A REPORT ON SOME PROGRAMS FOR FOR THE ANALYSIS OF VARIANCE

by E. J. Carney*

Biometrics Unit, Cornell University, Ithaca, N.Y.

BU-518-M

June, 1974

Analysis of variance programs are among the most numerous written for statistical data processing. As an example of their quantity, the IMSL package [11] has twelve subroutines related to this application, the BMD series [3] also has twelve, SAS [19] has five. There are also several large general purpose programs for analysis of variance; examples are AARDVARK [9,10,17] and MANOVA [2]. Why so many programs for this one application? One reason is that there are many different cases of analysis of variance. Several ways of classifying them are given below:

<u>by structure of data</u>

- balanced complete
- balanced incomplete
  - latin square
  - BIB
  - PBIB
  - lattice
  - etc.
- unbalanced, no missing cells
- unbalanced, missing cells
- hierarchical

<u>by model type</u>

- fixed
- random
- mixed

<u>by type of variable</u>

univariate
multivariate
covariance

The above classifications fall short of exhausting the possibilities. Of course not every combination of the above classes requires a different algorithm. The MANOVA program tries to do almost all of them. On the other hand, sometimes several different algorithms are possible for the same combination, and may give different results. This is certainly true for mixed models.

The purpose of this report is to review some of the available analysis of variance programs (especially those in packages implemented at Cornell and easily accessible to CAG users). The two main classes of algorithms used for analysis of variance on the computer may be called sums of squares algorithms and general linear hypothesis algorithms. Sums of squares algorithms are applicable to balanced complete data structures, some balanced incomplete structures, and to a few other situations, including hierarchical structures and, perhaps, random effects models. A complete data structure is one in which every "possible" combination of factor levels occurs, a combination being possible if, whenever it includes a level of a nested factor, it also includes the levels of all nesting factors which contain that level. A structure is balanced if equivalent occurring combinations of factor levels contain the same number of observations.

The balanced complete case has definite theoretical and computational advantages but may be impractical for many investigations, for example when the number of observations is not subject to experimental control, or when the number of treatment levels is larger than the number of homogeneous experimental units which can be obtained. Balanced complete structures typically result in a unique analysis of variance in which the estimates of effects and their corresponding sums of squares are uncorrelated. This uniqueness and orthogonality make computations easier, and the results may be interpreted with less ambiguity. A further computational advantage is that for balanced complete structures the same basic algorithm may be used for fixed, random or mixed models.

Sums of squares algorithms are not generally applicable to balanced incomplete structures. General linear hypothesis algorithms may be used for these, but often programs for these will overlook special features of the analysis [1, 7, 12] and will not provide all the output information which the user of a specialized balanced incomplete structure needs for interpretation of his results. More specialized programs are available for some incomplete structures, particularly lattice designs.

For unbalanced data general linear hypothesis algorithms may be employed. The basis of these algorithms is usually solution of the least squares normal equations, as in a multiple linear regression problem. For analysis of variance however, there are some complications. One of these is that the usual analysis of variance model is over-parameterized and the set of equations must somehow be reduced if an algorithm for solution of a full rank system of equations is to be used. A second and related problem is that the analysis of variance is no longer unique for unbalanced data, nor are the resulting estimates uncorrelated. The user of a general linear hypothesis program will have to determine which of several analyses of variance possible is the one he needs, how to specify that particular analysis and how it may be interpreted with the particular method in use by the program of reducing the model to full rank.

Programs for Balanced Complete Data

BMD01V

This program performs the usual one-was analysis of variance. It allows as an option the printing of group means and standard deviations. Inspection of group standard deviation may give some indication of whether the homogeneity of variance assumption is met by the data and also can serve to reveal gross errors in data input. This feature is not always present in AOV package programs. The data input format requires that all the observations for a single group be contiguous. FORTRAN type variable format cards are supplied by the user so that fields may be skipped. However, for analysis of variance on several variables this format may be awkward since a new data file will be required for each variable. No provision is made for planned or multiple comparison of group means with this program. The program BMD07V discussed later provides for these.

The program may be applied to balanced or unbalanced data, fixed or random models. However, for random models no variance component estimates not expected mean squares are given.

The program assumes the usual 1-way classification model

$$y_{ij} = \mu + \alpha_i + e_{ij}$$

with $i = 1,2,\ldots,a$; $j = 1,2,\ldots,n_i$, and $\sum_{i=1}^{a} \alpha_i = 0$. Program output includes the analysis of variance table:

	SUM OF SQUARES	DF	MEAN SQUARE	F RATIO
BETWEEN GROUPS	SS_B	a-1	MS_B	$F = \frac{MS_B}{MS_W}$
WITHIN GROUPS	SS_W	N-a	MS_W	
TOTAL	SS_T	N-1		

For the fixed effects model the printed F ratio is appropriate for testing the hypothesis that the groups all have the same mean.

