RELATIONSHIPS BETWEEN THE ESTIMABLE FUNCTIONS OF SAS GLM OUTPUT FOR UNBALANCED DATA AND THE HYPOTHESES TESTED BY TRADITIONAL-STYLE F-STATISTICS

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

ABSTRACT

The Statistical Analysis System (SAS) computer routine entitled General Linear Model (GLM) includes in its output four types of estimable functions that have certain arbitrariness (represented by the letter L) in their coefficients. This paper shows how such arbitrary estimable functions are derived from the known, general expressions for hypotheses tested by traditional-style F-statistics in analysis of variance calculations that are often made for unbalanced data (i.e., data having unequal numbers of observations in their subclasses).

1. INTRODUCTION

We follow the notation of Searle (1971) and refer to its pages in the form LM 164, for example, meaning page 164. There we find the general linear model represented familiarly as \( y = Xb + e \) where \( y \) is an \( N \times 1 \) vector of observations, \( b \) is a \( p \times 1 \) vector of parameters, \( X \) is an \( N \times p \) known matrix, and \( e \) is a vector of random error terms, assumed to have dispersion matrix \( \sigma^2 I \), where \( I \) is an identity matrix of order \( N \). A solution vector \( b^0 \) of the normal equations \( X'Xb^0 = X'y \) is represented by \( b^0 = GX'y \) where \( G \) is any matrix satisfying \( X'XGX'X = X'X \) and for \( r \) being the rank of \( X \), an unbiased estimator of \( \sigma^2 \) is

\[
\hat{\sigma}^2 = \frac{y'(I - XG'X)y}{(N - r)} \tag{1}
\]

For \( K' \) being a matrix of full row rank, \( s \) say, with each of the \( s \) functions of \( b \) in \( K'b \) being estimable (i.e., \( K'X = T'X \) for some matrix \( T' \)), the hypothesis

\[
H : K'b = m \tag{2}
\]

is said to be testable and, under normality assumptions,

\[
F = Q/\hat{\sigma}^2, \quad \text{with} \quad Q = (K'b^0 - m)'(K'G)K^{-1}(K'b^0 - m), \tag{3}
\]

is distributed as Snedecor's F-variable with \( s \) and \( N - r \) degrees of freedom, and can be used to test \( H \) (e.g., LM 190).

Correct use of this hypothesis testing procedure entails first formulating a hypothesis in the manner of (2) and then testing it using (3). In contrast, high-speed computing nowadays often results in easy calculation of F-statistics without prior formulation of the hypothesis to be tested. Hence we are frequently in the position of first calculating an F and then trying to understand what hypothesis it is testing. Not only is this logically backwards, but the hypothesis is often of neither use nor interest.

\(^1\) The text of this paper is essentially the same as that of "Arbitrary hypotheses in linear models with unbalanced data", paper BU-343 of the Biometrics Unit, Cornell University, that is being published in Communications in Statistics, in a special issue dealing with the analysis of variance of unbalanced data, edited by R. R. Hocking and F. M. Speed.
Nevertheless, users of computer packages, by their demands to those who develop packages, apparently con-
tinue wanting certain traditional F-statistics and sums of squares from their data, sums of squares that
are oftentimes easily denoted by the R(·)-notation [e.g., LM 246, Searle (1972), Speed and Rocking (1976)].
These are, for example, the sums of squares labeled as Types I, II and III in SAS output. Their inter-
pretation there is in terms of estimable functions that are the basis of the hypotheses being tested when
using those sums of squares as Q in the manner of (3). Although these estimable functions are more in-
formative than the succinct R(·)-symbols, the resulting hypotheses are only arbitrary restructuring of
the hypotheses already known as being tested by these sums of squares. Since these hypotheses often have
no practical value, they and the corresponding F-statistics should usually be given no attention whatever.
Users should start at (2) and then use (3).

