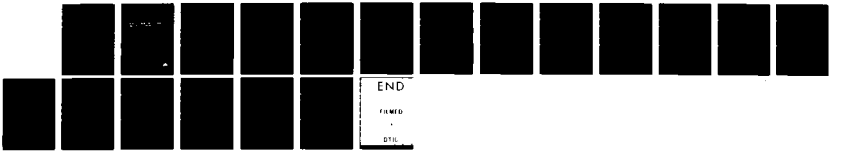
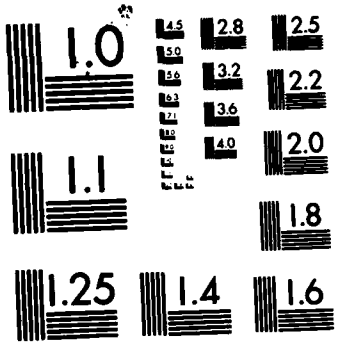
AD-A125 380 AN ALGORITHM FOR THE UNIVARIATE ANALYSIS OF VARIANCE IN EXPERIMENTS WITH REPEATED MEASURES(U) SCHOOL OF AEROSPACE MEDICINE BROOKS AFB TX W G JACKSON ET AL. 1/1

UNCLASSIFIED DEC 82 SAM-TR-82-37 F/G 12/1 NL






MICROCOPY RESOLUTION TEST CHART
NATIONAL BUREAU OF STANDARDS-1963-A

(12)

Report SAM-TR-82-37

# AN ALGORITHM FOR THE UNIVARIATE ANALYSIS OF VARIANCE IN EXPERIMENTS WITH REPEATED MEASURES

William G. Jackson, Jr., M.Stat.
Richard C. McNee, M.S.



December 1982

Final Report for Period August 1980 — August 1982



Approved for public release; distribution unlimited.

USAF SCHOOL OF AEROSPACE MEDICINE
Aerospace Medical Division (AFSC)
Brooks Air Force Base, Texas 78235

## NOTICES

This final report was submitted by personnel of the Advanced Analysis Branch, Data Sciences Division, USAF School of Aerospace Medicine, Aerospace Medical Division, AFSC, Brooks Air Force Base, Texas, under job order 7930-15-02.

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely Government-related procurement, the United States Government incurs no responsibility or any obligation whatsoever. The fact that the Government may have formulated or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication, or otherwise in any manner construed, as licensing the holder, or any other person or corporation; or as conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

The Office of Public Affairs has reviewed this report, and it is releasable to the National Technical Information Service, where it will be available to the general public, including foreign nations.

This report has been reviewed and is approved for publication.

William G. Jackson, Jr.

WILLIAM G. JACKSON, JR., M. Stat.
Project Scientist

Richard C. McNee

RICHARD C. MCNEE, M.S.
Supervisor

Roy L. DeHart

ROY L. DEHART
Colonel, USAF, MC
Commander

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

| REPORT DOCUMENTATION PAGE | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|
| 1. REPORT NUMBER: SAM-TR-82-37 | 2. GOVT ACCESSION NO.: AD-A125240 |
| | 3. RECIPIENT'S CATALOG NUMBER: |
| 4. TITLE *(and Subtitle)*: AN ALGORITHM FOR THE UNIVARIATE ANALYSIS OF VARIANCE IN EXPERIMENTS WITH REPEATED MEASURES | 5. TYPE OF REPORT & PERIOD COVERED: Final Report August 1980 - August 1982 |
| | 6. PERFORMING ORG. REPORT NUMBER: |
| 7. AUTHOR(s): William G. Jackson, Jr., M.Stat.<br>Richard C. McNee, M.S. | 8. CONTRACT OR GRANT NUMBER(s): |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS:<br>USAF School of Aerospace Medicine (BRA)<br>Aerospace Medical Division (AFSC)<br>Brooks Air Force Base, Texas 78235 | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:<br>62202F<br>7930-15-02 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS:<br>USAF School of Aerospace Medicine (BRA)<br>Aerospace Medical Division (AFSC)<br>Brooks Air Force Base, Texas 78235 | 12. REPORT DATE: December 1982<br>13. NUMBER OF PAGES: 14 |
| 14. MONITORING AGENCY NAME & ADDRESS *(if different from Controlling Office)*: | 15. SECURITY CLASS. *(of this report)*: Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE: |

16. DISTRIBUTION STATEMENT *(of this Report)*

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT *(of the abstract entered in Block 20, if different from Report)*

18. SUPPLEMENTARY NOTES

19 KEY WORDS *(Continue on reverse side if necessary and identify by block number)*

Analysis of variance
Repeated measurements
Missing data

20 ABSTRACT *(Continue on reverse side if necessary and identify by block number)*

A computing procedure is described for the univariate analysis of a repeated measurements experiment where the experimental units (frequently subjects) are arranged in a two-way classification with cell frequencies that can be disproportionate. The analysis is adjusted for missing values, provided their number and configuration do not violate certain limitations. Unlike some strategies for handling missing values in repeated measurements experiments, the

