DISTINCTIVE FEATURES OF OUTPUT FROM STATISTICAL COMPUTING PACKAGES FOR LINEAR MODEL CALCULATIONS ON UNBALANCED DATA

by

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

BU-752-M August, 1981

Abstract

A summary is given of certain features of the output from routines in BMDP, GENSTAT, SAS and SPSS that perform analysis-of-variance style calculations on unbalanced data.

1. Introduction

Computations for the analysis of variance, the estimation of variance components and the analysis of covariance of unbalanced data (having unequal numbers of observations in the subclasses) can be quite complicated. Fortunately, most widely-used statistical computing packages contain routines for doing many of these calculations. But, unfortunately, much of the output from these routines is not labeled unequivocally, and statisticians are often confused as to the utility of some computed values. The Annotated Computer Output project at Cornell is directed towards alleviating this situation.

A series of small hypothetical data sets of varying complexity with regard to unbalanced data, empty cells, covariates, and fixed and random effects are being processed on the analysis-of-variance style routines of BMDP, GENSTAT, SAS, SPSS and other packages. Output from each routine is then annotated with detailed notes and descriptions (designed for statisticians) explaining just exactly what

---

Paper No. BU-752-M in the Biometrics Unit.

it is that has been computed. The resulting document, of notes, data listings, hand-calculated results and annotated output for several data sets is called an Annotated Computer Output – ACO. The ACO for each routine is, in effect, a small manual describing what the routine calculates. These manuals are available from the Biometrics Unit, at Cornell University, for the routines shown in Table 1.

(SHOW TABLE 1)

2. Why the need for an ACO?

Although the ACO seems to be a useful idea, the need for it is the outcome of a curious set of circumstances. Statisticians should be seeing to it that computer packages compute only what statisticians want. Instead, to some extent, statisticians are having to figure out what it is that packages are computing. How has this situation arisen? Possibly the scenario has been somewhat as follows.

In the initial phases of developing a package, a statistician instructs a programmer what to program; or maybe even the statistician and programmer are one and the same person. But as a package progresses and expands, programmers learn, for example, that small changes here and there will, say, convert a program originally designed for a randomized complete block into one suitable for a many-factored factorial design; or matrix manipulations for fitting constants to 2-way cross classified data can easily be looped to handle many classifications; or regression can be adapted to analysis of variance. The catch is that in executing such programming generalizations on data, quite an array of subtleties can arise. Generally they are subtleties that are especially pertinent when the computations are applied to unbalanced data. This leads to the situation of then having calculated values among computer output that have innocent sounding names (like sum of squares due to the mean), but which are unrecognizable and unfamiliar. Worse still, for the same data analyzed on two different routines, output values

which by their labels would appear to be the same thing are, in fact, different values. Statisticians are today finding themselves more and more in this predicament. Table 2 shows examples of this.

3. Sums of squares "due to the mean"

The run-of-the-mill expression for the sum of squares due to the mean is generally accepted as being $N\bar{y}^2$. Table 2 shows that, for unbalanced data, at least four other values can be computed, some of which are explicitly produced on certain computer output. The occurrence of these values depends upon what kind of restrictions (if any) are imputed to the parameters of the linear model, whether or not a covariate is being used, and if it is, in what form: as $b(z_{ij} - \bar{z}_{..})$ or as bz_{ij}.

(SHOW TABLE 2)

Although tests of hypotheses "for the mean" are not widely used, the hypothesis tested by using each of the sums of squares in Table 2 as the numerator in an F-statistic is a useful means of identifying the utility of these sums of squares. In point of fact, there may be little utility in any of them, but what <u>is</u> important is to know that they are distinctly different and to know what those differences are. The lack of unequivocal and distinctive labeling on computer output provides no way of knowing these differences, which can and do occur not only with sums of squares due to the mean but also with those due to other (and often more important) factors in a model.

4. Models, Restrictions, Covariates and Hypotheses

We summarize features of the different computer packages in terms of the following linear models.