So long as traditional-style sums of squares are wanted and obtainable from computing facilities,
there needs to be adequate description of their utility, even if it is slight. Description is given in
Searle (1971, Chapters 6 and 7) for 1-way and 2-way classifications, in terms of explicit expressions for
the hypotheses being tested, using linear functions of parameters of the model with coefficients that are
functions of the numbers of observations in the subclasses of the data. Description can also be in terms
of estimable functions having coefficients that involve arbitrary values (as in SAS output). Relations-
ships between these two descriptions are illustrated in this paper using the 2-way crossed classification
model

\[ y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk} \]  

(4)

where, in the usual way, \( y_{ijk} \) is the k'th observation on the i'th level \((i = 1, \ldots, a)\) of the A-factor
and the j'th level \((j = 1, \ldots, b)\) of the B-factor, \( \mu \) is a general mean, \( \gamma_{ij} \) is an interaction effect and \( \epsilon_{ijk} \)
is the random error term. The number of observations in the cell defined by the i'th level of A and
the j'th level of B is denoted by \( n_{ij} \), with \( n_{ij} = 0 \) for each cell having no data, and otherwise \( k = 1, \ldots, n_{ij} \) in (4).

2. UTILITY OF SAS OUTPUT ESTIMABLE FUNCTIONS

The Type I, II, III and IV estimable functions in SAS output each consist of a function of L-values
alongside each parameter of the model. Multiplying those L-values by their corresponding parameters and
summing the products gives a (linear) estimable function, call it \( f \) say. Corresponding to \( f \), the output
also shows a sum of squares, SS(\( f \)) say. The utility of \( f \) is as follows. When there are \( s \) different L-
symbols in \( f \), the F-statistic that has SS(\( f \)) in its numerator, in the manner of \( Q \) of (3), tests a hypothe-
sis consisting of equating to zero \( s \) linearly independent (LIN) estimable functions derived from
\( f \) by using a sets of the \( s \) L-values in \( f \) that yield LIN functions of the parameters. We say that "\( f \) explains
SS(\( f \))".

The sums of squares SS(\( f \)) that are explained in this manner are as follows:

Type I : Sequential R(·)-values.

Type II : R(·)-values for each factor adjusted for all other factors that do not
contain it (e.g., not adjusted for interactions with the factor concerned, nor
for factors nested within it).

Type III: R(·)-values for each factor adjusted for all other factors, using the
"usual restrictions" of having elements in the model add to zero, e.g.,

\[ \sum_{i=1}^{a} \alpha_i = 0, \quad \sum_{j=1}^{b} \beta_j = 0, \]  

and so on.
In contrast to Types I - III, the Type IV procedure does not seek to provide estimable functions that explain some pre-ordained sum of squares such as R(·)-values, for that, after all, is not logically sound. Type IV uses the correct logic: it sets up a hypothesis, in the manner of (2), and calculates the F-statistic to test it, in the manner of (3). To do so, it examines the pattern of filled cells in the data and from this sets up some hypotheses that look reasonable (based on what would be done with balanced — all cells filled — data) and then calculates Q for (3). In some cases Q will be an R(·)-value, but in many cases it will not.

Further details of these descriptions are given in the examples which follow.

2. A NO-INTERACTION EXAMPLE

We begin with an example of (4) that excludes interactions,  \( \gamma_{ij} \), and with \( n_{ij} = 0 \) or 1 in all cells. The hypotheses tested when using two R(·)-values as Q in (3) are shown in Table I.

<table>
<thead>
<tr>
<th>R(·)-value</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>R(α</td>
<td>μ)</td>
</tr>
<tr>
<td>R(β</td>
<td>μ,α)</td>
</tr>
</tbody>
</table>

As a simple numerical example, we use that of LM 262 having the numbers of observations shown in Table II.

<table>
<thead>
<tr>
<th>( n_{ij} )-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>( i )</td>
</tr>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>( n_i )</td>
</tr>
</tbody>
</table>

Output from SAS for this example includes what is shown in Table III.
TABLE III

Output from SAS for Types I and II Estimable Functions
for the Example of Table II