For the random model the variance components may be estimated by

$$\hat{\sigma}^2_{\text{WITHIN}} = MS_W$$

$$\hat{\sigma}^2_{\text{BETWEEN}} = \frac{a-1}{\Sigma n_i - \frac{\Sigma n_i^2}{\Sigma n_i}} (MS_B - MS_W)$$

These are the unbiased "Analysis of Variance" or "Henderson's Method I" estimates. For the random model the printed F-ratio may be used in testing the hypothesis that $\hat{\sigma}^2_{\text{BETWEEN}} = 0$.

The calculations should be accurate enough for any but pathological data sets. The among groups sums of squares are computed around the mean of the first group, which will prevent loss of significant digits in this computation unless the means of the groups are widely divergent.

BMD07V

BMD07V performs a one-way analysis of variance, but also can be used as a multiple comparison test or for planned comparisons. The 1-way analysis of variance performed by the program is the same as for BMD01V.

In the case of multiple comparisons the program uses the Duncan [4] procedure unless the user supplies range cards for other procedures. Among choices of other methods available by specification of suitable range cards are Scheffe's S-method [15], Tukey's Multiple Range test [14,15], the Student-Newman-Keuls [14] procedure, the LSD (least significant difference) method [14], and Bonferonni t-statistics [6,14]. The necessary ranges to be supplied for each of these methods are given below:

1) Scheffe's S-method. If the number of groups is k the appropriate range value is $[2(k-1)F_{\alpha; k-1, v_e}]^{1/2}$ where $F_{\alpha,k-1,v}$ is the F-table entry for significance level α with numerator degrees of freedom k-1 and denominator degrees of freedom $v = \sum_{i=1}^{k} n_i - k$, the degrees of freedom for the within groups mean square. This range remains constant for number of means 2, 3, ..., k.

2) Tukey's Multiple Range test. The appropriate entry is $q_{\alpha;k,v}$, the upper α point of the studentized range for sample size = k and $v = \sum_{i=1}^{k} n_i - k$. This entry remains the same for all numbers of means 2, 3, ..., k.

3) Newman-Keuls procedure. The appropriate entry is $q_{\alpha;p,v}$, the upper α point of the studentized range for sample size p and $v = \sum_{i=1}^{k} n_i - k$. This entry varies with p = 2, 3, ..., k.

4) Least Significant Difference. The appropriate entry is $q_{\alpha;2,v} = (t_{\alpha/2,v})\sqrt{2}$ where $q_{\alpha,2,v}$ is the upper α point of the studentized range for a sample size of 2 and $v = \Sigma n_i - k$ degrees of freedom for the estimate of σ^2. This method presumes that the F test in the analysis of variance has been found significant, before the significance of particular comparisons is examined.

5) Bonferonni t-statistics. This method presumes that a fixed number of comparisons are of interest. In the case of comparing all pairs of k means the number would be $m = \frac{k(k-1)}{2} = \binom{k}{2}$. The appropriate range entry for all groups is then $(t_{\alpha/(2m), v})\sqrt{2}$ where $t_{\alpha/(2m); v}$ is the upper $\alpha/(2m)$ point of the student t-distribution with $v = \Sigma n_i - k$ degrees of freedom. These entries needed will usually not be given in most published tables of the t-distribution unless $\alpha/(2m)$ happens to be .10, .05, .025, .01 or .005. Tables with entries for many values of m are found in [6,14]. This test has a probability error rate at least as good as the S-method but will often be more powerful (i.e., have higher

probability of finding real differences to be significant). This method does not depend upon equal sample size nor even upon uncorrelated means.

The BMD07V algorithm allows the number of replications in a group to be unequal. However, the methods of Tukey and Newman-Keuls, above, and the Duncan method which is the default when no ranges are given use the values of the studentized range for the given α ("probability error rate", [14]) level, based upon the distribution of the range of identically and independently distributed normal variables. When group sizes differ the variances of the means are only approximately correct. Duncan [5] indicates by a geometric argument in the case of 3 means that the test will be conservative (i.e., the true α will be smaller; fewer differences between means will be judged significant than if the groups had been equally replicated). The extent of this conservatism may be guessed at by examining the following empirical results for several patterns of unequal replication (all based upon 10,000 pseudo random observations from $N(0,\frac{1}{n_i})$, $i = 1,2,\ldots 5$, comparing

$$|x_i - x_j|\sqrt{\frac{2n_i n_j}{n_i+n_j}}$$

with $1 \cdot q_{.05;5,\infty} = 3.858$ and computing the proportion of times (in 10,000) that <u>at least one</u> such weighted difference exceeds the table value 3.858).

Empirical Probability Error Rate
for 5 Means with Various Numbers
of Replications (Nominal $\alpha = .05$)

n_1	n_2	n_3	n_4	n_5	observed α	95% conf. interval
1	1	1	1	1	.0512	.0469 - .0555
1	1	1	1	100	.0451	.0410 - .0492
2	5	10	20	50	.0440	.0400 - .0480
1	1	50	50	50	.0366	.0329 - .0403

The above figures indicate the effect upon the probability error rate of unequal replications when means are compared 2 at a time, and thus are directly applicable to the Tukey procedure. The indicated conservatism also applied to the Duncan and Newman-Keuls procedures but in the case of these procedures the situation is further complicated by the fact that the means are tested in sets, starting with the set of all means. If a set is found to be homogeneous the means within the set are not tested in smaller sets. It is not completely clear how this procedural constraint affects the expected error rate of these procedures based upon the studentized range, but it would seem to have the effect of making them more conservative (and less powerful).

