QUIRKS IN LINEAR MODEL CALCULATIONS WITH UNBALANCED DATA[1/]

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

ABSTRACT

Completion over the past two years of Annotated Computer Output for a number of analysis of variance routines has revealed situations where linear model calculations for unbalanced data are sometimes a little surprising or, at best, somewhat difficult to understand. Such situations are illustrated with (i) faults in an algorithm for reparameterizing with Σ-restrictions, (ii) sums of squares for Σ-restricted models, (iii) least squares means and (iv) estimating variance components.

1. INTRODUCTION

The recent preparation of Annotated Computer Output (e.g., Searle et al., 1978, 1980) for a variety of statistical computer packages has highlighted certain quirks in linear model calculations with unbalanced data (data having unequal numbers of observations in the subclasses). Awareness of these quirks provides a basis for understanding relationships among output obtained from processing the same data on different computing procedures. This paper illustrates some of these relationships.

The illustrations are in terms of the two-way cross-classification model, specified by two factors which shall be called rows and columns. The model equation is either

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

or

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \qquad (2)$$

where $E(y_{ijk})$ is the expected value, over repeated sampling, of the k'th observation, y_{ijk}, in the i'th row and j'th column of the data. In both (1) and (2), μ is a general mean, α_i is the effect due to the i'th row, for $i = 1, \cdots, a$, and β_j is the effect due to the j'th column, for $j = 1, \cdots, b$; and in (2), γ_{ij} is the effect due to the interaction of the i'th row and j'th column. Thus (1) and (2) are called the no interaction model, and the interaction model, respectively.

Balanced data is where every one of the ab cells has the same number of observations, n say. Unbalanced data is where the cells have unequal numbers of observations, possibly including some empty cells; n_{ij} denotes the number of observations in the cell defined by row i and column j, and for $n_{ij} > 0$, the k of (1) and (2) takes values $k = 1, 2, \cdots, n_{ij}$. Empty cells correspond to $n_{ij} = 0$. Thus, in general, $n_{ij} \geq 0$. In the no interaction model (1), the values of n_{ij} are often just 0 or 1; and balanced data are just a special case of unbalanced data with every $n_{ij} = n$. With unbalanced data there is the necessary distinction between situations in which every cell contains at least one observation (the all-cells-filled case), and those in which some cells have no data (the some-empty-cells case).

2. Σ-RESTRICTED MODELS

Linear models that are not of full rank are often reparameterized to be of full rank by imposing restrictions on the parameters of the model. One popular form of such restrictions is that which is coming to be known as (e.g., Searle, et al. 1981) the Σ-restrictions. These define the effects for each factor so that they add to zero; for example, $\sum_{i=1}^{a} \alpha_i = 0$ and $\sum_{j=1}^{b} \beta_j = 0$. The Σ-restrictions have a long history in linear model theory for the analysis of balanced data. They can also be used with unbalanced data, whereupon a popular algorithm for incorporating them in the calculations can be faulty when used on data that have empty cells. The Σ-restrictions also lead to peculiarities in calculating certain sums of squares.

Consider a situation in which the numbers of observations in a data grid of 3 rows and 4 columns are as shown in Grid 1.

Grid 1: Numbers of Observations

3	-	1	2
2	2	-	-
-	2	2	4

This set of n_{ij}-values corresponds to the example in Table 7.6 of Searle (1971), and to Data Set 5 of the Annotated Computer output, Searle et al. (1978, 1980).

The Σ-restricted interaction model appropriate to Grid 1 is, akin to (2),

$$E(y_{ijk}) = \dot{\mu} + \dot{\alpha}_i + \dot{\beta}_j + \dot{\gamma}_{ij}, \qquad (3)$$

but with the Σ-restrictions

[1/] Paper No. BU-385 in the Biometrics Unit.

$$\left.\begin{array}{lr}
\dot\alpha_1 + \dot\alpha_2 + \dot\alpha_3 & = 0 \\
\dot\beta_1 + \dot\beta_2 + \dot\beta_3 + \dot\beta_4 & = 0 \\
\dot\gamma_{11} + \dot\gamma_{13} + \dot\gamma_{14} & = 0 \\
\dot\gamma_{21} + \dot\gamma_{22} & = 0 \\
\dot\gamma_{32} + \dot\gamma_{33} + \dot\gamma_{34} & = 0 \\
\dot\gamma_{11} + \dot\gamma_{21} & = 0 \\
\dot\gamma_{22} + \dot\gamma_{32} & = 0 \\
\dot\gamma_{13} + \dot\gamma_{33} & = 0 \\
\dot\gamma_{14} + \dot\gamma_{34} & = 0
\end{array}\right\}\text{rewritten as}\left\{\begin{array}{l}
\dot\alpha_3 = -\dot\alpha_1 - \dot\alpha_2 \\
\dot\beta_4 = -\dot\beta_1 - \dot\beta_2 - \dot\beta_3 \\
\dot\gamma_{11} = \dot\gamma_{11} \\
\dot\gamma_{13} = \dot\gamma_{13} \\
\dot\gamma_{14} = -\dot\gamma_{11} - \dot\gamma_{13} \\
\dot\gamma_{21} = -\dot\gamma_{11} \\
\dot\gamma_{22} = \dot\gamma_{11} \\
\dot\gamma_{32} = -\dot\gamma_{11} \\
\dot\gamma_{33} = -\dot\gamma_{13} \\
\dot\gamma_{34} = \dot\gamma_{11} + \dot\gamma_{13}
\end{array}\right. \qquad (4)$$

