DECIPHERING THE OUTPUT OF ANOVA PROGRAMS FOR UNEQUAL-SUBCLASS-NUMBERS DATA USING BENCHMARK DATA SETS[1/]

by

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

BU-671-M March, 1979

Abstract

A series of small, hypothetical data sets of increasing complexity regarding characteristics such as numbers of observations and empty cells are being used to ascertain the precise nature of output values (which are often not labeled unequivocally) from computer routines designed for analysis of variance of data having unequal numbers of observations in the cells. All possible analyses and calculations are known for the data sets and comparison with those of the computer output values reveals what those values exactly are. The different computing features of a variety of routines are easily identified by this means.

Introduction

Analysis of variance of data having unequal numbers of observations in the subclasses (hereafter called unbalanced data) is considerably more complicated than that of equal-subclass-numbers data (balanced data). Not only are the calculations more extensive and complicated, with a variety of alternatives for

---

[1/] Paper presented at Computer Science and Statistics: 12'th Annual Symposium, Waterloo, Ontario, May 1979.

Paper No. BU-671-M in the Biometrics Unit Mimeo Series.

partitioning a total sum of squares, for example, but interpretation is also more difficult. In recent years the calculation problem has largely (but not entirely) been overcome: there are now numerous computer packages that will do most of the necessary arithmetic. But the difficulty of interpreting the analyses still remains, and indeed it has become increasingly more evident (e.g., Speed et al. [1978]). Furthermore, insofar as computer packages are concerned, there is the prerequisite to interpreting calculations, of knowing exactly what calculations are represented by each individual computer output value. For example, in any particular package just exactly what is the sum of squares labeled "A" or "due to A"? With balanced data it can usually mean only one thing, but with unbalanced data it may mean one of several things, and its use can be and is different in different computer packages. Statisticians must therefore know not only what these different meanings might be, but also which of them occurs in each of the different computer packages that they use. Statisticians also need a vehicle for ascertaining from new programs what their output values are.

Reading program documentation is one method of ascertaining exactly what a program's output is; but in the case of analyses of variance of unbalanced data it is usually a most unsatisfying method. To the extent that users read a documentation in anticipation of learning statistics from it, their dissatisfaction with documentation is quite reasonable; after all, a user of statistical computer packages is meant to know the underlying statistics and should be reading documentation to find out solely what it is that a package does. A documentation does not have to be a statistics manual.

Regardless of documentation, there are at least two other methods for ascertaining precisely the mathematical description of computer output. One is to read program code — a quite impractical task for most people. The other is to use the routine on small, hypothetical data sets for which all possible analyses and

calculations are known exactly (preferably in rational fractions rather than decimals), or can be obtained with desk facilities. Comparing these known calculations with computer output provides a basis for ascertaining what the computer output is. For example, in the analysis of rows-by-columns data (A by B), exactly what is the sum of squares in a computer output that is labeled "A"? Is it $R(\mu,A)$, $R(A|\mu)$, $R(A|\mu,B)$, SSA_w, or, under some circumstances is it $R(A|\mu,B,AB)$?; i.e., is it the total sum of squares due to fitting a mean and rows, or that due to fitting rows adjusted for the mean, or due to fitting rows adjusted for the mean and columns, or that due to rows in the weighted squares of means analysis — or is it something else? Comparing known values of these possible interpretations with computer output reveals what that output is.

A possible weakness of this comparative method is, of course, that it is based solely on numbers and so, arising from idiosyncracies of input values in the hypothetical data sets, one might be led to conclusions about the meaning of output values that do not hold true in general. Using a series of data sets guards against this possibility.

Procedure

This method of comparing computer output with pre-calculated analyses of benchmark data sets is being used in a project at the Biometrics Unit, Cornell University. It is based on seven data sets, each consisting of a small amount of hypothetical data which, of themselves, have no intrinsic value other than being a vehicle for ascertaining what calculations are being done by different computer routines designed for computing analyses of variance of unbalanced data. The seven data sets represent, in some approximate sense, data of increasing complexity regarding features such as numbers of observations, numbers of empty cells, interactions and covariates. Their general characteristics are shown in Table 1.