BMD07V provides the option of coefficient cards for linear contrasts of the group means. The program computes the normalized contrasts, an estimate of their variance, the t-statistic for testing that an individual contrast is zero, and the cumulative probability of this t value.

It should be noted that these t probabilities may be compared with α or 1-α for a <u>1-sided</u> test of the hypothesis that a given contrast is zero; for a two-sided test the printed probability should be compared with $\frac{\alpha}{2}$ or $1-\frac{\alpha}{2}$. The program computes an approximation of the cross product matrix for the given contrasts (ignoring unequal replication) and also determines a "maximal orthogonal group" of contrasts. This latter feature appears to have a bug since in one case the group given in the output consisted of only two of three orthogonal contrasts available to it. (The contrasts chosen for the maximum group were 1,0,0,-1,0,0 and 0,1,-1,0,0,0. The contrast 0,0,0,0,1,-1 was also present but not included in the output of the maximum orthogonal group, although it is obviously orthogonal to the other two, and in fact the cross products in the printed cross product matrix were correctly computed as zero.) In other cases the program found a maximum orthogonal group correctly. The program also provides "All homogeneous subsets among the orthogonal group for Duncan's new multiple range test". This computation does not allow other options as does the comparison of pairs of means. As in the case of multiple comparisons, this procedure is not universally accepted, and in the unequal number situation is, at best, approximate. In any event it is unlikely that the pairwise comparison of contrasts has any usefulness for analysis of most experiments. It is suggested that this output be ignored. If a simultaneous test of the hypothesis that all of a given set of contrasts are zero is desired the studentized maximum modulus may be used [14,15].

SAS DUNCAN Procedure

The SAS procedure DUNCAN performs Duncan's new multiple range test. [5] However the user's manual gives little information about what is actually done. Presumably the program follows the methodology described in the reference [21]. This procedure is somewhat less convenient to use than BMD07V because the error mean square and its degrees of freedom must be supplied as input to the program. This feature requires that the analysis of variance be performed separately, so that two computer runs are necessary.

The program makes no provision for other comparisons than all pairs of means, nor other methods of comparison than Duncan's new multiple range test.

BMD02V Analysis of variance for factorial design

BMD02V performs analysis of variance on an n-way crossed design for up to 8 factors using a sum of squares algorithm. The program can be used for structures with nesting by pooling the proper sums of squares as explained in an appendix of the program description in the manual. To determine which sums of squares are to be pooled the following symbolic method may be used. The method is illustrated for a three factor example replicated twice (Snedecor 3d ed p. 365). The factors are as follows: Ashing at two levels, plants at three levels, and leaves nested in plants at four levels. In the BMD02V output these factors are treated as cross classification factors and identified by numbers 1,2,3, respectively. The analysis of variance table appears as follows:

SOURCE OF VARIATION	DEGREES OF FREEDOM	SUMS OF SQUARES	MEAN SQUARES
1	1	0.07760	0.07760
2	2	7.70945	3.85472
3	3	4.95461	1.65154
12	2	0.03822	0.01911
13	3	0.02459	0.00820
23	6	1.02710	0.17118
123	6	0.07624	0.01271
WITHIN REPLICATES*	24	0.09883	0.00412
TOTAL	47	14.00666	

*This labeling is poor...this line is the within cells or within subclass sums of squares. It is not within replicate.

To obtain the terms for pooling to obtain sums of squares for nested factors the following scheme may be used. Let A,P,L denote the factors ashing, plants, leaves, respectively. Terms involving nested factors are L(P), leaves within plants, and AL(P), the interaction of leaves and ashing within plants. Write L(P) → L(P+1) = LP + L to indicate that the LP interaction and main effect for leaves are to be pooled. Thus the sum of squares for leaves within plants is the sum of the sums of squares labeled 23, and 3: 1.02710 + 4.95461 = 5.98171. The degrees of freedom are also added: 6 + 3 = 9.

To obtain the sum of squares for the interaction of A and L within P the symbolic expression is AL(P+1) = ALP + AL indicating that the 13 and 123 interactions are to be pooled, along with their degrees of freedom.

The program will, optionally, print marginal means, or marginal means and cell means. However, since the program considers the model to be n-way crossed the means for nested factors will be incorrect. For example, in the above example the means for leaves are averaged over the four plants, but since leaves are nested in plants, averaging the responses for an arbitrarily selected leaf from each of several plants is not useful. To obtain the leaf-within-plant means one would need to obtain the cell means and average the appropriate ones.

