

The Computer Analysis of Factorial Experiments: In Memoriam—Frank Yates

John A. NELDER and Peter W. LANE

The responses to a recent paper by Dallal in this journal are evaluated by reference to the ideas of Frank Yates. It is concluded that much unnecessary complication has been introduced into the computer analysis of linear models by (1) the imposition of constraints on parameters, (2) neglect of marginality relations in forming hypotheses, and (3) confusion between the form of noncentrality parameters and hypotheses.

KEY WORDS: Constraint; Linear model; Marginality; Noncentrality parameter.

1. INTRODUCTION

Frank Yates died in June 1994 at the age of 92. He wrote a justly famous work, *The Design and Analysis of Factorial Experiments* (Yates 1937), universally known as "TC35." He also installed one of the very first electronic computers, the Elliott 401, at Rothamsted, and, on it and its successors, did pioneering work on statistical computing, including, of course the subject of TC35.

A recent paper in this journal (Dallal 1992) on the subject of the analysis of a small data set by modern packages gave rise to a series of invited responses by Lazaro, DeLong, Lam, Rogers, Fleury, and Kutner (1994) and by Searle (1994) on the methods employed by various packages and their interpretation. We think that we can safely say that FY (as he was known to all his colleagues) would have been more than a little critical of some of the points made, and we wish to say why we believe that to be so.

2. THE GREAT MIXED-MODEL MUDDLE

A substantial component of the literature on linear models can, we believe, be summed up by this title. One of the respondents to Dallal (Rogers) concludes with the following: "In testing a two-way model with interactions, the fixed-effects model uses the error mean square, but the mixed model uses the interaction mean square for testing the fixed main effect. As additional factors and higher-order interactions are added, the correspondence between mixed- and fixed-effects models breaks down." We reject both of these assertions.

The first assertion is equivalent to testing a main effect in the presence of interactions. We argue below that such a test corresponds to a hypothesis of no inferential interest. The second assertion, that fixed- and mixed-effects models

are basically different, is the subject of much of what follows here. Historically it is interesting to note that Yates (1975) described the contrast between fixed and random effects as "a distinction without a difference." Without following him entirely in this view, we assert that there are many more similarities than differences.

How then has the alternative view, as exemplified by Rogers, become established? We believe that it arises from the habit of imposing on the parameters in a linear model the same constraints as are imposed on the estimates of those parameters in order to get unique solutions to the least-squares equations. Thus it is assumed that we should impose the constraint $\sum \alpha_i = 0$ on the parameters α_i because we want to impose the constraint $\sum a_i = 0$ on the estimates a_i of the parameters. However, what looks like a nice symmetry between parameters and their estimates leads to unnecessary complications, including assertions such as those of Rogers.

Let us compare the two approaches in the analysis of an $N_A \times N_B$ two-way table indexed by factors A and B , and with a single element in each cell. Tables 1 and 2 show the expected values of mean squares for four models fitted to this table, with and without constraints on the parameters. Terms like σ_A^2 are variance components: σ_A^2 is the excess variance between rows (A margin) over that described by the interaction term $A \cdot B$. Terms like Σ_A are mean squares of parameters.

With constraints, the patterns of the entries in the four columns are all different, while without constraints they are all the same. The differences in Table 1 result from the inability to constrain independent random effects to add to zero, for obvious reasons. The similarity of the entries in Table 2 illustrates what Yates had in mind in making his remark: fixed effects produce terms that are sums of squares of parameters about their mean, while random effects produce corresponding variance components obtained by taking the expectations of these sums of squares. The appearance of the term Σ_{AB} in the expectation of the main effect of A , when A is fixed, shows us that what we should be assessing is not whether the A effects are null, given that the $A \cdot B$ effects are not, but rather whether they show excess variation over and above that displayed by $A \cdot B$. This observation leads naturally to another important simplifying idea, namely, that of marginality between terms in factorial models (Nelder 1977).