The dots above the symbols in (3) and (4) distinguish this model from the unrestricted model (2). The second column of equations in (4) is the Σ-restrictions of the first column rewritten in terms of the minimum number of parameters needed for the model, namely two $\dot\alpha$'s, three $\dot\beta$'s and two $\dot\gamma$'s. The first two equations in the $\dot\gamma$'s simply emphasize that all the $\dot\gamma$'s can be expressed in terms of two of them. This is further demonstrated by writing the $\dot\gamma$'s as in Grid 2.

Grid 2: $\dot\gamma$'s for Grid 1

$\dot\gamma_{11}$	-	$\dot\gamma_{13}$	$-\dot\gamma_{11} - \dot\gamma_{13}$
$-\dot\gamma_{11}$	$\dot\gamma_{11}$	-	-
-	$-\dot\gamma_{11}$	$-\dot\gamma_{13}$	$\dot\gamma_{11} + \dot\gamma_{13}$

Note in equations (4) and in Grid 2 that the Σ-restrictions for interaction effects apply only to those effects which occur in the data. Thus the first $\dot\gamma$-equation in (4) has no $\dot\gamma_{12}$ because the 1,2 cell is empty. This is not the same as including γ_{12} in the model and assuming it zero, as is done, implicitly, in RUMMAGE, for example.

Corresponding to (3) and (4), expected values of observations in the first row of Grid 1 are

$$E(y_{11k}) = \dot\mu + \dot\alpha_1 + \dot\beta_1 + \dot\gamma_{11}$$

$$E(y_{13k}) = \dot\mu + \dot\alpha_1 + \dot\beta_3 + \dot\gamma_{13}$$

and

$$E(y_{14k}) = \dot\mu + \dot\alpha_1 - \dot\beta_1 - \dot\beta_2 - \dot\beta_3 - \dot\gamma_{11} - \dot\gamma_{13}\,.$$

Equations such as these can be set out for all observations in Grid 1, whereupon if we write those equations as

$$E(\underset{\sim}{y}) = \dot{\underset{\sim}{X}}\dot{\underset{\approx}{b}}$$

the rows of $\dot{\underset{\sim}{X}}$ will be as given in Table 1.

Table 1: Rows of the $\dot{\underset{\sim}{X}}$-matrix for data of Grid 1, using the Σ-restrictions of (4).

No. of rows in $\dot{\underset{\sim}{X}}$	Column of $\dot{\underset{\sim}{X}}$							
	$\dot\mu$	$\dot\alpha_1$	$\dot\alpha_2$	$\dot\beta_1$	$\dot\beta_2$	$\dot\beta_3$	$\dot\gamma_{11}$	$\dot\gamma_{13}$
$n_{11} = 3$	1	1	.	1	.	.	1	.
$n_{13} = 1$	1	1	.	.	.	1	.	1
$n_{14} = 2$	1	1	.	-1	-1	-1	-1	-1
$n_{21} = 2$	1	.	1	1	.	.	-1	.
$n_{22} = 2$	1	.	1	.	1	.	1	.
$n_{32} = 2$	1	-1	-1	.	1	.	-1	.
$n_{33} = 2$	1	-1	-1	.	.	1	.	-1
$n_{34} = 4$	1	-1	-1	-1	-1	-1	1	1

2.1. Faults in an algorithm

In the first three lines of Table 1, and in the last two lines, the coefficient of each $\dot\gamma$ is the product of the coefficients of the $\dot\alpha$ and $\dot\beta$ having corresponding subscripts. For example, in the first and last lines the coefficient of $\dot\gamma_{11}$ is 1; in the first line this is the product of two 1's which are the coefficients of $\dot\alpha_1$ and $\dot\beta_1$; and similarly in the last line it is the product of two -1's. Except for the three "boxed" values, this product algorithm holds for all coefficients of $\dot\gamma_{ij}$'s in Table 1 — and it is an algorithm that has been known and used in computer programs for many, many years — certainly back to 1962, to this writer's knowledge. But notice that this algorithm does not hold for the "boxed" values in Table 1. For example, the first of these is -1 and, as the coefficient of $\dot\gamma_{11}$, the algorithm would have it be the product of 0 and 1 (the coefficients of $\dot\alpha_1$ and $\dot\beta_1$ in that line), which it is not. This is a fault of the algorithm: it does not apply universally, for all unbalanced data sets that have empty cells.