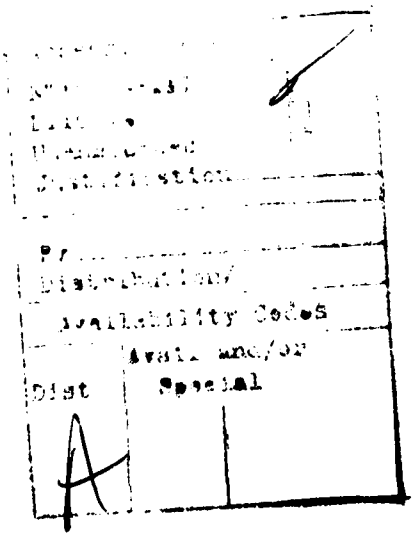
DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

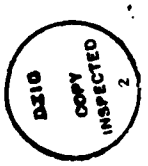
UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE *(When Data Entered)*

20. ABSTRACT (Continued)

method does not require the inclusion in the model of an explicit subject factor, meaning that the order of the matrix to be inverted does not depend on the number of subjects in the experiment. The algorithm has been incorporated into a SAS procedure, REP2W1F, which computes the full analysis with a single call and produces useful summary statistics (including least-squares means) particular to the design. The approach could be generalized to experiments where the number of treatment factors is other than two and where the repeated measures have a factorial arrangement of their own.



# AN ALGORITHM FOR THE UNIVARIATE ANALYSIS OF VARIANCE IN EXPERIMENTS WITH REPEATED MEASURES

## INTRODUCTION

In a typical repeated measurements experiment, subjects are randomly assigned to different treatment groups, and observations are made on the subjects at distinct points in time in order to estimate and test for the treatment effects. Under conditions given by Huynh [5], a univariate analysis of variance is appropriate for this purpose. If the treatment groups are of equal size and every subject has a data value at each time, the analysis can be readily obtained from packaged computer programs such as SAS ANOVA, SAS GLM, BMDP2V, BMDP4V, or BMDP8V [1,4,9]. One frequently encounters experimental situations in practice, however, where the cell sizes are unequal or where some of the subjects have incomplete data. The SAS ANOVA and BMDP8V programs were not designed for use in such cases, and there are situations where the other three programs all suffer from serious drawbacks, some of which have been discussed by Morris et al. [8].

For repeated measurements experiments with missing data, SAS GLM does not include any provision for computing least-squares means which are averaged across treatment groups of unequal size. The least-squares means for the highest order interactions can be computed provided an explicit factor for "between subjects within treatment groups" is included in the model. The computer core storage and central processing unit (CPU) time used grow rapidly as a function of the order of the X'X matrix, which can cause the analysis to become very expensive if the number of subjects is large, as demonstrated by the following. These examples were run using SAS GLM on an IBM 4341 Model 2 under a multiple virtual storage (MVS) operating system. The only differences are the number of subjects in the cells.

| Source | df | df | df |
|---|---|---|---|
| A | 1 | 1 | 1 |
| B | 2 | 2 | 2 |
| AxB | 2 | 2 | 2 |
| Between subjects (AxB) | 27 | 77 | 127 |
| Time | 3 | 3 | 3 |
| AxTime | 3 | 3 | 3 |
| BxTime | 6 | 6 | 6 |
| AxBxTime | 6 | 6 | 6 |
| Error | 81 | 231 | 381 |
| Core required (K bytes) | 256 | 354 | 504 |
| CPU time (min) | .30 | 1.62 | 5.04 |

If any subject has incomplete data, neither BMDP2V nor BMDP4V provides a method for computing properly adjusted least-squares means. Adjusted F-ratios can be easily obtained provided all treatment groups are of equal size by

including "between subjects within treatment groups" as a factor and analyzing the data as if it were a full factorial experiment [2]. If the number of subjects is large, however, we have the same problems discussed for SAS GLM. For situations where missing values are present and the cell sizes are unbalanced, the only way we have found to adjust the F-ratios in either BMDP program is to create dummy covariates. This can be a laborious exercise or a tricky programming problem, and it does not make these two programs generate the desired summary means because the adjusted means are computed as if the dummy covariates were true covariates.

This report describes a computing procedure for the univariate analysis of a repeated measurements experiment where the subjects are arranged in a two-way classification with cell frequencies that can be disproportionate. The analysis is adjusted for missing values, provided their number and configuration do not violate certain limitations. The method proceeds in such a way that the order of the X'X matrix does not depend on the number of subjects in the experiment (although it does depend on the number of missing values). The algorithm has been incorporated into a SAS procedure, REP2W1F, which produces useful summary statistics (including least-squares means) particular to the design. The approach could be generalized to experiments where the number of treatment factors is other than two and where the repeated measures have a factorial arrangement of their own.

## EXPERIMENTAL DESIGN AND MODEL EQUATION

In the particular situation we will consider here, N subjects are randomly assigned to treatment groups which are arranged in a two-way classification with factors A (having levels $A_i$, i=1,2,...a) and B (having levels $B_j$, j=1,2,...b). There are $n_{ij}$ (>0) subjects at the $i^{th}$ level of A and $j^{th}$ level of B, designated by the subscript $\ell$=1,2,...$n_{ij}$. Each subject is measured at t levels of a fixed Time factor ($T_k$, k=1,2,...t). Tests of the main effects A and B and of the interaction AxB are generally referred to as the between-subject portion of the analysis, whereas tests for a main effect due to Time or an interaction involving Time are referred to as belonging to the within-subject part of the analysis. The model equation for the case of no missing values appears as follows:

$$Y_{ijk\ell} = \mu_{ijk} + \varepsilon_{ijk\ell}, \qquad (1)$$

where $\mu_{ijk} = \mu + A_i + B_j + (AB)_{ij} + T_k + (AT)_{ik} + (BT)_{jk} + (ABT)_{ijk}$; $\mu$, $A_i$, $B_j$, $(AB)_{ij}$, $T_k$, etc., are the usual fixed main effects and interactions; and $\varepsilon_{ijk\ell}$ is a random error term.