Table 1. Characteristics of seven sets of hypothetical data used for ascertaining what calculations are being done by computer routines designed for analyses of variance of unbalanced data.

Data Set	Characteristics
	Balanced data
1	2-way crossed classification, 4 rows, 3 columns and 2 observations per cell.
	Unbalanced data, 2-way crossed classifications
2	4 rows, 3 columns, 0 or 1 observation per cell, no interaction.
3	2 rows, 3 columns, all cells filled.
4	2 rows, 3 columns, one empty cell.
5	3 rows, 4 columns, 4 empty cells.
	Covariance analysis, with 1 covariate
6	1-way classification, 3 groups, with 3, 2 and 2 observations.
7	2-way crossed classification, same layout as Data Set 5.

Although the data sets of Table 1 are by no means an exhaustive array for their intended purpose, they have proven to be varied enough to illustrate and verify numerous computing procedures in the routines that have been used to date: BMDP2V, GENSTAT ANOVA, SAS GLM, SAS HARVEY and SPSS ANOVA.

Results

Certain features of these routines that are very apparent from this kind of study are now listed, using as illustration the analysis of a 2-way crossed classification. The model is taken as

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (1)$$

with $i = 1, \cdots, a$, $j = 1, \cdots, b$, and $k = 1, \cdots, n_{ij}$ for $n_{ij} > 0$, and $n_{ij} = 0$ for cells having no data. Customary dot and bar notation is used for totals and means; e.g.,

$$y_{i..} = \sum_{j=1}^{b} \sum_{k=1}^{n_{ij}} y_{ijk}, \qquad n_{i.} = \sum_{j=1}^{b} n_{ij}, \qquad \bar{y}_{i..} = y_{i..}/n_{i.},$$

and reductions in sums of squares are exemplified by

$$R(\mu,\alpha,\beta,\gamma) = \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}\bar{y}^2_{ij.}, \tag{2}$$

$$R(\alpha|\mu) = \sum_{i=1}^{a} n_{i.}\bar{y}^2_{i..} - n_{..}\bar{y}^2_{...} \tag{3}$$

and

$$R(\alpha|\mu,\beta) = \underset{\sim}{u}'\underset{\sim\sim}{T}\underset{\sim}{u} \tag{4}$$

with

$$\underset{\sim}{u} = \left\{y_{i..} - \sum_{j=1}^{b} n_{ij}\bar{y}_{.j.}\right\} \quad \text{and} \quad \underset{\sim}{T} = \left\{\delta_{ii'}n_{i.} - \sum_{j=1}^{b} n_{ij}n_{i'j}/n_{.j}\right\}$$

for $i,i' = 1, \cdots, a-1$ (e.g., Searle [1971, p. 297]); and

$$SSA_w = \text{sums of squares for the } \alpha\text{-effects in the weighted squares of means analysis.} \tag{5}$$

Full details of these notations are in Searle [1971, Chapters 7 and 8].

We also use expressions like

$$R^*(\dot{\mu}|\dot{\alpha},\dot{\beta},\dot{\gamma})_\Sigma \equiv R(\mu,\alpha,\beta,\gamma) - R^*(\dot{\alpha},\dot{\beta},\dot{\gamma})_\Sigma \tag{6}$$

and

$$R^*(\dot{\alpha}|\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma \equiv R(\mu,\alpha,\beta,\gamma) - R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma \tag{7}$$

introduced in Searle, Speed and Henderson [1979]. They are sums of squares for a

model (1) in which the elements are $\dot{\mu}$, $\dot{\alpha}$, $\dot{\beta}$ and $\dot{\gamma}$ and satisfy what are often called the "usual" or Σ-restrictions, namely

$$\sum_{i=1}^{a} \dot{\alpha}_i = 0, \qquad \sum_{j=1}^{b} \dot{\beta}_j = 0, \qquad \sum_{i=1}^{a} \dot{\gamma}_{ij} = 0 \ \forall j \qquad \text{and} \qquad \sum_{j=1}^{b} \dot{\gamma}_{ij} = 0 \ \forall i . \quad (8)$$