The reason for the breakdown of the algorithm is as follows. Customary usage of the algorithm is based on replacing the last effect in each Σ-restriction by minus all the others therein; e.g., rewriting $\dot{\alpha}_1 + \dot{\alpha}_2 + \dot{\alpha}_3 = 0$ as $\dot{\alpha}_3 = -\dot{\alpha}_1 + \dot{\alpha}_2$ as in (4). The algorithm is always a correct representation of these replacements when all cells are filled. It is also correct when some cells are empty, providing they are not in the last row and column of the data. Then, so far as $\dot{\gamma}$'s are concerned, having empty cells is equivalent to having all cells filled but simply deleting the $\dot{\gamma}$'s corresponding to empty cells.

Having data in every cell of the last row and column ensures that in each Σ-restriction for the $\dot{\gamma}$'s, there is in that last row and column a $\dot{\gamma}$ (in the model for the data) that can be replaced by other $\dot{\gamma}$'s. For example, in Grid 1 the Σ-restriction for $\dot{\gamma}$'s in row 1 is $\dot{\gamma}_{11} + \dot{\gamma}_{13} + \dot{\gamma}_{14} = 0$, from which the $\dot{\gamma}_{14}$ of the last column can be replaced by $-\dot{\gamma}_{11} - \dot{\gamma}_{13}$. But in row 2 the Σ-restriction for $\dot{\gamma}$'s is $\dot{\gamma}_{11} + \dot{\gamma}_{12} = 0$, which contains no $\dot{\gamma}_{14}$ corresponding to the last column. Of course, some other, appropriate, replacement is readily ascertained in easy cases like this one, but a general algorithm using this procedure is usually tied to making all replacements from one row and one column, and so requires that that row and column have all cells filled. Presumably, this is why SAS HARVEY requires data to have (or be resequenced to have) the last row and column with all cells filled. Certain it is that when data corresponding to Grid 1 are processed by SAS HARVEY, no results are forthcoming and on investigation one finds that the three "boxed" values of Table 1 are being taken as zero, and not as shown in Table 1.

(Option 9 of SPSS ANOVA also requires one row and one column of the data to have all cells filled, and in this case they must be the first, not the last, row and column.)

2.2. Sums of squares

When rows and columns are called factors A and B respectively, and are presented to SAS GLM in the sequence A then B, the following well-known sums of squares are to be found among the output.

$$\begin{aligned} R(\alpha|\mu) &= R(\mu,\alpha) - R(\mu) \\ &= \text{sum of squares due to fitting } E(y_{ijk}) = \mu + \alpha_i \\ &\quad \text{over and above that due to fitting } E(y_{ijk}) = \mu \\ &= \text{SAS GLM Type I for A,} \end{aligned} \tag{5}$$

and

$$\begin{aligned} R(\alpha|\mu,\beta) &= R(\mu,\alpha,\beta) - R(\mu,\beta) \\ &= \text{sum of squares due to fitting } E(y_{ijk}) = \mu + \alpha_i + \beta_j \\ &\quad \text{over and above that due to fitting } E(y_{ijk}) = \mu + \beta_j \\ &= \text{SAS GLM Type II for A.} \end{aligned} \tag{6}$$

Furthermore, although $R(\alpha|\mu,\beta,\gamma)$ is a well defined symbol in this notation, it is always (in unrestricted models) identically zero, because

$$R(\alpha|\mu,\beta,\gamma) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\beta,\gamma) \tag{7}$$

where

$$\begin{aligned} R(\mu,\alpha,\beta,\gamma) &= \text{s.s. for fitting } E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}\bar{y}^2_{ij}. \end{aligned} \tag{8}$$

and

$$\begin{aligned} R(\mu,\beta,\gamma) &= \text{s.s. for fitting } E(y_{ijk}) = \mu + \beta_j + \gamma_{ij} \\ &= \sum_{i=1}^{a} \sum_{j=1}^{b} n_{ij}\bar{y}^2_{ij}. \end{aligned} \tag{9}$$

as in Searle (1971, p. 292 and 252, respectively). On substituting (8) and (9) into (7) it is clear that $R(\alpha|\mu,\beta,\gamma)$ is identically zero.