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Type I For A</th>
<th>Type I For B</th>
<th>Type II For A</th>
<th>Type II For B</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha_1$</td>
<td>$L_2$</td>
<td>0</td>
<td>$L_2$</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha_2$</td>
<td>$L_3$</td>
<td>0</td>
<td>$L_3$</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha_3$</td>
<td>$L_4$</td>
<td>0</td>
<td>$L_4$</td>
<td>0</td>
</tr>
<tr>
<td>$\alpha_4$</td>
<td>$-L_2-L_3-L_4$</td>
<td>0</td>
<td>$-L_2-L_3-L_4$</td>
<td>0</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>$-L_3/3 + L_4/6$</td>
<td>$L_6$</td>
<td>0</td>
<td>$L_6$</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>$-L_3/3 - L_4/3$</td>
<td>$L_7$</td>
<td>0</td>
<td>$L_7$</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>$+2L_3/3 + L_4/6 - L_6 - L_7$</td>
<td>0</td>
<td>$-L_6 - L_7$</td>
<td>0</td>
</tr>
</tbody>
</table>

Sum of Squares explained:

|         | $R(\alpha|\mu)$ | $R(\beta|\mu,\alpha)$ | $R(\alpha|\mu,\beta)$ | $R(\beta|\mu,\alpha)$ |

1/ Symbols are the same as in SAS output except for using subscripts and fractions: e.g., $L_3/3$ in place of 0.3333L3 of SAS.

2/ Types III and IV are the same as Type II in this example.

Relative to the hypotheses of Table I, three questions arise: where do the L's come from, how are the linear functions of them derived and how are they to be used?

Section 2 describes what sums of squares the different SAS output estimable functions explain. For the example of Table III, when the sequence in which the factors of the model are fitted is A then B, the Type I estimable function for A is explaining $R(\alpha|\mu)$ and that for B is explaining $R(\beta|\mu,\alpha)$. Type II estimable functions explain $R(\alpha|\mu,\beta)$ and $R(\beta|\mu,\alpha)$, respectively. Types III and IV in this case are the same as Type II.

The general use of the estimable functions is also described in Section 2. For Type I for A in Table III, $f$ is formed as

$$f = L_2\alpha_1 + L_3\alpha_2 + L_4\alpha_3 + (-L_2 - L_3 - L_4)\alpha_4$$


There are three L-symbols in $(5)$. Choose any three sets of values for them that gives three LIN forms of $(5)$: e.g.,

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>$L_3$</th>
<th>$L_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-1</td>
<td>0</td>
</tr>
</tbody>
</table>

gives

$$f_1 = \alpha_1 - \alpha_2 + \beta_1/3 + \beta_2/3 - 2\beta_3/3 .$$

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>$L_3$</th>
<th>$L_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

gives

$$f_2 = \alpha_1 - \alpha_4 .$$

<table>
<thead>
<tr>
<th>$L_2$</th>
<th>$L_3$</th>
<th>$L_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
</tbody>
</table>

gives

$$f_3 = \alpha_3 - \alpha_4 + \beta_1/6 + \beta_2/3 + \beta_3/6 .$$

Then the Type I sum of squares for A, which is $R(\alpha|\mu)$, is such that

$$R(\alpha|\mu)/2\sigma^2 \text{ tests } H: f_1 = 0 = f_2 = f_3 .$$
This is the sense in which \( f \) explains \( R(\alpha|\mu) \), except that in (5) any three sets of values of \( L_1, L_2 \) and \( L_3 \) can be used so long as the resulting \( f_1, f_2, f_3 \) are LIN functions of the \( \alpha \)'s and \( \beta \)'s. Equations (6) contain just one such example. Similar procedures apply to the other columns of Table III.

The origin of the \( \beta \)'s and the relationship of their occurrence in Table III to the hypotheses in Table I is now described. The starting point is to use the \( L_4 \)-values of Table II in the hypothesis corresponding to \( R(\alpha|\mu) \) in Table I. This shows that \( R(\alpha|\mu) \) is used for testing

\[
H: \left\{ \begin{align*}
\alpha_1 + \frac{1}{3}(\beta_1 + \beta_2 + \beta_3) \\
\alpha_2 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3) \\
\alpha_3 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3) \\
\alpha_4 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3)
\end{align*} \right\} \text{ all equal.}
\] (8)

To establish (5), note that the parametric functions in (8) are estimable and therefore any linear combination of them is also. Form a linear combination that is a contrast among the \( \alpha \)'s (i.e., in which the coefficients of the \( \alpha \)'s sum to zero) plus the resulting "mass" of \( \beta \)'s. Such a combination can be formed from the parametric functions in (8) by using arbitrary values \( L_2, L_3 \) and \( L_4 \) to create

\[
f = L_2[\alpha_1 + \frac{1}{3}(\beta_1 + \beta_2 + \beta_3)] + L_3[\alpha_2 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3)] \\
+ L_4[\alpha_3 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3)] + [-L_2 - L_3 - L_4][\alpha_4 + \frac{1}{2}(\beta_1 + \beta_2 + \beta_3)].
\] (9)

This simplifies to be (5).

