SOME DISTINCTIVE FEATURES OF OUTPUT FROM STATISTICAL COMPUTING PACKAGES FOR ANALYSIS OF COVARIANCE

by

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

BU-761-M February, 1982

Abstract

A summary is given of certain features of the output from routines in BMDP, GENSTAT, RUMMAGE, SAS and SPSS that perform analysis of covariance calculations.

1. INTRODUCTION

Output from statistical computing packages for analysis of covariance is not always labeled unequivocally. For example, we find for a completely random design (one-way classification) with one covariate and unequal numbers of observations in the classes, that the sum of squares label "mean" is given to at least five different values (Tables 3 and 4) computed by a current array of available computing packages. Ascertaining precisely what it is that an output label means, i.e., exactly what has been computed, is therefore an important problem for the statistician who wants to use a package for data analysis. One way of doing this is to process small, hypothetical data sets through the package and verify every output value from hand calculations. Verification consists of (perhaps repetitively) speculating on the nature of the package output values and checking that one's final conjecture is upheld for subsequent data sets. This procedure is not fully rigorous. However, since program documentation is often erroneous by omission or in fact, and since reading program code is a task beyond most of us, the method appears to be a useful and manageable one.

Paper No. BU-761-M in the Biometrics Unit, Cornell University.

The variety of calculations available in the analysis of variance and covariance of unbalanced data (having unequal numbers of observations in the subclasses) is quite large and yet, for balanced data (having equal numbers of observations in the subclasses), they nearly all reduce to the same set of well known calculations. (The simplest example of this is that $\sum_{i=1}^{a} n_i \bar{y}_{i.} / \sum_{i=1}^{a} n_i$ and $\sum_{i=1}^{a} \bar{y}_{i.}/a$ are different for unbalanced data but the same for balanced data.) Processing small sets of hypothetical, unbalanced data through a statistical computing package can therefore be a fruitful way of ascertaining not only what it is that any particular package calculates, but also what the differences between packages are. With these ends in mind the basis of this paper is the analysis of data from a completely randomized design (1-way classification) with one covariate. We use the data sets of Table 1, but implications drawn here extend to data from more complex situations.

(SHOW TABLE 1)

Our primary purpose is to highlight, illustrate, and explain differences that exist among different packages. Similarities among packages do, of course, abound, especially in output such as cell means, numbers of observations, predicted values, residuals, and so on. Our purpose is not to describe all output features of the packages, but rather to highlight some features that are distinctively different from one package to another. This is done, not for the purpose of suggesting what is preferable, but for the more important reason of emphasizing that the packages do compute different things, and for providing information on what some of those differences are. Statisticians using computer package output need to know that apparently-similar labels on output from different packages do not always mean the same thing. The conclusion is not necessarily that one package is better than another in any sense, but that packages do differ from one

another in certain prescribed ways. Only when one knows and understands the differences can one make appropriate decisions about which output values are suitable for the needs at hand.

2. LINEAR MODELS

Consider a completely randomized design with a classes and n_i observations y_{ij} in the i'th class, for $j = 1, 2, \cdots, n_i$ and $i = 1, \cdots, a$, with $n_i > 0$. We represent the corresponding observation on the covariate as z_{ij}. Then, with E representing expectation over repeated sampling, a without-covariate model for y_{ij} is

$$E(y_{ij}) = \mu + \alpha_i \tag{1}$$

where α_i is the effect due to the i'th class. And the covariate can be incorporated in the model so as to have either

$$E(y_{ij}) = \mu + \alpha_i + bz_{ij} \tag{2}$$

or

$$E(y_{ij}) = \mu + \alpha_i + b(z_{ij} - \bar{z}_{..}) , \tag{3}$$

where $\bar{z}_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n_i} z_{ij}/N$ for $N = n_. = \sum_{i=1}^{a} n_i$. (The symbol z is used for the covariate rather than the traditional x so that when writing a covariate model in matrix notation $\underset{\sim}{X}$ can be retained as the incidence matrix for terms other than the covariate.)

