HYPOTHESIS TESTING WITH TYPE IV SUMS OF SQUARES OF THE COMPUTER ROUTINE SAS GLM

by

G. F. S. Hudson and S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

BU-731-M January, 1981

Abstract

Suggestions are made as to the manner in which the Type IV estimable functions are determined by SAS GLM, and how this is affected by the sequencing of levels of the factors of the model.

Paper No. BU-731-M in the Biometrics Unit.

HYPOTHESIS TESTING WITH TYPE IV SUMS OF SQUARES OF THE COMPUTER ROUTINE SAS GLM

by

G. F. S. Hudson and S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

BU-731-M January, 1981

1. Introduction

Testing hypotheses in the general linear model involves estimable functions of parameters represented by $\underset{\sim}{b}$ in the equation $\underset{\sim}{y} = \underset{\sim\sim}{X}\underset{\sim}{b} + \underset{\sim}{e}$. For unbalanced data, the General Linear Model (GLM) procedure of the computing package Statistical Analysis System (SAS) produces four different types of estimable functions and the sums of squares associated with each. Searle (1980) details derivation of these estimable functions for Types I, II and III and suggests the concept involved in the derivation of Type IV sums of squares. Whereas Types I - III estimable functions are used for explaining pre-ordained [as Searle (1980) calls them] sums of squares, the Type IV sums of squares correspond to hypotheses specifically picked out from non-unique subsets of the filled subclasses in the data. We illustrate these hypotheses using some small sets of hypothetical data and show that, even with a given set of data, Type IV estimable functions and corresponding sums of squares are influenced by the arrangement of the data; and we suggest the manner in which that influence operates. Our illustrations are based on the 2-way, cross-classified, with-interaction model outlined in the following section. One Type IV sum of squares for rows for a particular data set is shown in Section 3, and the general non-uniqueness of Type IV sums of squares is illustrated in Section 4 by resequencing the rows. In Section 5 we do the same for columns.

A larger example with more empty cells is used in Section 6, and the paper ends with attempts at drawing general conclusions about the kinds of hypotheses tested by Type IV sums of squares.

2. The Two-way Cross-classified Model with Interaction

A suitable model equation for the two-way cross-classified model with interaction is

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

where μ is the overall mean, α_i for $i = 1, 2, \cdots, a$ is the effect due to the i'th level of the A-factor, β_j for $j = 1, 2, \cdots, b$ is the effect due to the j'th level of the B-factor, γ_{ij} is an interaction effect and e_{ijk}, the error term associated with the ijk'th observation, is assumed normally distributed, with zero mean variance σ_e^2 and with all such error terms uncorrelated. Then y_{ijk} is the k'th observation in the ij'th subclass ($k = 1, 2, \cdots, n_{ij}$) with expected value $E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$ and variance $\text{var}(y_{ijk}) = \sigma_e^2$.

Our illustrative data are shown in Table 1. As indicated there, rows represent levels of the A-factor and columns represent levels of the B-factor.

(SHOW TABLE 1)

These are the same data as used in Section 7.2 of Searle (1971) and as Data Set 5 in the Annotated Computer Output for SAS GLM, as in Searle and Henderson (1979).

3. A Type IV Sum of Squares for Rows

The SAS GLM output represents estimable functions by L-values alongside each parameter of the model. The Type IV estimable function labeled A in the output from the data of Table 1 is

$$f = L_2\alpha_1 + L_3\alpha_2 - (L_2 + L_3)\alpha_3 + \tfrac{1}{2}L_2(\gamma_{13} + \gamma_{14}) + L_3\gamma_{22} - L_3\gamma_{32} - \tfrac{1}{2}L_2(\gamma_{33} + \gamma_{34}) \qquad (2)$$

This is not a true contrast among the α's, because some of the γ's are involved. It is, in fact, a contrast among the α's plus a "mess" of other parameters, in

this case γ's. Searle (1980) names this an α-based contrast, and we use that name here. In any data set we have analyzed, using interaction models, Type IV estimable functions for rows are all α-based contrasts that involve γ's but no β's (with analogous results for columns).

There are two distinct L's in (2). Any two linearly independent pairs of numerical values may be assigned to these two L's, to generate two specific forms of the estimable function, f_1 and f_2, say. Then the numerator sum of squares calculated by SAS GLM for the F-statistic is for testing the hypothesis that the two LIN forms f_1 and f_2 equal zero. For example, setting $L_2 = 1$ and $L_3 = 0$, followed by $L_2 = 0$ and $L_3 = 1$ gives the hypothesis

$$H: \begin{Bmatrix} \alpha_1 - \alpha_3 + \frac{1}{2}(\gamma_{13} + \gamma_{14}) - \frac{1}{2}(\gamma_{33} + \gamma_{34}) \\ \alpha_2 - \alpha_3 + \gamma_{22} - \gamma_{32} \end{Bmatrix} = 0 . \qquad (3)$$

