Appl. Statist. (1974),
23, *No.* 1, p. 60.

Miscellanea

Multivariate Analysis and its Applications: A Report on the Hull Conference, 1973

By J. C. GOWER and K. V. MARDIA†

Rothamsted Experimental Station *University of Hull*

The General Applications Section organizes a Conference every two years. It was decided to poll members to see what topics, if any, they would prefer for the 1973 Conference, and we have set out below the number of replies in favour of several possibilities.

Results of Conference Questionnaire

Conference topic	*No. of votes*
1. Multivariate analysis	67
2. Measuring the effect of social and economic policies	30
3. Comparability with Europe	27
4. Analysis of large surveys	21
5. Management control	20
6. Manpower planning	20
7. Utilization of the earth's resources	11
8. Other suggestions	5
9. No Conference in 1973	3

Responses to Possible Conference Topics

WHY multivariate analysis should be so popular is uncertain, but the choice of conference was clear. The interest in multivariate analysis had been reflected earlier, in 1970, with the formation of the Multivariate Study Group, and it was decided that the General Applications Section and Multivariate Study Group should jointly organize the conference.

It is not easy to arrange a programme that interests a heterogeneous audience ranging from the knowledgeable to the not-so-knowledgeable and the activities of the M.V.S.G. had to some extent pre-empted certain topics, at least for London-based statisticians. It was thought that a series of papers on some of the main topics of current interest should be presented, allowing plenty of discussion time and, hopefully, presenting several case studies. The final programme, to some extent, consisted of a series of review papers that we hoped would be useful both as revision and as introductions to parts of a subject that may be little known to many statisticians. We think that there was room for a few more case studies. The word multivariate was interpreted fairly broadly so that, for example, regression problems and their associated linear models for contingency tables could be included. This programme was evidently popular as there were over 300 applications for a maximum of 250 places available at Hull University's Lawns Conference Centre.

† Now at University of Leeds.

First Day

After Professor Lewis's welcoming address, we were fortunate in having Professor M. S. Bartlett to read the first paper. As was appropriate from someone who himself was directly involved in the early development of distributional results in multivariate analysis, he spoke on "Some historical remarks and recollections in multivariate analysis". In 1928, Wishart had developed the so-called Wishart distribution of the terms of a variance/covariance matrix of samples from a multinormal distribution with zero correlations. This work was ultimately derived from Fisher's treatment of the bivariate case in 1918 when he was obtaining the distribution of the correlation coefficient in the null case, and so takes us back to the Biometric school's pre-occupation with the correlation coefficient. During the early 1930's, Hotelling and Wilks were concerned with multivariate generalizations of the t and F-distributions and this was very closely linked with Mahalanobis's development of D^2. In 1933, Professor Bartlett was concerned with the relationship of these new tests with the linear models of regression theory. Fisher's development of discriminant analysis with Hotelling's canonical correlation analysis (which includes discrimination when one set of variables are dummies) extended the class of models for which distributional results were required, and brought to light problems concerning the distribution of the latent roots of a covariance matrix. This culminated in 1939 with the publication of four papers all giving the simultaneous distribution of the latent roots of a covariance matrix in the null case (Fisher, Hsu, Roy, Girshick). This must be one of the most remarkable examples of parallel working in science. The subject has developed further since the end of the war and readers may refer to Bartlett (1974) for an up-to-date account of the latest developments, including likelihood ratio factorizations in multivariate regression and canonical analysis.

The next session was concerned with the recent interest in cluster analysis. Professor R. M. Cormack, whose review paper read to the Society in 1972 will be recalled by many readers, started off this session. He outlined many of the difficulties in cluster analysis, distinguishing problems where the populations are initially unknown from those where they are known, and appealing for classification criteria which yield classes with useful properties. One cluster analysis problem that is well defined, and in its univariate form goes back to K. Pearson at the end of the 19th century, is concerned with estimating parameters of a mixture of multivariate distributions. Different variants of the problem include those where the individual samples are to be assigned to their correct components and where the number of components is to be determined. D. Wishart's paper on mode analysis was concerned with this kind of problem and presented an improved version of his previous algorithm (Wishart, 1969). The method is concerned with constructing contours enclosing regions of equal sample density, and in this improved version the contours at different levels give a hierarchic classification. The algorithm will be submitted for publication in this journal. This was the first paper to include a case study, concerning organization of chains of command in a large office. The session closed with a lively discussion.