1-way classification

$$E(y_{ij}) = \mu + \alpha_i, \quad \text{for } \begin{cases} i = 1, \cdots, a \\ j = 1, \cdots, n_i \end{cases}$$

1-way classification with covariate

$$E(y_{ij}) = \mu + \alpha_i + b(z_{ij} - \bar{z}_{..}) \quad \text{for } \begin{cases} i = 1, \cdots, a \\ j = 1, \cdots, n_i \end{cases}$$

$$E(y_{ij}) = \mu + \alpha_i + bz_{ij} \quad \text{for } \begin{cases} i = 1, \cdots, a \\ j = 1, \cdots, n_i \end{cases}$$

2-way classification

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \quad \text{for } \begin{cases} i = 1, \cdots, a \\ j = 1, \cdots, b \\ k = 1, \cdots, n_{ij} \geq 0 \end{cases}$$

The preceding models are examples of unrestricted models. In contrast, consideration is also given to restricted models with either the Σ-restrictions such as $\Sigma\alpha_i = 0$ or the Σn-restrictions such as $\Sigma n_i\alpha_i = 0$. To distinguish restricted models from unrestricted models a dot is put above the parameter symbols; e.g.,

$$E(y_{ij}) = \dot{\mu} + \dot{\alpha}_i \quad \text{with} \quad \Sigma\dot{\alpha}_i = 0$$

$$E(y_{ij}) = \dot{\mu} + \dot{\alpha}_i + \dot{b}z_{ij} \quad \text{with} \quad \Sigma\dot{\alpha}_i = 0 \ .$$

The traditional use of a covariate, z_{ij} say, has been in the form $E(y_{ij}) = \mu + \alpha_i + b(z_{ij} - \bar{z}_{..})$, for example. But the form $E(y_{ij}) = \mu + \alpha_i + bz_{ij}$ is nowadays finding favor, and is also used by some computer routines. Table 3 shows which kind of restrictions and which form of the covariate is used by several different computer routines.

(SHOW TABLE 3)

The phrase "associated hypothesis" is used repeatedly herein. Consider $F = R(\cdot)/r\hat{\sigma}^2$ for r being the rank of the sum of squares $R(\cdot)$, and $\hat{\sigma}^2$ the error

mean square after fitting the model; for the model $E(\underset{\sim}{y}) = \underset{\sim}{X}\underset{\sim}{b}$, $\hat{\sigma}^2 = \underset{\sim}{y}'[\underset{\sim}{I} - \underset{\sim}{X}(\underset{\sim}{X}'\underset{\sim}{X})^{-}\underset{\sim}{X}']\underset{\sim}{y}/[N - \text{rank of } \underset{\sim}{X}]$. Whatever the hypothesis is that is tested by F, it is the "associated hypothesis" for $R(\cdot)$; i.e., it is the hypothesis associated with $R(\cdot)$.

5. Analysis of Variance

Distinctive features of analysis of variance output from several routines are summarized.

BMDP2V

No solution to normal equations.

Uses Σ-restrictions.

Sums of squares: each-after-all-others:

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

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

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

For all cells filled, weighted squares of means analysis.

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

SSE

Handles interactions only for all cells filled; aborts with interactions and empty cells.

GENSTAT ANOVA

Designed for balanced data; unbalanced data are generally treated as having missing cells which are estimated and then a balanced data analysis is made.

Yields solution to normal equation using Σn-restrictions.

GENSTAT REGRESSION

Can be manipulated to analyze unbalanced data.

Yields a solution of the normal equations.

Sums of squares: sequential, and each-after-all-others.

SAS GLM

Solves normal equations using generalized inverse.

Gives estimable function $f = \underset{\sim}{\ell}'\underset{\sim}{b}$ corresponding to each sum of squares, such that the associated hypothesis is $H: f_i = 0$ for $i = 1, \cdots, r$ for r linearly independent vectors $\underset{\sim}{\ell}_i$ with $f_i \equiv \underset{\sim}{\ell}_i\underset{\sim}{b}$.

Sums of squares:

Type I (Sequential)	Type II (Each after all others)
$R(\alpha\|\mu)$	$R(\alpha\|\mu,\beta)$
$R(\beta\|\mu,\alpha)$	$R(\beta\|\mu,\alpha)$
$R(\gamma\|\mu,\alpha,\beta)$	$R(\gamma\|\mu,\alpha,\beta)$

Type III (Σ-restrictions)	Type IV (Chooses hypotheses)
$R^*(\alpha\|\mu,\beta,\gamma)_\Sigma$	
$R^*(\dot{\beta}\|\dot{\mu},\dot{\alpha},\dot{\gamma})_\Sigma$	
$R(\gamma\|\mu,\alpha,\beta)$	

Least squares means. (See The American Statistician, 34:216-221)

$$PMM(\alpha_i) = \mu + \alpha_i + \sum_j \beta_j/b + \sum_j \gamma_{ij}/b$$

$EMM(\alpha_i)$ = b.l.u.e. of $PMM(\alpha_i)$.