Sums of squares of this nature can also be considered in Σ-restricted models, such as that illustrated in Section 2.1. Then we have

$$R^*(\dot{\alpha}|\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma = R^*(\dot{\mu},\dot{\alpha},\dot{\beta},\dot{\gamma})_\Sigma - R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma. \quad (10)$$

Dots above symbols signify that they are from a restricted model; the subscript Σ indicates that the restrictions are the Σ-restrictions, and the asterisk indicates that when a sub-model is involved it is based not on its own Σ-restrictions but on those of the full model from which it came. For this reason, for the 2-way classification model with interaction,

$$R^*(\dot{\mu},\dot{\alpha},\dot{\beta},\dot{\gamma})_\Sigma = R(\mu,\alpha,\beta,\gamma) \quad (11)$$

because both terms apply to the full model; but

$$R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma \neq R(\mu,\beta,\gamma) \quad (12)$$

because both terms apply to sub-models of their corresponding full models. Detailed explanation is given in Section 8 of Searle _et al._ (1981).

Consider the following example.

Table 2: Data

7,9	6	2
8	4,8	12

Normal equations for the full-rank, Σ-restricted, interaction model for these data turn out (_loc. cit._) to be

$$\begin{bmatrix} 8 & 0 & 1 & 1 & 1 & -1 \\ 0 & 8 & 1 & -1 & 1 & 1 \\ 1 & 1 & 5 & 2 & 1 & 0 \\ 1 & -1 & 2 & 5 & 0 & -1 \\ 1 & 1 & 1 & 0 & 5 & 2 \\ -1 & 1 & 0 & -1 & 2 & 5 \end{bmatrix} \begin{bmatrix} \hat{\dot{\mu}} \\ \hat{\dot{\alpha}}_1 \\ \hat{\dot{\beta}}_1 \\ \hat{\dot{\beta}}_2 \\ \hat{\dot{\gamma}}_{11} \\ \hat{\dot{\gamma}}_{12} \end{bmatrix} = \begin{bmatrix} 56 \\ -8 \\ 10 \\ 4 \\ 18 \\ 4 \end{bmatrix} \quad (13)$$

with solution

$$\frac{1}{72}\begin{bmatrix} 10 & 0 & -1 & -1 & -3 & 3 \\ 0 & 10 & -3 & 3 & -1 & -1 \\ -1 & -3 & 19 & -8 & -3 & 0 \\ -1 & 3 & -8 & 19 & 0 & 3 \\ -3 & -1 & -3 & 0 & 19 & -8 \\ 3 & -1 & 0 & 3 & -8 & 19 \end{bmatrix} \begin{bmatrix} 56 \\ -8 \\ 10 \\ 4 \\ 18 \\ 4 \end{bmatrix} = \begin{bmatrix} 7 \\ -10/6 \\ 1 \\ -1 \\ 10/6 \\ 10/6 \end{bmatrix} \quad (14)$$

From the data in Table 2, using (8),

$$R(\mu,\alpha,\beta,\gamma) = \sum_{i=1}^{2}\sum_{j=1}^{3} n_{ij}\bar{y}_{ij}^2 \quad (15)$$

$$= 2(8^2) + 6^2 + 2^2 + 8^2 + 2(6^2) + 12^2 = 448.$$

Then, using the algorithm (Searle, 1971, p. 170)

R = Σ(each solution element
x corresponding element of (16)
r.h.s. of normal equations),

we check the value of $R(\mu,\alpha,\beta,\gamma)$ from (13) and (14) as

$$\begin{aligned} R(\mu,\alpha,\beta,\gamma) &= 7(56) + (-10/6)(-8) + 1(10) \\ &\quad + (-1)4 + (10/6)(18 + 4) \quad (17) \\ &= 448, \end{aligned}$$

which agrees with (15).

To illustrate (10) we first have, from (11) and (15),

$$R^*(\dot{\alpha}|\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma = 448 - R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma. \quad (18)$$

Calculation of $R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma$ for this comes from deleting $\dot{\alpha}$'s and the $\dot{\alpha}$-equations from (13) to get

$$\begin{bmatrix} 8 & 1 & 1 & 1 & -1 \\ 1 & 5 & 2 & 1 & 0 \\ 1 & 2 & 5 & 0 & -1 \\ 1 & 1 & 0 & 5 & 2 \\ -1 & 0 & -1 & 2 & 5 \end{bmatrix} \begin{bmatrix} \bar{\dot{\mu}} \\ \bar{\dot{\beta}}_1 \\ \bar{\dot{\beta}}_2 \\ \bar{\dot{\gamma}}_{11} \\ \bar{\dot{\gamma}}_{12} \end{bmatrix} = \begin{bmatrix} 56 \\ 10 \\ 4 \\ 18 \\ 4 \end{bmatrix}, \quad (19)$$

$$\text{with solution} \quad \begin{bmatrix} 7 \\ \frac{1}{2} \\ -\frac{1}{2} \\ 1\frac{1}{2} \\ 1\frac{1}{2} \end{bmatrix}.$$

Applying (16) to this gives

$$\begin{aligned} R^*(\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma &= 7(56) + \tfrac{1}{2}(10) + (-\tfrac{1}{2})4 \\ &\quad + 1\tfrac{1}{2}(18) + 1\tfrac{1}{2}(4) \quad (20) \\ &= 428 \end{aligned}$$

so that in (18)

$$R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma = 448 - 428 = 20 \\ = \text{SAS GLM Type III s/s for A.} \tag{21}$$

This is not the same as $R(\alpha|\mu,\beta,\gamma)$. Neither it should be. $R(\alpha|\mu,\beta,\gamma)$ is for the unrestricted model and is identically zero.