None of the models (1), (2) or (3) has a restriction of the form

$$\sum_{i=1}^{a} \alpha_i = 0 \tag{4}$$

which can be attached to a model such as (1). So as to distinguish (1), (2) and (3) from their counterparts with (4) included, we refer to (1), (2) and (3) as

unrestricted models. We call (4) a Σ-restriction, and when it is used in conjunction with (1), (2) or (3) we call the resulting model a Σ-restricted model and write, for example, (1) and (4) as

$$E(y_{ij}) = \dot{\mu} + \dot{\alpha}_i \quad \text{with} \quad \sum_{i=1}^{a} \dot{\alpha}_i = 0 \,. \tag{5}$$

The notation of putting a dot above each parameter in (5) is done to clearly distinguish (5) as being a restricted model, in contrast to (1) which is an unrestricted model. Similarly

$$E(y_{ij}) = \dot{\mu} + \dot{\alpha}_i + \dot{b}z_{ij} \quad \text{with} \quad \sum_{i=1}^{a} \dot{\alpha}_i = 0 \tag{6}$$

and

$$E(y_{ij}) = \dot{\mu} + \dot{\alpha}_i + \dot{b}(z_{ij} - \bar{z}_{..}) \quad \text{with} \quad \sum_{i=1}^{a} \dot{\alpha}_i = 0 \tag{7}$$

are the Σ-restricted counterparts of (2) and (3). Although in (6) and (7) the restriction $\sum_{i=1}^{a} \dot{\alpha}_i = 0$ applies only to the $\dot{\alpha}_i$'s, in which notation the dot above the α emphasizes that $\dot{\alpha}_i$ of (6), for example, is to be distinguished from α_i of (2), the dot notation covers all parameters of the model so as to distinguish the whole of a restricted model from its unrestricted counterpart.

A different restriction which is sometimes used in place of the Σ-restriction is

$$\sum_{i=1}^{a} n_i \dot{\alpha}_i = 0 \,, \tag{8}$$

which we call the Σn-restriction. Used in place of $\sum_{i=1}^{a} \dot{\alpha}_i = 0$ in (5), (6) and (7) it defines a third set of models (for which we should perhaps use a further distinguishing symbol such as $\ddot{\alpha}_i$ — but we shall not).

3. CELL MEANS MODELS

Models (5), (6) and (7) are simply one way of overcoming the inherent difficulties of the overparameterized models (1), (2) and (3). A straightforward way of avoiding these difficulties is simply to use full-rank, cell means models from the outset:

$$E(y_{ij}) = \mu_i \,, \tag{9}$$

$$E(y_{ij}) = \mu_i + bz_{ij} \,, \tag{10}$$

and

$$E(y_{ij}) = \mu_i + b(z_{ij} - \bar{z}_{..}) \tag{11}$$

in place, respectively, of either (1), (2) and (3), or (5), (6) and (7). Certainly (9), (10) and (11) are much easier to understand than either the overparameterized models (1), (2) and (3) with the accompanying need for estimability considerations, or the restricted models such as (5), (6) and (7) with their need for knowing the effects of the restrictions. In this connection, the reader is referred to Hocking and Speed (1975) and Speed et al. (1978) for excellent discussion of cell means models.

Despite the advantages of cell means models (especially for unbalanced data), many current computer packages, whilst being able to handle cell means models very easily, are nevertheless designed with the long-standing overparameterized models in mind. Furthermore, they are certainly used this way on numerous occasions by users who are ofttimes more familiar with overparameterized models than with cell means models. Because the latter are, indeed, easier to understand, we prefer them; but because some computer packages contain features oriented to overparameterized and to restricted models our descriptions use all three kinds of model. We believe the reader will find that understanding some of these features is made easier by thinking in terms of cell means models.