Second Day

The second day of the conference began with two review papers on different kinds of regression models. Professor Sprent discussed the estimation of linear relationships and the problems arising from fitting linear structural and functional relationships. Mr P. R. Fisk outlined the econometricians' approach to linear models (see Sprent,

1969, and Fisk, 1967, for a substantial, if not complete, account of what was covered by these two papers). It was good for statisticians to be reminded that so much work had been done for models other than the Gauss linear model and the multinormal distribution. In reply to a questioner who seemed doubtful of the practical value of some of the economic linear models, Mr Fisk said that many econometricians were also sceptical.

The next session was entitled "Linear models for contingency tables"; an area which is well known, but because it has not been much discussed in textbooks, remains outside the general run of methods used by practising statisticians (the forthcoming Griffin monograph of Professor Plackett will soon fill the gap). The position is confused by the unhelpful, and often lengthy, discussions of partitioning χ^2 into interaction components and by the many special cases that have been considered in the journal literature. This is rather reminiscent of the days when special methods were advocated for estimating a single missing value in randomized block experiments, and for estimating two missing values either in the same or different blocks, not to mention different formulae for the different kinds of design. Dr J. A. Nelder showed how a variety of different models could be absorbed into a unified treatment which could be, and has been, implemented in a useful general computer program, the users of which could forget nine things they would have to consider with a more specialized approach (see Nelder and Wedderburn, 1973, and Nelder, 1974). Professor Plackett agreed with most of Nelder's remarks but put forward the case for using the iterative scaling procedure (ISP) for solving the likelihood equations, emphasizing the simplicity of the method which may be practicable even with desk calculators, and hence of special value in teaching. He gave several examples of the use of the method.

After lunch we had two case study papers. One, by Dr J. Markham, was about a study of the "Choice of mode for the journey to work in Dublin". This was based on a survey of how car owners journeyed to and from work and related to their socio-economic factors. Multiple regression and linear discriminant functions were used in an attempt to isolate the main features which it was hoped to use in planning the future road and public transport facilities in and around Dublin. Mr Settle was also concerned with a survey, this time of recreational activities in the N.W. of England, done under the auspices of the N.W. Sports Council. A statistical package had been used to fit multiple regression relationships between participant measures and social and other characters. Apart from technical computing troubles the fact that many of the variables were binary had also led to difficulties. Here was a clear case where linear contingency table models might be especially helpful.

The final session of the second day was concerned with aspects of factor analysis. Most statisticians are aware of the recent revolution in the computational side of factor analysis and its spin-off in terms of variance formulae, etc., all fully documented in the second edition of Lawley and Maxwell (1971). Rather than go into this again, Professor Maxwell preferred to discuss some data, based on scores obtained in two sub-samples each of 38 boys and 37 girls and on 10 sub-tests of the Wechsler Pre-school and Primary Scale of Intelligence. The original problem was to compare the factor loadings in one sub-sample with those of the other but it was found that the correlations in a group with high scores were much lower than those in a low-scoring group (Maxwell, 1972a). To interpret these results a model of the brain (Thomson, 1939; Maxwell, 1972b), which is supposed to consist of many components a proportion of which are used for each test, was invoked. Thomson showed that the expected

correlation between tests in this model naturally led to the factor analytical equations for a single general factor and Bartlett (1937) extended this model to account for several factors. A useful feature of all conferences is the opportunity they give for informal discussions outside the official sessions. It was either in such a discussion or during Professor Maxwell's talk, that he mentioned some work he and Dr Lawley had been doing on selecting variables in structural relationship linear models. Because of the errors in the "independent" variables it is well known that regression coefficients estimated by standard least-squares methods will be attenuated and this attenuation can be corrected for when estimates of the variances are available. By linking the structural relationship and factor models (Lawley and Maxwell, 1973) the variances concerned can be estimated as specific variances. Of course, another method, that many will prefer when its use is possible, is to estimate variances (and covariances too) from replicates of the observations of independent variables. However, the link between structural and factor models seems to be a new and intriguing one that may have some bearing on the next paper.

