Program Abstracts/Algorithms

Use of a generalized analysis of variance program in repeated measures designs having unequal group sizes

JAMES J. LYNCH

Fordham University, New York, New York 10458

Most computer center libraries contain a group of programs to perform repeated measures analysis of variance. When group sizes are equal, the form of the program is a straight-forward application of a general N-way analysis of variance program (Clifford, 1968). "Tailor-made" programs, however, must be written for analyses of variance in which group sizes are unequal. These are generally cookbook solutions, based on textbook cases.

A more general solution uses an N-way analysis of variance program, such as that performed by the subroutines AVDAT, AVCAL, and MEANQ of the Scientific Subroutine Package (SSP) (IBM, 1968), which are based on the decomposition algorithm of Hartley (1960). The analysis presented here is based on the unweighted means analysis described by Winer (1971). The computation is performed in three stages. First, within-subjects analyses of variance are performed to calculate all sums of squares required for error terms. Next, an analysis of variance is performed on the cell means of the data, and the resulting sums of squares are scaled up by the harmonic mean of the cell sizes. The final analysis of variance table is then composed and all F-ratios are calculated.

The FORTRAN IV coding to perform this procedure is listed in Table 1. The program was run on an IBM 360 Model 40. As presently composed, the program will analyze a repeated measures analysis of variance design of up to eight factors, with up to N-1 (maximum = 7) repeated measures. The major array is the X array, which is used for two purposes. The first segment, up to X(N), is used to store the data. The remaining segment, beginning at X(N + 1), is used as the work area for the SSP analysis of variance subroutines.

Parameters are read in the initial part of the program. The variable HD is a 16-character alphameric title of the problem being run. The number of between-subject factors is labeled NB, and the number of within-subject factors is labeled NW. The variables HD, NB, and NW should be typed in successive fields, on the first data card. The array LEVEL contains the number of levels of each factor, the within-subject factors appearing first. The first entry in LEVEL is the number of levels of the within-subject factor which has the most rapidly changing subscript. The rest of the within-subject factor levels follow in order. Then the levels of the between-subject factors are specified, in order, with the earliest entry being the number of levels of the between-subject factor having the most rapidly changing subscript. The variables in LEVEL should be typed in successive fields on the second data card. The variable FORMAT (named FMT) contains the FORMAT information for data input, and should specify the location of each data point of the data card. This is the third data card.

James J. Lynch is now at College of Mount St. Vincent, Riverdale, New York 10471.

The next card specified the names of the factors in the analysis of variance. Each is 16 alphameric characters long, and should be left justified. The names of the between-subject factors appear first. The first name corresponds to the between-subject factor whose subscript changes most rapidly, followed by the other factors in order. The within-subject factor names appear next. The first name corresponds to the within-subject factor whose subscript changes most rapidly followed by the other factor names in order.

The unidimensional array NSIZE contains the number of subjects in each between-subject group, listed in the order in which the groups are arranged for input. The values for NSIZE should be typed in successive fields on the next data card. Note the example in the test data at the end of the program listing for the specific way this is accomplished.

The data are read next. The program is arranged so that the data can be punched in a very natural way, with all the data for one subject on one card or on one group of successive cards. The data should, however, be nested on the card according to the factors of the design. Likewise, the blocks of subjects representing different between-subject groups should be nested according to the scheme specified in LEVEL.

The analysis then proceeds. Each between-subject block of data is moved into the working area of the X array, and an analysis of variance is performed, with subjects included as a pseudofactor. The sums of squares and the degrees of freedom are accumulated from each group in turn. This technique thus applies the definitional formulation of the within-subject error terms.

The second phase of the analysis begins with the computation and printing of the cell means along with the data in each cell. As each mean is computed, it is stored in the appropriate location of the working area of the X array. This array of means is then passed to the SSP subroutines, and the returned sums of squares are multiplied by the harmonic mean of the cell sizes.

In the third phase of the analysis, the between-subject F ratio(s) are calculated and printed, followed by the within-subject F ratio(s).