4. COMPUTER PACKAGES

The computer packages used in this study are listed in Table 2. Three, P1V,

(SHOW TABLE 2)

P2V and P4V, are from the BMDP package of the Department of Biomathematics of the University of California at Los Angeles, RUMMAGE is from the Department of Statistics at Brigham Young University in Provo, Utah, GENSTAT ANOVA is part of the GENSTAT package from Rothamsted Experimental Station of Harpenden, England, SAS GLM is the general linear model routine of the Statistical Analysis System of the SAS Institute in Raleigh, North Carolina, and SAS HARVEY is a user-supported procedure therein; and SPSS ANOVA and MANOVA are two routines in the statistical programs for social scientists emanating from SPSS Inc., 444 N. Michigan Avenue, Chicago, Illinois.

Table 2 summarizes several characteristics of these packages. It shows the form in which covariates are handled (either as z_{ij} or as $\delta_{ij} = z_{ij} - \bar{z}_{..}$), the restrictions used (if any), and the values given in the output as solutions to the normal equations for the one-way classification. Those packages for which no solutions are shown do not, in general, have solutions to normal equations as part of their output. Table 2 also shows which packages have adjusted class (treatment) means $A_i = \bar{y}_{i.} + \hat{b}(\bar{z}_{i.} - \bar{z}_{..})$ among their output, and whether or not the package handles intraclass slopes as in the model $E(y_{ij}) = \mu + \alpha_i + b_i z_{ij}$.

An immediate reaction to Table 2 is that no two of these nine computer packages are exactly the same. No doubt other features of the packages could have been listed, which would perhaps have made some packages look more alike, such as the output of means, standard deviations, analyses of variance and so on. But the object of Table 2 is to show features where there are important differences between the packages.

5. SUMS OF SQUARES DUE TO THE MEAN

In fitting the general linear model $E(\underset{\sim}{y}) = \underset{\sim\sim}{X}\underset{\sim}{b}$, the normal equations are $\underset{\sim}{X}'\underset{\sim\sim}{X}\underset{\sim}{b}^{o} = \underset{\sim}{X}'\underset{\sim}{y}$ where $\underset{\sim}{b}^{o}$ is any solution thereof, and the sum of squares due to fitting the model is

$$R(\underset{\sim}{b}) = \underset{\sim}{b}^{o\prime}\underset{\sim}{X}'\underset{\sim}{y} \tag{12}$$

which is the inner product of the vector of solutions and vector of right-hand-sides of the normal equations. Searle et al. (1981) call (12) the R-algorithm. It plays an important role in understanding how certain sums of squares in restricted models are calculated.

An extension to $R(\underset{\sim}{b})$ arises from partitioning $\underset{\sim\sim}{X}\underset{\sim}{b}$ so as to have $E(\underset{\sim}{y}) = \underset{\sim}{X}_1\underset{\sim}{b}_1 + \underset{\sim}{X}_2\underset{\sim}{b}_2$. Then the sum of squares for $\underset{\sim}{b}_1$ adjusted for $\underset{\sim}{b}_2$ is

$$R(\underset{\sim}{b}_1|\underset{\sim}{b}_2) = R(\underset{\sim}{b}_1\underset{\sim}{b}_2) - R(\underset{\sim}{b}_2) \ . \tag{13}$$

In the right-hand side of (13), each term can be calculated from (12) using first $[\underset{\sim}{X}_1 \ \underset{\sim}{X}_2]$ as $\underset{\sim}{X}$ and $[\underset{\sim}{b}_1' \ \underset{\sim}{b}_2']'$ as $\underset{\sim}{b}$ for calculating $R(\underset{\sim}{b}_1,\underset{\sim}{b}_2)$, and then $\underset{\sim}{X}_2$ as $\underset{\sim}{X}$ and $\underset{\sim}{b}_2$ as $\underset{\sim}{b}$ for $R(\underset{\sim}{b}_2)$.

