COMPSTAT 1982

part 1:
proceedings in
computational statistics

COMPSTAT 1982

5th Symposium held at Toulouse 1982

Part I: Proceedings in Computational Statistics

Edited by
H. Caussinus
P. Ettinger
R. Tomassone

Physica-Verlag · Wien 1982

CIP-Kurztitelaufnahme der Deutschen Bibliothek

COMPSTAT:
COMPSTAT : proceedings in computational statistics ;
. . . symposium. – Wien : Physica-Verlag

5. Held at Toulouse 1982.
Pt. 1. Proceedings. – 1982
ISBN 3-7051-0002-5

Printed in Germany by repro-druck „Journalfranz" Arnulf Liebing GmbH + Co., Würzburg

ISBN 3 7051 0002 5

Preface

Pour le cinquième congrès de la série, COMPSTAT 82 réunit environ 500 participants d'origines scientifiques et géographiques très variées, prouvant à l'évidence l'intérêt persistant de la communauté scientifique pour tous les problèmes de calculs statistiques.

Le Comité de Programme chargé de l'organisation scientifique du Congrès était composé de:
S. Apelt (République démocratique d'Allemagne) – Å. Björck (Suède) – H. Caussinus (France), Président – Y. Escoufier (France) – A. de Falguerolles (France), Secrétaire – J.W. Frane (U.S.A.) – J. Gordesch (République Fédérale d'Allemagne) – Th. Havránek (Tchéchoslovaquie) – N. Lauro (Italie) – C. Millier (France) – R.J. Mokken (Pays-Bas) – R. Tomassone (France) – D. Wishart (Royaume Uni)

Ce Comité a décidé d'augmenter le nombre des conférenciers invités, cherchant de la sorte une représentation des diverses écoles ainsi que l'introduction de nouveaux thèmes. La tâche la plus difficile a ensuite été de sélectionner une soixantaine de contributions parmi 250 soumissions. Là encore le Comité de Programme s'est efforcé de favoriser des voies qui semblaient les plus nouvelles et a essayé de maintenir une bonne répartition scientifique et géographique. Cependant, comme dans les précédents congrès COMPSTAT, il a donné la préférence aux propositions clairement marquées simultanément du double aspect Statistique et Calcul. Dans bien des cas, ces deux aspects sont très liés rendant en particulier difficile et peu pertinente toute classification fine des contributions. Pour cette raison, nous n'avons pas cherché à séparer cet ouvrage en Chapitres si ce n'est la distinction habituelle en communications invitées et libres.

Les contributions de bonne valeur qui n'avaient pu être retenues faute de place ont pu s'ajouter aux affiches (posters) proposées. Elles font l'objet d'affiches et communications courtes. Une nouveauté de COMPSTAT 82 est l'édition (par Physica Verlag) d'un volume complémentaire contenant les résumés de ces travaux.

Comme précédemment, des démonstrations de logiciels ont été organisées tout le long du Congrès afin de compléter le programme scientifique. Des résumés descriptifs de ces démonstrations ont été publiés par le Comité d'Organisation.

Fifth in the series, COMPSTAT 82 gathered about 500 participants from very diversified scientific and geographical origins, thus proving the continuing surge of interest in statistical computing.

The programme for COMPSTAT 82 was selected by an international committee, the members being:
S. Apelt (German Democratic Republic) – Å. Björck (Sweden) – H. Caussinus (France), President – Y. Escoufier (France) – A. de Falguerolles (France), Secretary – J.W. Frane (U.S.A.) – J. Gordesch (Federal Republic of Germany) – Th. Havránek (Czechoslovakia) – N. Lauro (Italy) – C. Millier (France) – R.J. Mokken (The Netherlands) – R. Tomassone (France) – D. Wishart (United Kingdom)

This committee decided to increase the number of invited speakers in order to achie the representation of different schools and the introduction of new themes. The mc difficult task was then to select about sixty papers out of 250 projects. Here also t committee tried to support novel tracks and to maintain a satisfactory scientific and ge graphical representation. Like in other COMPSTAT meetings, the committee gave its pre erence to papers simultaneously containing both statistical and computational aspect In many cases, these two aspects are very linked, thus making difficult and somewh irrelevant any refined classification of the contributions. Therefore this book is not c vided into chapters apart from the usual distinction between invited papers and contri uted papers.

