A CHARACTERIZATION OF THE GLM SUMS OF SQUARES

F. M. Speed, Louisiana State University
R. R. Hocking, Mississippi State University

The SAS procedure, GLM, which is heavily used by practicing statisticians, has generated considerable discussion at both SUGI meetings as well as regional and national meetings. The flexibility of GLM to provide four ANOVA tables has left some users in a quandry as to which table to use. Goodnight, by allowing the user to print out the estimable functions, has helped some statisticians in deciding which set of sums of squares are most desirable. Speed (1977) and Searle (1979) have presented papers at previous SUGI meetings in which they attempted to help clarify this situation. However, questions still persist as to which type of sums of squares to use.

The purpose of this paper is to present a characterization of the GLM sums of squares that should provide further insight as to their nature. This characterization is done via the constrained cell means model. The constraints necessary to generate each type of sums of squares are given. Hence, the user can determine if such constraints are realistic in light of the physical situation, and thus determine if a particular type is appropriate.

INTRODUCTION

The SAS procedure, GLM (1979), which has become a predominant tool among practicing statisticians, is responsible for a considerable amount of discussion on the relative merits of its four types of sums of squares. Indeed, the flexibility of GLM in providing the four ANOVA tables has caused a good number of statisticians to be confused as to which ANOVA table is most appropriate in a given situation. Even though GLM provides the estimable function being tested and even though several papers have discussed this issue [e.g. (1975), (1976) and (1980)], questions still persist as to which type to use.

The purpose of this paper is to present a characterization of the SAS Types I, II and III sums of squares. This is done by expressing each of these in a common setting, i.e. all three are considered in light of testing the same hypothesis in the cells means model with appropriate constraints. By utilizing different constraints, it is possible to generate each of the sums of squares. Since all three types are testing the *same* hypothesis, but under *different* conditions, these conditions can then be overviewed so as to determine their appropriateness. The condition, if any of them, which best reflects the physical situation would then dictate which type to use. Of course, we must be open to the possibility that none of the conditions are realistic and thus none of the ANOVA tables are for use.

This paper addresses two additional points. First, one method of judging which type is best is to consider the algebraic form of the hypothesis. While explicit expressions of the hypothesis are usually preferred, they are sometimes quite complex and difficult to assess. We provide a discussion of the algebraic expressions of the hypothesis being tested. Second, the paper concludes with a discussion of the relative merits of the *SAS Type IV* sum of squares. *Type IV* does not fit into the characterization framework when there are missing cells.

THE CELL MEANS MODEL

The cell means model is used as the foundation on which the characterization is based. This model, in matrix notation, is given by

$$Y = W\mu + e \tag{1}$$

Subject to the constraints

$$G\mu = 0 \tag{2}$$

where Y is the $nx1$ vector of observations, W is the nxp of rank p incidence matrix, μ is the vector of population means, G is the rxp of rank r constraint matrix; and e is the usual error term. If there are no constraints known about the cell means, the $G \equiv 0$.

We now summarize the essential features of estimation and hypothesis testing:

i) $\hat{\mu} = A(W'W)^{-1} W'Y$ (3)

ii) $\hat{\sigma}^2 = (Y - W\hat{\mu})' (Y - W\hat{\mu})/(n-p + r)$ (4)

iii) for testing $H_0 : H\mu = 0$, the numerator of square is

$$SS(H) = (H\hat{\mu})' [HA(W'W)^{-1} A'H']^{-1} H\hat{\mu} \tag{5}$$

where $A = [I - (W'W)^{-1} G (G(W'W)^{-1}G')^{-1} G']$

THE TWO-WAY CLASSIFICATION MODEL

We write the model as

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

where

$$i=1,\ldots, a;\ j=1,\ldots, b;\ k=0,\ldots, n_{ij}$$

For $a = b = 3$, we use the following layout to aid in the formulation of the hypothesis.

TABLE 1

μ_{ij} Layout

		J = 1	J = 2	J = 3	
I = 1		μ_{11}	μ_{12}	μ_{13}	$\bar{\mu}_{1.}$
I = 2		μ_{21}	μ_{22}	μ_{23}	$\bar{\mu}_{2.}$
I = 3		μ_{31}	μ_{32}	μ_{33}	$\bar{\mu}_{3.}$

Here, $\mu_{i.} = {}^1/_3 (\mu_{i1} + \mu_{i2} + \mu_{i3})$.

Table 3 contains the results of analyzing the data given in Table 2.

Table 2

Hypothetical Data Set No. 1

I \ J	1	2	3
1	1.8 1.9 2.0 2.2 2.1	11.9	6.0
2	3.9 4.1	6.0	9.8 10.0 9.9 10.2 10.1
3	3.9 4.1	1.8 2.1 1.9 2.2 2.0	13.9 14.0 14.1

Table 3

F Ratios and Probabilities
For Data Set No. 1

Source	df	Type I F	Type I PR	Type II F	Type II PR	Type III(IV) F	Type III(IV) PR
I	2	840	.000	108	.000	0.0	1.000
J	2	6153	.000	6153	.000	3287.0	.000
IJ	4	2292	.000	2292	.000	2291.0	.000

As is evident from Table 3, the user has three distinct choices for SSI, with F ratios ranging from 0.0 to 840. It is our intent to provide the user with guidelines by which the most appropriate, if any, sum of squares may be chosen. As was indicated earlier, we shall consider two procedures; *i*) the algebraic formulation approach and *ii*) the characterization approach.