The summary table presents a factor label (a letter) followed by the name of the factor. All effects are then listed by their label, followed by their sum of squares, degrees of freedom, mean square, and F ratio.

Modification of the program revolves around the size of the arrays specified in the DIMENSION statement. The size of the arrays LEVEL, ISTEP, KOUNT, LASTS, and MSTEP are determined by the number of factors in the overall analysis of variance, that is, the sum of NB and NW, which is specified as K in the program. The length of SUMSQ, SMEAN, and NDF are all 2**K-1, while the length of SSUMSQ, NNDF, and SSMEAN are all 2**NW-1. The length of NSIZE is the product of the levels of the between-subjects factors. The length of the X-array is the number of data points analyzed plus the largest area required for the generalized analysis of variance. The latter may be either one of the within-subjects analyses of variance or the between-subjects analyses of variance. Exact formulas for determining minimum array size for the generalized analysis of variance may be found in the IBM SSP Manual (1968). With the exception of the FMT array, all other arrays are for the purpose of labeling and are fixed in size. The variables I1 and I2 are used to pick the proper labels for the analysis of variance table when it is being printed. The constants associated with these

TABLE 1
FORTRAN CODE OF MAIN PROGRAM FOR UNWEIGHTED MEANS
REPEATED MEASURES ANALYSIS OF VARIANCE

```

      DIMENSION LEVEL(8),ISTEP(8),KOUNT(8),LASTS(8),MSTEP(8),
      1SUMSQ(255),SMEAN(255),NDF(255),SSUMSQ(127),NNDF(127),SSMEAN(127),
      2      NSIZE(100),FMT(20),HD(4),AST(18),SOURCE(32),LABELR(4,8),
      3LL(8),X(2000)
      DATA AST /17*4H***,4H** /,SOURCE /4H ,4H A,4H B,4H AB,
      14H C,4H AC,4H BC,4H ABC,4H D,4H AD,4H BD,4H ABD,4H CD,
      24H ACD,4H BCD,4H ABCD,4H ,4HE ,4HF ,4HEF ,4HG ,4HEG ,
      34HFG ,4HEFG ,4HH ,4HEH ,4HGH ,4HEFH ,4HGH ,4HEGH ,4HFGH ,
      44HEFGH/, LL /1HA,1HB,1HC,1HD,1HE,1HF,1HG,1HH/
C READ PARAMETERS
      1 READ(1,2) HD,NB,NW
      2 FORMAT(4A4,2I4)
      K=NB*NW
      READ(1,3) (LEVEL(I),I=1,K)
      3 FORMAT(20I4)
      READ(1,30) FMT
      READ(1,30) ((LABELR(I,J),I=1,4),J=1,K)
      30 FORMAT(20A4)
      NCELLS=1
      NNW=NNW+1
      NLEVEL=LEVEL(NNW)
      DO 12 I=NNW,K
      12 NCELLS=NCELLS*LEVEL(I)
      READ(1,3) (NSIZE(I),I=1,NCELLS)
      NBAREA=1
      NCL=1
      DO 11 I=1,K
      NCL=NCL*LEVEL(I)
      11 NBAREA=NBAREA*(LEVEL(I)+1)
      JNW=1
      NVW=1
      DO 50 I=1,NW
      JNW=JNW*(LEVEL(I)+1)
      50 NVW=NVW*LEVEL(I)
      N=0
      HMEAN=0.0
      DO 55 I=1,NCELLS
      N=N+NVW*NSIZE(I)
      55 HMEAN=HMEAN+1.0/NSIZE(I)
      HMEAN=NCELLS/HMEAN
C READ DATA
      READ(1,FMT) (X(I),I=1,N)
C INITIALIZE VARIABLES FOR WITHIN CELLS ANALYSIS
      LENGTH=2**(NW+1)-1
      DO 80 I=1,LENGTH
      SSUMSQ(I)=0.0
      80 NNDF(I)=0
      IX=0
C PERFORM WITHIN CELLS ANALYSIS
      DO 90 I=1,NCELLS
      NR=N+2
      LLEVEL(NW+1)=NSIZE(I)
      JNX=JNW*(NSIZE(I)+1)+N
      DO 82 J=NR,JNX,2
      X(J-1)=0.0
      82 X(J)=0.0
      NNW=NVW*NSIZE(I)
      DO 85 J=1,NNW
      85 X(N+J)=X(IX+J)
      CALL AVDAT (NW+1,LEVEL,NNW ,X(N+1),L,ISTEP,KOUNT)
      CALL AVCAL (NW+1,LEVEL,X(N+1),L,ISTEP,LASTS)