Suppose we have the normal equations for such a model, the Σ-restricted model: they yield the second terms in (6) and (7). From those equations delete $\dot{\mu}$ and the $\dot{\mu}$-equation, and for the remaining equations calculate the inner product of the solution vector and the vector of right-hand-sides. The result is what we call $R^*(\dot{\alpha},\dot{\beta},\dot{\gamma})_\Sigma$ of (6); and $R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma$ of (7) is obtained similarly. The overhead dots show that it is a restricted model, the Σ indicates that it is the Σ-restrictions being used, and the asterisk indicates they are being used throughout all of the calculations – for otherwise $R(\dot{\alpha},\dot{\beta},\dot{\gamma})$ would, by definition, be identical to $R(\alpha,\beta,\gamma)$ and $R(\mu,\alpha,\beta,\gamma)$. Further details of (6) and (7) are available in Searle, Speed and Henderson [1979]. With this notation we are able to summarize some features of the five computer routines considered to date.

1. In SPSS, the total main effects sum of squares is $R(\alpha,\beta|\mu)$. Without option 10, the succeeding items are $R(\alpha|\mu,\beta)$ and $R(\beta|\mu,\alpha)$, which do not sum to $R(\alpha,\beta|\mu)$; with option 10, those items are a partitioning of $R(\alpha,\beta|\mu)$, such as $R(\alpha|\mu)$ and $R(\beta|\mu,\alpha)$.

2. The SPSS Multiple Classification Analysis is based on a no-interaction model even if the input model being used contains interactions. "Unadjusted deviations" are simply deviations of marginal means, e.g., $\bar{y}_{i..} - \bar{y}_{...}$, and the corresponding ETA-value is $[R(\alpha|\mu)/SST_m]^{\frac{1}{2}}$. "Adjusted deviations" are simply solutions to normal equations based on the weighted Σ-restrictions like $\Sigma n_{i.}\dot{\alpha}_i = 0$. The resulting BETA-values are values like $(\Sigma n_{i.}\dot{\alpha}_i^{\circ 2}/SST_m)^{\frac{1}{2}}$, where SST_m is the total sum of squares corrected for the mean.

3. BMDP2V uses Σ-restrictions, e.g., equation (8).

4. BMDP2V does not print a solution to the normal equations.

5. BMDP2V calculates sums of squares of the nature illustrated in (6) and (7). This means that

$$\text{Sum of squares due to mean} = R^*(\dot{\mu}|\dot{\alpha},\dot{\beta},\dot{\gamma}) \neq n_{\cdot\cdot}\bar{y}^2_{\cdots} .$$

And for the case of all cells filled (i.e., every cell containing data)

$$\text{Sum of squares due to A} = R^*(\dot{\alpha}|\dot{\mu},\dot{\beta},\dot{\gamma}) = SSA_w ; \tag{9}$$

$$\text{Sum of squares due to B} = R^*(\dot{\beta}|\dot{\mu},\dot{\alpha},\dot{\gamma}) = SSB_w . \tag{10}$$

When there are empty cells in the data the second equalities in (9) and (10) do not hold.

6. BMDP2V cannot handle interaction models when data have any empty cells.

7. GENSTAT ANOVA is designed for balanced data and handles unbalanced data using "missing value" techniques; this requires the user to indicate which values in his data are "missing".

8. SAS HARVEY uses Σ-restrictions and yields many of the same sums of squares as does BMDP2V. But its calculation procedure is "indirect", using the "invert part of the inverse" rule for full rank models (see Searle [1971, p. 115] and Searle, Speed and Henderson [1979]).

9. SAS HARVEY outputs numerous sums of squares and products, and correlations, based on the columns of the $\underset{\sim}{X}$-matrix in $E(y) = \underset{\sim\sim}{X}b$; for analysis of variance models these outputs are of no use.

10. For unbalanced data and interaction models, SAS HARVEY will function only if there is at least one level of the A-factor and one level of the B-factor that has data in every cell; i.e., at least one row must have data in every column and one column must have data in every row.