The Algebraic Formulation Approach

For the sum of squares SS(I), the hypotheses associated with Types I, II and III are as follows:

$$\text{Type I:}\quad H_0:\ \frac{\sum n_{ij}\,\mu_{ij}}{n_{i\cdot}} = \frac{\sum_{i'j}\,\mu_{i'j}}{n_{i'\cdot}} \qquad (6)$$

for $i=1,\ldots,a$; $i'=1,\ldots,a$; $i \neq i'$.

$$\text{Type II:}\quad H_0:\ \sum_j n_{ij}\,\mu_{ij} = \sum_j \sum_i n_{ij}\,n_{i'j}\,\mu_{i'j}/n_{\cdot j} \qquad (7)$$

$$\text{Type III:}\quad H_0:\ \sum_j d_{ij}\,\mu_{ij} = \sum_j \sum_i d_{ij}\,d_{i'j}\,\mu_{i'j}/d_{\cdot j} \qquad (8)$$

where $d_{ij} = \begin{cases} \text{if } n_{ij} > 0 \\ \text{if } n_{ij} = 0 \end{cases}$

when there are no missing cells, i.e. all $n_{ij}>0$, then (8) reduces to

$$H_0:\ \bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot} \quad \begin{matrix} i=1,\ldots,a \\ i'=1,\ldots,a \\ i\neq i' \end{matrix}$$

which is the hypothesis being tested by all four types when the design is completely balanced.

While this is a very useful and informative method of considering the three types of sums of squares, it is possible that we may reject a method because the form of the hypothesis might appear rather complex. However, under certain conditions or under a non-singular transformation, the complex looking hypothesis may be quite reasonable. [Equation (7) is an example of this which will be discussed later.] Thus, we will consider the second approach.

The Characterization Approach

In this section, we use SS(i), i = 1,2,3, to denote the sum of squares associated with Type I, II and III respectively. In light of this, we now characterize SS(i) as testing

$$H_0:\ \bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$$

in the model

$$Y_{ijk} = \mu_{ij} + e_{ijk}$$

Subject to:

$$G_i\mu = 0$$

Thus, all three types can be thought of as testing the same hypothesis but under different conditions. If a particular condition is reasonable, then we have an appropriate test. If none of the conditions appears to be reasonable, the user must supply his own hypothesis.

For the two-way classification, the different conditions or G_i are:

1) For SS(1), G_1 represents no J and no IJ effects, or

$$G_1:\ \begin{Bmatrix} \mu_{\cdot j} - \mu_{\cdot j} = 0 \quad j = 1,2,3 \\ \mu_{\cdot j} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot} = 0 \quad \begin{matrix} i=1,\ldots,a \\ j=1,\ldots,b \end{matrix} \end{Bmatrix} \qquad (9)$$

2) For SS(2), G_2 represents no IJ effect, or

$$G_2:\ \mu_{ij} - \bar{\mu}_{i\cdot} - \bar{\mu}_{\cdot j} + \bar{\mu}_{\cdot\cdot} = 0 \quad \begin{matrix} i=1,\ldots,a \\ j=1,\ldots,b \end{matrix} \qquad (10)$$

3) For SS(3), G_3, in general, represents no interaction for the missing cells, i.e. if $n_{rs}=0$, then

$$\mu_{rs} - \bar{\mu}_{r\cdot} - \bar{\mu}_{\cdot s} + \bar{\mu}_{\cdot\cdot} = 0 \qquad (11)$$

If all $n_{ij} > 0$, then $G_3 \equiv 0$; that is with no missing cells SS(3) is testing $H_0: \bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$ without any conditions. This is different from

G_1 and G_2 in that their conditions are always present even if all $n_{ij} = n$.

Hence, for the data in Table 2, SS(3) is simply testing H_0: $\bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$, while SS(2) and SS(1) are testing the same hypothesis but under conditions of i) no IJ interaction and no J effect and ii) no IJ interaction, respectively. Clearly, there is no evidence to support these conditions. Thus, the choice would be the Type III sum of squares.

If we consider the data in Table 4 and the corresponding analysis in Table 5, we see that it is reasonable to assume that the interaction is small, if not zero. Since SS(2) is testing H_0: $\bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$ under the condition of no IJ interaction, it would seem that this is to be preferred over SS(3) while is testing the same hypothesis, but in an interaction model.

This is an example of where the use of the algebraic formulation approach might prove to be difficult for some people. The hypothesis being tested by SS(2) is given by (7). However, if there is no interaction among the cell means, then (7) reduces to testing H_0: $\bar{\mu}_{i\cdot} = \bar{\mu}_{i\cdot}$, which is not at all obvious. By using the characterization approach, it is quite clear that SS(2) is justified in this example.