3. MARGINALITY RELATIONS

In our two-way table there is an $N_A \times N_B$ parameter space, which we denote by γ , associated with the parameter set γ_{ij} , say. There are three subspaces of importance, which we denote by 1 , α , and β , associated with the parameter sets μ , α_i , and β_j . The first one corresponds to the intercept or grand-mean term, the second to the set of A effects, and the third to the set of B effects. The subspace 1

John A. Nelder is Visiting Professor of Statistics, Imperial College, London SW7 2BZ, England. Peter W. Lane is Senior Biometrician, IACR—Rothamsted, Harpenden AL5 2JQ, England. IACR receives grant-aided support from the Biotechnology and Biological Sciences Research Council of the United Kingdom. The authors appreciate the helpful comments of Donald A. Preece.

Table 1. Expected Mean Squares for Models With Constraints

Term	A fixed B fixed	A fixed B random	A random B fixed	A random B random
A	$\sigma^2 + N_B \Sigma_A$	$\sigma^2 + \sigma_{AB}^2 + N_B \Sigma_A$	$\sigma^2 + N_B \sigma_A^2$	$\sigma^2 + \sigma_{AB}^2 + N_B \sigma_A^2$
B	$\sigma^2 + N_A \Sigma_B$	$\sigma^2 + N_A \sigma_B^2$	$\sigma^2 + \sigma_{AB}^2 + N_A \Sigma_B$	$\sigma^2 + \sigma_{AB}^2 + N_A \sigma_B^2$
A · B	$\sigma^2 + \Sigma_{AB}$	$\sigma^2 + \sigma_{AB}^2$	$\sigma^2 + \sigma_{AB}^2$	$\sigma^2 + \sigma_{AB}^2$

lies wholly inside each of the subspaces α , β , and γ , while both α and β lie wholly inside γ . We say that 1 is marginal to α , β , and γ , and that α and β are marginal to γ . The residual subspace, defined by those contrasts among the γ_{ij} not included in 1, α , and β , we call $A \cdot B$ in the notation of Wilkinson and Rogers (1973).

Neglect of marginality relations leads to the introduction of hypotheses that, although well defined mathematically, are, we assert, of no inferential interest. The simplest example of such a hypothesis is that A is null although $A \cdot B$ is not. This hypothesis is illustrated by both Kutner (his first table) and Searle (his equations 7, 10, and 11) in their responses. It does not make inferential sense because it asserts that although the A effects for each level of B are entirely unrelated, nonetheless their average effects are null. Far from being a baseline state from which to measure deviations, a null margin in the presence of interactions is an extreme state, and thus of no inferential interest. This problem was discussed by Kempthorne (1975, p. 483) who stated that "The testing of main effects in the presence of interaction, without additional input, is an exercise in futility." Nelder (1977, p. 50) went further, stating that "I contend that such models [with an interaction and a null marginal effect] are of no practical interest." If this view is accepted, much complication disappears.

4. NONCENTRALITY PARAMETERS AND UNBALANCED DATA

The component of the expectation of a sum of squares in an analysis-of-variance table produced by the fixed terms of the model is usually known as the noncentrality parameter (and referred to here as the N-CP). When the data are unbalanced, the N-CP is a quadratic form in the relevant parameters with coefficients that depend on the numbers of observations in each cell. It is widely asserted in the literature that this is an undesirable state of affairs because, it is said, hypotheses to be tested should not depend upon the numbers in the cells, whose values may often be accidental.

We believe that there is a confusion here between the form of an N-CP and the hypothesis to be tested. When the n_{ij} are unequal, the form of the N-CP reflects the fact that there is variable information about parameters α_i , say, in the data, and the least-squares equations correctly allow for that. However, the hypothesis associated with this N-CP remains simply that $\alpha_i = 0$. There is no conflict between the two. Where it appears that some hypothesis is a function of the cell numbers, it will invariably be found that marginality constraints have been ignored. Once such uninteresting hypotheses are discarded, the apparent discrepancies vanish.