As the simplest illustration of complications that can arise with unbalanced data when using restricted models as part of the computing procedure, we consider what might loosely be called sums of squares due to the mean: we find that at least five different expressions can come under this rubric. True, "testing the mean" may often not be of much practical interest, although one case where it can be important is, to quote from Bryce (1982), in "testing the gain score in a pre-post type design where a pretest is made, the treatments are applied, and the post-test follows." Nevertheless, the complications we illustrate here are symptomatic of those that can arise from the same kind of causes in sums of squares that are more involved than just those concerned with the mean.

Consider

$$R(\mu|\underset{\sim}{\alpha}) = R(\mu,\underset{\sim}{\alpha}) - R(\underset{\sim}{\alpha}) \ . \tag{14}$$

From fitting (1) we get $R(\mu,\underset{\sim}{\alpha}) = \Sigma n_i \bar{y}_{i.}^2$, and the sum of squares for fitting $E(y_{ij}) = \alpha_i$ is $R(\underset{\sim}{\alpha}) = \Sigma n_i \bar{y}_{i.}^2$. Hence (14) is identically zero. This is an example of Nelder's (1974) concept of marginality, and as such $R(\mu|\underset{\sim}{\alpha})$ is a sum of squares that has no use. But certain variants of it in restricted models are not zero. For example, consider $R(\dot{\mu}|\dot{\underset{\sim}{\alpha}})_\Sigma$, the same kind of sum of squares as (14) but for the Σ-restricted model (5). Using the notation of Searle et al. (1981), we denote it as $R^*(\mu|\dot{\underset{\sim}{\alpha}})_\Sigma$ so that in writing

$$R^*(\dot{\mu}|\dot{\underset{\sim}{\alpha}})_\Sigma = R^*(\dot{\mu},\dot{\underset{\sim}{\alpha}})_\Sigma - R^*(\dot{\underset{\sim}{\alpha}})_\Sigma \tag{15}$$

analogous to (14), the asterisk emphasizes that the Σ-restrictions of model (5) which are implicit in the first term on the right-hand side of (15) are also being used in the second term there.

Example. For Data 1b of Table 1 the model equation and normal equations for the Σ-restricted model (5) without covariate are, respectively,

$$E\begin{bmatrix}74\\68\\77\\76\\80\\85\\93\end{bmatrix} = \begin{bmatrix}1&1&0\\1&1&0\\1&1&0\\1&0&1\\1&0&1\\1&-1&-1\\1&-1&-1\end{bmatrix}\begin{bmatrix}\dot{\mu}\\\dot{\alpha}_1\\\dot{\alpha}_2\end{bmatrix}, \quad \text{and} \quad \begin{bmatrix}7&1&0\\1&5&2\\0&2&4\end{bmatrix}\begin{bmatrix}\hat{\dot{\mu}}\\\hat{\dot{\alpha}}_1\\\hat{\dot{\alpha}}_2\end{bmatrix} = \begin{bmatrix}553\\41\\-22\end{bmatrix} . \tag{16}$$

Solutions to the normal equations are $\hat{\dot{\mu}} = 80$, $\hat{\dot{\alpha}}_1 = -7$ and $\hat{\dot{\alpha}}_2 = -2$. Hence, by (12)

$$R^*(\dot{\mu},\dot{\underset{\sim}{\alpha}})_\Sigma = 80(553) - 7(41) - 2(-22) = 43,997 \ .$$