Although (5) includes a contrast among \( \alpha \)'s, it also contains \( \beta \)'s, and so is not a true \( \alpha \)-contrast. Instead, we name it an \( \alpha \)-based contrast, and use this name repeatedly in the sequel.

Not only is (5) estimable, but from its origin (9) it is a contrast among the estimable functions that are part of the hypothesis (8). Furthermore, (8) can be re-expressed as equating to zero three LIN contrasts of those estimable functions; i.e., of three LIN forms of (5). This is precisely what (7) is. Thus it is that the arbitrary estimable function \( f \) is derived for Table III and used for explaining \( R(\alpha|\mu) \) — explaining, in the sense of Section 2. The crux of the matter is deriving \( f \) from the form of the hypotheses given in Table I. Once obtained, \( f \) is used in the manner just illustrated, the general manner of its use being as described in Section 2. This applies for all the \( f \)'s to be discussed here. The remainder of the paper is therefore devoted to deriving \( f \)'s. Each of them is used in the manner just demonstrated.

The Type I \( f \) for \( B \) in Table III is

\[
f = L_6\beta_1 + L_7\beta_2 + (-L_6 - L_7)\beta_3.
\] (10)

It explains \( R(\beta|\mu,\alpha) \), for which from Table I the corresponding hypothesis is

\[
H: \beta_1 = \beta_2 = \beta_3.
\] (11)

A \( \beta \)-based contrast of the parametric functions in (11) is \( L_6\beta_1 + L_7\beta_2 + (-L_6 - L_7)\beta_3 \), which is precisely \( f \) of (10). It is also in this case, of course, a true \( \beta \)-contrast.
Hypotheses corresponding to three R(·)-values in the 2-way crossed classification with interaction model are shown in Table IV.

**TABLE IV**

Hypotheses Corresponding to Three R(·)-values in the 2-way Crossed Classification with Interaction. [See LM 305-311.]

<table>
<thead>
<tr>
<th>R(·)-value</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>R(α</td>
<td>μ)</td>
</tr>
<tr>
<td>R(β</td>
<td>μ,α)</td>
</tr>
<tr>
<td>R(γ</td>
<td>μ,α,β)</td>
</tr>
</tbody>
</table>

As a numerical example, we assume the data have the \( n_{ij} \)-values of Table V and show how the Type I, II and III estimable functions are derived from the hypotheses of Table IV.

**TABLE V**

\( n_{ij} \)-values

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Type I for A explains \( R(α|μ) \). Using Table V in Table IV yields the corresponding hypothesis as

\[
H : \begin{cases} 
\alpha_1 + \frac{1}{4} (2\beta_1 + \beta_2 + \beta_3 + 2\gamma_{11} + \gamma_{12} + \gamma_{13}) \\
\alpha_2 + \frac{1}{4} (2\beta_1 + \beta_2 + \beta_3 + 2\gamma_{21} + 2\gamma_{22} + \gamma_{23}) 
\end{cases} \text{ equal.} \quad (12)
\]

An \( \alpha \)-based contrast of the parametric functions in (12) is...
\[ f = L_2(x_1 + \frac{1}{2}(2\beta_1 + \beta_2 + \beta_3 + 2\gamma_{11} + \gamma_{12} + \gamma_{13})) \]
\[ + (-L_2)[(\alpha_2 + \frac{1}{2}(\beta_1 + 2\beta_2 + \beta_3) + \gamma_{21} + 2\gamma_{22} + \gamma_{23})] \]

which simplifies to
\[ f = L_2\alpha_1 - L_2\alpha_2 + \frac{1}{2}L_2\beta_1 - \frac{1}{4}L_2\beta_2 + \frac{1}{2}L_2\gamma_{11} + \frac{1}{4}L_2\gamma_{12} + \frac{1}{4}L_2\gamma_{13} \]
\[ - \frac{1}{4}L_2\gamma_{21} - \frac{1}{2}L_2\gamma_{22} - \frac{1}{4}L_2\gamma_{23} , \]

which is the Type I estimable function for \( A \).