Good contributions which could not be retained by lack of space were added to su mitted posters. They are presented in the format of short communications and poster A COMPSTAT 82 novelty is the editing (by Physica Verlag) of a supplementary volun containing summaries of these contributions.

Like in former COMPSTAT, demonstrations of statistical software and computi facilities have been organized. Summaries of these demonstrations have been publishe by the organizing committee.

Toulouse, juin 1982

H. Caussin
P. Etting
R. Tomasson

Contents

1. Invited Papers

2. Contributed Papers

1. Invited Papers

On Some Problems in Analysing Non-orthogonal Designs

T. Caliński, Academy of Agriculture, Poznań, Poland

SUMMARY: Discussion on some controversial problems in analysing and interpreting data from non-orthogonal designs is reviewed and certain suggestions are made on how to overcome the difficulties. Geometric approach is adopted to show possible reconcilations.

KEYWORDS: non-orthogonal designs, unbalanced data, non-orthogonal analysis of variance, linear models.

1. Introduction

Association of a linear model with the experimental data is fundamental for the statistical analysis in many fields of research. Though the origin of the methodology dates back to Gauss (1809) and the literature on the subject is vast, not many statistical techniques have caused so much controversy as the least-squares fit of a linear model. In fact, as far as the analysis of variance is applied to a properly balanced experimental design there is a common agreement on how the analysis should be performed and how its results can be interpreted (although there may be disagreements when distinguishing between the fixed effects and the random effects models - the matter that will not be discussed here). The troubles start when the data are unbalanced and the usual orthogonal analysis of variance can not be applied uniquely. As Bock and Brandt (1980) put it, "to move from balanced to unbalanced designs in analysis of variance is not only to lose the ease of computation of the orthogonal solution, but also the intuitively appealing equivalence of the observed marginal means to the least-squares estimates of effects and, perhaps more important, also the uniqueness of the additive partition of the total sum of squares".

As the unity of method and clarity of interpretation disappear, various diverging approaches are adopted and controversies emerge.

In this paper attention is drawn to certain controversial issues that may influence preparations of analysis of variance programs. An attempt is made to clarify some of the problems by using the geometric approach to linear models.

2. Some of the Controversies

As an illustration to one of the most controversial problems, let us look at the results obtained when applying five different analysis of variance programs to the same set of data coming from an unbalanced 2x5 (Sex x Religion) classification originally analysed by Francis (1973) and then subsequently discussed by several authors. The results are reproduced in the table, given here in the form due to Aitkin (see his contribution to the discussion on Nelder's,1977, paper), who added the last column. Aitkin explains the various results as follows: "Column (a) arises from a computing method which ignored the non-orthogonality, substracting the unadjusted main

ANOVA RESULTS FROM FRANCIS AND AITKIN

Source	d.f.	Sums of squares obtained by different programs				
		(a)	(b)	(c)	(d)	(e)
Mean	1	Not given	7305.78	12982.73	12982.73	12982.73
Sex	1	43.58	11.17	43.58	28.71	28.71
Religion	4	69.36	57.74	54.49	69.36	54.49
SxR	4	- 4.61 (!)	10.25	10.25	10.25	10.25
Error	1300	2988.95	2988.95	2988.95	2989.00	2988.95

effects SS from the among cell SS. Columns (c) and (d) are hierarchical analyses, giving respectively sex, religion adjusted for sex and interaction adjusted for sex and religion, and religion, sex adjusted for religion, and interaction adjusted for religion and sex. Column (b) arises from a "regression" analysis in which each effect is adjusted for all others in the model: thus sex is adjusted for religion and the interaction, and religion is adjusted for sex and the interaction (and the mean is adjusted for all effects!)". As to the column (e); "Here the main effects are adjusted for each other as in column (b), but are not adjusted for the interaction".