Mr H. J. L. Herne had been dissatisfied with the performance of orthogonal regression methods (i.e. regressing on principal components) and had been investigating the alternative use of factor analysis. It is true that components are scale dependent and maximum likelihood factors are not, but this seemed insufficient reason to expect factors to behave satisfactorily and it was no surprise to hear that they had not worked very well with the example discussed. But with data more appropriately analysed by a structural relationship model, this approach might be justified and useful. It is certainly worth close examination.

Dr P. Slater talked about repertory grids which are used in some psychological investigations. The basic data refer to the attitudes of a single test subject, and is in the form of a two-way table classified by Elements (which are people) and Constructs (which are attitude scales such as cheerful/miserable). The subject is asked to rank the people (who are relatives and close friends) on several scales. The traditional way of analysing such tables involves the extraction of various latent roots and vectors very reminiscent of principal components analysis and only tenuously connected with factor analysis (Slater, 1965). Dr Slater outlined the difficulties that arise when one extends data to a three-way structure obtained by questioning the same individual on many occasions. Some of the methods discussed in the multidimensional scaling session next day are appropriate here.

The Conference dinner was a great success and we were especially grateful to Professor David Kendall for giving both an entertaining and instructive after-dinner speech, that held the audience's attention throughout. Professor Kendall described an experiment at Cambridge in which specialists had first outlined problems arising in their work which had then been put into the form of a mathematical/statistical model by somebody else. The experiment included various thought-provoking topics such as (i) bird navigation, (ii) seeds and Saharan dust in the trade winds and continental drift, (iii) critical branching and structure of living matter, (iv) family size and inheritance and (v) mathematical aspects of archaeological seriation. Particularly interesting was the model in which a fulmer crossing the Atlantic was modelled as a light spherical body falling in a viscous fluid according to Young's law. We also heard of the speaker's work in mapping manorial field boundaries and fitting the South American continent into the West African coastline. Although few of the models had much to do with multivariate analysis, they should be a lesson to us all to widen our horizons when analysing and interpreting data. The students at Cambridge, who we

were glad to hear responded enthusiastically to this experiment, are very fortunate to have such a lively and wide-ranging introduction to the problems of mathematical modelling.

Final Day

The final day of the Conference began too early for many of the participants who had seized the opportunity to prolong their informal discussions long after the official close of the dinner. Professor J. Aitchison who opened the session on "Diagnosis and identification problems" had not helped matters by setting people a diagnostic task which kept one of us, at least, from his bed far too long. Three diseases were under consideration which were described by the same six tests or symptoms. A table of the values found for subjects with known diseases was given and we were asked to assign five other subjects to their appropriate group. The probabilities of belonging to each of the three groups can be represented in a triangle diagram, and we were asked to assign the five samples to the appropriate place in the diagram. The best result was presented by Miss R. N. Ward who received a prize, appropriate from a Scottish professor. Professor Aitchison had done a number of tests of this kind to ascertain how various groups of individuals such as doctors, statisticians, students, etc. behaved when given gradually more information with which to make a diagnosis. Certainly the value of an objective statistical method of diagnostic discrimination seemed to be well justified. Some account of this work is given by Taylor and Aitchison (1971) and the statistical techniques are described by Geisser (1964) and Dunsmore (1966). Dr J. A. Anderson discussed a case study of logistic discrimination arising in dental work, which has been submitted for publication in *Applied Statistics* (Anderson, 1974). Mr and Mrs M. W. Clark presented another discrimination case study in which an attempt had been made to distinguish between wind-blown and water-deposited fossil sands. As in most practical studies they had met difficulties, especially with problems of heteroscedasticity.