This equals $R(\mu,\underset{\sim}{\alpha})$, as would be expected because the $(\dot{\mu},\dot{\underset{\sim}{\alpha}})$-model is just a re-parameterization of the $(\mu,\underset{\sim}{\alpha})$-model. But the asterisk on $R^*(\dot{\underset{\sim}{\alpha}})_\Sigma$ in (15) means, not that we start with an α-model and (if necessary) reparameterize it using the Σ-restriction, but that in fact we take just the $\dot{\underset{\sim}{\alpha}}$ part of the already-restricted $(\mu,\dot{\underset{\sim}{\alpha}})$-model, i.e., just the $\dot{\underset{\sim}{\alpha}}$ parts of equations (16), and apply (12) to them. This means using just

$$\begin{bmatrix} 5 & 2 \\ 2 & 4 \end{bmatrix}\begin{bmatrix} \hat{\dot{\alpha}}_1 \\ \hat{\dot{\alpha}}_2 \end{bmatrix} = \begin{bmatrix} 41 \\ -22 \end{bmatrix} \quad \text{with solution} \quad \begin{bmatrix} 13 \\ -12 \end{bmatrix},$$

and in applying (12) getting $R^*(\dot{\underset{\sim}{\alpha}})_\Sigma = 13(41) - 12(-22) = 797$. Then (15) gives

$$R^*(\dot{\mu}|\dot{\underset{\sim}{\alpha}})_\Sigma = 43,997 - 797 = 43,200 . \tag{17}$$

This result is important in several respects. First, it is not identically zero as is $R(\mu|\underset{\sim}{\alpha})$ of (14). Second, it differs from $R(\mu) = N\bar{y}^2_{\cdot\cdot} = 7(79^2) = 43,687$. Thus the sum of squares labeled "mean" from one computer package may be $R(\mu)$ and from another it may be $R^*(\dot{\mu}|\dot{\underset{\sim}{\alpha}})_\Sigma$, and by the labeling of the output, it may be difficult to distinguish one from another, especially since the distinction does not make itself evident with balanced data; only with unbalanced data does it become readily apparent. Finally, this is the simplest example of where an $R(\cdot|\cdot)$ in an unrestricted model is identically zero and its counterpart $R^*(\cdot|\cdot)_\Sigma$ in a restricted model is not. Another example occurs in the two-way crossed classification with main effect factors denoted by α and β and their interaction by γ. Then, like (14) and (15), respectively,

$$R(\underset{\sim}{\alpha}|\mu,\underset{\sim}{\beta},\underset{\sim}{\gamma}) \equiv 0 \quad \text{whereas} \quad R^*(\dot{\underset{\sim}{\alpha}}|\dot{\mu},\dot{\underset{\sim}{\beta}},\dot{\underset{\sim}{\gamma}})_\Sigma \neq 0 . \tag{18}$$

Indeed, as Searle *et al.* (1981) show, when all cells contain data, $R^*(\dot{\underset{\sim}{\alpha}}|\dot{\mu},\dot{\underset{\sim}{\beta}},\dot{\underset{\sim}{\gamma}})_\Sigma$ is the sum of squares for the α-factor in the weighted squares of means analysis.

Table 3 shows an array of values like that in (17), for the two sets of data in Table 1, using Σ-restrictions and Σn-restrictions, both without a covariate, and with a covariate used either as z or as $\delta = z - \bar{z}_{..}$. The illustration of

(SHOW TABLE 3)

line 3 of Table 3 between (16) and (17) can be repeated for lines 4 through 8, where line 4 is for using the Σn-restriction on the no-covariate model, and lines 5 through 8 show the four cases of combining each of the Σ- and Σn-restrictions with each way of treating the covariate, as $b(z_{ij} - \bar{z}_{..})$ and as bz_{ij}.

A noticeable feature of Table 3 is, apart from line 2, the five different values in its last column, for the unbalanced data; i.e., five different values that might be called a sum of squares due to the mean. In contrast, there are only two different values for the balanced data.