Type I for \( B \), assuming the order of fitting the factors is \( A, B, AB \), explains \( \mathbb{R}(B|A,0) \). For the corresponding hypothesis in Table IV write \( \psi_j \) as
\[ \psi_j = \sum_{j'=1}^{b} c_{j,j'} \beta_{j'} + \sum \lambda_{j,j'} Y_{i,j} , \]
defining
\[ c_{j,j'} = c_{j,j'}^{(r)} - \sum_{i=1}^{a} n_{i,j} n_{i,j'}/n_1 . \]
and
\[ \lambda_{j,j'} = \delta_{j,j'} n_{i,j} - n_{i,j} n_{i,j'}/n_1 . \quad (13) \]
with \( \delta_{j,j'} \) being the Kronecker delta: \( \delta_{j,j'} = 0, j \neq j', \delta_{j,j'} = 1 \). Then the \( c_{j,j'} \)’s are those of equation (17), LM 367, and for each \( \psi_j \) have the property
\[ (\text{coefficients of all } \beta_{j'}'s) = 0; \text{ i.e., } \sum_{j'=1}^{b} c_{j,j'}^{(r)} = 0 \forall j . \quad (14) \]

And by their definition in (13), the \( \lambda \)'s for each \( \psi_j \) have the properties

for every \( i \):
\[ (\text{coefficients of } Y_{i,j}) = 0; \text{ i.e., } \sum_{j'=1}^{b} \lambda_{j,j'} = 0 \forall i, j \quad (15) \]

and

for every \( j' \):
\[ (\text{coefficients of } Y_{i,j'}) = c_{i,j'} \text{ i.e., } \sum_{i=1}^{a} \lambda_{i,j,j'} = c_{i,j'} \forall j,j' . \quad (16) \]

Now, all of the \( c \)'s and \( \lambda \)'s are known values. But equation (14) means for each \( j \) that \( b-1 \) of the \( c \)'s determine the other; and, when all cells of the data have observations, as in Table V, (15) and (16) mean that \( (a-1)(b-1) \) of the \( \lambda \)'s determine the others. And these statements are true \( (b-1) \) times, namely for \( (b-1)\psi_j \)'s - since \( \sum_{j=1}^{b} \psi_j = 0 \) - see LM 308. These equalities are the basis for development of \( f \). First
write out the $\phi_j$'s using (14), (15) and (16):

\[
\begin{array}{ccccccc}
\phi_1 & \phi_2 & \phi_3 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\end{array}
\]

\[
\begin{array}{ccccccc}
\phi_1 & \phi_2 & \phi_3 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\end{array}
\]

\[
\begin{array}{ccccccc}
\phi_1 & \phi_2 & \phi_3 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\beta_1 & \beta_2 & \beta_2 & \gamma_{11} & \gamma_{12} & \gamma_{13} \\
\end{array}
\]

All these values are known - from (13). But from this table it is clear that they are all determined by just the values $c_{11}$, $c_{12}$, $\lambda_{1,11}$, $\lambda_{1,12}$ and $c_{21}$, $c_{22}$, $\lambda_{2,11}$, $\lambda_{2,12}$. But these into a matrix and write

\[
\begin{bmatrix}
\phi_1 \\
\phi_2 \\
\end{bmatrix} =
\begin{bmatrix}
\begin{array}{cccc}
c_{11} & c_{12} & \lambda_{1,11} & \lambda_{1,12} \\
c_{21} & c_{22} & \lambda_{2,11} & \lambda_{2,12} \\
\end{array}
\end{bmatrix}
\begin{bmatrix}
1 & 0 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\beta_1 \\
\beta_2 \\
\beta_3 \\
\gamma_{11} \\
\gamma_{12} \\
\gamma_{13} \\
\gamma_{21} \\
\gamma_{22} \\
\gamma_{23} \\
\end{bmatrix}
\]

