COMPUTING FOR UNBALANCED REPEATED MEASURES EXPERIMENTS

Kenneth Berk, Illinois State University

ABSTRACT

Repeated measures experiments involve two or more intended measurements per subject. If the within-subjects design is the same for each subject, and if no data are missing, then the analysis is not very hard and there are readily available programs that do the analysis automatically. However, if the data are not balanced, with the same arrangement for each subject, then the analysis becomes much more difficult.

Beginning with procedures which are not optimal but are fairly simple, we move on to unbalanced linear model analysis, and then normal maximum likelihood procedures. We discuss ML and REML estimators for the mixed model and also maximum likelihood estimators for a model which allows arbitrary within-subjects covariances. We give a program which uses SAS MATRIX to do generalized least squares based on the output of BMDPAM, which gives a maximum likelihood estimate for the covariance matrix.

INTRODUCTION

We consider the basic repeated measures situation, with one between-subjects factor and one within-subjects factor. If the within-subjects data pattern is the same for all subjects, then the analysis is fairly straightforward, and programs like SPPS WANOVA and BMDP2V are available. We give a program which uses SAS MATRIX to do generalized least squares based on the output of BMDPAM, which gives a maximum likelihood estimate for the covariance matrix.

TWO SIMPLE METHODS

Allen and Cady (1982) suggest taking differences of measurements from adjacent times. If there are three repeated measures, then we form for each subject the Y1 = TIME2 - TIME1 and Y2 = TIME3 - TIME2 differences. The two variables Y1 and Y2 must be analyzed separately because of different missing value patterns. The main effect of TIME is significant if the mean of either Y1 or Y2 differs significantly from zero. The interaction of TIME with a between-subject factor A is significant if A is significant for either Y1 or Y2.

Another fairly easy method involves fitting a curve such as a parabola for each subject, assuming that such a curve is appropriate. Then the parameters (coefficients) of the curve are used in a between-subjects analysis. For example, to test for interaction of time and between-subject groups, we would do a multivariate analysis of variance of the linear and quadratic coefficients between groups.

This method has been advocated by Sanders (1978) and others.

UNBALANCED LEAST SQUARES

There are three equivalent methods for testing the interaction between time and groups of subjects. The basic idea is to fit the linear model (in SAS GLM notation)

\[ Y = A + T + S(A) + A \times T, \]

where A is the between-subjects effect, T is the time effect, A \( \times \) T is their interaction, S(A) is the effect of subjects nested within A, and the error terms are normal with zero mean, are independent between subjects, and have the same distribution for each subject. Schwertman (1978) has shown that the GLM analysis is legitimate if it is valid when the data are complete; that is, it is legitimate if the symmetry conditions hold: orthornormal contrasts of the repeated measure have equal variance and are uncorrelated.

Back in the old days before GLM, Box (1947) showed how to do the GLM analysis using a balanced analysis of covariance with a covariate, where m is the number of missing values. The covariates are indicator variables, the ith covariate being one at the ith missing value and zero elsewhere. With arbitrary values such as zero in place of the missing data, the analysis is done with adjustment for the covariates. The analysis can be easily carried out with BMDP2V (Dixon, 1973). There are also SAS PROCs by Jackson and McNee (1982).

Rubin (1972, 1976) described a procedure which gives results equivalent to those from the unbalanced linear model and analysis of covariance procedures. His method involves doing m balanced ANOVA's and then inverting an \( m \times m \) matrix. Hamilton, McCray and Palmer (1983) have implemented the 1972 version, which just fills in missing data and therefore gives correct error sums of squares but gives effect sums of squares which are larger than from the linear model and ANCOVA methods.

The above methods share some deficiencies. Take, for example, the situation with no between-subjects factors and just two repeated measures per subject. Subjects with incomplete data have only one measurement, and these subjects will be ignored in each of the methods. Intuitively, a good procedure should take advantage of inter-block (between subject) information. In particular, if there is only the first measure for some subjects and only the second measure for some other subjects, then the difference of the means for the two sets should somehow enter the analysis.