      CALL MEANG (NW+1,LEVEL,X(N+1),G,SUMSQ,NDF,SMEAN,MSTEP,KOUNT,LASTS)
      DO 90 J=1,LENGTH
      SSUMSQ(J)=SSUMSQ(J)+SUMSQ(J)
      90 NNDF(J)=NNDF(J)+NDF(J)
      95 IX=IX+NNW
      DO 100 I=1,LENGTH
      100 SMEAN(I)=SSUMSQ(I)/NNDF(I)
C PERFORM BETWEEN CELLS ANALYSIS
      LENGTH=2**K-1
      LEVEL(NW+1)=NLEVEL
      NNW=N+NBAREA
      NR=N+2
      DO 20 I=NR,NNW,2
      X(I-1)=0.0
      20 X(I)=0.0
C GENERATE CELL MEANS
      WRITE(3,155) HD
      155 FORMAT(1H1,52X,15HLISTING OF DATA,/,53X,15HWITH CELL MEANS,/,52X,
      14A4,/,6X,10HCELL MEANS,19X,12HDATA IN CELL)
      IX=0
      IN=N
      DO 60 I=1,NCELLS
      FNO=NSIZE(I)
      DO 7 J=1,NNW
      XX=0.0
      IX=IX+J*NNW
      NNW=NSIZE(I)
      DO 6 II=1,NNW
      IX=IX+NNW
      SUMSQ(II)=X(IX)
      6 XX=XX+X(IX)
      XX=XX/FNO
      X(IN+J)=XX
      7 WRITE(3,65) XX,(SUMSQ(M),M=1,NNW)
      65 FORMAT(1H0,3X,E14.7,7X,13F7.2,/, (25X,13F7.2 ))
      IX=IX+NSIZE(I)*NNW
      60 IN=IN+NNW
C PERFORM BETWEEN CELLS ANALYSIS OF VARIANCE
      CALL AVDAT (K,LEVEL,NCL,X(N+1),L,ISTEP,KOUNT)
      CALL AVCAL (K,LEVEL,X(N+1),L,ISTEP,LASTS)
      CALL MEANG (K,LEVEL,X(N+1),G,SUMSQ,NDF,SMEAN,MSTEP,KOUNT,LASTS)
C MULTIPLY SUMS OF SQUARES BY HARMONIC MEAN, AND CALCULATE CORRECTED MEAN SQUARE
      DO 57 I=1,LENGTH
      SUMSQ(I)=HMEAN*SUMSQ(I)
      57 SMEAN(I)=SUMSQ(I)/NDF(I)
C INITIALIZE VALUES FOR ANOVA TABLE CONSTRUCTION
      WRITE(3,160) HD,AST
      160 FORMAT(1H1,47X,26HANALYSIS OF VARIANCE TABLE,/,52X,4A4,/,25X,
      118A4)
      WRITE(3,200) (LL(J),(LABELR(I,J),I=1,4),J=1,K)
      200 FORMAT(1H0, 16X,3(9X,14HLABEL SOURCE),/, (28X,1A1,3X,4A4,
      13X,1A1,3X,4A4,3X,1A1,3X,4A4))
      WRITE(3,161) AST,AST
      161 FORMAT(1H0,24X,18A4,/,30X,
      1 60HSOURCE SUM OF SQUARES DF MEAN SQUARE F RAT
      210 ,/,25X,18A4,/,27X,16HBETWEEN SUBJECTS,/)
      NBE=2**NB-1
      NERR=2**NW
C CALCULATE BETWEEN-SUBJECT F-RATIOS
      II=1
      I2=17
      ERROR=SSMEAN(NERR)
      J=0
      DO 120 I=1,NBE
      J=J+NERR
      I1=I1+1
      IF(I1-I7) 119,118,118
      118 I1=1
      I2=I2+1
      119 FRATIO=SMEAN(J)/ERROR
      120 WRITE(3,131) SOURCE(I1),SOURCE(I2),SUMSQ(J),NDF(J),SMEAN(J),FRATIO
      WRITE(3,130) SSUMSQ(NERR),NNDF(NERR),SSMEAN(NERR)
C CALCULATE WITHIN-SUBJECT F-RATIOS
      WRITE(3,121)
      121 FORMAT(1H0,26X,15HWITHIN SUBJECTS)
      NBE=NBE+1
      NNW=NERR-1
      DO 140 I=1,NNW
      WRITE(3,171)
      171 FORMAT(1H )
      INERR=NERR+1
      ERROR=SSMEAN(INERR)
      KN=I-NERR
      DO 150 J=1,NBE
      KN=KN+NERR
      I1=I1+1
      IF(I1-I7) 149,148,148
      148 I1=1
      I2=I2+1
      149 FRATIO=SMEAN(KN)/ERROR
      150 WRITE(3,131)SOURCE(I1),SOURCE(I2),SUMSQ(KN),NDF(KN),SMEAN(KN),
      1FRATIO
      140 WRITE(3,130) SSUMSQ(INERR),NNDF(INERR),SSMEAN(INERR)
      131 FORMAT(1H ,27X,2A4,3X,E14.7,2X,14,3X,E14.7,3X,E14.7)
      150 FORMAT(1H ,27X,5HERROR,6X,E14.7,2X,14,3X,E14.7,3X,E14.7)
      WRITE(3,122) AST
      122 FORMAT(1H0,24X,18A4)
      999 CALL EXIT
      END
C SAMPLE PROBLEM FROM: WINER,B.J. STATISTICAL PRINCIPLES IN EXPERIMENTAL
C DESIGN, SECOND ED. NY:MCGRAW-HILL, 1971.
      WINER,1971,P.601 1 1
      5 5
      (40F2.0)
      A B
      5 5 4
      5 6 9 61014101518 81216 3 5 8 1 3 8121826 91018102216 315 8 71610 52012