Multidimensional scaling methods which produce scales from distance-like data are currently fashionable. Mr J. C. Gower reviewed the subject. He outlined four different classes of problems: (a) multidimensional scaling, (b) individual scaling, which has arisen from a belated recognition in the psychological world that individuals may behave differently from the average, (c) multidimensional unfolding and (d) Procrustes rotation. Two kinds of criteria are currently used to fit models of these kinds; the so-called metric and non-metric criteria. However, their respective merits are secondary to a proper understanding of the models they are used to fit, which are likely to have much wider use than in the restricted fields to which they have so far been confined, see Gower (1974). Mr M. O. Hill read a paper on correspondence analysis. This is another "well-known" procedure that has been rediscovered on several occasions but is not adequately covered in textbooks. In France, the method is known as "l'analyse factorielle des correspondences", and it is nice to know that this term which has been appearing somewhat cryptically in the literature for the past few years can be unified with other independent discoveries of essentially the same method (see Hill, 1974).

Mr E. M. L. Beale presented a joint paper with Mr R. J. A. Little (1974) on missing values in multivariate analysis. Using simulation techniques on some alternative methods of estimating missing values, it was found that the best method consists of estimating the parameters of the parent normal population by maximum likelihood. Until recently, it presented a formidable computational problem but the

method of Orchard and Woodbury (1972) removes this difficulty. An alternative derivation of this iterative procedure, not depending on normality assumptions, was also presented. The speaker showed how one can estimate approximate standard errors of regression coefficients from incomplete data. He then discussed the facilities available in their program for missing values, and concluded with the application to some school examination data. The program is restricted to a few hundred incomplete observations which is, of course, enough for many practical problems.

The session ended with a talk by Professor K. V. Mardia who reviewed various techniques of assessing multivariate normality and studies regarding the effect of non-normality on the standard tests of multivariate theory. Significant developments in the area have included techniques of assessing normality based on the Mahalanobis "angles" and distances; a technique involving some measures of skewness and kurtosis (Mardia, 1970, 1974a) was presented in detail. Following Mardia (1971), the effect of non-normality on the general multivariate linear model was discussed. It was shown how the measures of multivariate skewness and kurtosis help in interpreting the existing Monte Carlo studies related to the effect of non-normality on the one-way analysis of variance and on tests of equality of covariance matrices. The material was presented in the context of an important practical problem (Mardia, 1974b). In the discussion, it was emphasized that the subject is of much more importance than its univariate counterpart since there are only a few multivariate non-parametric tests of any practical value.

Concluding Remarks

Throughout the Conference an exhibition of books and calculating machines was held. Many members found these useful and we understand that, at least, the sale of books was very satisfactory—we do not know about the calculating machines.

The Conference building itself is a modern structure whose designer clearly had a hexagon fixation, giving the impression that one was working in an open-plan beehive. Some people found it a little off-putting, but the effect was not serious. The Lawns Conference Centre was an agreeable setting for our meeting and we must thank Professor Lewis and the University of Hull for the excellent facilities they provided.

References

*These references are to papers read at the conference.

*Anderson, J. A. (1974). Diagnosis by logistic discriminant function: further practical problems and results. Submitted for publication in *Applied Statistics*.

Bartlett, M. S. (1937). The statistical conception of factors. *Brit. J. Psychol.*, **28**, 97–104.

*—— (1974). Some historical remarks and recollections on multivariate analysis. *Sankhyā* (in press).

*Beale, E. M. L. and Little, R. J. A. (1974). Missing values in multivariate analysis. Submitted for publication in *J. R. Statist. Soc.* B.

Cormack, R. M. (1971). A review of classification (with discussion). *J. R. Statist. Soc.* A, **134**, 321–367.