We assume that the components of model equation 1 meet a number of conditions. The fixed main effects and interactions are assumed to follow the so-called $\Sigma$-restrictions [12]. These are the same as the "usual constraints" in which the sum over any subscript (i, j, k, or $\ell$) of any set of fixed main effects or interactions containing that subscript is 0. For example, $\sum_i A_i = 0$

and $\sum_i (AB)_{ij} = \sum_j (AB)_{ij} = 0$. Provision for weighted restrictions (such as $\sum_i w_i A_i = 0$) could also have been made, although in the absence of substantial prior knowledge about a given set of data, the $\Sigma$-restrictions are usually reasonable.

The remaining assumptions involve the errors. If $\underline{\varepsilon}_{ij\ell}$ represents a t-vector of errors for the $\ell^{th}$ subject in treatment group (i,j), then we assume $\underline{\varepsilon}_{ij\ell}$ has been randomly drawn from a multivariate normal population with mean vector $\underline{0}$ and covariance matrix $\Sigma_{ij}$. For the between-subject portion of our analysis to be valid, we assume in addition that the variance of the sum over time, $\sum_k \varepsilon_{ijk\ell}$, is the same for all subjects in the experiment. The within-subject portion of the analysis is valid if and only if the errors have the property of multisample sphericity [5], which means that if C is a (t-1)x(t) orthogonal contrast matrix, then there is a constant $\lambda$ for which the equation $C\Sigma_{ij}C' = \lambda I_{t-1}$ holds independently of i and j. ($I_{t-1}$ is the identity matrix of order t-1).

It frequently occurs that some of the subjects do not have complete data, a circumstance that complicates the analysis. The analysis described here assumes that the pattern of missing values is random and independent of any treatment effects. At present, there is no universally accepted method for handling all parts of the repeated measurements analysis of variance when missing values occur under this condition [2,3]. To compute the between-subject portion of the analysis, we require that every treatment combination (i,j) has at least one subject with complete data, and that at least one such treatment combination has two or more subjects with complete data. If we denote by $r_{ij}$ the number of subjects with incomplete data in the (i,j) cell, this says that $(n_{ij}-1) - r_{ij}$ is nonnegative for every cell and positive for at least one cell. If this condition is met, the algorithm produces the between-subject analysis we would have if subjects with missing values were completely excluded. We have programmed REP2W1F so that if the condition is not met, a message is printed informing the user that the between-subject analysis cannot be computed, and the program continues with the within-subject analysis.

The within-subject portion of the analysis is obtained using dummy covariates to adjust the tests for missing values. The resulting within-subject analysis is equivalent to that suggested by Schwertman [10]. He proved that for repeated measurements experiments with missing values, where the required assumptions for a univariate analysis are otherwise met, F-tests of null hypotheses involving the within-subject parameters can be constructed using ordinary general linear regression techniques if explicit effects for subjects within treatment groups ($S_{ij\ell}$) are added to model equation 1. The sum of squares for each hypotheses $H_0$ is calculated by taking the difference between the residual sum of squares for the reduced model under $H_0$ and for the full model. The difficulty with implementing Schwertman's idea directly is that as the number of subjects increases, the order of the X'X matrix to be stored and inverted also increases, leading to core storage and CPU time requirements that can become very expensive, or even impossible, to meet. The algorithm presented here avoids the use of explicit subject effects and the resulting problem.

If the total number of missing values in the experiment is $m > 0$, our technique generates for each one a dummy covariate $Z_{vijk\ell}$ (v=1,2,...m) which equals -1 if observation $(i,j,k,\ell)$ is the $v^{th}$ missing value, and 0 otherwise. Model equation 1 is then altered as follows:

$$Y_{ijk\ell} = \mu_{ijk} + \sum_{v}(\gamma_v)(Z_{vijk\ell}) + \varepsilon_{ijk\ell}, \qquad (2)$$

where $\mu_{ijk}$ is the same as in equation 1; $Y_{ijk\ell} = 0$ if the observation is missing; $Z_{vijk\ell}$ is the dummy covariate and $\gamma_v$ the regression coefficient associated with the $v^{th}$ missing observation (v=1,2,...m).

It should be noted that our assumptions concerning multisample sphericity of the error term are not altered in equation 2 by the presence of missing data. If $Y_{ijk\ell} = 0$ represents the $v^{th}$ missing value, then we have

$$0 = \mu_{ijk} + (-1)(\mu_{ijk} + \varepsilon_{ijk\ell}) + \varepsilon_{ijk\ell}, \qquad (3)$$

where $(\mu_{ijk} + \varepsilon_{ijk\ell}) = \gamma_v$ and $(-1) = Z_{vijk\ell}$ in equation 2.