With this formulation the advantage claimed for Types III and IV sums of squares, namely, that their N-CP's are symmetric functions of the parameters, disappears. The effect of using them in significance tests is simply to lose power. It cannot be stressed too strongly that the hypotheses for the balanced and unbalanced cases are exactly the same; changes in the N-CP's are a reflection of changed amounts of information about the parameters in the data.

5. THE CASE AGAINST TYPES III AND IV SUMS OF SQUARES

Many experimenters, faced with output that includes these sums of squares, have found difficulty in understanding what they are supposed to be telling them about their data. We suspect that many statisticians, on being asked for advice, have also been uncertain of their value. The arguments above show that when interactions are present, the hypotheses that correspond to them are of the uninteresting kind that break marginality constraints, and that even when not, the form of the N-CP causes loss of power in any test.

There remains one argument that has been put forward to justify their use, and that is that in certain cases a Type III sum of squares is the same as the weighted sum of squares of means given by Yates (1934). We believe that this argument is without substance. In 1934, to fit a main-effects model to a substantial two-way table with

Table 2. Expected Mean Squares for Models Without Constraints

Term	A fixed B fixed	A fixed B random	A random B fixed	A random B random
A	$\sigma^2 + \Sigma_{AB} + N_B \Sigma_A$	$\sigma^2 + \sigma_{AB}^2 + N_B \Sigma_A$	$\sigma^2 + \sigma_{AB}^2 + N_B \sigma_A^2$	$\sigma^2 + \sigma_{AB}^2 + N_B \sigma_A^2$
B	$\sigma^2 + \Sigma_{AB} + N_A \Sigma_B$	$\sigma^2 + \sigma_{AB}^2 + N_A \sigma_B^2$	$\sigma^2 + \sigma_{AB}^2 + N_A \Sigma_B$	$\sigma^2 + \sigma_{AB}^2 + N_A \sigma_B^2$
A · B	$\sigma^2 + \Sigma_{AB}$	$\sigma^2 + \sigma_{AB}^2$	$\sigma^2 + \sigma_{AB}^2$	$\sigma^2 + \sigma_{AB}^2$

unbalanced numbers in the cells was a major task, involving many hours of handle-turning on a Brunsviga hand calculator or the like. By contrast, a weighted sum of squares of means was relatively quick to calculate, and gave a rough-and-ready assessment of the size of row or column effects. Today we have no need of that shortcut.

There is, however, a modern interpretation of the weighted means from which Yates's sum of squares is formed. They comprise, in the formulation of Lane and Nelder (1982), a predictive margin for the table. Such predictive margins can be used to answer what-if questions of the form "What would the mean response be for factor A if the weights in the cells were those of factor B as a whole?" Such predictive margins are of undoubted use, and are often the quantities of interest formed at the end of an analysis. However, they belong to the predictive stage and not to the model-selection stage. There is no question at this stage of testing hypotheses; what we are interested in is predicting the margins under whatever weighting system is chosen, and giving estimates of uncertainty via covariance matrices, standard errors of contrasts, and so on. A sum of squares formed from a predictive margin rarely will be of any inferential use, and is certainly irrelevant at the model-selection stage, when parsimonious models are being sought.

To summarize this section, we see no place for Types III and IV sums of squares in making inferences from the use of linear models.

6. DALLAL'S DATA

The data set which began this discussion, although small in size, is not without interest. For convenience we repeat it in Table 3. Two things stand out: first, the response variable, which we call y , has a range of values that exceeds 20:1. This suggests that it is most unlikely that an analysis of variance done on the original scale will be satisfactory. The variance is unlikely to be independent of the mean, and systematic effects may well not be additive. Second, the data values within the cells show a pattern that is not reflected in the indexing system. The A2 values in each column show the same pattern, three low values followed by three high values, with no overlap whatever. One possible explanation is that this was once a 3×3 table with three observations in each cell that has been collapsed into a 2×3 table for purposes of illustrating some arithmetic. If this explanation is correct, then violence has