Nobody would support method (a), though the existence of such a program was in a way "fortunate", by initiating examination of some mysteries of the various programs (see Francis, 1973). According to their attitude towards the remaining four methods, statisticians can be clustered into three schools: (i) of the "hierarchical" partition (c) and (d), (ii) of the "full-regression" approach (b), and (iii) of the "experimental design" solution (e). Some representatives of the schools are: Bock (1963), Searle (1971), Overall and Klett (1972), O'Brien (1976), Aitkin (see Nelder, 1977) of school (i); Afifi and Azen (1972), Francis (1973), Kutner (1974) and Speed and Hocking (1976) of school (ii); Rao (1973), Nelder (1974, 1976, 1977) and *Gianola (1975) of school (iii). (Also see Bock and Brandt, 1980.)*

Another important problem that splits statisticians is the choice between the non-full rank overparameterized model (termed Model A by Hocking and Speed, 1975) and the full rank cell means model (termed by them Model B). Model A represents the present conventional approach to linear models. But Urquhart, Weeks and Henderson (1973), giving a historical account of the models, trace back Model B to the pioneers (particularly to the early works of Yates, 1933, 1934). They argue that the Model B approach "clarifies what common techniques are really estimating without the necessity of introducing estimability or imposing restrictions", and in a more recent paper Urquhart and Weeks (1978) conclude that "the cell means model gives the researcher the flexibility of specifying precisely the functions of interest". Even more outspoken critics of the Model A approach are Bryce, Scott and Carter (1980), who express their objections as follows: "In the early 1950's the use of the new common overparameterized model came into vogue. With it came concepts of estimability, testability, generalized inverse, non-unique solutions, etc. While these served to broaden the understanding of mathematical statisticians and supplied them with endless topics for papers in technical journals, the practicioner of statistical science either threw away data until balance was achieved or applied regression or other methods without knowing the hypotheses being tested or even realizing that different approaches implied different hypotheses".

This brings us to one of the main issues in computations of the non-orthogonal analysis of variance, to the "know-what-you-are-testing" problem.

Several Statisticians have examined what kind of hypotheses are really tested when applying the various analysis of variance procedures implemented in widely available computer programs (see, e.g., Francis, 1973, Kutner, 1974, Speed and Hocking, 1976, Speed, Hocking and Hackney, 1978, Hocking, Hackney and Speed, 1978, Frane, 1980, Goodnight, 1980, Searle, 1980, and Bock and Brandt, 1980). Their opinions show that we are far from a common agreement on how the analysis should proceed and how results can be interpreted. In particular, the gap between the overparameterized modelers and the cell means modelers has not been bridged, though most seem to agree that the only logical procedure is first to set out a hypothesis of interest and then to test it. This view is strongly advocated by Searle(1980), from the position of the Model A camp, and by Hocking, Speed and Coleman (1980), from the Model B camp. This contradicts with the, at present, prevailing practice of "producing" estimable functions and testable hypotheses by the programmes themselves, often without much control and understanding by the user.

It seems that a common agreement would have a chance to be reached only if we could look at the two models (A and B) as two sides of the same think, one rather supporting than contradicting the other. Moreover, it might be helpful to return to the basic principles and have a fresh look at the linear model, without the inherent prejudices of the two camps. Since both, A and B, models are algebraic in nature, the reconcilation might be sought in the geometric approach. This was already tried in the past (see Corsten, 1958). In fact, the use of geometric approach in the theory of linear models has a long and interesting history, as revealed recently by Herr (1980). Unfortunately, this approach has not yet become very popular.

3. The Principles of Geometric Approach

In this section we would like to draw attention to some features of the geometric approach to linear models that can help in profound understanding of the least-squares estimation and hypotheses testing. We do not want to go into details, as they have already been exposed in a number of papers (see, e.g., Monlezun and Speed, 1980.). We will only indicate certain principle ideas, leaving their explorations to those who become interested.

In the classical, algebraic, approach to linear models the attention is *focused on a parameter space, and the parameter vector is estimated* (not necessarily uniquely) by minimizing a quadratic form. The estimate of the mean (expectation) for the observation vector can then be obtained by transforming an estimate of the parameter vector with the use of the design matrix. In the geometric approach, on the contrary, attention is *focused on* observation space, where the mean vector of observations is estimated directly by orthogonally projecting the vector of observations onto an appropriate subspace. From this estimate the parameter vector, or its linear function, can be estimated by an appropriate linear transformation, provided such a transformation exists, i.e. the parameter vector, or its function, is identifiable.

To use some notation, suppose y is an $n \times 1$ vector of observations, in the *n*-dimentional Euclidean space E^n, with a mean vector $E(y) = \theta$. Taking $e = y - \theta$, the model can be written $y = \theta + e$. But the model is not set completely until we do two thinks more: define a p-dimensional subspace of E^n, Ω, in which θ is supposed to lie, and assume some distribution for e. As for the latter, the usual assumption is that it is multivariate normal with a covariance matrix $\sigma^2 I$ (though further generalizations are possible). As to Ω, its choice is usually not a straightforward matter. It has to be defined in such a way to reflect the basic structure of the experimental data and the interest of the researcher.