There are limitations on the number and configuration of missing observations we can have for any treatment combination (i,j) and still be able to compute the within-subject analysis. Within each cell, when all subjects in that cell have complete data, there are $(n_{ij}-1)(t-1)$ degrees of freedom for the residual sum of squares. Each missing observation in the cell costs one degree of freedom, so that if there are $m_{ij}$ missing values in the (i,j) cell, the number of degrees of freedom the cell contributes to the residual sum of squares is $(n_{ij}-1)(t-1) - m_{ij}$. Our algorithm requires that this quantity be nonnegative within every cell and that it be positive for at least one cell. In addition, certain configurations of missing values violate our requirements for the within-subject analysis. Two examples would be missing data for 1) all times on a given subject or 2) all subjects within a cell for a particular time. Either too many missing values or the wrong configuration of missing values will result in our trying to invert a singular matrix. The matrix inversion routines in REP2W1F recognize singularities, and when one is found, the program immediately prints a message to indicate what has happened and terminates the analysis.

When both the between- and within-subject analyses can be computed, REP2W1F produces an analysis of variance table which includes all the source lines shown in Table 1. The use of dummy covariates to adjust source lines in the table for missing values has previously been recommended by Jarrett [7] based on the ideas of Wilkinson [13]. The approach as they described it involves building and testing effects in a hierarchical manner instead of making use of the Σ-restrictions to test main effects and first-order interactions adjusted for all other effects in the model. As a result, their tests for effects due to A, B, Time, AxTime, and BxTime would differ from those to be presented here.

TABLE 1. SOURCE LINES AND DEGREES OF FREEDOM FOR THE ANALYSIS OF VARIANCE BASED ON MODEL EQUATIONS 1 AND 2

| | Source | Degrees of freedom |
|---|---|---|
| Between subjects | A | a-1 |
| | B | b-1 |
| | AxB | (a-1)(b-1) |
| | Error(a) | $\sum_{ij}[(n_{ij}-1) - r_{ij}]$ |
| Within subjects | Time | t-1 |
| | AxTime | (a-1)(t-1) |
| | BxTime | (b-1)(t-1) |
| | AxBxTime | (a-1)(b-1)(t-1) |
| | Error(b) | $\sum_{ij}[(n_{ij}-1)(t-1) - m_{ij}]$ |

## THE MODEL IN MATRIX NOTATION

We now introduce the matrix notation we use in describing the computing algorithm, which will make it possible to present the derivations in a more compact form. Starting with equations 1 and 2, let $\underline{Y} = (Y_{ijk\ell})$ denote the vector of observations (including zeros for missing values), sorted so that k is the fastest moving index, $\ell$ the next fastest, and i the slowest. Similarly, define the error vector $\underline{\varepsilon} = (\varepsilon_{ijk\ell})$ so that it conforms with $\underline{Y}$.

In writing the parameter vector $\underline{\beta}$, we eliminate certain of the main effects and interactions which are redundant due to the $\Sigma$-restrictions. For the main effects, this is accomplished by arbitrarily selecting the effect with the largest subscript for deletion, and since, for example, $A_a = -(A_1+A_2+\ldots A_{a-1})$, these deleted effects are not needed in the calculations to follow. Effects for a-1 levels of factor A, b-1 levels for factor B, and t-1 levels of Time will thus be included in $\underline{\beta}$, and any interaction effect which involves the $a^{th}$ level of A, the $b^{th}$ level of B, or the $t^{th}$ level of Time can be deleted.

It is convenient to write $\underline{\beta}$ as a partitioned vector with $\underline{\beta}' = (\underline{\beta}_1' | \underline{\beta}_2' | \underline{\beta}_3')$ where $\underline{\beta}_1$ contains the between-subject parameters, $\underline{\beta}_2$ contains the within-subject parameters, and $\underline{\beta}_3 = (\gamma_v)$ contains the regression coefficients associated with the missing values. A total of (a)(b) elements are contained in $\underline{\beta}_1$, including (in order) $\mu$, a-1 main effects from factor A, b-1 main effects from factor B, and (a-1)(b-1) first-order interaction effects from AxB. There are (a)(b)(t-1) elements in $\underline{\beta}_2$, the first t-1 of which are the main effects for Time, followed by (a-1)(t-1) effects for AxTime, (b-1)(t-1) effects for BxTime, and (a-1)(b-1)(t-1) effects for AxBxTime. In $\underline{\beta}_3$ are m elements, the number of missing observations. Shown in equation 4 for the case of a=2, b=3, and t=2 are $\underline{\beta}_1'$ and $\underline{\beta}_2'$ .

$$\underline{\beta}_1' = (\mu\ A_1\ B_1\ B_2\ (AB)_{11}\ (AB)_{12})$$

and

$$\underline{\beta}_2' = (T_1\ (AT)_{11}\ (BT)_{11}\ (BT)_{21}\ (ABT)_{111}\ (ABT)_{121}). \tag{4}$$

We also use the notation $J_t$ for a t-vector of 1's; $I_t$ for an identity matrix of order t; and $K_t$ for the (t)x(t-1) partitioned matrix defined as

$$K_t = \left(\frac{I_{t-1}}{-J'_{t-1}}\right). \tag{5}$$

The Kronecker product operator ⊗ is defined in the following illustration. If $U = (u_{ij})$ is a (2)x(3) matrix and V is a matrix, then the Kronecker product of U and V is the partitioned matrix

$$U \otimes V = \begin{pmatrix} u_{11}V & u_{12}V & u_{13}V \\ u_{21}V & u_{22}V & u_{23}V \end{pmatrix} \tag{6}$$

Equation 2 can now be rewritten in matrix notation as

$$\underline{Y} = (D\otimes J_t \mid D\otimes K_t \mid Z)\begin{pmatrix} \underline{\beta}_1 \\ \underline{\beta}_2 \\ \underline{\beta}_3 \end{pmatrix} + \underline{\varepsilon}, \tag{7}$$

where D is a matrix discussed in the following paragraph, and $Z = (Z_{vijk\ell})$ is the matrix of dummy covariates (needed only if missing values occur).