The specific pairs of values used for the L's, and the resulting expression of the hypothesis, does not affect calculation of the F-statistic corresponding to (2). Whatever form different from (3) is obtained will just be linear combinations of the equations in (3). For example, suppose we used $L_2 = 2$, $L_3 = 1$ and $L_2 = 3$, $L_3 = 2$. This gives, from (2),

$$H^*: \begin{Bmatrix} 2\alpha_1 + \alpha_2 - 3\alpha_3 + \gamma_{13} + \gamma_{14} + \gamma_{22} - \gamma_{32} - (\gamma_{33} + \gamma_{34}) \\ 3\alpha_1 + 2\alpha_2 - 5\alpha_3 + 1\frac{1}{2}(\gamma_{13} + \gamma_{14}) + 2(\gamma_{22} - \gamma_{32}) - 1\frac{1}{2}(\gamma_{33} + \gamma_{34}) \end{Bmatrix} = 0 \qquad (4)$$

which, in terms of the functions in (3) can be written as

$$H^*: \begin{Bmatrix} 2[\alpha_1 - \alpha_3 + \frac{1}{2}(\gamma_{13} + \gamma_{14}) - \frac{1}{2}(\gamma_{33} + \gamma_{34})] + [\alpha_2 - \alpha_3 + \gamma_{22} - \gamma_{32}] \\ 3[\alpha_1 - \alpha_3 + \frac{1}{2}(\gamma_{13} + \gamma_{14}) - \frac{1}{2}(\gamma_{33} + \gamma_{34})] + 2[\alpha_2 - \alpha_3 + \gamma_{22} - \gamma_{32}] \end{Bmatrix} = 0 .$$

Therefore H and H^* are equivalent and have the same F-statistic. The general statement of this result is given in the Appendix.

Hypotheses in the 2-way classification with interaction are easily handled and understood if written in terms of cell means

$$\mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

for the cells containing data. Then

$$E(\underset{\sim}{y}) = \underset{\sim}{X}\underset{\sim}{\mu} \quad \text{for} \quad \underset{\sim}{\mu} = \{\mu_{ij}\} .$$

This is the cell means model encouraged, for example, by Searle, Speed and Henderson (1981). It has the simple properties:

$$\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{D} = \text{diag}\{n_{ij}\}$$

$$\underset{\sim}{G} = \underset{\sim}{D}^{-1} = \text{diag}\{1/n_{ij}\}$$

and

$$\underset{\sim}{\mu}^{o} = \underset{\sim}{\bar{y}} = \{\bar{y}_{ij.}\} ,$$

the vector of observed cell means. Any hypothesis $H : \underset{\sim}{K}'\underset{\sim}{b} = \underset{\sim}{0}$ is equivalent to $H : \underset{\sim}{M}'\underset{\sim}{\mu} = 0$, where rows of $\underset{\sim}{M}'$ are proportional (often equal) to the coefficients of γ's in the rows of $\underset{\sim}{K}'$. This is shown by Searle (1971, Section 7.2f) who uses the symbol $\underset{\sim}{L}$ where we here use $\underset{\sim}{M}$ to avoid confusion with other uses for $\underset{\sim}{L}$ (see Appendix). Then the F-statistic for testing

$$H : \underset{\sim}{M}'\underset{\sim}{\mu} = \underset{\sim}{0} \tag{5}$$

is

$$F = Q/r\hat{\sigma}^2 \tag{6}$$

with

$$Q = \underset{\sim}{\bar{y}}'\underset{\sim}{M}(\underset{\sim}{M}'\underset{\sim}{D}^{-1}\underset{\sim}{M})^{-1}\underset{\sim}{M}'\underset{\sim}{\bar{y}} \tag{7}$$

and

$$r = \text{rank of } \underset{\sim}{M}', \text{ of full row rank.}$$

Example. For the data of Table 1

$$\mu = [\mu_{11} \quad \mu_{13} \quad \mu_{14} \quad \mu_{21} \quad \mu_{22} \quad \mu_{32} \quad \mu_{33} \quad \mu_{34}] \tag{8}$$

$$\underset{\sim}{X}'\underset{\sim}{X} = \underset{\sim}{D} = \text{diag}\{ 3 \quad 1 \quad 2 \quad 2 \quad 2 \quad 2 \quad 2 \quad 4 \} \tag{9}$$

$$\underset{\sim}{G} = \underset{\sim}{D}^{-1} = \text{diag}\{ \tfrac{1}{3} \quad 1 \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{4} \} \tag{10}$$

and

$$\underset{\sim}{\mu}^{o\prime} = \underset{\sim}{\bar{y}}' = [10 \quad 12 \quad 9 \quad 9 \quad 13 \quad 8 \quad 15 \quad 12 \,]. \tag{11}$$

The hypothesis (3) is

$$H: \begin{cases} \mu_{13} + \mu_{14} - \mu_{33} - \mu_{34} = 0 \\ \mu_{22} - \mu_{32} = 0 \end{cases}. \tag{12}$$

It can also be stated equivalently as

$$H: \begin{cases} \mu_{13} + \mu_{14} = \mu_{33} + \mu_{34} \\ \mu_{22} = \mu_{32} \end{cases}. \tag{13}$$

In the form of (12), $\underset{\sim}{M}'$ for (5) based on (8) is

$$\underset{\sim}{M}' = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 & -1 & -1 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix} \tag{14}$$

and so, using (10) and (11), Q of (7) is

$$Q = [-6 \quad 5] \begin{bmatrix} 2\frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} -6 \\ 5 \end{bmatrix} = 6^2(4/9) + 5^2 = 41 \,. \tag{15}$$

This is the SAS GLM Type IV sum of squares obtained by processing the data of Table 1, as shown in Searle and Henderson (1979, p. 54).

The hypothesis in (12) is shown schematically in Table 2,

(SHOW TABLE 2)

with + and - signs in the cells of the data grid. The + and - signs correspond to μ_{ij}'s on the left and right, respectively, of equalities in the statement of a hypothesis such as (13). Those in the upper left corner of their cells refer to the first equality listed in such a hypothesis statement, and encircled symbols in the upper right corner refer to the second equality listed. Cells containing data but not involved in the hypothesis are indicated by check marks (✓). This schematic representation of a hypothesis is useful in understanding how a hypothesis is constituted, and in clarifying the differences among the Type IV contrasts SAS GLM can produce with the same data, as demonstrated in the following sections. We use this representation repeatedly.