ML AND REML

Normal maximum likelihood procedures estimate the covariance matrix within subjects and
use it to weight the data, allowing all of the observations to enter with estimated weights. This takes advantage of interblock information in the estimation of repeated measures contrasts.

The usual mixed model can be used, with

\[ y = \mu + a + s(a) + t + e, \]

where \( s(a) \) and \( e \) represent independent random effects with variances \( \sigma_a^2 \) and \( \sigma_e^2 \) respectively. The other terms represent fixed effects, \( \mu \) the mean, \( a \) the between-subjects effect, \( t \) the repeated measure, and \( s(a) \) at the interaction of \( a \) and \( t \).

Estimation of the parameters for the fixed effects, \( \mu \) and \( \gamma \), and at along with the variance components \( \sigma_a^2 \) and \( \sigma_e^2 \) for the random effects is not easy for unbalanced designs, but the normal likelihood can be maximized using the scoring and Newton-Raphson techniques as described by Jennrich and Sampson (1978). Their procedure is available in BMDPSV, which does one scoring step and then Newton-Raphson iterations to maximize the likelihood for the general mixed model. The program does likelihood ratio tests for the fixed effects and random effects by fitting also appropriate submodels. The tests are performed with p-values given corresponding to the asymptotic chi-squares (-2 log(likelihood ratio)) for the likelihood ratios. The SAS VARCOMP procedure also estimates the variance components, but does not print the fixed effects, and it does no tests for the fixed effects. The fixed effects can be computed externally, but this requires a generalized least squares computation based on weights constructed from the random effects.

However, there is a method available, called restricted maximum likelihood (REML), which does yield estimates that agree with anova estimates for balanced data. It maximizes the likelihood for the residuals from the fixed effects and thereby takes account of degrees of freedom. The BMDPSV program computes REML estimates, again using a scoring step followed by Newton-Raphson iterations. Instead of likelihood ratio tests, Wald tests are used for the fixed effects computed by the REML procedure. That is, if \( b \) is the column vector of coefficients corresponding to an effect such as the interaction of groups with repeated measures, and \( V \) is the estimated covariance matrix for these coefficients, then the Wald statistic is \( b'V^{-1}b \). Asymptotically, this will have a chi-square distribution, but BMDPSV attempts to take account of the finiteness of the sample by printing the corresponding F-ratio along with a p-value.

There are some weaknesses shared by the ML and REML estimates for the mixed model. The general maximum likelihood mixed model procedures require the inversion of a matrix of order equal to the total of the numbers of levels of random effects (other than the residual) in the model. In the present context this means inverting a matrix of order equal to the number of subjects, a substantial task if there are several hundred subjects. On the other hand, it is possible to program the calculations much more efficiently in our special repeated measures situation. Also, there is a recently announced program by Robinson, Thompson, and Digby which has handled as many as 1000 subjects (bulls).

The assumptions of the mixed model are quite restrictive, requiring equal variances for all of the observations on a subject and equal covariances between each pair of observations on a subject, regardless of the time interval between observations. There are many situations in which these assumptions are clearly not satisfied. Then the multivariate normal model may be appropriate.

THE MULTIVARIATE APPROACH

The ML and REML approaches by way of the mixed model for repeated measures can be regarded as special cases of generalized least squares, using an estimated within-subjects covariance matrix. The mixed model assumes a very special structure for this matrix, requiring all of the diagonal elements to be equal (to the sum of the two variance components) and requiring all of the off-diagonal elements to be equal (to the subject variance component). If we are uncomfortable about this strong assumption, then we can estimate a more general likelihood, and in particular, we can assume only that the observations for each subject are multivariate normal and require only that the covariance matrix be symmetric positive definite. What is nice about this model is that there is a relatively simple iterative procedure to estimate the covariance matrix, published originally by Orchard and Woodbury (1972), and later called the EM-algorithm by Dempster, Laird and Rubin (1977). Furthermore, this procedure is available in the program BMDPAM. The output of this program can be fed to a program which uses the estimated covariance matrix to do generalized least squares, as is shown in the accompanying listing. This shows BMDPAM being used as a SAS procedure (PROC BMDP PROGRAM = BMDPAM), with the covariance matrix being sent to a file that is read by SAS. The generalized least squares is fairly easy to do in SAS MATRX.