The D matrix is the design matrix we would associate with $\underline{\beta}_1$ for an experiment with only one reading per subject (that is, no repeated measures). It has one row for each subject in the experiment and one column for each of the (a)(b) elements of $\underline{\beta}_1$. This means that D does not change as the number of repeated measures increases. Our requirement that $(n_{ij}-1)(t-1) - m_{ij}$ be nonnegative within every treatment combination (i,j) means that $n_{ij} > 0$ and D is of full column rank. An example of the D matrix for the case of a=2, b=3, and $n_{ij}$=2 for all cells is given in Table 2.

Because any pair of error elements from distinct subjects are independent, the variance of the vector $\underline{\varepsilon}$ is block diagonal. The matrices down the diagonal we denote by $\Sigma_{ij\ell}$ ($\ell$=1,2,...$n_{ij}$), where $\Sigma_{ij\ell}$ is the covariance matrix for the errors of the $\ell^{th}$ subject in the cell (i,j). Huynh and Feldt [6] showed that because of multisample sphericity, each $\Sigma_{ij\ell}$ may be written as:

$$\Sigma_{ij\ell} = (\underline{\alpha}_{ij\ell}J'_t) + (J\ \underline{\ }_{ij\ell}) + \nu_{ij}, \tag{8}$$

where $\underline{\alpha}_{ij\ell}$ is a t-vector which may change from subject to subject, and $\lambda$ is a constant which is the same across all subjects.

TABLE 2. EXAMPLE OF THE MATRIX DENOTED BY D IN EQUATION 7 FOR THE CASE OF a=2, b=3, AND $n_{ij}$=2 FOR ALL CELLS

| Subscripts of Y | | | | | Elements of D | | | | |
|---|---|---|---|---|---|---|---|---|---|
| i | j | $\ell$ | $\beta_1$ = | $\mu$ | $A_1$ | $B_1$ | $B_2$ | $(AB)_{11}$ | $(AB)_{12}$ |
| 1 | 1 | 1 | | 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 1 | 2 | | 1 | 1 | 1 | 0 | 1 | 0 |
| 1 | 2 | 1 | | 1 | 1 | 0 | 1 | 0 | 1 |
| 1 | 2 | 2 | | 1 | 1 | 0 | 1 | 0 | 1 |
| 1 | 3 | 1 | | 1 | 1 | -1 | -1 | -1 | -1 |
| 1 | 3 | 2 | | 1 | 1 | -1 | -1 | -1 | -1 |
| 2 | 1 | 1 | | 1 | -1 | 1 | 0 | -1 | 0 |
| 2 | 1 | 2 | | 1 | -1 | 1 | 0 | -1 | 0 |
| 2 | 2 | 1 | | 1 | -1 | 0 | 1 | 0 | -1 |
| 2 | 2 | 2 | | 1 | -1 | 0 | 1 | 0 | -1 |
| 2 | 3 | 1 | | 1 | -1 | -1 | -1 | 1 | 1 |
| 2 | 3 | 2 | | 1 | -1 | -1 | -1 | 1 | 1 |

## COMPUTATIONAL ALGORITHM

### Between Subjects

If no missing values occur in the data to be analyzed, we compute the average value across Time for each subject, using the notation $\bar{Y}_{ij\cdot\ell} = \frac{1}{t}\Sigma_k Y_{ijk\ell}$. When these values are considered in terms of the $\Sigma$-restrictions on the within-subject parameters, we obtain equation 9:

$$\bar{Y}_{ij\cdot\ell} = \mu + A_i + B_j + (AB)_{ij} + \bar{\varepsilon}_{ij\cdot\ell}. \qquad (9)$$

In our matrix notation, if there were N subjects in the experiment, equation 9 could be rewritten as:

$$\frac{1}{t}(I_N \otimes J_t')\underline{Y} = \frac{1}{t}(I_N \otimes J_t')(D \otimes J_t \mid D \otimes K_t)\begin{pmatrix}\underline{\beta}_1\\ \underline{\beta}_2\end{pmatrix} + \frac{1}{t}(I_N \otimes J_t')\underline{\varepsilon}. \qquad (10)$$

Since $J_t'K_t = 0$, equation 10 reduces to

$$\frac{1}{t}(I_N \otimes J_t')\underline{Y} = \frac{1}{t}(D \otimes J_t'J_t)\underline{\beta}_1 + \frac{1}{t}(I_N \otimes J_t')\underline{\varepsilon}. \qquad (11)$$

Equation 11 has the familiar general form $\underline{y} = X\underline{\beta}_1 + \underline{e}$ for linear models. The assumptions on the variances of the errors, namely independence between subjects and homogeneity of $\bar{\varepsilon}_{ij\cdot\ell}$ across subjects, make the between-subject analysis straightforward using well-known methods for obtaining tests of main effects and interactions, all mutually adjusted for each other [12].

