tested in the Tilburg University ICL 2960 computer. Conversion to other computer systems should require no reprogramming. The core requirement is approximately 67K bytes and the execution is reasonably fast. The total source deck contains 368 cards, including 64 comment cards. The data set reference numbers for the card reader, line printer, and card punch are 5, 6, and 7, respectively.

Availability. The listings of the source program are available at no charge from the author: Martijn P. F. Berger, Psychological Laboratory, Katholieke Hogeschool, Postbus 91 53, 5000 LE Tilburg, The Netherlands.

REFERENCE NOTE

1. Dear, R. E. *A principal-component missing data method for multiple regression models.* System Development Corporation, Technical Report SP-86, 1959.

REFERENCES

Buck, S. F. A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, Series B, 1960, **22**, 302-307.

Frane, J. W. Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*, 1976, **41**, 409-415.

Glasser, M. Linear regression analysis with missing observations among the independent variables. *Journal of the American Statistical Association*, 1964, **59**, 834-844.

Gleason, T. C., & Staelin, R. A proposal for handling missing data. *Psychometrika*, 1975, **40**, 229-252.

Kaiser, H. F., & Dickman, K. W. A FORTRAN program for inverting a positive definite matrix. *Educational and Psychological Measurement*, 1972, **32**, 179-180.

Koopman, R. F. Fast regression estimates of missing data. *Psychometrika*, 1976, **41**, 277.

Timm, N. H. The estimation of variance-covariance and the correlation matrices from incomplete data. *Psychometrika*, 1970, **35**, 417-438.

Wilks, S. S. Moments and distributions of estimates of population parameters from fragmentary samples. *Annals of Mathematical Statistics*, 1932, **3**, 163-195.

(Accepted for publication December 28, 1978.)