Table 4

Hypothetical Data Set No. 2

		J = 1	J = 2	J = 3
I	1	1.0 2.0 1.5 2.5 3.0	8.5 11.5	8.0
	2	.5 1.5	10.0	3.5 10.5 4.5 9.5 7.0
	3	2.7 3.3	9.7 13.4 12.3 11.0 8.6	8.2 9.3 9.5

Table 5

F Ratios and Probabilities
For Data Set No. 2

Source	df	Type I F	Type I PR	Type II F	Type II PR	Type III(IV) F	Type III(IV) PR
I	2	11.2	.001	1.8	.200	1.4	.284
J	2	37.2	.000	37.2	.000	33.9	.000
IJ	4	.05	.995	.05	.995	.05	.995

As a final example, let us consider the data in Table 6 and the analysis in Table 7. Here G_1 and G_2 are as before. However, since n_{11} and n_{23} are zero, then using (11), G_3 becomes

$$G_3: \quad \left\{ \begin{array}{l} \mu_{11} - \bar{\mu}_{1\cdot} - \bar{\mu}_{\cdot 1} + \bar{\mu}_{\cdot\cdot} = 0 \\ \mu_{23} - \bar{\mu}_{2\cdot} - \bar{\mu}_{\cdot 3} + \bar{\mu}_{\cdot\cdot} = 0 \end{array} \right\} \qquad (12)$$

Thus SS(1), SS(2) and SS(3) are all testing H_0 : $\bar{\mu}_{i\cdot} = \bar{\mu}_{i'\cdot}$ but under their different conditions. There appears to be no justification for assuming no J and no IJ effects. Thus, SS(1) and SS(2) are not feasible choices. Likewise, there is little justification for assuming no (1,1) and (2,3) interaction simply because $n_{11} = n_{23} = 0$. In fact, there is strong evidence that there is interaction for all of the other cells. The presence or absence of interaction is a property of the population parameters and is in no way connected to sample size. Thus, we feel that the prudent choice has to be "none of the above", i.e. do not use any of the three types of sums of squares. The user should construct meaningful hypotheses and test them using the *Estimate* or *Contract* statements in GLM.

Using the algebraic approach, we see that SS(3) is testing (after a transformation)

$$H_0: \left\{ \begin{array}{l} \mu_{12} + 2\mu_{13} + \mu_{21} = \mu_{22} + \mu_{31} + 2\mu_{33} \\ 2\mu_{13} + 7\mu_{21} + 2\mu_{22} + 6\mu_{23} = 2\mu_{12} + \\ 7\mu_{31} + 8\mu_{33} \end{array} \right\} \qquad (13)$$

Thus, by considering H_0 as given by (13), the user would come to the same conclusion that the hypothesis tested by SS(3) is not appropriate. However, we caution again that, when utilizing the algebraic approach, the user does not reject a type without considering other forms of the hypothesis. It might be that what looks quite messy will become clear upon a non-singular transformation.

Table 6

Hypothetical Data No. 3

		J = 1	J = 2	J = 3
I	1		11.9 12.1	6.0
	2	3.9 4.1	6.0	9.8 9.9 10.2 10.1 10.0
	3	3.9 4.1		13.9 14.0 14.1

Table 7

F Ratios and Probabilities
For Data Set No. 3

Source	df	Type I F	Type I PR	Type II F	Type II PR	Type III(IV) F	Type III(IV) PR
I	2	400	.000	726	.000	790	.000
J	2	3704	.000	3704	.000	2730	.000
IJ	4	1296	.000	1296	.000	1296	.000

SAS TYPE IV

We conclude with a discussion of the SAS Type IV sum of squares. If all $n_{ij} > 0$, then Types III and IV are identical and nothing more need be said. However, if some $n_{ij} = 0$ then Types III and IV are different. The hypotheses tested by Type IV are, in general, quite simple in nature and easy to interpret. Type IV strives to find "balanced" subsets to analyze. For example, using the data in Table 6, Type IV is testing, for *I*;

$$H_0: \quad \begin{matrix} \mu_{13} = \mu_{33} \\ \mu_{21} + \mu_{23} = \mu_{31} + \mu_{33} \end{matrix}$$

One major drawback to Type IV is that different hypotheses are tested by just renaming the levels of the factors. This lack of uniqueness and the fact that the hypothesis are, in general, not of interest lends us to recommend the users test their own hypothesis rather than relying on Type IV.

REFERENCES

GLM. SAS USER'S Guide - 1979.

Hocking, R. R. and Speed, F. M. - 1975. "A Full Rank Analysis of Some Linear Model Problems." JASA. 70

Hocking, R. R., Speed, F. M., and Coleman, A. T. - 1980. "Hypotheses to be Tested with Unbalanced Data." Communications in Statistics.

Searle, S. R. - 1979. "Relationships between the Estimable Functions of SAS GLM Output for Unbalanced Data and the Hypothesis Tested by Traditional-Style F-Statistics." SUGI Proceedings.

Speed, F. M. - 1977. "SAS-GLM Procedure and Analysis of Variance." SUGI Proceedings.

Speed, F. M. and Hocking, R. R. - 1976. "The Use of the R( )-notation with Unbalanced Data." American Statistician.