$R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma$ is for the Σ-restricted model. It is the type III output from SAS GLM. When all cells are filled, it is also the output of BMDP2V and, providing at least one row and column of the data have all cells filled, it is also the output from SAS HARVEY and SPSS ANOVA option 9.

When all cells are filled, $R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma$ is the numerator sum of squares for the F-statistic that tests $H: \alpha_i + \sum_j \gamma_{ij}/b$ all equal. When some cells are empty, the hypothesis tested is messy. Examples are given in Searle et al. (1981).

3. WEIGHTED SQUARES OF MEANS ANALYSIS

The weighted squares of means analysis is an analysis that is available only for data wherein all cells are filled. The sum of squares for rows in this analysis (Searle, 1971, p. 370) is

$$SSA_w = \sum_{i=1}^{a} w_i(\bar{x}_{i.} - \bar{x}_{[1]})^2 \tag{22}$$

where

$$1/w_i = \sum_{j=1}^{b} (1/n_{ij})/b^2, \quad x_{ij} = \bar{y}_{ij.},$$

$$\bar{x}_{i.} = \sum_{j=1}^{b} x_{ij}/b \quad \text{and} \quad \bar{x}_{[1]} = \sum_{i=1}^{a} w_i\bar{x}_{i.} \Big/ \sum_{i=1}^{a} w_i .$$

For the example, $w_1 = w_2 = 18/5$, so that $\bar{x}_{[1]} = \frac{1}{2}(16/3 + 26/3) = 7$ and so

$$SSA_w = (18/5)[(16/3 - 7)^2 + (26/3 - 7)^2] = 20. \tag{23}$$

Comparing (23) with (21) suggests that

$$R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma = SSA_w \tag{24}$$

which, for all cells filled, is indeed the case, as proven in Searle et al. (1981, Appendix B). This is why the hypothesis corresponding to $R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma$ is $H: \alpha_i + \bar{\gamma}_{i.}$ all equal. That is also the hypothesis when using SSA_w as the numerator sum of squares of an F-statistic.

4. "INDIRECT" CALCULATION OF SUMS OF SQUARES

SAS HARVEY calculates certain sums of squares by what is sometimes called an "indirect" calculation procedure or, more descriptively, the "invert-part-of-the-inverse" procedure. It is attributable to Henderson (1959). For a full rank model $E(\mathbf{y}) = \mathbf{Xb}$, in which the solution to the normal equations is

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}, \tag{25}$$

the procedure is as follows. Suppose the model is partitioned as

$$E(\mathbf{y}) = \mathbf{X}_1\mathbf{b}_1 + \mathbf{X}_2\mathbf{b}_2 . \tag{26}$$

Then

$$\begin{bmatrix} \hat{\mathbf{b}}_1 \\ \hat{\mathbf{b}}_2 \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1'\mathbf{X}_1 & \mathbf{X}_1'\mathbf{X}_2 \\ \mathbf{X}_2'\mathbf{X}_1 & \mathbf{X}_2'\mathbf{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{T}_{11} & \mathbf{T}_{12} \\ \mathbf{T}_{21} & \mathbf{T}_{22} \end{bmatrix} \begin{bmatrix} \mathbf{X}_1'\mathbf{y} \\ \mathbf{X}_2'\mathbf{y} \end{bmatrix} \tag{27}$$

so defining the matrix of T_{ij}'s as the inverse of the matrix of $\mathbf{X}_i'\mathbf{X}_j$'s. Then, although the definition of $R(\mathbf{b}_1|\mathbf{b}_2)$ is

$$R(\mathbf{b}_1|\mathbf{b}_2) = R(\mathbf{b}_1,\mathbf{b}_2) - R(\mathbf{b}_2), \tag{28}$$

a calculation formula using terms in (27) is

$$R(\mathbf{b}_1|\mathbf{b}_2) = \hat{\mathbf{b}}_1'\mathbf{T}_{11}^{-1}\hat{\mathbf{b}}_1 . \tag{29}$$

Derivation of (29) is shown in Searle (1971, p. 115) and a new extension of it to models not of full rank is given at equation (38) of Searle et al. (1981).