4. Resequencing Rows

As indicated on SAS GLM output, Type IV estimable functions are not unique for a given data set. Various hypotheses can be tested, each of which yields its own sum of squares. For example, having the rows in different sequences yields Type IV estimable functions and sums of squares different from those just shown. This is not the same as assigning different numerical values to the L's in a Type IV estimable function such as f of (2). Any pair of f's thus created involves the same model parameters (in the cell means model, the same μ_{ij}'s) as every other pair of f's. In contrast, different Type IV estimable functions due to relabeling rows, for example, involve different combinations of model parameters (or μ_{ij}'s). This we now illustrate for the example of Table 1.

A relabeling of rows that affects Type IV hypotheses and sums of squares is shown in Table 3. Its first section contains Table 2, as a basis for comparison.

(SHOW TABLE 3)

The other three sections show the rows that were interchanged to achieve re-sequencing, the resultant hypothesis and the numerator sum of squares of the corresponding F-statistic. The hypotheses are shown schematically in the manner of Table 2, but with the rows interchanged and showing new row numbers on the left of each diagram and old row numbers on the right. Also introduced is notation for the cell means after relabeling the rows. For example, in the second section of Table 3, rows 1 and 2 have been interchanged, so that cell means μ_{1j} before the interchange is now denoted as μ'_{2j}, and what was μ_{2j} is now μ'_{1j}. Each hypothesis is shown in terms of both μ_{ij}'s and μ'_{ij}'s. A second and third prime is used in the last two sections of Table 3.

Scrutiny of Table 3 reveals that in each section, the last row of the data is salient to the specification of the hypothesis being tested. This is evident from the last row having no check marks, the presence of which would indicate cells having data that are not used in the hypothesis. There are such cells in the data, but in Table 3, no matter what particular row is used as the last row, every cell in that row is used to specify the SAS Type IV hypothesis. This is the reason that SAS GLM produces different Type IV sums of squares, for rows, for different row sequences of the same data. In every case, SAS GLM compares cell means in the last row with cell means in each and all of the other rows. If interactions were ignored, the comparisons would be true contrasts and the hypothesis being tested would, in each case, be $\alpha_1 = \alpha_2 = \alpha_3$.

As indicated following (2), the hypothesis corresponding to a Type IV sum of squares for rows is based on what is being called an α-based contrast. It is a contrast among α's plus a "mess" of γ's; and since these α-based contrasts all involve the last row, the cells of that row which contain data determine the γ's that are involved in the hypothesis. This in turn determines the exact nature of the hypothesis and the corresponding sum of squares. When used as the last row, rows with different patterns of filled cells can therefore lead to different hypotheses and different sums of squares. Conversely, as in the second section of Table 3, when interchanging rows does not involve the last row, the Type IV hypothesis and sum of squares is not affected. But when the last row gets changed and the data are unbalanced, the Type IV α-based contrast and corresponding sum of squares depends specifically on the pattern of filled cells in the last row. For example, in the third section of Table 3, wherein rows 1 and 3 have been interchanged, the estimable function for rows generated by SAS GLM is

$$f = L_2\alpha''_1 + L_3\alpha''_2 - (L_2 + L_3)\alpha''_3 + \tfrac{1}{2}L_2(\gamma''_{13} + \gamma''_{14}) + L_3\gamma''_{21} - L_3\gamma''_{31} - \tfrac{1}{2}L_2(\gamma''_{33} + \gamma''_{34}) \ . \quad (16)$$

Using $L_2 = 1$, $L_3 = 0$ and $L_2 = 0$, $L_3 = 1$, the hypothesis corresponding to (16) is

$$H: \begin{Bmatrix} \alpha''_1 - \alpha''_3 + \tfrac{1}{2}(\gamma''_{13} + \gamma''_{14}) - \tfrac{1}{2}(\gamma''_{33} + \gamma''_{34}) \\ \alpha''_2 - \alpha''_3 + \gamma''_{21} - \gamma''_{31} \end{Bmatrix} = 0 \ ,$$

or, equivalently,

$$H: \begin{cases} \mu''_{13} + \mu''_{14} = \mu''_{33} + \mu''_{34} \\ \mu''_{21} = \mu''_{31} \end{cases} \quad \text{or} \quad H: \begin{cases} \mu_{33} + \mu_{34} = \mu_{13} + \mu_{14} \\ \mu_{21} = \mu_{11} \end{cases} \quad (17)$$

By formulating the latter statement in (17) as $\underset{\sim}{M}'\underset{\sim}{\mu} = \underset{\sim}{0}$, we have for (7)

$$\underset{\sim}{M}' = \begin{bmatrix} 0 & -1 & -1 & 0 & 0 & 0 & 1 & 1 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Hence, using this and (10) and (11) in (7) we have

$$Q = [-6 \quad 1] \begin{bmatrix} 2\frac{1}{4} & 0 \\ 0 & 5/6 \end{bmatrix}^{-1} \begin{bmatrix} -6 \\ 1 \end{bmatrix} = 36/2\tfrac{1}{4} + 6/5 = 17\tfrac{1}{5},$$

as shown in Table 3. This is the value calculated by SAS GLM when rows 1 and 3 of the data in Table 1 are interchanged.