```

variables need not be adjusted unless the problem exceeds eight factors. The variable FORMAT array FMT may be adjusted for the size of the actual variable FORMAT used.

The computational technique presented in Table 1 allows the researcher maximum flexibility in the analysis of repeated-measures designs. The program presented allows analysis of those data sets in which, for reasons unrelated to the research, the number of subjects is not the same in all between-subject groups.

The computation of regression estimate factor scores for singular and nonsingular data

RICHARD J. HOFMANN

Miami University, Oxford, Ohio 45056

Let $\tilde{Z}$ be an N by p matrix of p standard score observations on N people such that $\tilde{Z}'\tilde{Z} = \tilde{R}$, the intercorrelation matrix of the p variables. Let $\tilde{A}$ be an oblique primary structure, or oblique reference pattern or an orthogonal structure matrix, p by m , where m designates the number of factors. Assume $\tilde{R}$ to be either singular or nonsingular. Resolve $\tilde{R}$ completely into p eigenvalues $\tilde{M}^2$ and associated unit length eigenvectors $\tilde{Q}$ such that $\tilde{R} = \tilde{Q}\tilde{M}^2\tilde{Q}'$. Then, take the nonzero eigenvalues of $\tilde{M}^2$ and compute their reciprocals. Call the matrix of nonzero reciprocals $\tilde{M}^{-2*}$. Then the inverse of a nonsingular intercorrelation matrix, $\tilde{R}^{-1}$, is just $\tilde{Q}\tilde{M}^{-2*}\tilde{Q}'$. However, for a singular matrix $\tilde{R}$, the product $\tilde{Q}\tilde{M}^{-2*}\tilde{Q}'$ has all of the algebraic properties of a generalized inverse (Graybill, 1969) and is just $\tilde{R}^-$. Thus, the regression factor scores F (Mulaik, 1972) may be generally defined as

