

MULTIVARIATE ANALYSIS

© J. N. R. Jeffers 1993

Statistical inference

1. Has the population about which you wish to make inferences been carefully defined?
2. Are the data to be analysed a fair sample of the defined population?
3. If the answer to either of these questions is 'No', is it worth continuing with the analysis?
4. Is the number of cases included in the data sufficient to provide a reasonable estimate of the multivariate parameters? The number of cases should always exceed the number of variables, and, ideally, should be at least four times the number of variables.
5. Have the variables included in the data been chosen specifically with some hypothesis in mind?

Choice of variables

6. Are any of the variables linearly dependent on one or more of the others? If so, those variables should be omitted.
7. Are any of the variables included in the data set ratios of other variables? If so, it will usually be preferable to use the original data with a suitable transformation.
8. Are the distributions of the separate variables approximately symmetrical? Have you looked at histograms, stem-leaf diagrams or box-plots for each of the variables?
9. Are there any outliers among the values for the variables? If so, should these outliers be removed, or do they indicate important events which should be included in the analysis?

10. Are the relationships between the variables linear? Have you looked at scatter plots of the relationships between each pair of variables?
11. Are the variances of the variables approximately the same for all values of the variables? Again, heteroscedacity can often be detected by scatter plots of the relationships between each pair of variables.
12. Have you considered appropriate transformations of the variables to cope with problems of outliers, non-linear relationships, or heteroscedacity? Have you tested the effect of using the transformations that you have chosen? Is there a good theoretical reason for the choice of one transformation in preference to others?
13. Can the variables be divided into logical groups that would help to simplify the interpretation of the results, or reflect the hypotheses underlying the analysis? What is the practical significance of those groups?
14. Are there any missing values in the data matrix? If so, it may be necessary to eliminate variables or cases until there are no missing values. Replacement of missing values may only be possible in very special cases.
15. Are all of the variables quantitative, i.e. the numbers used to measure the quantity have equal intervals? The data may, however, be either continuous or discrete. If some or all of the data are nominal (objects, events or behaviours that can be sorted into classes or groups) or ordinal (ranked or ordered from lowest to highest on the basis of some quality) special methods of analysis may have to be used.

Choice of cases

16. Do the cases for which the variables have been recorded fall into a priori groups?
17. Has any structure been imposed upon the cases for which the variables have been recorded, e.g. as the plots of an experimental design, or the sampling units in a stratified random, multi-stage or cluster sampling design?
18. Do some of the cases have disproportionate numbers of missing values? If so, can the analysis be done in such a way as to minimise the effect of those missing values?

19. If the number of cases is large, would there be some advantage in analysing the data as several independent sets by taking cases at random from the total number? Alternatively, would there be an advantage in developing the analysis on a working sample and then testing any derived hypotheses on the remainder?

Choice of method

20. If there are no logical divisions of the variables into groups, and no a priori groupings of the cases, is your main interest in the relationships between the variables? If so, principal component analysis is likely to be the most appropriate method.
21. If there are no logical divisions of the variables into groups, and no a priori groupings of the cases, is your main interest in the relationships between the cases? If so, principal co-ordinate analysis, cluster analysis, or multi-dimensional scaling are likely to be the most appropriate methods.
22. If there are no logical divisions of the variables into groups, and no a priori groupings of the cases, and you are equally interested in the relationships between the variables and between the cases, reciprocal averaging or correspondence analysis are likely to be the most appropriate methods.
23. Do the variables consist of one or more dependent variables, and several variables from which you wish to predict each of the dependent variables? If so the most appropriate method is multiple regression analysis.
24. If there are several logically distinct groups of variables and you are interested in the relationships between them, the most appropriate method is likely to be a canonical correlation analysis. Alternatively, it may be useful to calculate the principal components for each of the groups and then to examine the correlations between the calculated component scores.
25. If there are two a priori groups of cases, and you wish to discriminate between them, the most appropriate method is a discriminant function.
26. If there are more than two a priori groups of cases, and you wish to discriminate between them, the most

appropriate methods are canonical variate analysis or a multiple discriminant analysis.

27. If there are two or more groups of cases and you need to test whether there are significant differences between the variables for those groups, the most appropriate method is the multivariate analysis of variance.
28. If the variables consist of the presence or absence of a large number of characteristics, e.g. species, symptoms, identification marks, etc., the most appropriate methods are likely to be association analysis or twinspan.
29. If your primary interest is in deriving a set of rules to explain the variation contained by a data set, the most appropriate method is likely to be a genetic algorithm such as BEAGLE or GAFFER, or a rule-induction algorithm such as First Class Fusion.

Interpretation of results

30. If your primary concern is with the extent to which there is redundancy in the variability described by the variables included in the analysis, the number of orthogonal dimensions revealed by analyses such as principal components, canonical correlations, and discriminant analysis will be of major importance.
31. The proportions of the total variability accounted for by the various multivariate techniques provides a guide as to their success in explaining the underlying relationships.
32. It is extremely dangerous to rely on interpretations of the coefficients derived from multiple regression analysis, canonical correlation, or discriminant analysis. If there is marked collinearity between the original variables, even the signs attached to the coefficients can be misleading.
33. The ability to discriminate between two or more a priori groups of cases is a measure of an ability to classify new cases into those groups, but the possibility of there being other groups not included in the original data set must be borne in mind.
34. The results of a cluster analysis are strongly dependent on the clustering algorithm and the metric used in the analysis. Wherever possible, use more than one method, and check that the metric is appropriate to the original

data.

35. Where multivariate methods give rise to combinations of the original observations these combinations can sometimes be interpreted in physical terms. This sort of interpretation is called reification, and requires considerable skill and experience, including detailed knowledge of the scientific background to the data being analysed.
36. Reduction of the dimensionality of the variation contained in the data, followed by plotting of the cases on the new dimensions is called ordination and may reveal new information about the relationships between the cases. Care must be taken, however, in plotting points from three or more dimensions on to a two-dimensional surface, not to assume that two points that are close together on that plane are necessarily close on all other planes.
37. Various graphical devices such as biplots, Chernoff faces, and Andrews' Function Plots have been developed to aid in the interpretation of multivariate analyses.

Recommended texts

Afifi, A.A. and Clark, V. 1996

Computer-aided multivariate analysis. Chapman & Hall, London.

Howard, P.J.A. 1991

An introduction to environmental pattern analysis. Parthenon Publications, Carnforth.

Jeffers, J.N.R. 1992

Microcomputers in environmental biology. Parthenon Publications, Carnforth.

Jolliffe, I.T. 1986

Principal component analysis. Springer-Verlag, New York.

Kendall, M. 1980

Multivariate analysis. Griffin, London.

Krzanowski, W.J. 1996

Principles of multivariate analysis: a user's perspective. Clarendon Press, Oxford.

Lunn, A.D. and McNeil, D.R. 1991

Computer-interactive data analysis. Wiley, New York.