Of course, in our repeated measures situation we need to get a covariance matrix that is pooled across groups, and the BMDP manual indicates that the program has this capability, but requesting the feature causes the program to say that it cannot deal with groups. However, the program can be tricked into giving a covariance matrix that is pooled across \( g \) groups. All that is needed is to include for each subject a set of \( (g-1) \) indicator (dummy) variables which separate the groups, along with the usual dependent variables, the repeated measures for each subject. Then the result will be a joint covariance matrix for the indicator variables together with the dependent variables. If this matrix is adjusted for (swept on) the indicator variables, then it yields the covariance matrix of the dependent variables.
variables pooled across groups.

It should be mentioned that BMDPAM does not exactly give multivariate normal maximum likelihood estimates, but uses an internal divisor of (n-1) instead of (n-g) if the program is used at the end of the program. Indeed, there is no need for a restriction to repeated measures. By coding only between-subjects contrast matrices, one can do a multivariate analysis of variance with incomplete data. It should be mentioned that Hosking (1980) and Hoffman (1981) in their Ph.D. theses gave procedures which seem likely to give similar results to those of the procedure described above.

SUMMARY

Beginning with two simple methods, we have described also the least squares method and two methods based on the mixed model for repeated measures analysis. In addition, we have given a procedure which combines BMDPAM and SAS MATRIX to do multivariate analysis of variance for incomplete data. This allows the multivariate approach to repeated measures as a special case.

ACKNOWLEDGMENT

Most of this work was supported by a National Institute of Mental Health grant 37188 in the Department of Biomathematics at UCLA. It is intended that a more detailed version will be available soon as a BMDP technical report.

REFERENCES


Ekbohm, G. (1976), "On comparing means in the
paired case with incomplete data on both responses," Biometrika 63, 299-304.


DATA COONS; *BIOMETRICS(1957), P.402;
  INPUT A B1 B2 B3 B4;
  CARDS;
  0  75  70 122 114
  0  69  51  91 125
  0  66  68 102  84
  0  56  50  72
  1  42  59  39  82
  1  75  66  50  46
  1 106  74  77
  1  62 102  34  56