Useful summary statistics printed by REP2W1F for the between-subject analysis include the table of AxB means (the arithmetic means within each (i,j) cell) and the marginal means for A and B (the unweighted averages of these cell means).

If missing values do occur for some subjects, an appropriate between-subject analysis can still be computed by limiting this part of the analysis to only those subjects with complete data, which REP2W1F does, provided every (i,j) cell has at least one subject with complete data, and at least one cell has two or more subjects with complete data. When these last two conditions are not met, REP2W1F prints a message to this effect and moves on to the within-subject analysis. If we use BMDP2V or BMDP4V with dummy covariates to compute the entire analysis in a single call of the program, the between-subject portion of the analysis is equivalent to that just described. A generalized least-squares [11] analysis, making use of available data from subjects with values missing and in which tests of the between-subject parameters are adjusted for the within-subject factors, would require additional knowledge of, or assumptions about, the covariance matrix of the error vector $\underline{\varepsilon}$.

## Within Subjects

Consider a transformation of the Y values in which the last observation for each of the N subjects is subtracted from each of the first t-1 levels, and the last value then dropped. This is equivalent to multiplying each side of equation 7 by the matrix $I_N \otimes K_t'$, giving:

$$(I_N \otimes K_t')\underline{Y} = (I_N \otimes K_t')(D \otimes J_t \,|\, D \otimes K_t \,|\, Z)\begin{pmatrix}\underline{\beta}_1\\ \underline{\beta}_2\\ \underline{\beta}_3\end{pmatrix} + (I_N \otimes K_t')\underline{\varepsilon}. \qquad (12)$$

This transformation leads to two simplifications: first, $(I_N \otimes K_t')(D \otimes J_t) = (I_N\ D)\otimes(K_t'J_t)$, but $K_t'J_t = \underline{0}$, which means that the coefficients multiplying the elements of $\underline{\beta}_1$ in equation 12 are all 0's; second, $(I_N \otimes K_t')(D \otimes K_t) = D \otimes K_t'K_t$. Equation 12 therefore reduces to

$$(I_N \otimes K_t')\underline{Y} = (D \otimes K_t'K_t \,|\, (I_N \otimes K_t')Z)\begin{pmatrix}\underline{\beta}_2\\ \underline{\beta}_3\end{pmatrix} + (I_N \otimes K_t')\underline{\varepsilon}. \qquad (13)$$

Now we introduce the following notation:

$$V = K_t'K_t = I_{t-1} + J_{t-1}J_{t-1}',$$

$$\underline{Y}^* = (I_N \otimes K_t')\underline{Y},$$

$$Z^* = (I_N \otimes K_t')Z,$$

$$\underline{\varepsilon}^* = (I_N \otimes K_t')\underline{\varepsilon},$$

and

$$\sigma^2 = \lambda. \tag{14}$$

If equation 13 is rewritten in this new notation, the result is

$$\underline{Y}^* = (D \otimes V|Z^*)\begin{pmatrix}\underline{\beta}_2\\ \underline{\beta}_3\end{pmatrix} + \underline{\varepsilon}^*. \tag{15}$$

The form of equation 15 is the familiar linear model.

Now we need to find the variance of $\underline{\varepsilon}^*$. We have $\text{Var}(\varepsilon^*) = (I_N \otimes K_t')\text{Var}(\underline{\varepsilon})(I_N \otimes K_t)$, where $\text{Var}(\underline{\varepsilon})$ is block diagonal with the matrices $\Sigma_{ij\ell}$ from equation 8 down the diagonal. $\text{Var}(\underline{\varepsilon}^*)$ is also block diagonal, with the matrices $K_t'\Sigma_{ij\ell}K_t$ on the diagonal. Finally, using equation 8 we have

$$K_t'\Sigma_{ij\ell}K_t = K_t'((\underline{\alpha}_{ij\ell}J_t') + (J_t\underline{\alpha}_{ij\ell}) + \lambda I_t)K_t = \lambda K_t'K_t = \sigma^2 V, \tag{16}$$

since $J_t'K_t = \underline{0}'$ and $K_t'J_t = \underline{0}$.

Equations 14, 15, and 16 combine to give equation 17.

$$\text{Var}(\underline{\varepsilon}^*) = \sigma^2(I_N \otimes V), \tag{17}$$

where V is known and $\sigma^2$ is unknown. Because of equations 15 and 17, the normal equations to be solved in obtaining the generalized least-squares estimates [11] of $\underline{\beta}_2$ and $\underline{\beta}_3$ are as follows:

$$\left[\begin{array}{c|c} D'D \otimes V & (D' \otimes I_{t-1})Z^* \\ \hline Z^{*\prime}(D \otimes I_{t-1}) & Z^{*\prime}(I_N \otimes V^{-1})Z^* \end{array}\right] \begin{bmatrix} \hat{\underline{\beta}}_2 \\ \hat{\underline{\beta}}_3 \end{bmatrix} = \left[\begin{array}{c} D' \otimes I_{t-1} \\ \hline Z^{*\prime}(I_N \otimes V^{-1}) \end{array}\right] \underline{Y}^*. \quad (18)$$

You can reduce computational effort by observing that $\underline{Y}^*$ and $Z^*$ can be partitioned as shown in equation 19, with each pair $(\underline{Y}_n^*, Z_n^*)$ associated with one subject.