As illustration of (29) we calculate $R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma$ from (14):

$$R^*(\dot\alpha|\dot\mu,\dot\beta,\dot\gamma)_\Sigma = (-10/6)(10/72)^{-1}(-10/6) = 20. \tag{30}$$

5. SUMS OF SQUARES FOR THE MEAN

$R(\mu) = N\bar{y}^2$ is generally described as the correction for the mean, or as the sum of squares due to the mean. It is often interpreted (wrongly) as being a sum of squares for testing $H: \mu = 0$, whereas its correct interpretation (see Searle, 1971, pages 104 and 178) is that of testing $H: E(\bar{y}) = 0$. Moreover, although $N\bar{y}^2$ is the most usual value calculated as a sum of squares for the mean, some computing procedures do calculate other values; e.g., SAS HARVEY uses $R^*(\dot\mu|\dot\alpha,\dot\beta,\dot\gamma)_\Sigma$.

For the data of Table 2,

$$R(\mu) = N\bar{y}^2 = 8(7^2) = 392 .$$

And applying the "invert-part-of-the-inverse" calculation of (29) to (13) and (14) gives

$$R^*(\dot\mu|\dot\alpha,\dot\beta,\dot\gamma)_\Sigma = 7(10/72)^{-1}7 = 352.8 .$$

This is also the calculation in BMDP2V.

The hypotheses tested by using these sums of squares as numerators of F-statistics are,

for $R(\mu)$:

$$H: \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \frac{1}{8}(3\beta_1 + 3\beta_2 + 2\beta_3)$$
$$+ \frac{1}{8}(2\gamma_{11} + \gamma_{12} + \gamma_{13} + \gamma_{21} + 2\gamma_{22} + \gamma_{23}) = 0$$

and

for $R^{\#}(\dot{\mu}|\dot{\alpha}, \dot{\beta}, \dot{\gamma})_{\Sigma}$:

$$H: \dot{\mu} = 0; \text{ i.e.,}$$

$$H: \mu + \frac{1}{2}(\alpha_1 + \alpha_2) + \frac{1}{3}(\beta_1 + \beta_2 + \beta_3)$$
$$+ \frac{1}{6}(\gamma_{11} + \gamma_{12} + \gamma_{13} + \gamma_{21} + \gamma_{22} + \gamma_{23}) = 0.$$

Each of these hypotheses is less straightforward when there are empty cells in the data.

Of course, in other models the calculation based on full rank reparameterization using Σ-restrictions can give further different values. Thus, for the no-interaction, Σ-restricted model, the normal equations are equations (13) with the $\dot{\gamma}$'s and $\dot{\gamma}$-equations deleted, namely

$$\begin{bmatrix} 8 & 0 & 1 & 1 \\ 0 & 8 & 1 & -1 \\ 1 & 1 & 5 & 2 \\ 1 & -1 & 2 & 5 \end{bmatrix} \begin{bmatrix} \tilde{\mu} \\ \tilde{\alpha}_1 \\ \tilde{\beta}_1 \\ \tilde{\beta}_2 \end{bmatrix} = \begin{bmatrix} 56 \\ -8 \\ 10 \\ 4 \end{bmatrix} \quad \text{with solution}$$

$$\frac{1}{22(27)} \begin{bmatrix} 77 & 0 & -11 & -11 \\ 0 & 81 & -27 & 27 \\ -11 & -27 & 152 & -64 \\ -11 & 27 & -64 & 152 \end{bmatrix} \begin{bmatrix} 56 \\ -8 \\ 10 \\ 4 \end{bmatrix} = \frac{1}{11} \begin{bmatrix} 77 \\ -15 \\ 16 \\ -16 \end{bmatrix}.$$

Hence, on applying (29)

$$\begin{aligned} R^{\#}(\dot{\mu}|\dot{\alpha}, \dot{\beta})_{\Sigma} &= 7[77/22(27)]^{-1}7 \\ &= 378 \neq 352.8 = R^{\#}(\dot{\mu}|\dot{\alpha}, \dot{\beta}, \dot{\gamma})_{\Sigma} \\ &\neq 392 \quad = R(\mu) = N\bar{y}^2 . \end{aligned}$$

Thus, although $N\bar{y}^2$ is the same for every model, the $R^{\#}(\dot{\mu}|\cdot)_{\Sigma}$ calculation can yield different values when using different models on the same data.

6. POPULATION AND ESTIMATED MARGINAL MEANS

The term "least squares mean" has been in the literature for at least twenty years, but it has never been carefully defined. As it stands, the expression "least squares mean" has no implicit meaning because "least squares" is an estimation procedure and those two words do not function informatively as an adjective to "mean", be it a population mean or a sample average. SAS GLM output perpetuates this confusion, using the term inconsistently for both a function of parameters and for an estimate. To clarify matters, Searle et al. (1980) suggest new terms: population marginal mean (PMM), the parameter function, and estimated marginal mean (EMM), the estimate thereof.