Technically, the definition of Ω can be accomplished by imposing any of the two types of relations, $A\theta = 0$ or $\theta = X\beta$, with appropriately chosen matrices A or X. In the first case, Ω is termed the null space of A, written $\Omega = N[A]$; in the second, it is called the range cpace of X, written $\Omega = R[X]$. (For details see Seber, 1980.) Depending on circumstances, one or the other way may be more convenient. For example, *in a simple 1-way classification* of data, with 2 classes, we have initially the model $y_{ij} = \theta_{ij} + e_{ij}$ ($i = 1,2,$; $j = 1,2, \ldots, n_i$). But usually it will be natural to assume that $\theta_{11} = \theta_{12} = \ldots = \theta_{in_i} = \mu_i$ (say), for 1,2. This can be imposed on the model either by the $n_1 + n_2 - 2$ relations $\theta_{ij} - \theta_{in_i} = 0$ ($i = 1,2$; $j = 1,2, \ldots, n_i - 1$) or by the $n = n_1 + n_2$ relations $\theta_{ij} = \mu_i$ ($i = 1,2$; $j = 1,2, \ldots, n_i$). The form of the $(n - 2)\times n$ matrix A and that of the $n \times 2$ matrix X are obvious. Also it is evident that the vector β is here $(\mu_1,\mu_2)'$. Both definitions of Ω are in this case simple, and evidently equivalent.

In a more complex case of a 3×4 crossed classification (as discussed by Searle, 1971, Chapter 7) with n_{ij} observations in the (i,j)-cell, we can first ignore the fact that the classes cross and define the model as for a 1-way classification with 12 classes (cells). If the vector θ is written

$$\theta = (\theta_{11}',\theta_{12}',\theta_{13}',\theta_{14}',\theta_{21}',\theta_{22}',\theta_{23}',\theta_{24}',\theta_{31}',\theta_{32}',\theta_{33}',\theta_{34}')'$$

where $\theta_{ij}' = (\theta_{ij1}, \theta_{ij2}, \ldots, \theta_{ijn_{ij}})$, then the $(n-12) \times n$ matrix A is of the form $A = \mathrm{diag}\{A_{11},A_{12},A_{13},A_{14},A_{21},A_{22},A_{23},A_{24},A_{31},A_{32},A_{33},A_{34}\}$ where the $(n_{ij} - 1) \times n_{ij}$ submatrix A_{ij} represents $n_{ij}-1$ contrasts within the (i,j)-cell ($i = 1,2,3$; $j = 1,2,3,4$). Using the other method of defining Ω, we could use as matrix X the matrix

$$W = \begin{bmatrix} 1_{11} & 0 & \cdots & 0 \\ 0 & 1_{12} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1_{34} \end{bmatrix},$$

where 1_{ij} is the $n_{ij} \times 1$ vector of ones ($i = 1,2,3$; $j = 1,2,3,4$). As the corresponding vector β we would then have the 12×1 vector $\mu = (\mu_{11},\mu_{12}, \ldots, \mu_{34})'$, where $\mu_{ij} = \theta_{ij1} = \theta_{ij2} = \ldots = \theta_{ijn_{ij}}$ ($i = 1,2,3$; $j = 1,2,3,4$) (We have changed here from X to W and from β to μ to keep with the usual notation of that approach.) In either of the two ways of defining Ω we have arrived at the so called full rank cell means model (Model B). Usually the more convenient definition $\Omega = R[W]$, instead of $\Omega = N[A]$, is used in this case. To get the non-full rank overparameterized model (Model A) we would use for this example the common $n \times (1+3+4+3\times4)$ matrix X that corresponds to the vector $\beta = (\mu,\alpha_1,\alpha_2,\alpha_3,\beta_1,\beta_2,\beta_3,\beta_4,\gamma_{11},\gamma_{12}, \ldots, \gamma_{34})'$. Now the subspace for θ is defined as $\Omega' = R[X]$. But since the matrix W is equal to the last 12 columns of X and since $p = \mathrm{rank}\ X = \mathrm{rank}\ W$ (= 12), $\Omega = \Omega'$. Hence the two models are equivalent (assuming the same distribution for the random part e) and the overparameterization of Model A is evident. The only appar-