Data input for the BMD02V program is somewhat awkward because single observations of the same cell must be separated in the data. Since data usually are not recorded in this way this arrangement will be inconvenient. There is no sequence checking with this program and therefore a careful check should be made that the data are in correct order. This will be facilitated if the factor levels are punched on the data cards. Although these cannot be read by the program they will aid in keeping and checking the proper sequence by hand or by using a sorter. The card fields containing the level indicators can be skipped by using the X format code on the variable format card supplied with the input.

BMD02V may, optionally, be used to obtain linear, quadratic and cubic components for selected factors.

BMD08V Analysis of Variance

BMD08V performs analysis of variance on general balanced complete structures. Unlike BMD02V the analysis is performed for nested and/or crossed factors and for fixed, random, mixed or finite population models. The number of levels in the population for each factor is part of the input to the program. If this number is left blank it is taken to be infinite and the factor is considered random. If the population number of levels equals the sample number of levels the factor is considered to be a fixed factor.

The Cornell Office of Computer Services has added a warning to the BMD08V output which indicates the cell means and/or mean squares from this program have been found to be in error in some cases. There were no errors noted in the test case run in this investigation. This error method applies to the July, 1969 version of the program. The latest version was revised in February, 1971.

The chief unusual feature of the BMD08V program is its capability for obtaining the analysis of variance and expected mean squares for finite population numbers of levels. I know of no other general use program with that capability.

Output from BMD08V includes sums of squares and mean squares for each factor and interaction, F-test statistics (when there is a mean square having the same expected value under the null hypothesis), the coefficients of each variance component in the expected mean squares, estimates of the variance components (1971 version only) and, optionally, cell means and estimates of effects.

SAS ANOVA Procedure

This program computes a general analysis of variance using a sums of squares algorithm. The procedure allows unequal numbers, but except for hierarchical (i.e., completely nested) models, and, perhaps, random effects models*, the analysis obtained can only be approximately correct [8]. Unfortunately the SAS manual description of this procedure does not make this very clear, although the output from this procedure when the data are unbalanced includes a warning message.

Use of the procedure requires some general knowledge of the SAS system, but this is easily mastered. The program is easy to use, and may be combined with the SAS MEANS procedure if marginal or cell means are desired. The MEANS procedure produced (optionally) other statistics including standard deviations and extreme values which are useful as checks for errors in the data, outliers, etc. It may be necessary or desirable to use the SORT procedure in conjunction with the ANOVA program.

Programs for Unbalanced Data

BMD10V (BMDX64) General Linear Hypothesis

The BMD10V General Linear Hypothesis program to some extent supplants the BMD05V program which had been the standard BMD general linear hypothesis program. BMD10V is more automatic than the BMD05V program in that it computes the values of dummy variables for the analysis, while these have to be supplied as input to BMD05V. This has the advantage of simplifying data preparation and input, but the disadvantage that the way dummy variables are computed, and hence the tests of hypotheses which can be easily made, are not completely controlled by the user. A further disadvantage is that BMD10V may give incorrect answers if there are missing cells in the data. (These wrong answers are usually

*No single method of estimation of variance components for unbalanced data can be viewed as correct. If one obtains the expected mean squares for the mean squares computed by ANOVA for his model, these can be set equal to the computed mean squares and (perhaps) solved for unbiased estimate of the variance components.

easily detected being accompanied by such anomalies as too many or too few degrees of freedom in the output.)

As noted in the introduction more than one analysis of variance is possible with unbalanced data. BMD10V automatically provides an F test for each model term by computing the difference in the sum of squares when the model has all terms included and when the single term is left out. In case of covariance a sum of squares is given for each covariate left out singly, and also for all left out of the model simultaneously. Other tests of hypotheses can be specified by the user so long as they are stated in terms of the dummy variables computed by the program. Theoretically any testable linear hypothesis can be stated in terms of the dummy variables computed by the program. However, if the model involves nested factors it may be difficult to determine how the hypothesis the user wishes to test should be stated in terms of the dummy variables computed by the program, and use of BMD05V or some other program would seem to be desirable.

The description of the manufacture of dummy variables by BMD10V to be found in the manual was found by this reader to be incomprehensible. The following description is believed to be equivalent to the operations performed by the program.

Dummy Variables for BMDX64 (BMD10V)

The dummy variables for BMDX64 are based upon a set of "design variables" for each subscript which vary in value as the subscripts vary. These design variables may be submitted as part of the program input, their values being supplied on a DESIGN card for each cell. If DESIGN cards are not submitted the design variables are manufactured by the program. If the model has p subscripts, the k^{th} subscript having n_k levels, each design card must specify the value of $\sum_{k=1}^{n}(n_k-1)$ values, $t_{k\ell}$; $k=1,2,\ldots,p$, $\ell=1,2,\ldots,n_k-1$; in the order $t_{11}, t_{12}, \ldots, t_{1\,n_1-1}, \ldots, t_{p\,n_p-1}$. Consider, for example, the following model:

$$E(y_{ijk}) = \mu + a_i + b_{ij};\ i=1,2;\ j=1,2,3.$$

For each of the six cells a DESIGN card may be submitted with $n_1-1 + n_2-1 = 2-1 + 3-1 = 3$ values, t_{11}, t_{21}, t_{22}. As an instance, it might be desired to give the following values, indicating a contrast of a effects and linear and quadratic contrasts of b effects, averaged over the levels of a.