SAS HARVEY

Uses Σ-restrictions (see notes for BMDP2V).

Normal equations solutions called "CONSTANT ESTIMATES".

Sums of squares calculated by "invert part of the inverse" rule.

(See The American Statistician, 35:16-33, 1981):

For full rank model $E(\underset{\sim}{y}) = \underset{\sim}{X}_1\underset{\sim}{\beta}_1 + \underset{\sim}{X}_2\underset{\sim}{\beta}_2$

$$\begin{bmatrix} \hat{\underset{\sim}{\beta}}_1 \\ \hat{\underset{\sim}{\beta}}_2 \end{bmatrix} = \begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{X}_1 & \underset{\sim}{X}_1'\underset{\sim}{X}_2 \\ \underset{\sim}{X}_2'\underset{\sim}{X}_1 & \underset{\sim}{X}_2'\underset{\sim}{X}_2 \end{bmatrix}^{-1} \begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{y} \\ \underset{\sim}{X}_2'\underset{\sim}{y} \end{bmatrix} \equiv \begin{bmatrix} \underset{\sim}{T}_{11} & \underset{\sim}{T}_{12} \\ \underset{\sim}{T}_{21} & \underset{\sim}{T}_{22} \end{bmatrix} \begin{bmatrix} \underset{\sim}{X}_1'\underset{\sim}{y} \\ \underset{\sim}{X}_2'\underset{\sim}{y} \end{bmatrix}$$

$$R(\underset{\sim}{\beta}_1 | \underset{\sim}{\beta}_2) = \hat{\underset{\sim}{\beta}}_1'\underset{\sim}{T}_{11}^{-1}\hat{\underset{\sim}{\beta}}_1 .$$

Least squares means (see notes for SAS GLM).

With empty cells, there must be (in a row-by-column case) at least one row and one column that have all cells filled; and data must be sequenced so that these are the last row and column. If this requirement is not met, routine aborts.

SPSS ANOVA

Its various options yield sequential and each-after-all-others sums of squares, with and without restrictions.

Output labeled "Multiple Classification Analysis" (based on no-interaction models, even when interactions are part of the model for calculating sums of squares) yields differences between class means, and also elements of solutions to the normal equations based on Σn-restrictions. Ratios of the form $\sqrt{R(\alpha|\mu)/SST_m}$ and $\sqrt{\Sigma n_i \alpha_i^o / SST_m}$ are also computed.

With empty cells, the same requirement is needed as in SAS HARVEY except that data must be sequenced so as to have a full row and column coming first, not last.

6. Analysis of Covariance

All the features described in Section 5, Analysis of Variance, apply here too. All of the routines considered handle single slope models, e.g., $E(y_{ij}) = \mu + \alpha_i + bz_{ij}$ for which the estimated b is calculated:

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

Other features of the routines are indicated in Table 4.

TABLE 4: Features of Output from Computer Routines when Used for Analysis of Covariance

Feature	Computing Routine					
	BMDP1V	BMDP2V	GENSTAT ANOVA	SAS GLM	SAS HARVEY	SPSS ANOVA
k > 1 classifications	No	Yes	Yes	Yes	Yes	Yes
Sums of cross-products	No	No	No	No	One[1/]	No
Form of covariate	z_{ij}	z_{ij}	$z_{ij} - \bar{z}_{..}$	z_{ij}	$z_{ij} - \bar{z}_{..}$	$z_{ij} - \bar{z}_{..}$
Multiple slopes	Yes	No	No	Yes	No	No
Restrictions on model[2/]	Σ	Σ	Σn	None	Σ	Σn
Solution to normal equations	No	No	Yes	Yes	Yes	Yes
Adjusted means	Yes	Sometimes[3/]	Yes	EMM's[4/]	EMM's	No

1/ Only $\Sigma\Sigma_{ij}(y_{ij} - \bar{y})(z_{ij} - \bar{z}_{..})$.

2/ Σ denotes Σ-restrictions, e.g., $\Sigma\alpha_i = 0$; Σn denotes Σn-restrictions, e.g., $\Sigma n_i\alpha_i = 0$.

3/ When model contains all interactions (and all cells are filled). This includes the 1-way classification.

4/ See SAS GLM in Section 5.

7. Variance Components Estimation

Generally speaking, routines for estimating variance components are different from those for analysis of variance and covariance.

BMDP1V

Does not estimate variance components.

BMDP2V

Does not estimate variance components.