Table 3. Data From a 2×3 Two-Factor Experiment

	B1	B2	B3
A1	3.81	3.42	3.55
	4.64	3.57	3.71
	4.09	3.55	3.66
A2	.22	.36	.37
	.33	.27	.31
	.36	.26	.28
	1.08	.83	.70
	1.33	.90	.89
	1.15	.93	.93

Table 4. Analysis of Variance on the Original Scale

Source of variation	df	s.s.	m.s.	v.r.	F pr.
A	1	59.1157	59.1157	482.80	<.001
B	2	.5717	.2858	2.33	.121
A · B	2	.2822	.1411	1.15	.335
Residual	21	2.5713	.1224		
Total	26	62.5409			

been done to the data. Even if not correct, it suggests that some unnamed factor has been influencing the variation, and that its omission may make any analysis of doubtful value.

The analysis on the original scale gives the anova shown in Table 4. There appears to be some sign of interaction in that the A · B mean square is greater than the residual. The result of analyzing $\log(y)$ is shown in Table 5. Now all sign of interaction has disappeared; in fact the interaction is significantly small.

Note that our analysis for y gives the same results as columns II and III in Lazaro's Table 1. The other columns in that table can be obtained by treating the means in each cell as having equal weight. This means, however, changing the assumptions about the error structure of the data. The justification given seems to be that it is a way of coping with lack of balance. However, these data are not unbalanced and give an orthogonal anova. It is the introduction of the superfluous Type III sums of squares that gives rise to this latter analysis. Such an analysis would be relevant if the data were of the split-plot type in which the main-plot component greatly exceeded the within-plot component. Here, the change from one form of analysis to another would be the consequence of changing the model of the error structure, so that the least-squares equations change accordingly.

The analyses in this section were done with Genstat (Genstat 5 Committee 1993), which has several useful features for the analysis of designed experiments. First, and most importantly, it distinguishes between block factors, which define the basic structure of the experimental units, and treatment factors, which define the experimental treatment combinations applied. This distinction makes it possible to describe the experimental structure of all the designs in, for example, Cochran and Cox (1957) by two simple model formulas, one for blocks and one for treatments. From these, together with the design matrix, the analysis-of-variance table is derived by the algorithm, together with all the distinct standard errors that are required for interpretation of the tables of means. Other facilities

Table 5. Analysis of Variance on the Log Scale

Source of variation	d.f.	s.s.	m.s.	v.r.	F pr.
A	1	22.7111	22.7111	74.17	<.001
B	2	.1324	.0662	.22	.807
A · B	2	.0007	.0004	.00	.999
Residual	21	6.4299	.3062		
Total	26	29.2742			

include functions in the treatment formula for breaking down factors into individual degrees of freedom, a facility for generating pseudofactors, which can render designs like lattices generally balanced, and full access to all the components of an analysis, so that information from experiments can be combined.

We conclude that more questions need to be asked about how these data came into being, before anything sensible can be done about their analysis.

7. WHAT SHOULD WE TEACH STUDENTS?

First, we should teach them about data. Thus a student faced with Dallal's data should be taught to ask: (1) What are the units of measurement of the response? (2) What are the indexing factors? (3) What are their levels? (4) Did the data come from a randomized experiment and, if so, how was it randomized? Second, we should teach them about models, starting with linear models, and shorn of the unnecessary complexities discussed above. The basic notion, we suggest, is that of the sequential fitting of a nested sequence of linear models, with the first differences of the residual sums of squares leading to the anova table (Type I sums of squares in SAS notation). There should be no introduction of constraints on parameters to match those imposed on estimates, and the latter should be explained as existing solely to give unique solutions to the least-squares equations and arising from the earlier fitting of marginal terms. The fact that estimable functions have values independent of the particular set of constraints imposed should be illustrated in detail.