Suppose in the 2-way classification that μ_{ij} is the population mean of the cell defined by the i'th row and j'th column. For the

$$\text{no interaction model: } \mu_{ij} = \mu + \alpha_i + \beta_j \qquad (31)$$

and for the

$$\text{with interaction model: } \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}. \qquad (32)$$

Then the population marginal mean (PMM) corresponding to the i'th row is defined as

$$\text{PMM}(\alpha_i) = \sum_{j=1}^{b} \mu_{ij}/b = \bar{\mu}_{i.} \,. \qquad (33)$$

This embodies the main idea of the undefined term "least squares mean", that it is a population marginal mean; but, in contrast, PMM(α_i) is clearly defined with the following characteristics.

(i) PMM(α_i) is a function of population parameters;

(ii) PMM(α_i) does not involve the n_{ij}'s of the data, and is not in any way contingent upon empty cells in the data;

(iii) There is one PMM(α_i) for every row in the data, and correspondingly one PMM(β_j) for every column, and one PMM(γ_{ij}) for every cell.

Having defined a PMM, we can then consider its estimability:

(a) Because PMM(α_i) is a function of parameters, it is estimable when that function is estimable. Hence a PMM is estimable only when every μ_{ij} in the PMM is estimable; a PMM is not estimable, otherwise.

(b) When a PMM is estimable, its best linear unbiased estimator (b.l.u.e.) is the same function of the b.l.u.e.'s of the μ_{ij}'s as the

PMM is of the μ_{ij}'s themselves. For example, if

$$\text{PMM}(\alpha_i) = \sum_{j=1}^{b} \mu_{ij}/b \quad \text{is estimable}$$

then

$$\text{EMM}(\alpha_i) = \sum_{j=1}^{b} \hat{\mu}_{ij}/b = \text{b.l.u.e. of PMM}(\alpha_i) \quad (34)$$

where EMM is the acronym for estimated marginal mean and $\hat{\mu}_{ij}$ is the b.l.u.e. of μ_{ij}. Thus EMM(α_i) exists only when $\hat{\mu}_{ij}$ exists for j = 1, ..., b. This is true in general: an EMM, being the b.l.u.e. of a PMM, exists only when $\hat{\mu}_{ij}$ exists for every μ_{ij} in the PMM. Hence estimability of PMM's and existence of EMM's depends upon estimability of μ_{ij}'s.

Three cases must be distinguished in the 2-way crossed classification:

1. Without interaction model, where every μ_{ij} is estimable, with $\hat{\mu}_{ij} = \mu^o + \alpha_i^o + \beta_j^o$, for μ^o, α_i^o and β_j^o being solutions to the normal equations (Searle, 1971, Section 7.4). For every row and column the corresponding EMM(α_i) and EMM(β_j), respectively, exist.

2. With interaction model, for all cells filled, where every μ_{ij} is estimable ($\hat{\mu}_{ij} = \bar{y}_{ij.}$) and so every EMM exists.

3. With interaction model, and some cells empty, where $\hat{\mu}_{ij} = \bar{y}_{ij.}$ only for the cells that contain data. EMM(α_i) and EMM(β_j) exist only for rows and columns, respectively, that have no empty cells.

Example. Suppose data are available corresponding to Grid 3.

Grid 3

n_{11}	n_{12}
n_{21}	-

.

From (33)

$$\text{PMM}(\alpha_1) = \tfrac{1}{2}(\mu_{11} + \mu_{12})$$

and $$(35)$$

$$\text{PMM}(\alpha_2) = \tfrac{1}{2}(\mu_{21} + \mu_{22}) \; .$$

In the no-interaction model, it will be found (Searle, loc. cit.) that solutions to the normal equations are

$$\mu^o = 0 \quad \alpha_1^o = \bar{y}_{12.} \qquad \beta_1^o = \bar{y}_{11.} - \bar{y}_{12.}$$

$$\alpha_2^o = \bar{y}_{12.} + \bar{y}_{21.} - \bar{y}_{11.} \qquad \beta_2^o = 0 \; .$$

Hence

$$\hat{\mu}_{11} = \bar{y}_{11.}, \qquad \hat{\mu}_{12} = \bar{y}_{12.},$$

$$\hat{\mu}_{21} = \bar{y}_{21.} \quad \text{and} \quad \hat{\mu}_{22} = \bar{y}_{12.} + \bar{y}_{21.} - \bar{y}_{11.}$$

and so

$$\text{EMM}(\alpha_1) = \tfrac{1}{2}(\bar{y}_{11.} + \bar{y}_{12.})$$

and $$(36)$$

$$\text{EMM}(\alpha_2) = \bar{y}_{21.} - \tfrac{1}{2}(\bar{y}_{11.} - \bar{y}_{12.}) \; .$$