Derivation of the hypothesis and of $Q = 26\frac{1}{5}$ in the last section of Table 3 (interchanging rows 2 and 3 of Table 1), follows precisely the same principles as those just described for interchanging rows 1 and 3, of the third section. But there is one very noticeable consequence here: there are four cells containing data that are not involved in the hypothesis, and by their location in the resequenced data grid, as cells 1,3 and 1,4, and 2,3 and 2,4, one would in practice find them quite useful for comparing rows 1 and 2 (in the presence of averaged interactions). Yet SAS Type IV calculations do not do this.

5. Resequencing Columns

All that has been said about resequencing rows applies in precisely the same fashion to resequencing columns – as one would expect, because the data grid of Table 1 can easily be redefined with columns as rows and vice versa. The results of two resequencings of columns, through interchanging columns 1 and 4 and 2 and 4, are shown in Table 4, set out in essentially the same manner as is Table 3. Each statement of a hypothesis involves three equations, and in the schematic repre-

(SHOW TABLE 4)

sentation of each hypothesis, using + and - signs in the cells of the data grid, the third equation of each hypothesis is schematically represented by the boxed + and - signs in the lower right corner of appropriate cells. The hypotheses corresponding to the resequencings of columns are again expressed in terms of μ'_{ij}'s and μ''_{ij}'s, for which it is important to note the μ'_{ij}'s and μ''_{ij}'s in Table 4 are not the same as those of Table 3.

Just as the pattern of filled cells in the last row is a determining factor for the nature of a Type IV hypothesis for rows, so also for columns: the Type IV hypothesis is based on comparing the last column with each of the others. In general, this comparison uses all the filled cells of the last column (and as many other filled cells as possible), except that filled cells in the last column are not used when they are the only filled cells in their rows (see Section 5). If interactions are ignored, each hypothesis reduces to $H: \beta_1 = \beta_2 = \beta_3 = \beta_4$. When interchanging columns leads to the last column being changed, the Type IV estimable function and sum of squares for columns change also.

Example. The Type IV estimable function for columns, for data of Table 1, is

$$\begin{aligned} f = {} & L_5\beta_1 + L_6\beta_2 + L_7\beta_3 - (L_5 + L_6 + L_7)\beta_4 + L_5\gamma_{11} + \tfrac{1}{2}L_7\gamma_{13} \\ & - (L_5 + \tfrac{1}{2}L_7)\gamma_{14} + L_6\gamma_{32} + \tfrac{1}{2}L_7\gamma_{33} - (L_6 + \tfrac{1}{2}L_7)\gamma_{34} . \end{aligned} \tag{18}$$

Setting each L-value in (18) equal to one in turn, with the remaining L-values equal zero, the hypothesis is formulated as

$$H: \left\{ \begin{array}{l} \beta_1 - \beta_4 + \gamma_{11} - \gamma_{14} \\ \beta_2 - \beta_4 + \gamma_{32} - \gamma_{34} \\ \beta_3 - \beta_4 + \tfrac{1}{2}(\gamma_{13} + \gamma_{33}) - \tfrac{1}{2}(\gamma_{14} + \gamma_{34}) \end{array} \right\} = 0 , \tag{19}$$

or, expressed in terms of population cell means, as

$$H: \begin{cases} \mu_{11} = \mu_{14} \\ \mu_{32} = \mu_{34} \\ \mu_{13} + \mu_{33} = \mu_{14} + \mu_{34} \end{cases} . \qquad (20)$$

Using (7) to calculate the numerator sum of squares Q corresponding to $H: \underset{\sim}{M}'\underset{\sim}{\mu} = 0$ of (20) we have

$$\underset{\sim}{M}' = \begin{bmatrix} 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \\ 0 & 1 & -1 & 0 & 0 & 0 & 1 & -1 \end{bmatrix} .$$

This, with (10) and (11) leads to

$$Q = [1 \quad -4 \quad 6] \begin{bmatrix} 5/6 & 0 & \frac{1}{2} \\ 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{4} & 2\frac{1}{4} \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ -4 \\ 6 \end{bmatrix} = 46\frac{23}{28} ,$$

as shown in the first section of Table 4, and also in Searle and Henderson (1979, p. 54).

Similarly, when columns 1 and 4 are interchanged, the Type IV estimable function in terms is

$$f = L_5\beta_1' + L_6\beta_2' + L_7\beta_3' - (L_5 + L_6 + L_7)\beta_4' + L_5\gamma_{11}' + L_7\gamma_{13}' - (L_5 + L_7)\gamma_{14}' + L_6\gamma_{22}' - L_6\gamma_{24}' . \qquad (21)$$

Equivalent forms of the hypothesis available from this are

$$H:\begin{cases}\beta'_1 - \beta'_4 + \gamma'_{11} - \gamma'_{14} = 0\\ \beta'_2 - \beta'_4 + \gamma'_{22} - \gamma'_{24} = 0,\\ \beta'_3 - \beta'_4 + \gamma'_{13} - \gamma'_{14} = 0\end{cases} \quad \text{or} \quad H:\begin{cases}\mu'_{11} = \mu'_{14}\\ \mu'_{22} = \mu'_{24},\\ \mu'_{13} = \mu'_{14}\end{cases} \quad \text{or} \quad H:\begin{cases}\mu_{14} = \mu_{11}\\ \mu_{22} = \mu_{21}.\\ \mu_{13} = \mu_{11}\end{cases} \tag{22}$$

For the latter formulation, $\underset{\sim}{M}'$ of (7) is

$$\underset{\sim}{M}' = \begin{bmatrix} -1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix},$$

so that with (10) and (11) we get from (7)

$$Q = [-1 \quad 4 \quad 2] \begin{bmatrix} 5/6 & 0 & 1/3 \\ 0 & 1 & 0 \\ 1/3 & 0 & 4/3 \end{bmatrix}^{-1} \begin{bmatrix} -1 \\ 4 \\ 2 \end{bmatrix} = 22$$

as shown in Table 4.