11. SAS GLM, basing its calculations on a generalized inverse of the coefficient matrix of the normal equations for the unrestricted model (1), calculates four types of sums of squares and arbitrary forms of estimable functions that can be used to explain hypotheses corresponding to each. For each sum of squares, the estimable function has the form $f = \underset{\sim}{\ell}'\underset{\sim}{b}$ where $\underset{\sim}{\ell}'$ is a vector whose elements are linear functions of r arbitrary values, for r being the degrees of freedom of the sum of squares; and the hypothesis that is tested by using the sum of squares as the numerator of an F-statistic is then $H: f_i = 0$ for $i = 1, 2, \cdots, r$, where the individual f_i are r linearly independent forms of f obtained from using r sets of the r arbitrary values that are the basis of $\underset{\sim}{\ell}'$ in $f = \underset{\sim}{\ell}'\underset{\sim}{b}$.

The four types of sums of squares are as follows:

Type 1: $R(\alpha|\mu)$, $R(\beta|\mu,\alpha)$, $\cdots$, for fitting factors sequentially.

Type 2: $R(\alpha|\mu,\beta)$, $R(\beta|\mu,\alpha)$, $\cdots$, for fitting each factor adjusted for all others (but not adjusted for interactions of other factors with, nor for factors nested within, the factor concerned).

Type 3: Implicitly uses the Σ-restrictions like (8), as does BMDP2V and SAS HARVEY.

Type 4: Based on "contrasts" derived from non-unique, balanced subsets of filled cells of the data.

Further details of this classification are available in Searle [1979].

12. SAS GLM output has a vector labeled "ESTIMATE" following the table of sums of squares. It is the solution to the normal equations $\underset{\sim}{X}'\underset{\sim}{X}\underset{\sim}{b}^{o} = \underset{\sim}{X}'\underset{\sim}{y}$, corresponding to the generalized inverse $\underset{\sim}{G}$ of $\underset{\sim}{X}'\underset{\sim}{X}$ that is used (customarily one that corresponds to restrictions of setting certain individual effects to zero). Following the "ESTIMATE" vector is a vector labeled "T FOR HO PARAMETER = 0".

This is a t-statistic, but its use is not always a test of H : parameter = 0; it is, only when there is no B following the output ESTIMATE value. Otherwise it is a test of $\underset{\sim}{h}_i'\underset{\sim}{b} = 0$ where, for the i'th element in the ESTIMATE vector h_i is the i'th row of $\underset{\sim}{\underset{\sim}{G}}\underset{\sim}{X}'\underset{\sim}{X}$. An example, for a no-interaction 2-way classification model with four α-effects is that corresponding to α_1^o the t-statistic tests $H : \alpha_1 - \alpha_4 = 0$.

Availability of Annotated Output

Output generated by processing the data sets of Table 1 through each computer routine has been extensively annotated with illustrations, comments and descriptions that expand upon the preceding results. The resulting document for each routine includes the data sets and their basic analyses, and is available as an Annotated Computer Output (ACO), in 8 x 11 format, for the following routines: BMDP2V, GENSTAT ANOVA, SAS GLM, SAS HARVEY, and SPSS ANOVA. The ACO's are obtainable ($5 each) from the Biometrics Unit, 339 Warren Hall, Cornell University, Ithaca, New York, 14853.

The project is expected to continue, using other computer routines and updated versions of routines as they appear. Only routines for fixed effects models have been considered to date, but those for variance components estimation are likely to become part of the project also. Suggestions for improvements and extensions to the project, and to the annotated outputs themselves, will be welcomed.

References

Searle, S. R. [1971]. Linear Models. Wiley and Sons, New York.

Searle, S. R. [1979]. Arbitrary hypotheses in linear models with unbalanced data. Communications in Statistics (in press).

Searle, S. R., Speed, F. M., and Henderson, H. V. [1979]. Some computational and model equivalences in analyses of variance of unequal-subclass-numbers data. Paper No. BU-608-M in the Biometrics Unit, Cornell University, Ithaca, N. Y.

Speed, F. M., Hocking, R. R., and Hackney, O. P. [1978]. Methods of analysis of linear models with unbalanced data. J. Amer. Stat. Assoc., 73, 105-112.