$$\underline{Y}^* = \begin{bmatrix} \underline{Y}_1^* \\ \underline{Y}_2^* \\ \vdots \\ \underline{Y}_N^* \end{bmatrix} \qquad Z^* = \begin{bmatrix} Z_1^* \\ Z_2^* \\ \vdots \\ Z_N^* \end{bmatrix} . \quad (19)$$

This produces normal equations given by

$$\left[\begin{array}{c|c} D'D \otimes V & (D' \otimes I_{t-1})Z^* \\ \hline Z^{*\prime}(D \otimes I_{t-1}) & \sum_n (Z_n^{*\prime} V^{-1} Z_n^*) \end{array}\right] \begin{bmatrix} \hat{\underline{\beta}}_2 \\ \hat{\underline{\beta}}_3 \end{bmatrix} = \left[\begin{array}{c} (D' \otimes I_{t-1})\underline{Y}^* \\ \hline \sum_n (Z_n^{*\prime} V^{-1} \underline{Y}_n^*) \end{array}\right] \quad (20)$$

In equation 20 the only matrices which must be stored and whose orders depend on N are $\underline{Y}^*$ and $Z^*$; the normal equations shown are in the form we use for computational purposes in REP2W1F. Conceptually, however, it is helpful in the remaining discussion to back up to equation 18 and rewrite it as

$$W\hat{\underline{\theta}} = U\underline{Y}^* \quad (21)$$

where

$$W = \left[\begin{array}{c|c} D'D \otimes V & (D' \otimes I_{t-1})Z^* \\ \hline Z^{*\prime}(D \otimes I_{t-1}) & Z^{*\prime}(I_N \otimes V^{-1})Z^* \end{array}\right],$$

$$\hat{\underline{\theta}} = \begin{bmatrix} \hat{\underline{\beta}}_2 \\ \hat{\underline{\beta}}_3 \end{bmatrix},$$

and

$$U = \left[\begin{array}{c} D' \otimes I_{t-1} \\ \hline Z^{*\prime}(I_N \otimes V^{-1}) \end{array}\right].$$

The within-subject parameter estimates can be computed as in equation 22 provided W is nonsingular.

$$\underline{\hat{\theta}} = W^{-1}U\underline{Y}^{*}. \tag{22}$$

The order of the matrix W is (a)(b)(t-1) + m, which does not involve N. This is where the savings in computer core storage and CPU processing-time requirements are achieved by REP2W1F relative to SAS GLM. If m = 0 (meaning there are no missing values), then W = D'D ⊗ V can be shown to be a nonsingular matrix under the condition we have imposed that $(n_{ij}-1)(t-1)$ be nonnegative within every cell (i,j) and positive for at least one cell. Under this condition, the rank of D'D, which is (a)(b), equals the order of D'D, so that D'D is nonsingular. Additionally, V is nonsingular. Since the eigenvalues of D'D ⊗ V are equal to $\delta_r v_s$, where $\delta_r$ is the $r^{th}$ eigenvalue of D'D and $v_s$ is the $s^{th}$ eigenvalue of V, D'D ⊗ V has (a)(b)(t-1) nonzero eigenvalues and is thus nonsingular.

If m > 0, the possibility exists that W might be singular. This will happen if $(n_{ij}-1)(t-1) - m_{ij} < 0$ for one of the cells (i,j). It can also happen for certain configurations of missing values such as all times missing for a given subject or all subjects missing within a cell for a particular time. The matrix inversion routine in REP2W1F uses Cholesky decomposition in finding $W^{-1}$. When the lower triangular matrix T such that W = TT' has been found, the elements along the diagonal of T are compared. If the ratio of the smallest to largest elements is less than $10^{-5}$, the matrix is declared to be singular. In such cases a message is printed and the particular analysis terminates.

Provided W is nonsingular, the quantitites needed to test linear hypotheses of the form $C\underline{\theta} = \underline{0}$ are straightforward [11, pp. 110-112].

$$SSE = \underline{Y}^{*}'(I_N \otimes V^{-1})\underline{Y}^{*} - (U\underline{Y}^{*})'W^{-1}(U\underline{Y}^{*}) \tag{23}$$

with degrees of freedom (N-(a)(b))(t-1) - m,

$$\hat{\sigma}^2 = SSE \;/\; \{(N-(a)(b))(t-1) - m\}, \tag{24}$$

$$Q = \text{estimated variance of } \hat{\theta} = \hat{\sigma}^2(W^{-1}U(I_N \otimes V)U'W^{-1}), \tag{25}$$

$$SSH(H_0\colon C\underline{\theta} = \underline{0}) = (C\underline{\hat{\theta}})'(CQC')^{-1}(C\underline{\hat{\theta}}), \tag{26}$$

where C is required to have full row rank equal to h, which is also the degrees of freedom for SSH. REP2W1F employs efficient algorithms equivalent to these formulas to calculate the sums of squares for lines in the within-subject portion of Table 1. To illustrate in more detail what is being computed, suppose a=2, b=3, t=2, and m=2, so that

$$\underline{\beta}_2' = (T_1\ (AT)_{11}\ (BT)_{11}\ (BT)_{21}\ (ABT)_{111}\ (ABT)_{121}) \tag{27}$$