6. Two Other Examples

In the data of Table 1, each row and column has sufficient filled cells to allow direct comparisons among rows and among columns using relatively "logical" subsets of the data. Conceivably, the pattern of filled cells may be so sparse that the comparisons used are less obvious. This is the case with the Example 2 of Table 5, for which α-based contrasts and β-based contrasts are shown in Table 6.

(SHOW TABLE 5)

As with Example 1, discussed in preceding sections, data in the last row and column are involved in the Type IV α-based and β-based contrasts, respectively.

However, in the last row of Table 5, there is only one cell containing data, cell (4,4). The only direct comparison involving row 4 is therefore that with row 2, because the only other cell in the fourth column is (2,4). The α-based contrasts involving rows 1 or 3 with row 4 must be indirect. Because the only direct Type IV α-based contrast is $\mu_{24} = \mu_{44}$, this must be, and indeed is, incorporated in the remaining, indirect α-based contrasts.

Although there are two filled cells in the last column, only the data in one of these, i.e., cell (2,4), can be used to test β-based contrasts, because cell (4,4) is the only filled cell in its row. Hence, the data in cell (2,4) is crucial to testing both α-based and β-based contrasts. The population cell mean μ_{24} appears in all contrasts in Table 6, and $\bar{y}_{24\cdot}$, the estimate of μ_{24}, will contribute to each element in $\underset{\sim}{M}'\bar{\underset{\sim}{y}}$ of (7).

(SHOW TABLE 6)

Whereas in Example 2, the Type IV procedure relies heavily on one of the filled cells of the data, a simple extension of those data illustrates a case where it omits a logical subset of the filled subclasses. Consider adding data to those of Table 5 as in the footnote thereto: for cell (3,4) have $\bar{y}_{34\cdot} = 7$ and $n_{34} = 2$. Surprisingly, the α-based contrasts shown in the upper part of Table 7 ignore the useful subset of filled cells (1,1), (1,2), (2,1), (2,2). This is because Type IV hypotheses are based first on comparing each row with the last; and in this case this can be done without using those cells. However, this feature of always using whichever row is coded as last can be used, in this example, to coax out a Type IV

(SHOW TABLE 7)

hypothesis based on all the filled cells. Interchanging rows 1 and 4 achieves this, as shown in the lower half of Table 7, wherein the subset of data omitted prior to interchanging rows is involved in the very logical contrast $\mu'_{21} + \mu'_{22} = \mu'_{41} + \mu'_{42}$.

So, although the non-uniqueness of SAS GLM Type IV sums of squares can be a source of confusion, knowledge about how the Type IV estimable functions are derived gives a general understanding of what the Type IV sums of squares represent. The general philosophy is appropriate: comparisons of cell means (corresponding to filled cells in the data) that test hypotheses about row (or column) effects in the presence of interactions. The exact comparisons depends upon the pattern of filled cells and upon which row (and column) in the data is sequenced as the last one. Hence, although the general procedure is appropriate, specifics depend upon data. In practice this is precisely what the experimenter whose data are being analyzed should do — and Homo sapiens usually does a better job of this than computers do.

References

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

Searle, S. R. (1980). Arbitrary hypotheses in linear models with unbalanced data. Communications in Statistics - Theory and Methods, A9, 181-200.

Searle, S. R. and Henderson, H. V. (1979). Annotated Computer Output for Analyses of Unbalanced Data: SAS GLM. Paper No. BU-641-M in the Biometrics Unit, Cornell University.

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, 16-33.

Appendix

In the linear model $\underset{\sim}{y} \sim N(\underset{\sim\sim}{X}\underset{\sim}{b}, \sigma^2\underset{\sim\sim}{I})$, the F-statistic for testing the hypothesis $H: \underset{\sim\sim}{K'}\underset{\sim}{b} = \underset{\sim}{0}$ is (e.g., Searle, 1971, Section 5.5c) $F(H) = Q/\hat{s\sigma}^2$ for $Q = \underset{\sim}{b}^{o\prime}\underset{\sim\sim}{K}(\underset{\sim\sim}{K'GK})^{-1}\underset{\sim\sim}{K'}\underset{\sim}{b}^{o}$ with s = rank of $\underset{\sim\sim}{K'}$ for $\underset{\sim\sim}{K'}$ being of full row rank and $\underset{\sim\sim}{K'}\underset{\sim}{b}$ estimable, and with $\underset{\sim}{b}^{o} = \underset{\sim\sim}{GX'}\underset{\sim}{y}$ for $\underset{\sim\sim}{X'XGX'X} = \underset{\sim\sim}{X'X}$. Recalling that any matrix $\underset{\sim}{A}_{p\times q}$ of rank s can be factored as $\underset{\sim}{A} = \underset{\sim\sim}{BC}$ where $\underset{\sim}{B}$ has full column rank s and $\underset{\sim}{C}$ has full row rank s, let this factoring for $\underset{\sim\sim}{K'}$ be $\underset{\sim\sim}{K'} = \underset{\sim\sim}{L'T}$. Then, because $\underset{\sim\sim}{K'}$ has full row rank, $\underset{\sim\sim}{L'}$ is non-singular and $\underset{\sim}{T}$, of course, has full row rank. Therefore the hypothesis $H: \underset{\sim\sim}{K'}\underset{\sim}{b} = \underset{\sim}{0}$ can also be written as $H: \underset{\sim\sim}{L'T}\underset{\sim}{b} = \underset{\sim}{0}$, equivalent to

$$H: f_i = 0 \quad \text{for } i = 1, \cdots, s \tag{A1}$$

with

$$f_i = \underset{\sim i\sim\sim}{\ell'}\underset{\sim}{Tb} \tag{A2}$$

for

$$\underset{\sim i}{\ell'},\ i = 1, \cdots, s, \text{ being s linearly independent vectors of order s.} \tag{A3}$$