From (14) and (15) it is clear that $\phi_1$ and $\phi_2$ are contrasts, so that for arbitrary $P_1$ and $P_2$ any linear combination of them $f = P_1\phi_1 + P_2\phi_2$ is also. Then

\[
f_1 = P_1\phi_1 + P_2\phi_2 = [P_1 \ P_2]
\begin{bmatrix}
\begin{array}{cccc}
c_{11} & c_{12} & \lambda_{1,11} & \lambda_{1,12} \\
c_{21} & c_{22} & \lambda_{2,11} & \lambda_{2,12} \\
\end{array}
\end{bmatrix}
\begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 \\
0 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\beta \\
\gamma \\
\end{bmatrix}
\]

where $\beta$ and $\gamma$ represent the vectors of $\beta$'s and $\gamma$'s in (17), respectively. Since $P_1$ and $P_2$ are arbitrary, write

\[
L_b = P_1c_{11} + P_2c_{21} \quad \text{and} \quad L_3 = P_1c_{12} + P_2c_{22}
\]

and then

\[
f = [L_b \ L_3]
\begin{bmatrix}
1 & 0 & 0 & 0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 1 \\
\end{bmatrix}
\begin{bmatrix}
\beta \\
\gamma \\
\end{bmatrix}
\]

203
This is the Type I estimable function for \( B \). Its specific value is obtained using Table V in (13) to obtain

\[
\begin{bmatrix}
  c_{11} & c_{12}
\end{bmatrix}^{-1}
\begin{bmatrix}
  \lambda_{1,11} & \lambda_{1,12}
\end{bmatrix}
\begin{bmatrix}
  \frac{7}{4} & -\frac{1}{4} \\
  -\frac{1}{2} & \frac{3}{4}
\end{bmatrix}
\begin{bmatrix}
  60 & 6 \\
  6 & 39
\end{bmatrix}
= \begin{bmatrix}
  0.606 & -0.060 \\
  0.393 & 0.393
\end{bmatrix},
\]

so that

\[
f = \begin{bmatrix}
  L_4 & L_5
\end{bmatrix}
\begin{bmatrix}
  1 & 0 & 0.606 & -0.606 \\
  0 & 1 & 0.606 & 0.3939 \\
  1 & 1 & 0.606 & 0.393 \\
  0 & 0 & 0 & 1
\end{bmatrix}
\begin{bmatrix}
  \beta \\
  \gamma
\end{bmatrix}
= \begin{bmatrix}
  -0.5455 & 0.3939 & 0.6061 \\
  0.4545 & -0.606 & 0.0606
\end{bmatrix}
\begin{bmatrix}
  \beta \\
  \gamma
\end{bmatrix}
\]  

which, when expanded, is precisely the estimable form found in SAS output.

Type I for \( AB \) explains \( R(\gamma, \alpha, \beta) \) and from Table IV one form of the hypothesis is

\[
H: \begin{cases}
Y_{11} - Y_{12} - Y_{21} + Y_{22} = 0 \\
Y_{12} - Y_{13} + Y_{22} + Y_{23} = 0
\end{cases}
\]

Any linear combination of the parametric functions here is a contrast, and its form can be written as

\[
z = L_7(Y_{11} - Y_{12} - Y_{21} + Y_{22}) + (L_7 + L_8)(Y_{12} - Y_{13} + Y_{22} + Y_{23})
= L_7Y_{11} + L_7Y_{12} + (L_7 + L_8)Y_{13} - L_8Y_{21} + L_8Y_{22} + (L_7 + L_8)Y_{23},
\]  

which is the SAS estimable function for Type I for \( AB \).

Type II estimable functions for \( A, B \) and \( AB \) explain \( R(\alpha | \mu, \beta), R(\beta | \mu, \alpha) \) and \( R(\gamma | \mu, \alpha, \beta) \), respectively.