Will handle repeated measures designs (which implicitly involve random effects in a linear model).

BMDP3V

Maximum Likelihood (ML) and Restricted Maximum Likelihood (REML) estimation of variance components.

Iterative procedures, using objective function $\log_e|\underset{\sim}{V}| + N + N\log_e 2\pi$.

When zero'th iterate of REML has all values positive, they are MINQUE0 estimates – see SAS VARCOMP.

Mixed and random models.

Balanced and unbalanced data.

Crossed and/or nested classifications.

Estimates fixed effects using Σ-restrictions and estimated components.

Estimates sampling large sample variances and covariances of estimated fixed effects and components.

BMDP8V

Estimation by analysis-of-variance method only, from balanced data, mixed or random models, crossed and/or nested classifications. Not for unbalanced data.

GENSTAT

Does not estimate variance components.

SAS HARVEY

Does not handle interactions.

Estimation by Henderson's Method III only, using just the each-after-all-others sums of squares; e.g., $R(\alpha|\mu,\beta)$, $R(\beta|\mu,\alpha)$ and SSE.

Uses Σ-restrictions to calculate sums of squares, and to derive expected values of those sums of squares.

Aside from the inability to handle interactions (noted above), mixed and random models can be used, with balanced or unbalanced data.

SAS NESTED

Estimates variance components only for models which are nested and random, with balanced or unbalanced data. Crossed classifications and/or fixed effects cannot be handled.

SAS RANDOM

Derives expected values of mean squares for all Types I - IV sums of squares of SAS GLM.

Balanced or unbalanced data can be used.

Handles random and mixed models, as specified by user.

Expected values involve variance components and quadratic forms of fixed effects; matrices of the quadratic forms are part of the output.

Type III sums of squares are based on Σ-restrictions, but their expected values (like all others in this routine) are derived

on the basis of unrestricted models. This can lead to a calculated Type III sum of squares being the same as a SAS HARVEY sum of squares (which also uses Σ-restrictions), but having different expectation – because in taking expectations HARVEY uses Σ-restrictions but RANDOM does not.

SAS VARCOMP

Estimation is by several methods:

Henderson Method III, using a sequential set of sums of squares, e.g., $R(\alpha|\mu)$, $R(\beta|\mu,\alpha)$, $R(\gamma|\mu,\alpha,\beta)$ and SSE.

Maximum Likelihood, using objective function $\log_e|\underset{\sim}{V}|$.

MINQUE0, which is Rao's minimum norm, unbiased estimation (MINQUE), using à priori values $\sigma_e^2 = 1$ and other components zero.

ML for nested classifications appears to have a programming error. (ML equations have analytic solution, but VARCOMP iterated divergently.)

SPSS ANOVA

Does not estimate variance components.

References

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, 34:216-221.

TABLE 1: Annotated Computer Output Available and in Preparation, August 1981

Analysis of Variance	Estimation of Variance Components	Analysis of Covariance (in preparation)
BMDP2V	BMDP: 2, 3 and 8V	BMDP: 1 and 2V
GENSTAT	SAS VARCOMP	GENSTAT
SAS GLM	SAS HARVEY	SAS GLM
SAS HARVEY	SAS RANDOM	SAS HARVEY
RUMMAGE		SPSS
SPSS		

TABLE 2: SUMS OF SQUARES THAT COULD BE CALLED "DUE TO THE MEAN".

Two data sets (balanced and unbalanced), for the 1-way classification, with two kinds of restrictions, $\Sigma\alpha_i = 0$ and $\Sigma n_i\alpha_i = 0$, with and without a covariate, using the covariate in two different ways, as $b(z_{ij} - \bar{z})$ and as bz_{ij}.