Expressions (A1), (A2) and (A3) are effectively the manner in which SAS GLM output is used for describing a hypothesis corresponding to any of the Type I, II, III or IV sums of squares. The specific output is (A2), in the form of $\underset{\sim\sim}{\ell'T}$ as a vector of coefficients which are linear functions of s arbitrary L's (elements of $\underset{\sim}{L}$) printed alongside the parameters. For example, the hypothesis output corresponding to the Type IV row sum of squares for the data of Table 1 is

MU	0	A ⊗ B	11	0
A1	L_2		13	$\frac{1}{2}L_2$
A2	L_3		14	$\frac{1}{2}L_2$
A3	$-L_2 - L_3$		21	0
B1	0		22	L_3
B2	0		32	$-L_3$
B3	0		33	$-\frac{1}{2}L_2$
B4	0		34	$-\frac{1}{2}L_2$.

(A4)

The corresponding $f = \underset{\sim}{\ell}'\underset{\sim}{T}\underset{\sim}{b}$ of (A2) is shown in (2), for which $s = 2$, $\underset{\sim}{\ell}' = [L_2 \;\; L_3]$ and, corresponding to

$$\underset{\sim}{b}' = [\mu \;\; \alpha_1 \;\; \alpha_2 \;\; \alpha_3 \;\; \beta_1 \;\; \beta_2 \;\; \beta_3 \;\; \beta_4 \;\; \gamma_{11} \;\; \gamma_{13} \;\; \gamma_{14} \;\; \gamma_{21} \;\; \gamma_{22} \;\; \gamma_{32} \;\; \gamma_{33} \;\; \gamma_{34}] ,$$

$\underset{\sim}{T}$ of (A2) is

$$\underset{\sim}{T} = \begin{bmatrix} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}$$

An important consequence of expressing a hypothesis in the form of (A1), (A2) and (A3) is that in (A3) it does not matter what set of s linearly independent (LIN) vectors is used for the $\underset{\sim}{\ell}_i'$-vectors. Any set can be used, and for all such sets the F-statistic is the same. Proof of this is as follows.

Let $\underset{\sim}{L}'$ be some particular set of s LIN row vectors $\underset{\sim}{\ell}_i'$ for $i = 1, \cdots, s$. Then stating the hypothesis as $H : \underset{\sim\sim}{LT}\underset{\sim}{b} = \underset{\sim}{0}$, the corresponding F-statistic is

$$F(H) = Q/s\hat{\sigma}^2 \quad \text{for} \quad Q = \underset{\sim}{b}^{o\prime}\underset{\sim}{T}'\underset{\sim}{L}(\underset{\sim}{L}'\underset{\sim\sim}{TGT}'\underset{\sim}{L})^{-1}\underset{\sim}{L}'\underset{\sim\sim}{T}\underset{\sim}{b}^{o} .$$

Since $\underset{\sim}{L}$ is, by definition, non-singular,

$$Q = \underset{\sim}{b}^{o\prime}\underset{\sim}{T}'\underset{\sim\sim}{LL}^{-1}(\underset{\sim\sim}{TGT}')^{-1}\underset{\sim}{L}'^{-1}\underset{\sim}{L}'\underset{\sim\sim}{T}\underset{\sim}{b}^{o} = \underset{\sim}{b}^{o\prime}\underset{\sim}{T}'(\underset{\sim\sim}{TGT}')^{-1}\underset{\sim\sim}{T}\underset{\sim}{b}^{o} ,$$

which is immediately seen not to depend on $\underset{\sim}{L}$, whatever its value; i.e., F(H) does not depend on any particular values used for row vectors $\underset{\sim}{\ell}_i'$ in (A1) and (A2).

Example. The data of Table 1 are those of the example in Section 7.2 of Searle (1971). The matrix $\underset{\sim}{G}$ used there is the diagonal matrix

$$\underset{\sim}{G} = \text{diag}\left\{ 0\underset{\sim}{1}_8' \,;\, \tfrac{1}{3} \quad 1 \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{2} \quad \tfrac{1}{4} \right\} \qquad \text{(A5)}$$

and the corresponding $\underset{\sim}{b}^{o}$ is

$$\underset{\sim}{b}^{o\prime} = \left[0\underset{\sim}{1}_8' \,;\, 10 \quad 12 \quad 9 \quad 9 \quad 13 \quad 8 \quad 15 \quad 12 \right]. \qquad \text{(A6)}$$