Type III estimable functions are based implicitly on using the usual constraints

\[
\begin{align*}
\sum_{i=1}^{a} \alpha_i &= 0, & \sum_{j=1}^{b} \beta_j &= 0, \\
\sum_{i=1}^{a} \gamma_{i,j} &= 0, & \sum_{j=1}^{b} \gamma_{i,j} &= 0
\end{align*}
\]

when solving the normal equations. A particular use of (21) in calculating \( R(\alpha | \mu, \beta, \gamma) \) turns out to yield [as discussed, for example, in Searle (1977, pp. 23–25) and Speed et al. (1976)] the sum of squares for \( A \) in the weighted squares of means analysis (SSA \( W \) in EM 359–372). The hypothesis associated with this is

\[H: \alpha_1 + \bar{\gamma}_1, \text{ equal } Y 1. \]  

For Table IV, with \( i = 1, 2 \), an \( \alpha \)-based contrast of the parametric functions in \( H \) is
which is the Type III estimable function for $A$. That for $B$ is derived in similar manner and that for $AB$ is the same as Types I and II for Table IV. Equation (22) does, of course, in the presence of constraints (21) reduce to $f = L_2(a_1 - a_2)$.

Type IV estimable functions for Table IV are the same as Type III. This is true for all data sets in which all cells contain data.

5. SOME CELLS EMPTY, WITH INTERACTION

As a numerical example we use the set of $n_{ij}$'s shown in Table VI.

TABLE VI

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>$n_{ij}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>-</td>
<td>2</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>3</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>10</td>
</tr>
</tbody>
</table>

Types I and II estimable functions for data of this nature are derived in exactly the same manner as were those in Section 4.

Type III estimable functions correspond to weighted squares of means sum of squares when all cells are filled - which they are not, in Table VI. In that case, for columns, for example, the $f_i$'s are linear combinations of the parameters that

(i) are $\beta$-based contrasts,

(ii) involve no $\alpha$'s, and

(iii) are orthogonal to $\gamma$-contrasts that explain $R(\gamma|\lambda, \alpha, \beta)$.

From the pattern of filled cells in Table VI, suitable contrasts satisfying (i) and (ii) are

$$
\beta_1 - \beta_2 + \frac{1}{2}(\gamma_{11} + \gamma_{21}) - \frac{1}{2}(\gamma_{12} + \gamma_{22})
$$

$$
\beta_1 - \beta_2 + \gamma_{11} - \gamma_{13}
$$

$$
\beta_2 - \beta_3 + \gamma_{12} - \gamma_{13}
$$

These are not LIN. But any $b - 1 = 2$ LIN combinations of them can be used. Using coefficients $m_1$, $m_2$ and $m_3$, let a general linear combination of them be

$$
f = m_1[\beta_1 - \beta_2 + \frac{1}{2}(\gamma_{11} + \gamma_{21}) - \frac{1}{2}(\gamma_{12} + \gamma_{22})] + m_2[\beta_1 - \beta_2 + \gamma_{11} - \gamma_{13}] + m_3[\beta_2 - \beta_3 + \gamma_{12} - \gamma_{13}].
$$
Now use (iii): the only contrast available for explaining $R(\gamma|\alpha, \beta, \theta)$ is $\gamma_{11} - \gamma_{12} - \gamma_{21} + \gamma_{22}$. Coefficients of the parameters in this and in $f$ are as follows:

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$b_1$</th>
<th>$b_2$</th>
<th>$b_3$</th>
<th>$\gamma_{11}$</th>
<th>$\gamma_{12}$</th>
<th>$\gamma_{13}$</th>
<th>$\gamma_{21}$</th>
<th>$\gamma_{22}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f$</td>
<td>$m_1 + m_2$</td>
<td>$-m_1 + m_3$</td>
<td>$-m_2 - m_3$</td>
<td>$\frac{1}{2}m_1 + m_2$</td>
<td>$-\frac{1}{2}m_1 + m_3$</td>
<td>$-m_2 - m_3$</td>
<td>$\frac{1}{2}m_1 - \frac{1}{2}m_3$</td>
<td>$\frac{1}{2}m_1$</td>
</tr>
<tr>
<td>$\gamma$-contrast</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>-1</td>
<td>0</td>
<td>-1</td>
<td>1</td>
</tr>
</tbody>
</table>

Orthogonality of the two therefore implies

$$\frac{1}{2}m_1 + m_2 + \frac{1}{2}m_1 - m_3 - \frac{1}{2}m_1 - \frac{1}{2}m_1 = 0; \text{ i.e., } m_2 = m_3.$$