cell		t_{11}	t_{21}	t_{22}
i	j			
1	1	1	1	1
1	2	1	0	-2
1	3	1	-1	1
2	1	-1	1	1
2	2	-1	0	-2
2	3	-1	-1	1

If no design variables are submitted, they are constructed by the program as contrasts between each of the first n_k-1 levels and the last level. For the above example the values would be as follows:

cell		t_{11}	(t_{12})	t_{21}	t_{22}	(t_{23})
i	j					
1	1	1	1	1	0	1
1	2	1	1	0	1	1
1	3	1	1	-1	-1	1
2	1	-1	1	1	0	1
2	2	-1	1	0	1	1
2	3	-1	1	-1	-1	1

The values $t_{k\,n_k} = 1$ are automatically produced whether or not the design variables are part of the input to the program.

Each analysis of variance term (main effect, interaction) is described in the output by a DUMVAR card. This card, in addition to giving a name to the term, indicates, for each subscript, the multiplier for the number of degrees of freedom for the term associated with that·subscript. In the case of non-nested effects this multiplier is the number of degrees of freedom for the main effect corresponding to that subscript. If a term is nested the multiplier for subscripts corresponding to nesting factors will be the number of levels of the factor; for subscripts associated with the term but not with nesters it will be the number of degrees of freedom for the corresponding main effect; for subscripts not found in the description of a term the entry should be 0 (or left blank). For the simple nested example above the DUMVAR cards could appear as follows:

```
DUMVAR    MU
DUMVAR    A     1
DUMVAR    B     2     2
```

If the same data were to be analyzed using a two-way crossed model with interaction these cards might be:

```
DUMVAR    MU
DUMVAR    A     1
DUMVAR    B           2
DUMVAR    AB    1     2
```

Let d_k be the entry for the k^{th} subscript on a given DUMVAR card. For each such card there will be as many dummy variables manufactured for the term as there are degrees of freedom; that is, a number equal to the product of the non-zero (non-blank) entries on the card. These dummy variables (for a given model term, i.e., DUMVAR card) may be represented by

$$u_{\ell_1 \ldots \ell_p}; \; \ell_k = 1,2,\ldots,d_k; \text{ or } \ell_k = 0 \text{ if } d_k = 0.$$

These dummy variables are taken to be in the order of the DUMVAR cards, and within a term, in lexographical order with earlier subscripts moving more rapidly. The value computed for each dummy variable, for each cell is computed as follows:

$$u_{\ell_1 \dots \ell_p} = \prod_{k=1}^{p} v_{k\,\ell_k},$$

where $v_{k\,\ell_k} = t_{k\,\ell_k}$ for $\ell_k \neq 0$, $v_{k\,\ell_k} = 1$, for $\ell_k = 0$. Thus, for example, if a dummy variable u_{12012} were to arise for some five factor model, its value would be: $t_{11} \cdot t_{22} \cdot 1 \cdot t_{41} \cdot t_{52}$. Some examples are given below.

Dummy Variables for $E(y_{ijk}) = \mu + a_i + b_{ij}$; $i=1,2$; $j=1,2,3$

Model term:		μ	a	b			
"name":		u_{00}	u_{10}	u_{11}	u_{21}	u_{12}	u_{22}
"formula":		$1 \cdot 1$	$t_{11} \cdot 1$	$t_{11} \cdot t_{21}$	$t_{12} \cdot t_{21}$	$t_{11} \cdot t_{22}$	$t_{12} \cdot t_{22}$
i	j						
1	1	1	1	1	1	0	0
1	2	1	1	0	0	1	1
1	3	1	1	-1	-1	-1	-1
2	1	1	-1	-1	1	0	0
2	2	1	-1	0	0	-1	1
2	3	1	-1	1	-1	1	-1

Dummy Variables for $E(y_{ijk}) = \mu + a_i + b_j + (ab)_{ij}$

Model term:		μ	a	b		ab	
"name":		u_{00}	u_{10}	u_{01}	u_{02}	u_{11}	u_{12}
"formula":		$1 \cdot 1$	$t_{11} \cdot 1$	$1 \cdot t_{21}$	$1 \cdot t_{22}$	$t_{11} \cdot t_{21}$	$t_{11} \cdot t_{22}$
i	j						
1	1	1	1	1	0	1	0
1	2	1	1	0	1	0	1
1	3	1	1	-1	-1	-1	-1
2	1	1	-1	1	0	-1	0
2	2	1	-1	0	1	0	-1
2	3	1	-1	-1	-1	1	1

It may be noted that in the case of the two-way crossed classification model immediately above the dummy variables generated are those which would be obtained by applying constraints of the form $\sum_{i=1}^{I} a_i = 0$, and replacing a_I by $-a_1 - a_2 \dots -a_{I-1}$ and treating the b_j's similarly, etc. in the familiar manner. This is the case, generally, with crossed classification models, but nested models are not treated in the usual way. With the two factor nested model above the usual device of setting $a_2 = -a_1$ and $b_{i3} = -(b_{i1} + b_{i2})$, i = 1,2 would give the following dummy variables:

μ^*	a^*_1	b^*_{11}	b^*_{12}	b^*_{21}	b^*_{22}
1	1	1	0	0	0
1	1	0	1	0	0
1	1	-1	-1	0	0
1	-1	0	0	1	0
1	-1	0	0	0	1
1	-1	0	0	-1	-1

There is no way of inducing the BMDX64 program to generate the above dummy variables for a two factor model. The model could, however, be specified as a one-way classification, and then whatever contrasts it was desired to obtain could be specified using design variables and hypothesis cards. In data in which cells are missing it is also necessary, or at least convenient, to treat the data in this way. If the model has missing cells and is specified in the usual way using multiple DUMVAR cards the results from the program are likely to be incorrect.

BMD05V General Linear Hypothesis Program

This program requires that "design variables" be supplied for each cell in the data structure. However, the "design variables" for BMD05V are the actual dummy variables for the analysis instead of the raw material from which they are manufactured as in the case of BMD10V described above. The hypotheses to be tested must also be supplied as input. While working out dummy variables may be inconvenient, there is the advantage that the dummy variables can be made to match the hypotheses it is desired be tested. There is the added advantage that the knowledge required to form the dummy variables and hypotheses may be related to that needed for correct interpretation of the data. In this latter regard it should be noted that supplying dummy variables constitutes a reparameterization of the model, and the hypotheses cards supplied to the program allow one to compare models with these new parameters included or excluded. The F-tests supplied in the program output compare each hypothesis model with the full model, thus testing the contribution of the new parameters set to zero by that hypothesis card. The appropriateness of these tests depends upon 1) the nature of the investigation which has lead to the data; 2) the manner in which the dummy variables were constructed. Other tests may be obtained by comparing models other than the full model with each other, by subtracting the residual for the model having some parameters unequal to zero from a model in which these parameters are zero.

An example may be instructive. Consider an experiment with two crossed factors each at two levels and a third factor nested within combinations of the other two, with numbers of observations as shown.

	B_1		B_2	
		n_{ijk}		n_{ijk}
A_1	C_{111}	2	C_{121}	3
	C_{112}	2	C_{122}	3
	C_{113}	2		
A_2	C_{211}	2	C_{221}	3
	C_{212}	2	C_{222}	3
	C_{213}	2		

This data is peculiar in that the number of levels of the C-factor varies over the B factor, while the number of observations for the A B 2 X 2 is constant. The result is that sums of squares algorithms give correct results for this program even though it is only "accidentally" balanced. BMD05V also gives correct results of course, and would even if the number of observations in one of the cells was changed. If these numbers were arbitrary the sums of squares programs would no longer give exact results (some "sum of squares" might be found to be negative, for example) and BMD10V would also fail, for, with respect to its method of finding dummy variables, there is a missing cell.

A linear model for this data structure is

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \gamma_{ijk} + e_{ijkl}$$

$$\Sigma\alpha_i = 0,\ \Sigma\beta_j = 0,\ \sum_i(\alpha\beta)_{ij} = \sum_j(\alpha\beta)_{ij} = 0$$

$$\sum_k\gamma_{ijk} = 0,\ i = 1,2;\ j = 1,2.$$

Specifying that the model terms obey the given linear constraints insures that they are uniquely defined. These constraints may be used to reparameterize the model as

$$y = \mu^* + \alpha^*_i + \beta^*_j + (\alpha\beta)^*_{ij} + \gamma^*_{ijk} + e_{ijk}$$

where $\mu = \mu^*$, $\alpha_1 = \alpha^*_1$, $\alpha_2 = -\alpha^*_1$, $\beta_1 = \beta^*_1$, $\beta_2 = -\beta^*_1$, $(\alpha\beta)_{11} = (\alpha\beta)^*_{11}$, $(\alpha\beta)_{12} = -(\alpha\beta)^*_{11} = (\alpha\beta)_{21}$, $(\alpha\beta)_{22} = (\alpha\beta)^*_{11}$, $\gamma_{111} = \gamma^*_{111}$, $\gamma_{112} = \gamma^*_{112}$, $\gamma_{113} = -\gamma^*_{111} - \gamma^*_{112}$, $\gamma_{121} = \gamma^*_{121}$, $\gamma_{122} = -\gamma^*_{121}$, $\gamma_{211} = \gamma^*_{211}$, $\gamma_{212} = \gamma^*_{212}$, $\gamma_{213} = -\gamma^*_{211} - \gamma^*_{212}$, $\gamma_{221} = \gamma^*_{221}$, $\gamma_{222} = -\gamma^*_{222}$. These relationships are all obtained from the model constraints. The reparameterized model has only 10 parameters, μ^*, α^*_1, β^*_1, $(\alpha\beta)^*_{11}$, γ^*_{111}, γ^*_{112}, γ^*_{121}, γ^*_{211}, γ^*_{212}, γ^*_{221}. This is exactly the same number as the number of cells having observations. The "design variables" (i.e., dummy variables) for the BMD05V design cards are as follows:

ijk	μ^*	α^*_1	β^*_1	$(\alpha\beta)^*_{11}$	γ^*_{111}	γ^*_{112}	γ^*_{121}	γ^*_{211}	γ^*_{212}	γ^*_{221}
111	1	1	1	1	1	0	0	0	0	0
112	1	1	1	1	0	1	0	0	0	0
113	1	1	1	1	-1	-1	0	0	0	0
121	1	1	-1	-1	0	0	1	0	0	0
122	1	1	-1	-1	0	0	-1	0	0	0
211	1	-1	1	-1	0	0	0	1	0	0
212	1	-1	1	-1	0	0	0	0	1	0
213	1	-1	1	-1	0	0	0	-1	-1	0
221	1	-1	-1	1	0	0	0	0	0	1
222	1	-1	-1	1	0	0	0	0	0	-1

The BMDO5V program will generate three hypotheses automatically. For this model they will be written as

	μ^*	α^*_1	β^*_1	$(\alpha\beta)^*_{11}$	γ^*_{111}	γ^*_{112}	γ^*_{121}	γ^*_{211}	γ^*_{212}	γ^*_{221}
1	0	0	0	0	0	0	0	0	0	0
2	1	1	1	1	1	1	1	1	1	1
⋮										
last	1	0	0	0	0	0	0	0	0	0

The first of these corresponds to a model with all parameters equal to zero. The "residual" for this model will be the total sum of squares. The second hypothesis corresponds to the full (reparameterized) model and its residual will be the within subclasses cum of squares. The last hypothesis corresponds to a model with all parameters but the overall mean set to zero, and the difference between its residual and that for hypothesis 2 will be the "among subclasses" sum of squares for what is sometimes called [20] the "preliminary analysis of variance".

Suppose, for our own reasons, we wished to test the hypothesis that the γ^*'s were 0 after fitting μ^*, and then that the two by two factorial (α^*_1, β^*_1, $(\alpha\beta)^*_{11}$) parameters were 0 after fitting μ^* and the γ^*'s. If we supply an additional hypothesis (number 3) 1 0 0 0 1 1 1 1 1 1, corresponding to a model with α^*_1, β^*_1, $(\alpha\beta)^*_{11}$ equal to zero we can obtain the desired analysis as follows:

Source	Method of obtaining SS from BMD05V output
γ^{*}'s after μ^{*}:	Subtract the residual for hypothesis 3 from that for the last (all zero but μ^{*}) hypothesis.
α_1^{*}, β_1^{*}, $(\alpha\beta)_{11}^{*}$ after μ^{*} and γ^{*}'s	Subtract the residual for the full model (hypothesis 2) from the hypothesis 3 residual.
Residual	Residual from hypothesis 2.

The degrees of freedom for these sums of squares may be obtained by subtracting the degrees of freedom for corresponding residuals given in the BMD05V output. They are, of course, equal to the number of parameters found in one model but not in the other for the two residuals subtracted. In the present case there will be 6 degrees of freedom for γ^{*}'s after μ^{*} since there are 6 γ^{*}'s, and 3 degrees of freedom for the 2 X 2 factorial after μ^{*} and γ^{*}'s, corresponding to the three parameters α_1^{*}, β_1^{*}, $(\alpha\beta)_{11}^{*}$. Degrees of freedom for the full model will be given with the residual for hypothesis 2 and for the data described here will be 14. F-ratios for the tests described above would be formed by dividing the mean squares from each of the first two sources by that for the residual. (Assuming a fixed effect model, if some effects are random the situation is more complicated [16,18].

The program output includes some F-ratios. In each case the numerators of these F ratios are obtained by subtracting the full model residual from that for the particular hypothesis and dividing by the number of parameters set to zero in that hypothesis. It is thus a test of the hypothesis that the parameters set to zero by that hypothesis are zero after fitting all other parameters. These tests may or may not be appropriate.

It should be emphasized that all the hypotheses generated by or supplied to the program are stated in terms of the reparameterized model implied by the "design variables" given to the program; and the interpretation of tests and estimates must reflect this fact. Testing that a certain group of parameters are equal in the reparameterized model, may or may not be equivalent to testing that a corresponding group of parameters in the original over-parameterized model are equal, depending upon the way in which the dummy variables are assigned [18].

SAS REGR Procedure

In the SAS system the REGR procedure is used for general linear hypothesis analysis of variance problems as well as for multiple linear regression problems. While any multiple linear regression program can be used to solve analysis of variance problems by including dummy variables for the classification variables, this is not necessary with REGR, since the program will manufacture dummy variables for the classification variables if these are identified by a CLASSES statement. The method used by SAS is similar to that used in the example in the description of BMD05V above of obtaining the "design variables" for input to that program. It differs from the method employed by BMD10V in the handling of nested factors, and seems more straight-forward. SAS REGR also differs from BMD10V in that its treatment of missing cells is more sophisticated. In BMD10V missing cells are ignored and this may lead to incorrect output. In the SAS program missing cells are detected and cause the elimination of some dummy variables from the model. This results in output which is correct, but which may be difficult to interpret. A warning message is generated.