A first pair of LIN values for $\underset{\sim}{\ell}_i' = [L_2 \quad L_3]$ implicit in (A4) is [1 0] and [0 1] giving

$$\underset{\sim}{L'T} = \left[\begin{array}{cccccccc:cccccccc} 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 & 0 \end{array}\right] \qquad \text{(A7)}$$

corresponding to (3). Notice that the right-hand sub-matrix of (A7) is, apart from a scalar factor of $\frac{1}{2}$ in the first row, the same as $\underset{\sim}{M}'$ in (13). In (A7) the left-hand part of the partitioning corresponds to the terms $0\underset{\sim}{1}_8'$ in (A5) and (A6). Then

$$Q = (\underset{\sim}{L'}\underset{\sim\sim}{T}\underset{\sim}{b}^{o})'(\underset{\sim}{L'}\underset{\sim\sim\sim\sim}{TGT'}\underset{\sim}{L})^{-1}\underset{\sim}{L'}\underset{\sim\sim}{T}\underset{\sim}{b}^{o} = [-3 \quad 5]\begin{bmatrix} 9/16 & 0 \\ 0 & 1 \end{bmatrix}^{-1}\begin{bmatrix} -3 \\ 5 \end{bmatrix}$$

$$= [-3 \quad 5]\begin{bmatrix} 16/9 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} -3 \\ 5 \end{bmatrix} = 9(16/9) + 25 = 41 ,$$

as in (15). A second pair of LIN values for $\underset{\sim}{\ell}_i'$ in (A4) is [2 1] and [3 2], gives

$$\underset{\sim}{L'}\underset{\sim}{T} = \left[\begin{array}{cccccccc:cccccccc} 0 & 2 & 1 & -3 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & -1 & -1 & -1 \\ 0 & 3 & 2 & -5 & 0 & 0 & 0 & 0 & 0 & 1\tfrac{1}{2} & 1\tfrac{1}{2} & 0 & 2 & -2 & -1\tfrac{1}{2} & -1\tfrac{1}{2} \end{array}\right]$$

corresponding to (4). Hence

$$Q = (\underset{\sim}{L'}\underset{\sim\sim}{Tb^{\circ}})'(\underset{\sim}{L'}\underset{\sim\sim\sim}{TGT'}\underset{\sim}{L})^{-1}\underset{\sim}{L'}\underset{\sim\sim}{Tb^{\circ}} = [-1 \quad 1]\begin{bmatrix} 3\frac{1}{4} & 5\frac{3}{8} \\ 5\frac{3}{8} & 9\frac{1}{16} \end{bmatrix}^{-1}\begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

$$= \left[3\tfrac{1}{4} + 9\tfrac{1}{16} + 10\tfrac{3}{4}\right] / \left[\tfrac{13}{4}\left(\tfrac{145}{16}\right) - \tfrac{43^2}{64}\right] = \frac{369}{16}\frac{64}{36} = 41 ,$$

the same as already obtained.

Table 1

Cell means $\bar{y}_{ij.}$ and numbers of observations (n_{ij})

for Example 1

i \ j	1	2	3	4
1	10(3)		12(1)	9(2)
2	9(2)	13(2)		
3		8(2)	15(2)	12(4)

Table 2

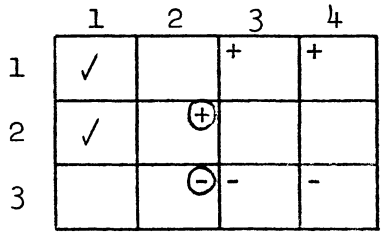
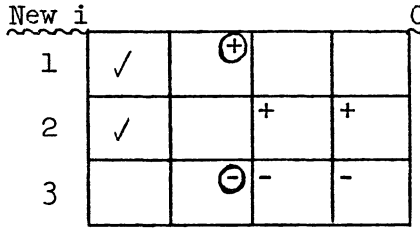
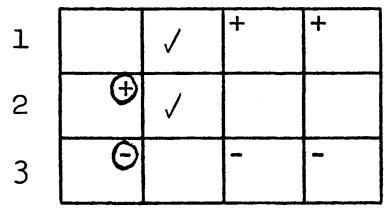
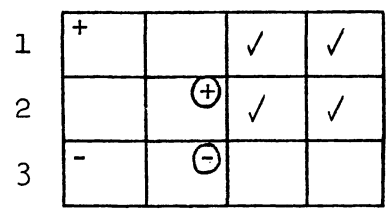
Schematic representation of hypothesis (13)

$$H:\begin{cases}\mu_{13} + \mu_{14} = \mu_{33} + \mu_{34} \\ \mu_{22} = \mu_{32}\end{cases}$$

	1	2	3	4
1	✓		+	+
2	✓	⊕		
3		⊖	-	-

Table 3

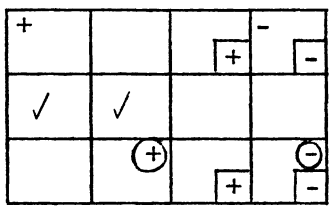
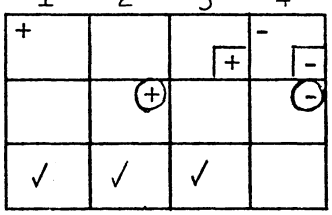
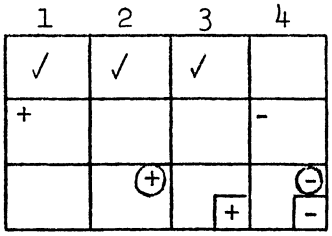
Type IV hypotheses and sums of squares for rows, for data of Table 1 with rows resequenced

Resequencing of rows	Hypothesis[1]		Sum of squares for rows
	Schematic representation, and in terms of new μ_{ij}'s	In terms of old μ_{ij}'s	
None		$\mu_{13} + \mu_{14} = \mu_{33} + \mu_{34}$ $\mu_{22} = \mu_{32}$	41
Rows 1 and 2 interchanged	New i / Old i 	$\mu_{13} + \mu_{14} = \mu_{33} + \mu_{34}$ $\mu_{22} = \mu_{32}$	41
	$\mu'_{23} + \mu'_{24} = \mu'_{33} + \mu'_{34}$ $\mu'_{12} = \mu'_{32}$		
Rows 1 and 3 interchanged		$\mu_{33} + \mu_{34} = \mu_{13} + \mu_{14}$ $\mu_{21} = \mu_{11}$	$17\frac{1}{5}$
	$\mu''_{13} + \mu''_{14} = \mu''_{33} + \mu''_{34}$ $\mu''_{21} = \mu''_{31}$		
Rows 2 and 3 interchanged		$\mu_{11} = \mu_{21}$ $\mu_{32} = \mu_{22}$	$26\frac{1}{5}$
	$\mu'''_{11} = \mu'''_{31}$ $\mu'''_{22} = \mu'''_{32}$		

1/ Old μ_{ij}'s means μ_{ij}'s before resequencing rows.
New μ_{ij}'s means μ_{ij}'s after resequencing rows, denoted μ'_{ij}, μ''_{ij} and μ'''_{ij}.