$$F = \tilde{Z}\tilde{Q}\tilde{M}^{-2*}\tilde{Q}'\tilde{A}$$

A program for latent partition analysis

RICHARD J. HOFMANN

Miami University, Oxford, Ohio 45056

Wiley (1967) has developed a multivariate procedure for the analysis of categorical or nominal data. His procedure, latent partition analysis, was formulated to study the "relationships" between two or more categorizations of the same set of items. Given certain assumptions, his procedure is *analogous* to an independent cluster solution of an iterated principal axis factor solution.

Input. Data that are determined by individual's sortings or placings of items into particular categories are used as input; ordinal data may also be thought of within the framework of categorical data (Likert-type data).

Output. Program output includes such matrices as the estimated latent category matrix (ELC) and the symmetric confusion matrix. The ELC matrix, analogous to a primary factor pattern matrix, indicates item membership in "hypothetical categories," assuming the hypothetical categories are mutually exclusive. The

REFERENCES

- Clifford, T. Simplified computational programming for analysis-of-variance designs with correlated observations. *Psychological Bulletin*, 1968, 69, 439-440.
Hartley, H. O. Analysis of variance. In A. Ralston and H. S. Wilf (Eds.), *Mathematical methods for digital computers*. New York: Wiley, 1960. Pp. 221-230.
IBM Scientific Subroutine Package. White Plains: IBM, 1968, Manual H20-0205-3.
Winer, B. J. *Statistical principles in experimental design*. (2nd ed.) New York: McGraw-Hill, 1971.

regardless of either the obliquity of the factors or the nature of the determinant of the variable intercorrelation matrix.

A series of subroutines have been written to compute factor scores according to this general algorithm. The subroutines require as input, the matrices $\tilde{A}$, $\tilde{R}$, and $\tilde{Z}$. To facilitate storage, $\tilde{Z}$ and $\tilde{R}$ are initially required to compute the product $\tilde{Z}\tilde{Q}\tilde{M}^{-2*}\tilde{Q}'$ which is stored on a scratch tape so that it may be used repeatedly with a variety of $\tilde{A}$ matrices regardless of whether or not $\tilde{R}$ is singular.

The subroutines are written in FORTRAN IV utilizing the principles of dynamic storage and single subscripting. They have been implemented on the IBM Model 370-168 G-level compiler as well as certain RCA, CDC, and UNIVAC models.

Routine listings as well as documentation illustrative data and computations are available from the author.

REFERENCES

- Graybill, F. A. *Introduction to matrices with applications in statistics*. Belmont: Wadsworth Publishing Company, 1969.
Mulaik, S. A. *The foundations of factor analysis*. New York: McGraw Hill, 1972.

off-diagonal elements of the confusion matrix, analogous to a primary intercorrelation matrix, indicate the probability of two distinct items from two distinct hypothetical categories being placed in the same category by a particular sorter.

Capacity. The program requires 38094 bytes of core as it is presently written. Presently programmed in FORTRAN IV to accommodate 40 items and 20 hypothetical categories with no limit on the number of sorters the capacity is easily enlarged by changing one statement in the program. With the use of an IBM 370-168 G-level compiler, a test problem of 10 items, 14 sorters and three hypothetical categories required 5 sec to compile and run.

Availability. Copies of this paper, a source listing, and documented output from an illustrative example can be obtained by writing to Richard J. Hofmann, Department of Educational Psychology, Miami University, Oxford, Ohio 45056.

REFERENCES

- Wiley, D. E. Latent partition analysis. *Psychometrika*, 1967, 32, 183-193.