6. HYPOTHESES

We have illustrated how, for the same unbalanced data set, different values can be calculated for the label "sum of squares due to the mean". Knowing how these values are calculated unfortunately provides no real understanding of their usefulness (if any). For this reason, Table 4 shows the hypothesis tested were each sum of squares to be used as the numerator of an F-statistic under the customary normality assumptions. The sums of squares in lines 1 through 4 are

(SHOW TABLE 4)

for models without covariate and the corresponding hypotheses are in terms of the unrestricted model $E(y_{ij}) = \mu_i = \mu + \alpha_i$; and for lines 5 through 8, for sums of squares for models with covariate, the hypotheses are in terms of the unrestricted model $E(y_{ij}) = \mu_i + bz_{ij} = \mu + \alpha_i + bz_{ij}$. In each of the six restricted models

of lines 3 - 8, the hypothesis being tested is $H : \dot{\mu} = 0$. But, because from one restricted model to another $\dot{\mu}$ is not always the same function of the parameters of an unrestricted model, the hypothesis $H : \dot{\mu} = 0$ which applies in each restricted model takes on different forms when expressed in terms of unrestricted parameters. For example, for line 3, $\dot{\mu} = \Sigma\mu_i/a = \mu + \Sigma\alpha_i/a$, and so $H : \dot{\mu} = 0$ is equivalent to $H : \Sigma\mu_i/a = 0$.

The hypotheses shown in Table 3 may or may not be of any practical value in themselves. Nevertheless, they do provide a means of understanding what the sums of squares can be used for, and what the distinctions between them are.

7. EXPLICIT FORMULAE

The R- and R^*-notation used in Tables 3 and 4 generalizes easily from the sums of squares due to the mean shown in those tables to sums of squares for any factor or covariate in a multi-factor situation. An example is $R^*(\dot{\alpha}|\dot{\mu},\dot{\beta},\dot{\gamma})_\Sigma$ shown in (18) and discussed at some length in Searle et al. (1981). Even though in being either an $R(\cdot)$ or an $R^*(\ )$, such sums of squares can always be calculated as illustrated between (16) and (17), the corresponding quadratic functions of the observations are not, in general, easily derived for unbalanced data. However, for the sums of squares of Tables 3 and 4 these functions are readily available and are shown in Table 4. They can be verified as being numerator sums of squares for F-statistics for testing the hypotheses shown in Table 4. Their availability permits ready observation that for balanced data (when $\bar{y} = \Sigma n_i\bar{y}_{i.}/N$ and $\tilde{\bar{y}} = \Sigma\bar{y}_{i.}/a$ are equal, as are $\bar{z}$ and $\tilde{\bar{z}}$, and when $h = \Sigma 1/a^2 n_i = 1/N$) the formulae on lines 7 and 8 become equal, and the rest reduce to $N\bar{y}^2$.

Conclusion

Computer output for analysis of covariance is not all that (by its labeling) it is made out to be. Values with labels that appear to be the same can be quite different because they do in fact represent different calculations.

Acknowledgments

Grateful thanks for providing partial support of the work reported here go to Bristol Laboratories, Syracuse, New York; Ciba-Geigy Corporation, Summit, New Jersey; McNeil Pharmaceutical, Spring House, Pennsylvania; Searle Research and Development, Chicago, Illinois; Sterling-Winthrop Research Institute, Rensselaer, New York; and the Upjohn Company, Kalamazoo, Michigan. Without this support the work would not have been accomplished. Editorial and referees' comments on an early draft of the paper led to appreciable improvements.

References

Bryce, G. R. (1982). Personal communication.

Hocking, R. R. and Speed, F. M. (1975). A full rank analysis of some linear model problems. Journal of the American Statistical Association, 70, 706-712.

Nelder, J. A. (1974). A reformulation of linear models. Journal of the Royal Statistical Society (A), 140, 48-77.

Searle, S. R., Speed, F. M. and Henderson, H. V. (1981). Some computational and model equivalences in analyses of variance of unequal-subclass-numbers data. The American Statistician, 35, 16-33.

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, 51-54.

Speed, F. M., Hocking, R. R. and Hackney, O. P. (1978). Methods of analysis of linear models with unbalanced data. Journal of the American Statistical Association, 73, 105-112.