The idea of marginality and the constraints it puts on sequential analysis will limit the sequences that have to be considered to produce a parsimonious model. Thus for two factors (leading to three possible terms) there are just two possible sequences: $A, B, A \cdot B$ or $B, A, A \cdot B$. With three factors the seven terms in the full factorial model produce not $2^7 = 128$ possible sequences, but just 48, when marginality constraints are taken into account. The presence of partial or total orthogonality further limits the distinct set of sequences. Model selection involves a strategy for picking sets of these sequences in an adaptive way. An obvious idea is to introduce the main-effect model, followed by the introduction, and selection from, two-factor interactions. Heuristic rules include not accepting an apparently large three-factor interaction if all of its two-factor marginal interactions are negligible.

Ideas of model checking are vital, and modern computing facilities should make them easy to carry out. They should be introduced in parallel with the model-selection strategies. For teaching the design of experiments and the corresponding forms of anova we recommend the Genstat approach because it gives a single framework within which the whole large class of generally balanced designs can be placed.

If the consequence of such a course is to make the students very suspicious of packages that purport to give the experimenter THE analysis of his or her data, then it will have succeeded.

8. CONCLUSIONS

Much unnecessary complication has been introduced into the treatment of linear models in statistical packages by:

- (1) the assumption that, because constraints must be placed on the estimates of parameters in linear models in order to get unique solutions to the least-squares equations, the same constraints must be placed on the parameters themselves;
- (2) ignoring the fact that hypotheses that break marginality relations between terms in a factorial model are of no inferential interest; and
- (3) confusing the form of a noncentrality parameter in a sum of squares with the corresponding hypothesis.

These points are developed further, with special reference to computing, in Nelder (1994).

To answer the specific query raised by Dallal, there should indeed be no difference between the two analyses he presents. If he were to eschew the use of Type III sums of squares in favor of Type I, as we recommend, the difference would disappear whatever package was used. However, we also suggest that the data he presented are not suited to this analysis anyway.

[Received October 1994. Revised May 1995.]

REFERENCES

- Cochran, W. G., and Cox, G. M. (1957), *Experimental Designs* (2nd ed.), New York: John Wiley.
- Dallal, G. E. (1992), "The Computer Analysis of Factorial Experiments With Nested Factors," *The American Statistician*, 46, 240.
- Genstat 5 Committee (1993), *Genstat 5 Release 3 Reference Manual*, Oxford, England: Clarendon Press.
- Kempthorne, O. (1975), "Fixed and Mixed Models in the Analysis of Variance," *Biometrics*, 31, 473-486.
- Lanc, P. W., and Nelder, J. A. (1982), "Analysis of Covariance and Standardization as Instances of Prediction," *Biometrics*, 38, 613-621.
- Lazaro, C. G., DeLong, D., Lam, M.-L., Rogers, W. H., Fleury, P. J., and Kutner, M. (1994), "Invited Response to Dallal, G. E. (1992)," *The American Statistician*, 48, 140-147.
- Nelder, J. A. (1977), "A Reformulation of Linear Models," *Journal of the Royal Statistical Society, Ser. A*, 140, 48-63.
- (1994), "Inference From Linear Models: Back to Basics," *Statistics and Computing*, 4, 221-234.
- Searle, S. R. (1994), "Analysis of Variance Computing Package Output for Unbalanced Data From Fixed-Effects Models With Nested Factors," *The American Statistician*, 48, 148-153.
- Wilkinson, G. N., and Rogers, C. E. (1973), "Symbolic Description of Factorial Models for Analysis of Variance," *Applied Statistics*, 22, 392-399.
- Yates, F. (1934), "The Analysis of Multiple Classifications With Unequal Numbers in the Different Classes," *Journal of the American Statistical Association*, 29, 51-66.
- (1937), *The Design and Analysis of Factorial Experiments*, Technical Communication No. 35, Commonwealth Bureau of Soils, Farnham Royal, England: Commonwealth Agricultural Bureaux.
- (1975), "The Early History of Experimental Design," *A Survey of Statistical Design and Linear Models*, ed. J. N. Srivastava, Amsterdam: North Holland, pp. 581-592.