Table 4

Type IV hypotheses and sums of squares for columns, for data of Table 1 with columns resequenced

Resequencing of columns	Hypothesis[1]		Sum of squares for columns
	Schematic representation, and in terms of new μ_{ij}'s	In terms of old μ_{ij}'s	
None		$\mu_{11} = \mu_{14}$ $\mu_{32} = \mu_{34}$ $\mu_{13} + \mu_{33} = \mu_{14} + \mu_{34}$	$46\frac{23}{28}$
Columns 1 and 4 interchanged	New j: 1 2 3 4  Old j: 4 2 3 1 $\mu'_{11} = \mu'_{14}$ $\mu'_{22} = \mu'_{24}$ $\mu'_{13} = \mu'_{14}$	$\mu_{14} = \mu_{11}$ $\mu_{22} = \mu_{21}$ $\mu_{13} = \mu_{11}$	22
Columns 2 and 4 interchanged	New j: 1 2 3 4  Old j: 1 4 3 2 $\mu''_{21} = \mu''_{24}$ $\mu''_{32} = \mu''_{34}$ $\mu''_{33} = \mu''_{34}$	$\mu_{21} = \mu_{22}$ $\mu_{34} = \mu_{32}$ $\mu_{33} = \mu_{32}$	$65\frac{1}{2}$

1/ Old μ_{ij}'s means μ_{ij}'s before resequencing columns.
New μ_{ij}'s means μ_{ij}'s after resequencing columns, denoted μ'_{ij} and μ''_{ij}.

Table 5: Examples 2 and 3

Example 2

Cell means $\bar{y}_{ij.}$ and numbers of observations (n_{ij})

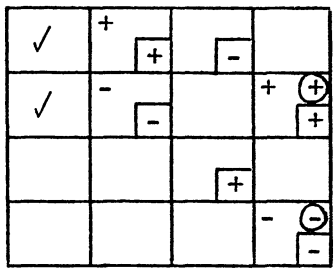
i \ j	1	2	3	4
1	5(2)	7(2)	6(3)	
2	9(4)	5(3)		6(2)
3			7(1)	
4				4(4)

Example 3: include $\bar{y}_{34} = 7$ with $n_{34} = 2$.

Table 6

Type IV contrasts for Example 2 of Table 5

α-based contrasts

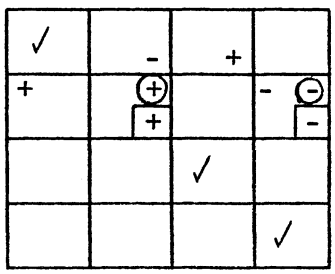


$$\mu_{12} + \mu_{24} = \mu_{22} + \mu_{44}$$

$$\mu_{24} = \mu_{44}$$

$$\mu_{12} + \mu_{24} + \mu_{33} = \mu_{13} + \mu_{22} + \mu_{44}$$

β-based contrasts



$$\mu_{21} = \mu_{24}$$

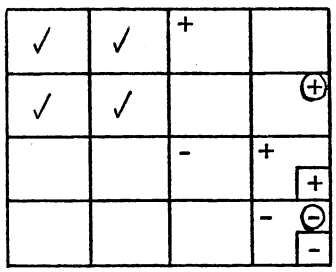
$$\mu_{22} = \mu_{24}$$

$$\mu_{13} + \mu_{22} = \mu_{12} + \mu_{24}$$

Table 7

Type IV contrasts and sums of squares for Example 3 of Table 5

α-based contrasts



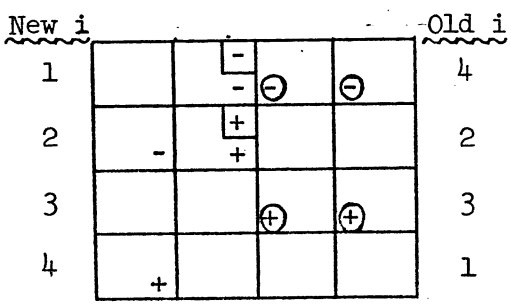
$$\mu_{13} + \mu_{34} = \mu_{33} + \mu_{44}$$

$$\mu_{24} = \mu_{44}$$

$$\mu_{34} = \mu_{44}$$

sum of squares: $14\frac{1}{4}$

α-based contrasts with rows 1 and 4 interchanged



$$\mu_{33} + \mu_{44} = \mu_{13} + \mu_{34}$$

$$\mu_{21} + \mu_{22} = \mu_{11} + \mu_{21}$$

$$\mu_{33} = \mu_{13}$$

sum of squares: $15\frac{21}{76}$

$$\mu'_{14} + \mu'_{33} = \mu'_{34} + \mu'_{43}$$

$$\mu'_{21} + \mu'_{22} = \mu'_{41} + \mu'_{42}$$

$$\mu'_{33} = \mu'_{43}$$