Dunsmore, I. R. (1966). A Bayesian approach to classification. *J. R. Statist. Soc.* B, **28**, 568–577.

Fisk, P. (1967). *Stochastically Dependent Equations*. London: Griffin.

Geisser, S. (1964). Posterior odds for multivariate normal classification. *J. R. Statist. Soc.* B, **26**, 69–76.

*Gower, J. C. (1974). An introduction to multidimensional scaling. To be submitted for publication in *Applied Statistics*.

*Hill, M. O. (1974). Correspondence analysis: a neglected multivariate method. *Appl. Statist.*, **23**, No. 3 (in press).

LAWLEY, D. N. and MAXWELL, A. E. (1971). *Factor Analysis as a Statistical Method*, 2nd ed. London: Butterworth.

—— (1973). Regression and factor analysis. *Biometrika*, **60**, 331–338.

MARDIA, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, **57**, 519–530.

—— (1971). The effect of non-normality on some multivariate tests and robustness to non-normality in the linear model. *Biometrika*, **58**, 105–121.

—— (1974a). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā* (in press).

*—— (1974b). Assessment of multivariate normality and robust tests. *Appl. Statist.* (to appear).

MAXWELL, A. E. (1972a). The WPPSI: a marked discrepancy in the correlations of the subjects for good and poor readers. *Br. J. math. Psychol.*, **25**, 273–291.

—— (1972b). Factor analysis: Thomson's sampling theory recalled. *Br. J. math. Psychol.*, **25**, 1–21.

*NELDER, J. A. (1974). Log linear models for contingency tables: a generalization of classical least squares. *Appl. Statist.*, **23**, No. 3 (in press).

NELDER, J. A. and WEDDERBURN, R. M. W. (1972). Generalized linear models. *J. R. Statist. Soc.* A, **135**, 370–384.

ORCHARD, T. and WOODBURY, M. A. (1972). A missing information principle: theory and applications. In *Proc. Sixth Berk. Symp. on Math. Stat. and Prob.*, **1**, pp. 697–715.

PLACKETT, R. L. (1974). *The Analysis of Categorical Data.* London: Griffin. Griffin monograph (in press).

SLATER, P. (1965). The use of the repertory grid technique in the individual case. *Br. J. Psychiat.*, **111**, 965–975.

SPRENT, P. (1969). *Models in Regression and Related Topics.* London: Methuen.

TAYLOR, T. R., AITCHISON, J. and MCGIRR, E. M. (1971). *Brit. Med. J.*, **3**, 35–40.

THOMSON, G. H. (1939). *The Factorial Analysis of Human Ability.* London: University of London Press. (Subsequent editions 1948, 1951.)

WISHART, D. (1969). Mode analysis: a generalisation of nearest neighbour which reduces chaining effects. In *Numerical Taxonomy*, (A. J. Cole, ed.), pp. 282–311. London: Academic Press.

*—— (1974). An improved multivariate mode-seeking cluster method. To be submitted for publication in *Applied Statistics.*

Appl. Statist. (1974),
23, *No.* 1, p. 66.

A Note on the Analysis of Gap-acceptance in Traffic

By ALAN J. MILLER

Division of Mathematical Statistics, C.S.I.R.O., Sydney, Australia

SUMMARY

In a recent paper, Ashton (1971) has proposed two methods for the analysis of gap-acceptance data. It is not clear what parameters of driver behaviour and of the distribution of offered gaps are being estimated in these two methods. A model is proposed for the gap-acceptance behaviour of drivers. Using this model it is shown that the asymptotic values of Ashton's estimators are functions of both the parameters describing driver behaviour and those describing the stream of offered gaps. Consequently it is not valid to use these methods to compare driver behaviour for different traffic flows. References are given to recent work on alternative methods of gap-acceptance analysis.

Keywords: TRAFFIC; GAP-ACCEPTANCE; PROBIT ANALYSIS