TABLE 1. TWO SETS OF HYPOTHETICAL DATA FOR THE 1-WAY CLASSIFICATION

	Balanced Data			
	y	z	y	z
	29	2	39	3
Data 1a	30	5	40	11
	31	8	41	7

	Unbalanced Data					
	y	z	y	z	y	z
	74	3	76	2	85	4
Data 1b	68	4	80	4	93	6
	77	2				

TABLE 2. DISTINCTIVE FEATURES OF ANALYSIS OF COVARIANCE OUTPUT FROM SEVERAL COMPUTING PACKAGES

Note: In the 1-way classification, for either $E(y_{ij}) = \mu + \alpha_i + bz_{ij}$ or $E(y_{ij}) = \mu + \alpha_i + b(z_{ij} - \bar{z}_{..})$, all packages compute

$$\hat{b} = \Sigma\Sigma(y_{ij} - \bar{y})(z_{ij} - \bar{z})/\Sigma\Sigma(z_{ij} - \bar{z})^2 .$$

$\bar{y}$ and $\bar{z}$ denote $\bar{y}_{..}$ and $\bar{z}_{..}$, respectively.

Computing Package	Covariate $z \equiv z_{ij}$ $\delta \equiv z_{ij} - \bar{z}$	Restrictions[1]	Solutions to normal equations for the 1-way classification		Adjusted means[2]	Intra-class slopes[3]
			For μ	For α_i		
BMD: P1V	z	Σ	–	–	A_i	Yes
P2V	z	Σ	–	–	Some[4]	No
P4V	z	Σ or Σn[5]	–	–	No	No
RUMMAGE	z or δ[6]	Σ	$\bar{A} - b\bar{z}$ or $\bar{A}$[6]	c[7]	A_i	Yes
GENSTAT ANOVA	δ	Σn	$\bar{y}$	$A_i - \bar{y}$	A_i	No
SAS: GLM	z	None	$\bar{y}_{a.} - \hat{b}\bar{z}_{a.}$	$A_i - A_a$	A_i[8]	Yes
HARVEY	δ	Σ	$\bar{A}$	$A_i - \bar{A}$	A_i[8]	No
SPSS: ANOVA	δ	Σn	$\bar{y}$	$A_i - \bar{y}$	No	No
MANOVA	z	Σn[9]	$\bar{A} - \hat{b}\bar{z}$	$A_i - \bar{A}$	$A_i + d$[10]	Yes

1/ Σ denotes Σ-restrictions: $\Sigma\alpha_i = 0$; Σn denotes Σn-restrictions: $\Sigma n_i\alpha_i = 0$.

2/ Adjusted mean for class i is $A_i = \bar{y}_{i.} - \hat{b}(\bar{z}_{i.} - \bar{z})$; and $\bar{A} = \Sigma A_i/a$.

3/ Model: $E(y_{ij}) = \mu + \alpha_i + b_i z_{ij}$.

4/ Calculated only for models with highest-order interactions and all cells filled.

5/ User-specified cell weights of unity (n_i) are equivalent to Σ (Σn).

6/ δ can be used only when $\bar{z}$ is available as input.

7/ Estimates of user-supplied contrasts among α_i's.

8/ Labeled "Least Squares Means"; see Searle et al. (1980).

9/ Specifying METHOD = SSTYPE(UNIQUE) uses Σ; default is Σn.

10/ $d = \hat{b}(\Sigma\bar{z}_{i.}/a - \bar{z})$.

TABLE 3. SUMS OF SQUARES THAT COULD BE CALLED "DUE TO THE MEAN"