Two analysis of variance outputs are produced and printed by SAS REGR with sums of squares, F-ratios and F-tail probabilities for each. These are called the Sequential SS and the Partial SS. The partial SS are the same as those automatically produced by BMD05V and BMD10V; that is they test the contribution of each factor or interaction group of parameters (of the reparameterized model) to the fit by computing the difference in the regression sums of squares when these particular terms are in and not in the model, and all other parameters are present. The sequential SS computes the increment in the regression sum of squares as each group of parameters (of the reparameterized model) is added to the model in the order in which the factors and interactions are specified in the MODEL statement for the problem. Several model statements may be used if needed.

The input required for use of SAS REGR is fairly convenient and simple. For the example given above for BMD05V the following might be used:

```
DATA TESTSET;
INPUT A 1 B 2 C 3 Y 6-10;
CARDS;
111  13.41    )
111  14.06    )
 .            )
 .            )  data cards
 .            )
222   7.92    )
PROC REGR; CLASSES A B C;
MODEL Y = C(A B) A B A*B;
```

The Sequential SS for this model will be:

γ^*'s after μ^*

α_1^* after μ^*, γ^*'s

β_1^* after μ^*, α_1^*, γ^*'s

$(\alpha\beta)_{11}^*$ after μ^*, α_1^*, β_1^*, γ^*'s

The Partial SS will be for:

γ^*'s after μ^*, α_1^*, β_1^*, $(\alpha\beta)_{11}^*$

α_1^* after μ^*, β_1^*, $(\alpha\beta)_{11}^*$, γ^*'s

β_1^* after μ^*, α_1^*, $(\alpha\beta)_{11}^*$, γ^*'s

$(\alpha\beta)_{11}^*$ after μ^*, α_1^*, β_1^*, γ^*'s

If the data are balanced both of these analyses will produce the same 4 numbers. If the data are not balanced the first three will be different for the two sets of output.

Additional optional output may be obtained from REGR of particular interest, a table showing confounding of factors will be produced if /CONF is added to the model statement before the semicolon.

References

1. Cochran, W. G. and G. Cox (1957). Experimental Designs, 2d ed., John Wiley, New York.

2. Cramer, E. M. (1973). MANOVA, A Computer program for univariate and multivariate analysis of variance. L. L. Thurstone Psychometric Laboratory Report Series, 1973, No. 124. University of North Carolina, Chapel Hill.

3. Dixon, W. J., ed., BMD Biomedical Computer Programs, University of California Press, Berkeley, 1973.

4. Duncan, D. B. (1955). Multiple Range and Multiple F Tests, Biometrics, 11, 1-42.

5. Duncan, D. B. (1957). Multiple Range Tests for Correlated and Heteroscedastic Means, Biometrics, 13, 164-176.

6. Dunn, O. J. (1959). Confidence Intervals for the Means of Dependent, Normally Distributed Variables, J. Am. Statist. Assoc., 54, 613-621.

7. Federer, W. T. (1955). Experimental Design Theory and Application, Macmillan, New York.

8. Francis, I. (1973). Comparison of Several Analysis of Variance Programs, J. Am. Statist. Assoc., 68, 860-865.

9. Hemmerle, W. J. (1964). Algebraic Specifications of Statistical Models for Analysis of Variance Computations, Journal of the Association for Computing Machinery, 11, 234-239.

10. Hemmerle, W. J., E. J. Carney, and A. B. D'Silva (1969). Multivariate AARDVARK Reference Manual, Computer Laboratory, University of Rhode Island, Kingston, R. I.

11. International Mathematical & Statistical Libraries, Inc. (1973). Computer Subroutine Libraries in Mathematics and Statistics (Abilities). Houston, Texas.

12. Kempthorne, O. (1952). *Design and Analysis of Experiments*, John Wiley, New York.

13. Kramer, C. Y. (1956). Extension of multiple range tests to group means with unequal numbers of replications, *Biometrics*, *12*, 307-310.

14. Miller, R. G., Jr. (1966). *Simultaneous Statistical Inference*, McGraw-Hill, New York.

15. Scheffé, H. (1953). A Method for Judging All Contrasts in the Analysis of Variance, *Biometrika*, *40*, 87-104.

16. Scheffé, H. (1959). *Analysis of Variance*, John Wiley, New York.

17. Schlater, J. E. and W. J. Hemmerle (1966). Statistical Computations Based Upon Algebraically Specified Models, *Communications of the ACM*, *9*, 865-869.

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

19. Service, J. (1972). *A User's Guide to the Statistical Analysis System*, Student Supply Stores, North Carolina State University, Raleigh.

20. Snedecor, G. W. (1937). *Statistical Methods*, 3d ed., Iowa State University Press, Ames, Iowa.

21. Steele, R. G. D., and J. H. Torrie (1960). *Principles and Procedures of Statistics*, McGraw-Hill, New York.