Behavior Research Methods & Instrumentation
1975, Vol. 7 (5), 487

A note on a FORTRAN IV program performing a set of data analysis checks

ROLF LANGEHEINE
Institut für Soziologie, Universität Kiel
D-23 Kiel (West Germany), Olshausenstr. 40-60

Reviewing the literature, one will find more and more studies applying multivariate data-analysis techniques. This holds true for the social sciences as well as for other disciplines, e.g., biology and medicine, to name a few. Researchers, however, very often do not take into account that the respective statistical routines impose certain restrictions on the data analyzed. To give an example, if you perform a metric factor analysis, the following requirements should be met:

(1) Product moment correlations analyzed should be based on: (a) variables which are normally distributed and have been measured on an interval scale, and (b) variables showing a linear relationship.

(2) The correlation matrix should show a significant deviation from the identity matrix.

A program (DC1) has been designed, therefore, to perform a series of such checks which are by far more extensive than those of known data analysis systems (e.g., SPSS, Nie, Bent, & Hull, 1970). A detailed description of the topics covered by the program is given in Langeheine (1975).

Description. Given a set of variables which have been measured on a set of objects, the program does perform the following computations:

(1) Simple statistics: Mean, standard deviation, minimum, maximum, centiles C_{10}, C_{25}, C_{50}, C_{75}, and C_{90}.

(2) Deviation from normal and/or uniform distribution: Skewness, and kurtosis.

(3) Irregularities of the distribution: Kolmogorov-Smirnov test.

(4) Histogram plot providing information of absolute frequencies, expected frequencies, and density.

(5) Product-moment correlations, where each coefficient and the matrix as a whole are submitted to tests of significance.

(6) Multiple linear correlation: Multiple correlation coefficients, beta weights, b weights, and the Cooley-Lohnes regression factor structure coefficients.

(7) Curvilinear relationships (the heart of the program): A check may be performed by eta square as well as by a polynomial curve fitting approach.

(8) Data transformation: Linear transformations based on z scores as well as the nonlinear transformation of normalized T scores which is similar to that known as McCall T-scale scores.

(9) In case normalized T scores are computed, an automatic check will be performed by polynomial regression on raw scores and normalized T scores. A special option is available if T-scale scores have been computed: The whole analysis performed so far may be repeated on T-scale scores (i.e., go to 1).

All possible significance tests are performed, and exact probabilities associated with the respective statistics are given. An option is available furthermore to handle missing data. The number of interval classes used in some computations may be either specified by the user, or a default option of the program may be chosen.

Having this information at hand, the researcher may decide how to proceed in the analysis, e.g., he may reasonably argue that a metric analysis is adequate or he may come to the conclusion that the requirements to perform such an analysis are not met and continue analyzing by some nonmetric or parametric procedure. At any rate, he has available a sounder basis which might cut down speculations.

Input. Normal input is by cards. The program may be easily modified, however, to process input from other media, e.g., disk or tape.

Running time. The program has been run on the following systems with seven sets of test data (number of variables/number of objects: 6/22, 7/22, 5/10, 3/21, 19/142, 2/15, 7/72): (1) IBM 370/168 taking 48 sec including compilation; (2) PDP-10 taking 190 sec including compilation; (3) TR440 taking 118 sec including compilation. TR440 Telefunken is similar to IBM 360/65 concerning running time.

General figures are difficult to provide as running time depends heavily on options used. In total, the program did work sufficiently on about 200 data sets so far.

Availability. Rolf Langeheine, Institut für Soziologie, Universität Kiel, D-23 Kiel (West Germany), Olshausenstr. 40-60. A detailed documentation is given in the comment of the program. Copies can be loaded on magnetic tape or DEC-tape. The cost is $30.

REFERENCES

LANGEHEINE, R. DC1—Data Check 1: A FORTRAN-IV Program. *EDV in Medizin und Biologie*, 1975, in press.

NIE, N. H., BENT, D. H., & HULL, C. H. *Statistical package for the social sciences*. New York: McGraw-Hill, 1970.