Notation: $\bar{y} \equiv \bar{y}_{..}$ and $\bar{z} \equiv \bar{z}_{..}$

Sum of Squares	Computer Package [1]	Calculated Values: Balanced Data, Data 1a	Calculated Values: Unbalanced Data, Data 1b
1. $R(\mu)$		$7350 = N\bar{y}^2$	$43,687 = N\bar{y}^2$
2. $R(\mu\|\alpha) = R(\mu,\alpha) - R(\alpha) = 0$	–	0	0
3. $R^*(\dot{\mu}\|\dot{\alpha})_\Sigma$ [2]	BMDP2V, 4V [4]	7350	$43,200 \neq N\bar{y}^2$
4. $R^*(\dot{\mu}\|\dot{\alpha})_{\Sigma n}$	BMDP4V [4]	7350	43,687
5. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_\delta)_\Sigma$ [3]	RUMMAGE SAS HARVEY	7350	$42,712\frac{49}{66} \neq N\bar{y}^2$
6. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_\delta)_{\Sigma n}$		7350	43,687
7. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_\Sigma$ [3]	BMDP2V, 4V [4] RUMMAGE	$1288\frac{62}{133} \neq N\bar{y}^2$	$2,557\frac{59}{86} \neq N\bar{y}^2$
8. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_{\Sigma n}$	BMDP4V [4] SPSS MANOVA [5]	$1288\frac{62}{133} \neq N\bar{y}^2$	$2,627\frac{55}{58} \neq N\bar{y}^2$

1/ BMDP1V, 2V (since 1981) and 4V, and GENSTAT and SAS GLM produce no sum of squares "due to the mean".

2/ Σ denotes Σ-restrictions: $\Sigma\dot{\alpha}_i = 0$; Σn denotes Σn-restrictions: $\Sigma n_i\dot{\alpha}_i = 0$.

3/ b_δ represents using $z_{ij} - \bar{z}$ as the covariate, and b_z represents using z_{ij}.

4/ User-specified cell weights of unity (n_i) in BMDP4V produces lines 3 and 7 (lines 4 and 8).

5/ Specifying METHOD = SSTYPE(UNIQUE) produces line 7.

TABLE 4. SUMS OF SQUARES (from Table 3), EXPLICIT FORMULA AND ASSOCIATED HYPOTHESIS

Sum of Squares		Associated
R( )-form	Explicit Formula[1]	Hypothesis[2]
		Model: $E(y_{ij}) = \mu_i = \mu + \alpha_i$
1. $R(\mu)$	$N\bar{y}^2$	$H: \Sigma n_i\mu_i/N = 0$
2. $R(\mu\|\alpha)$	0	None
3. $R^*(\dot{\mu}\|\dot{\alpha})_\Sigma$	$\tilde{y}^2/h$	$H: \Sigma\mu_i/a = 0$
4. $R^*(\dot{\mu}\|\dot{\alpha})_{\Sigma n}$	$N\bar{y}^2$	$H: \Sigma n_i\mu_i/N = 0$
		Model: $E(y_{ij}) = \mu_i + bz_{ij}$
5. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_\delta)_\Sigma$	$[\tilde{y} - \hat{b}(\tilde{z} - \bar{z})]^2/[h + (\tilde{z} - \bar{z})^2/S_z]$	$H: \Sigma\mu_i/a + b\bar{z} = 0$
6. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_\delta)_{\Sigma n}$	$N\bar{y}^2$	$H: \Sigma n_i\mu_i/N + b\bar{z} = 0$
7. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_\Sigma$	$(\tilde{y} - \hat{b}\tilde{z})^2/(h + \tilde{z}^2/S_z)$	$H: \Sigma\mu_i/a = 0$
8. $R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_{\Sigma n}$	$(\bar{y} - \hat{b}\bar{z})^2/(1/N + \bar{z}^2/S_z)$	$H: \Sigma n_i\mu_i/N = 0$

1/ Notation: $\bar{y}$, $\bar{z}$ and $\hat{b}$, as in Table 3.

$\tilde{y} = \Sigma\bar{y}_{i.}/a$ and $\tilde{z} = \Sigma\bar{z}_{i.}/a$.

$S_z = \Sigma\Sigma(z_{ij} - \bar{z})^2$ and $h = [\Sigma(1/n_i)]/a^2$.

2/ Hypothesis tested when sum of squares is used as numerator of an F-statistic.