and

$$\underline{\beta}_3' = (\gamma_1\ \gamma_2).$$

The adjusted hypothesis sums of squares to test for an overall main effect due to Time and for interactions involving Time are shown in equation 26 with C given as:

$$\text{Time:}\quad C = (\,1\,0\,0\,0\,0\,0\,0\,0\,),$$

$$\text{AxTime:}\quad C = (\,0\,1\,0\,0\,0\,0\,0\,0\,),$$

$$\text{BxTime:}\quad C = \begin{bmatrix} 0\,0\,1\,0\,0\,0\,0\,0 \\ 0\,0\,0\,1\,0\,0\,0\,0 \end{bmatrix},$$

and

$$\text{AxBxTime:}\quad C = \begin{bmatrix} 0\,0\,0\,0\,1\,0\,0\,0 \\ 0\,0\,0\,0\,0\,1\,0\,0 \end{bmatrix}. \tag{28}$$

The sum of squares for Error(b) is SSE in equation 23.

All of these sums of squares are the same as the Type III and Type IV sums of squares from SAS GLM when "subjects within treatment groups" is included as a term in the model for that program. These are also the same as the sums of squares given by BMDP2V and BMDP4V when dummy covariates are used in these programs to adjust for missing values.

Adjusted means for the within-subject portion of the analysis are obtained after inserting the estimates of the missing values (from $\hat{\underline{B}}_3$) back into their respective locations in the original data vector $\underline{Y}$. The AxBxTime means are computed from the reconstituted $\underline{Y}$ vector. These are the same as the AxBxTime least-squares means produced by SAS GLM. The marginal means for Time, AxTime, and BxTime in REP2W1F are the unweighted averages of these adjusted AxBxTime means. The SAS GLM procedure can be manipulated to give these means, provided we do not have both unbalanced treatment groups and missing data simultaneously. The BMDP4V program can generate these means if the cell sizes are unequal, but none of the BMDP programs print useful summary means unless all subjects have complete data.

Standard errors for the adjusted means are also produced by REP2W1F. Morris et al. [8] have discussed the shortcomings of SAS GLM and BMDP2V in this regard. The fact that REP2W1F is written to analyze data from one particular design makes it possible to compute the standard errors of interest (although in the case of missing data, these are only approximate).

## CONCLUSION

We have described the method of analysis incorporated in REP2W1F, a SAS procedure for the univariate analysis of a repeated measurements experiment where the subjects are arranged in a two-way classification with treatment groups that may be unequal in size. The program assumes a Σ-restricted model [12] to produce analysis of variance F-tests and least-squares means that are adjusted for incomplete data. The computer resources required by REP2W1F to obtain least-squares means when we have missing values do not escalate nearly as rapidly as a function of the number of subjects in the experiment as SAS GLM. For example, in our introduction we showed a situation where SAS GLM required 504K bytes of storage and 5.04 minutes of CPU time. The same analysis using REP2W1F consumed 202K bytes and 0.23 minutes. When any data are missing, BMDP2V and BMDP4V do not produce any useful summary least-squares means at all. The REP2W1F program produces standard errors which can be used to place confidence intervals on, or test for differences between, selected least-squares means. As pointed out by Morris et al. [8], SAS GLM and BMDP2V are deficient in this regard. Our studies of BMDP4V indicate that it is also deficient. A generalization of the algorithm in REP2W1F to situations where the number of treatment factors is other than two, or where the repeated measures have a factorial arrangement of their own, would be straightforward.

## REFERENCES


1. BMDP Statistical Software. Berkeley, Calif.: University of California Press, 1981.

2. Frane, J. W. The univariate approach to repeated measures--foundation, advantages, and caveats (preliminary version). BMDP Technical Report No. 69, May 1980.

3. Frane, J. W. Some remarks regarding BMDP repeated-measures software for the 1981 annual meeting of the American Statistical Association. BMDP Technical Report No. 79, Aug 1981.

4. Freund, R. J., and R. C. Littell. SAS for linear models--A guide to the ANOVA and GLM procedures. Cary, N.C.: SAS Institute, Inc., 1981.

5. Huynh, H. Some approximate tests for repeated measurements designs. Psychometrika 43:161-175 (1978).

6. Huynh, H., and L. S. Feldt. Conditions under which mean square ratios in repeated measurements designs have exact F-distributions. J Am Stat Assoc 65:1582-1589 (1970).

7. Jarrett, R. G. The analysis of designed experiments with missing observations. Appl Stat 27:38-46 (1978).

8. Morris, D. D., et al. Using SAS for estimation and hypothesis testing in an unbalanced repeated measure design. Proc Seventh Annual SUGI Conference. Cary, N.C.: SAS Institute, Inc., 1982.

9. SAS User's Guide. Raleigh, N.C.: SAS Institute, Inc., 1979.

10. Schwertman, N. C. A note on the Geisser-Greenhouse correction for incomplete data split-plot analysis. J Am Stat Assoc 73:393-396 (1978).

11. Searle, S. R. Linear models. New York: John Wiley and Sons, Inc., 1971.

12. Searle, S. R., et al. Some computational and model equivalences in analysis of variance of unequal-subclass-numbers data. Am Stat 35:16-33 (1981).

13. Wilkinson, G. N. The analysis of variance and derivation of standard errors for incomplete data. Biometrics 58:369-384 (1958).

can be easily obtained provided all treatment groups are of equal size by



END

FILMED

3-83

DTIC