In the interaction model

$$\hat{\mu}_{11} = \bar{y}_{11.}, \qquad \hat{\mu}_{12} = \bar{y}_{12.},$$

$$\hat{\mu}_{21} = y_{21.} \quad \text{and} \quad \hat{\mu}_{22} \text{ does not exist.}$$

Therefore

$$\text{EMM}(\alpha_1) = \tfrac{1}{2}(\bar{y}_{11.} + \bar{y}_{12.})$$

and $$(37)$$

$$\text{EMM}(\alpha_2) \quad \text{does not exist.}$$

SAS GLM gives results (36) and (37). In contrast, SAS HARVEY does not indicate the nonexistence of EMM(α_2) in (37). Instead, it gives the EMM(α_2) of (36). Details of this example, and of other examples involving a nested factor, a mixed model, a 3-way classification, and covariance are shown in Searle et al. (1980).

7. VARIANCE COMPONENTS ESTIMATION

Annotated Computer Output for variance components (ACO σ^2) are now available for four routines: SAS VARCOMP, SAS HARVEY, SAS RANDOM, and BMDP-V. Some salient features of output from these routines are as follows.

1. SAS VARCOMP calculates three different kinds of estimates.

(a) Henderson Method 3, using the sub-method based on the order in which factors are presented, e.g., using $R(\alpha|\mu)$, $R(\beta|\mu,\alpha)$ and $R(\gamma|\mu,\alpha,\beta)$.

(b) MINQUE(0), being the minimum norm quadratic unbiased estimators (MINQUE), using prior values of zero for all variance components except the error component. Although computable for large data sets, there is evidence that the resulting estimates have much larger sampling variances than do other estimates (Quaas and Bolgiano, 1977).

(c) ML, maximum likelihood, which always yields non-negative estimates but which sometimes changes the model in so doing.

2. SAS HARVEY

(a) cannot handle interaction models.

(b) uses Henderson's Method 3, the sub-method based on each factor after all others, e.g., $R(\alpha|\mu,\beta)$ and $R(\beta|\mu,\alpha)$.

(c) uses Σ-restrictions both for calculating sums of squares and, more importantly, for taking expected values of those sums of squares.

3. SAS RANDOM calculates, for random and/or mixed models as specified by the user, expected mean squares for all Types I, II, III and IV sums of squares as used in SAS GLM. Although some of the Type III sums of squares are the same as in SAS HARVEY, based as they are on Σ-restrictions, their expectations are not always the same. Expected values in SAS HARVEY are based on Σ-restricted models whereas those of SAS RANDOM are not.

(SAS NESTED is available only for models that are both completely nested and completely random.)

4. In the BMDP-V routines

(a) P2V can do calculations for repeated measures experiments, which are nothing more than mixed models.

(b) P3V calculates REML (restricted maximum likelihood) and ML estimates and also does what none of the other routines do: it estimates the fixed effects of a mixed model and it gives estimated sampling variances and covariances of estimated components.

(c) P8V calculates ANOVA (analysis of variance) estimates for balanced data, using Σ-restricted models. REML estimates from P3V (for balanced data) equal ANOVA estimates using unrestricted models, and so they are not the same as the ANOVA estimates from P8V.

REFERENCES

Henderson, C. R. (1959). Design and analysis of animal husbandry experiments. Chapter 1 of Techniques and Procedures in Animal Science Research, 1st Ed., American Society of Animal Science.

Quaas, R. L. and Bolgiano, D. C. (1977). Sampling variances of the MIVQUE and Method 3 estimators of the sire component of variance. Variance Components and Animal Breeding, Eds. L. D. Van Vleck and S. R. Searle, Animal Science Department, Cornell University, Ithaca, New York, 99-106.

Searle, S. R. (1971). Linear Models. Wiley and Sons, New York.

Searle, S. R. (1979). Relationships between the estimable functions of SAS GLM output for unbalanced data and hypotheses tested by traditional F-statistics. Proceedings, SUGI 4th Annual Conference, SAS Institute, Raleigh, North Carolina, 196-207.

Searle, S. R. and Grimes, B. A. (1980). Annotated computer output for variance components (ACO σ^2): SAS VARCOMP. Paper No. BU-708-M in the Biometrics Unit, Cornell University, Ithaca, New York.

Searle, S. R. and Henderson, H. V. (1978). Annotated computer output for analyses of unbalanced data: ACO SAS GLM. Paper No. BU-641-M in the Biometrics Unit, Cornell University, Ithaca, New York.

Searle, S. R., Speed, F. M. and Henderson, H. V. (1981). Computational and model equivalences in analyses of variance of unequal-subclass-numbers data. The American Statistician, 35, No. 1 (in press).

Searle, S. R., Speed, F. M. and Milliken, G. A. (1980). Population marginal means in the linear model: an alternative to least squares means. The American Statistician, 34, 216-221.