(24)

The degrees of freedom for $f$ are 8, so there will be 2 arbitrary values for the $m$'s; choose them so that the $b_1$ and $b_2$ part of $f$ is $L_1b_1 + L_2b_2$, i.e.,

$$m_1 + m_2 = L_1 \quad \text{and} \quad -m_1 + m_3 = L_2.$$

(25)

Using (24) in (25) gives $m_1 = \frac{1}{2}(L_4 - L_5)$ and $m_2 = \frac{1}{2}(L_4 + L_5) = m_3$, and putting these into (21) gives

$$f = L_1b_1 + L_2b_2 - (L_4 + L_5)b_3 + \frac{1}{2}(3L_4 + L_5)\gamma_{11} + \frac{1}{2}(L_4 - L_5)\gamma_{12} - (L_4 - L_5)\gamma_{21} - (L_4 - L_5)\gamma_{22}$$

which is the Type III estimable function that will be yielded by SAS.

Type IV estimable functions do not have the purpose of explaining some pre-ordained sum of squares, as do Types I, II, and III. Type IV functions are in the nature of $\alpha$-based contrasts, derived from non-unique, balanced subsets of filled cells of the data. It is their non-uniqueness which gives rise to the NOTE that follows each SAS output. The choice of which balanced subsets can be used is arbitrary but limited by the pattern of filled cells. Note that the balanced subsets are subsets of the filled cells and not subsets of balanced data. Thus for data with the characteristics of Table VI the possible balanced subsets for an $\alpha$-based contrast are (i) cells 11, 21, (ii) cells 12, 22 or (iii) cells 11, 12, 21 and 22. Obviously, in some general sense, (iii) is the most efficient, for which the corresponding $\alpha$-based contrast is $\alpha_1 - \alpha_2 + \frac{1}{2}(\gamma_{11} + \gamma_{12} - \gamma_{21} - \gamma_{22})$. A general form of this is

$$f = L_2[\alpha_1 - \alpha_2 + \frac{1}{2}(\gamma_{11} + \gamma_{12} - \gamma_{21} - \gamma_{22})]$$

which is the estimable function to be found in SAS output.

For two $\beta$-based contrasts, possible balanced subsets of filled cells are either

(i) cells 11, 12 and 11, 13,

or (ii) cells 21, 22 and either 11, 13, or 12, 13, or 11, 12, 13

or (iii) cells 11, 12, 21, 22 and either 11, 13 or 12, 13 or 11, 12, 13.
Using (1), two \( \beta \)-based contrasts are \( \beta_1 - \beta_3 + \gamma_{11} - \gamma_{13} \) and \( \beta_2 - \beta_3 + \gamma_{12} - \gamma_{13} \), a linear combination of these being

\[
f = L_4(\beta_1 - \beta_3 + \gamma_{11} - \gamma_{13}) + L_5(\beta_2 - \beta_3 + \gamma_{12} - \gamma_{13})
\]

\[
= L_4 \beta_1 + L_5 \beta_2 - (L_4 + L_5)\beta_3 + L_4 \gamma_{11} + L_5 \gamma_{12} - (L_4 + L_5)\gamma_{13},
\]

which will be found among the SAS output as a Type IV estimable function for \( B \).

In the case of Types III and IV, the sums of squares explained by the \( f \)'s are those derived from \( Q \) of (3).

\section*{Conclusion}

We have illustrated how the four types of estimable functions in SAS output are derived and how they can be used. It is to be noted, however, that none of them negate the logical procedure of first setting out a hypothesis of interest in the form of (2) and then testing it using (3). If a hypothesis of interest can be formulated in terms of a SAS output arbitrary estimable function [e.g., \( H : \alpha_1 = \alpha_2 = \alpha_3 \) can be formulated in terms of \( L_1 \alpha_1 + L_2 \alpha_2 - (L_1 + L_2)\alpha_3 \) by using \( L_1 = 1, L_2 = -1 \) and \( L_1 = 1, L_2 = 0 \)], then the corresponding \( F \)-statistic will be a test of that hypothesis. Otherwise (2) and (3) should be used as indicated.

\section*{Bibliography}


