Regression estimation of missing data

KEVIN E. O'GRADY

University of New Mexico, Albuquerque, New Mexico 87131

Missing data are not uncommon in studies that employ multivariate statistical techniques (e.g., multiple regression, canonical correlation, principal components analysis, factor analysis, multivariate analysis of variance, discriminant function analysis). Various proposals have been offered to manage such situations. A useful approach, both in terms of its computational convenience and its accuracy, involves regression estimates of the missing values (Gleason & Staelin, 1975; Koopman, 1976). This method involves developing a series of regression equations for the total sample, in which each variable is predicted by the remaining variables in the set. Variables missing for some observations can be estimated through the application of the total sample regression equation, by using available observation values. The program described in this paper is intended to serve this purpose.

Estimation Procedure. The estimation procedure occurs in two stages. First, a missing data correlation matrix is constructed. This is accomplished by "plugging with means"; that is, each missing data value is replaced by its respective mean, and the correlation matrix is estimated as if all observations have complete data (Wilks, 1932). The second stage involves the regression estimation of the missing values. For each observation with missing data, the solution of the following matrix equation is undertaken:

$$y = -x * \{p^{-1} * [R_{xy} * (R_{yy})^{-1}] * A\} + C,$$

where y is a vector of data points whose values are to be estimated (i.e., missing data points); x is a vector of data points whose values are available (i.e., nonmissing data points); p^{-1} is a diagonal matrix containing the standard deviations of the y s; R_{xy} is a submatrix of the partitioned matrix of the inverse of the correlation matrix, R^{-1} , where

$$R = \begin{matrix} R_{yy} & R_{yx} \\ R_{xy} & R_{xx} \end{matrix}.$$

R_{yy} is likewise a submatrix of R^{-1} ; A is a diagonal matrix containing the standard deviations of the x s; C is a vector of intercept constants. This equation estimates the missing value(s) for an observation by applying the unstandardized regression weights, that is,

$$-P^{-1} * [R_{xy} * (R_{yy})^{-1}] * A$$

to the values that are present for that same observation, then adding the intercept constants.

Input. Input to the program consists of: (1) a title card, containing any 80 alphanumeric characters; (2) a parameter card, indicating the number of variables, number of observations, number of input and output format cards, number of iterations in the estimation stage, the value that represents missing data in the input data set, and various other program options; (3) variable format card(s) describing the input data set; (4) variable format card(s) indicating how the output data set is to be structured; (5) the input data. The input data may reside on some medium other than cards that follow immediately in the input card stream.

Output. The printed output includes the job title, program options selected, the pattern of missing data, percentage of missing data, means, standard deviations, and intercorrelations of the variables, the root-mean-square off-diagonal correlation, the determinant of the correlation matrix, Bartlett's (1950) test of independence with accompanying probability level, and the multiple correlation of each variable with every other variable. It is also possible to have the output data listed on the printed output, with estimated values underlined. In addition, the output data set can be written onto an alternate output device (e.g., tape, disk), punched onto cards, or both.

Technical Information. The program is written in ANSI standard FORTRAN IV, and should be compatible with any mainframe or minicomputer FORTRAN compiler. All computations are in single-precision arithmetic, and core storage requirements on an IBM 3032 are approximately 135 KB. This includes 35 KB for program storage and 100 KB for array storage. Array storage consists entirely of a singly subscripted array stored in COMMON. This array is subdivided at execution time according to the various program options chosen, through the use of "integer pointers" that allocate various portions of the array. Thus, the program is quite flexible in terms of the variety of problems it can handle. At present, the program can handle problems in which (number of output format cards * 20) + [number of variables * (number of variables + 1)/2] + (4 * number of variables) + (maximum size of the product of number of missing data values * number of present data values across all observations) + (2 * the maximum number of missing data values across all observations) does not exceed 25,000. This limitation can be altered easily by changing three statements in the MAIN routine.

Moreover, four peripheral storage devices are used as

binary scratch units, and one peripheral device can be allocated for permanent storage of the output data set. The unit numbers for these devices, as well as the card reader, printer, punch, and an alternate device that might contain the input data set, are assignable within the BLOCK DATA subprogram. Currently, these devices are assigned the numbers 1, 2, 3, 10, 4, 5, 6, 7, and 0, respectively.

Availability. Reprints of this paper and an annotated source listing can be obtained at no charge from Kevin E. O'Grady, Department of Psychology, University of New Mexico, Albuquerque, New Mexico 87131.

REFERENCES

- BARTLETT, M. S. Tests of significance in factor analysis. *British Journal of Psychology* (Statistics Section), 1950, **3**, 77-85.
- GLEASON, T. C., & STAELIN, R. A proposal for handling missing data. *Psychometrika*, 1975, **40**, 229-252.
- KOOPMAN, R. F. Fast regression estimates of missing data. *Psychometrika*, 1976, **41**, 227.
- 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 February 12, 1982.)