Model	Line	Sums of Squares	Associated Hypotheses in Unrestricted Models	Balanced Data, 2 groups				Unbalanced Data, 3 groups					
				y	z	y	z	y	z	y	z	y	z
				29	2	39	3	74	3	76	2	85	4
				30	5	40	11	68	4	80	4	93	6
				31	8	41	7	77	2				
No covariate	Line		Model: $E(y_{ij}) = \mu + \alpha_i$										
$E(y_{ij}) = \mu + \alpha_i$	1.	$R(\mu)$	$H: \mu + \Sigma n_i\alpha_i/N = 0$	$7350 = N\bar{y}^2$				$43687 = N\bar{y}^2$					
	2.	$R(\mu\|\alpha) = R(\mu,\alpha) - R(\alpha) \equiv 0$	None	0				0					
	3.	$R^*(\dot{\mu}\|\dot{\alpha})_\Sigma$ 1/ [BMDP2V]	$H: \mu + \Sigma\alpha_i/a = 0$	$7350 = N\bar{y}^2$				$43200 \neq N\bar{y}^2$					
	4.	$R^*(\dot{\mu}\|\dot{\alpha})_{\Sigma n}$ 2/	$H: \mu + \Sigma n_i\alpha_i/N = 0$	$7350 = N\bar{y}^2$				$43687 = N\bar{y}^2$					
Covariate $b(z_{ij} - \bar{z}_{..})$			Model: $E(y_{ij}) = \mu + \alpha_i + bz_{ij}$										
$E(y_{ij}) = \mu + \alpha_i + b(z_{ij} - \bar{z}_{..})$	5.	$R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_{z-\bar{z}})_\Sigma$ [SAS HARVEY]	$H: \mu + \Sigma\alpha_i/a + b\bar{z} = 0$	$7350 = N\bar{y}^2$				$42712\frac{49}{66} \neq N\bar{y}^2$					
	6.	$R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_{z-\bar{z}})_{\Sigma n}$	$H: \mu + \Sigma n_i\alpha_i/N + b\bar{z} = 0$	$7350 = N\bar{y}^2$				$43687 = N\bar{y}^2$					
Covariate bz_{ij}	7.	$R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_\Sigma$ [BMDP2V]	$H: \mu + \Sigma\alpha_i/a = 0$	$1288\frac{62}{133} \neq N\bar{y}^2$				$2557\frac{59}{86} \neq N\bar{y}^2$					
$E(y_{ij}) = \mu + \alpha_i + bz_{ij}$	8.	$R^*(\dot{\mu}\|\dot{\alpha},\dot{b}_z)_{\Sigma n}$	$H: \mu + \Sigma n_i\alpha_i/N = 0$	$1288\frac{62}{133} \neq N\bar{y}^2$				$2627\frac{55}{58} \neq N\bar{y}^2$					

1/ Σ denotes Σ-restrictions: $\Sigma\dot{\alpha}_i = 0$.

2/ Σn denotes Σn-restrictions: $\Sigma n_i\alpha_i = 0$.

TABLE 3: Restrictions, and Form of Covariate, Used by Different Computer Routines for Analysis of Variance Calculations

Form of Covariate	Restrictions on the Model		
	None	Σ-restrictions[1]	Σn-restrictions[1] (same as Σ-restrictions for balanced data)
z_{ij}	SAS GLM (Types I and II)	BMDP2V SAS GLM (Type III)	
$z_{ij} - \bar{z}_{..}$	–	SAS HARVEY	SPSS ANOVA GENSTAT ANOVA

TABLE 3a: The 1-way classification with one covariate: relationships in the output of four computer routines of (i) solutions to normal equations and of (ii) estimated marginal means (least squares means) to adjusted treatment means

$$\bar{y}_{A,i} = \bar{y}_{i.} - \hat{b}(\bar{z}_{i.} - \bar{z}_{..})$$

Computer Routine	Restrictions on model[1]	Form of Covariate	Relationships to $\bar{y}_{A,i}$: Solution α_i^o from Normal Equations	Relationships to $\bar{y}_{A,i}$: Estimated Marginal Means (Least Squares Means)
GENSTAT ANOVA	Σn	$z_{ij} - \bar{z}_{..}$	$\bar{y}_{A,i} - \bar{y}_{..}$[2]	Not calculated
SAS GLM	None	z_{ij}	$\bar{y}_{A,i} - \bar{y}_{A,a}$[3]	$\bar{y}_{A,i}$
SAS HARVEY	Σ	$z_{ij} - \bar{z}_{..}$	$\bar{y}_{A,i} - \sum_{i=1}^{a} \bar{y}_{A,i}/a$[4]	$\bar{y}_{A,i}$
SPSS	Σn	$z_{ij} - \bar{z}_{..}$	$\bar{y}_{A,i} - \bar{y}_{..}$[2]	Not calculated

[1] Σ denotes $\Sigma\alpha_i = 0$, and Σn denotes $\Sigma n_i\alpha_i = 0$; [2] Solution for μ is $\bar{y}_{..}$;

[3] Solution for μ is $\bar{y}_{a.} - \hat{b}\bar{z}_{a.}$, where i=1,···,a; [4] Solution for μ is $\frac{1}{a}\Sigma_i \bar{y}_{A,i}$.