Behavior Research Methods & Instrumentation
1975, Vol. 7 (5), 480

MISDAT: A Fortran IV program for the estimation of missing data in repeated measures design

DAVID J. STANG
Queens College, Flushing, New York 11367

and

VICTOR REZMOVIC
Syracuse University, Syracuse, New York 13210

Missing data has long presented an important, difficult problem in data analysis, one which has unfortunately received little attention. Three solutions have been previously advanced to handle this problem in repeated measures and randomized complete block designs. In the first technique, one discards all replications in which some data are missing. Unfortunately, this neat solution is often very uneconomical as well as being theoretically questionable, since the loss of data is likely to be nonrandom. With the second method (e.g., Yates, 1933), one estimates the missing values by variously weighting means from adjacent cells. As Winer (1971) observes, this method is of limited utility because it fails to take into account trends across treatments. Finally, one may estimate the missing values using a regression equation. Although elegant, this procedure is rarely used because of the enormity of the task. MISDAT estimates missing data using a computer-implemented regression technique.

Description. The present technique assumes that there is no interaction between replications (blocks or subjects) and treatments; i.e., the averaged group curve is assumed to legitimately describe the shape of the individual curves.

It is helpful to imagine a data matrix in which rows represent replications, blocks, or subjects and in which columns represent treatments or levels of the independent variable. The method then consists of predicting each missing value X_{ij} by means of the formula

$$X_{ij} = G + A_i + L(Y_j - I) + Q(Y_j - I)^2 - ((N^2 - 1)/12)),$$

where G is the grand mean of all original (experimentally obtained) values in the matrix, A_i is the average difference between the original values in row i and their respective column means, L is the linear regression coefficient, Y_j is the level of the independent variable for column j or the mean of column j, I is the mean of all Y, Q is the quadratic regression coefficient, and N is the total number of columns.

This equation, derived in part from Winer (1971, p. 185) must be recalculated for each missing value. The regression coefficients are determined using a technique for trend analysis which handles equally or unequally spaced levels of the independent variable (Stang & O'Connell, 1973). MISDAT returns the original data matrix, with missing values estimated, in the original input format. After estimating the missing values, the analysis of variance, multiple regression, etc., may be then carried out in the usual fashion, except that 1 df should be substracted from the error sum of squares and total sum of squares for each missing value.

Language. MISDAT is written in Fortran IV and was developed on a DEC System 10.

Availability. A listing of the program is available free of charge from David J. Stang, Department of Psychology, Queens College, Flushing, New York 11367.

REFERENCES

STANG, D. J., & O'CONNELL, E. J. TREND: A Fortran IV subroutine for trend analysis. *Behavioral Science*, 1973, **18**, 77.

WINER, B. J. *Statistical principles in experimental design.* 2nd ed. New York: McGraw-Hill, 1971.

YATES, F. The analysis of replicated experiments when the field results are incomplete. *Empire Journal of Experimental Agriculture*, 1933, **1**, 129-142.