*IF A HAS MORE THAN 2 LEVELS THEN USE DESIGN
  IN PROC MATRIX TO FORM DUMMY VARIABLES FOR A;
PROC PRINT;
PROC BMDP PROG=BMDPAM;
  VAR A B1-B4;
  PARMCARDS;
  CONTROL COLUMN=72./END
  /PROB TITLE IS 'RUN PAM TO FEED MATRIX'.
  /INPUT UNIT IS 3. CODE IS 'COONS'.
  /ESTIMATE TYPE IS ML.
  METHOD IS REGR.
  IPRINT eASE=8.
  MATRICES ARE CORR,CQUA,EST,DIS.
  ISAVE UNIT IS 15.
  /END
  NEW.
  CODE IS COONSCOV.
  CONTENT IS COVA.
  /END
PROC CONVERT BMDP=BMDPSAVE OUT=COVA;
PROC PRINT;
*DEFINE D TO INDICATE THE MISSING DATA;
*DEFINE M TO SHOW THE SUBJECT MISSING DATA PATTERN
FOR SORTING;
DATA COONS; SET COONS;
  ARRAY B(I) B1-B4;
  M=0;
  DO OVER B;
    IF B=. THEN 0=1;
    ELSE M=0;
    M=2*M + B;
  END;
  DROP 1;
PROC SORT; BY M;
PROC PRINT;
*TO DO GENERALIZED LEAST SQUARES SPLIT TO
GET ONE OBSERVATION PER CASE;
DATA SPLIT; SET COONS;
  ARRAY B(T) B1-B4; ARRAY DD(T) D1-D4;
  DO OVER B;
    Y=B; D=DD; OUTPUT;
  END;
  DROP B1-B4 D1-D4;
PROC PRINT;
PROC MATRIX ;
*EXTRACT THE COVARIANCE MATRIX FROM WHAT
EMDPAM PROVIDES;
FETCH SIG DATA=COVA ;
REPS=NCOL(SIG )-2; RI=REPS+1; RM1=REPS-1;
SIG=SIG(1:RI,1:RI);
DO I=2 TO RI;
  DO K=1 TO (I-1);
    SIG(K,I)=SIG(I,K);
  END;
SIG=SWEEP(SIG,1)(2,RI,2,RI),
PRINT SIG,
FETCH ATYD DATA=SPLIT(KEEP=A T Y D M),
*A DEFINES THE GROUPS, T DEFINES THE REPEATED
MEASURE, Y THE DEPENDENT VARIABLE;
*D TELLS WHERE THE MISSING DATA ARE
(1 INDICATES MISSING);
*M GIVES THE MISSING DATA PATTERN FOR EACH SUBJECT;
NN=NROW(ATYD);
*FORM CELL-MEAN DUMMIES:
T=DESIGN(ATYD(,,2));
A=ATYD(.1);
A =DESIGN(A);
GRPS=NCOL(A);
AT=A @ T;
X=AT ;
SUBS=NN/REPS,
*ADJUST SIGMA FOR DEGREES OF FREEDOM;
SIG=SIG#((SUBS-1)#(SUBS-GRPS));
*XY HAS THE CELL INDICATOR VARIABLES AND Y;
XY=X || ATYD( ,3);
P=NCOL(XY);
D=ATYD( ,4);
PRINT XY P SUBS NN;
M=ATYD( ,5);
PREVM=1; SSCP=J(P,P,0);
NR=NN;
*FORM X'INV(S)X AND X'INV(S)Y BY ADDING IN
ONE SUBJECT AT A TIME;
DO J=1 TO SUBS,
  *FOR EACH SUBJECT, FIND THE INDICES OF THE
NONMISSING DATA;
  DO=D(1:REPS , ); PRINT DO I;
  IF DO NE J(REPS,1) THEN DO; PRINT I DO;
    IND=1:REPS)'#(J(REPS,1)-DO );
    IND=IND(LOC(IND),);
  *DO THE INDICES DIFFER FROM THE INDICES FOR
  THE PREVIOUS SUBJECT?;
    IF M(1,1) NE PREVM THEN SIGINV=INV(SIG(IND,IND));
  *INCREMENT THE WEIGHTED XX,XY SSCP;
    XYI=XY(IND , );
    SSCP=SSCP + XYI 'SIGINV*XYI;
END;
  IF I + SUBS THEN DO;
    PREVM=M(1,1);
    SPLIT;
    XY=XY(RI:NR , );
    D=D(RI:NR , );
    M=M(RI:NR , );
    NR=NR-REPS;
END;
END;
*OBTAIN COEFFICIENTS $B$ AND THEIR ESTIMATED COVARIANCE MATRIX $V$;

\textsc{Sweep} = \textsc{sweep}(\textsc{scpc}, 1:(p-1));
$B$ = \textsc{sweep}(1:(p-1), p);
$V$ = \textsc{sweep}(1:(p-1), l:(p-1));

\textsc{Print} \textsc{scpc} \textsc{sweep} \textsc{b} \textsc{v};

*NOW TESTS;

*H$_1$ TESTS THE OVERALL MEAN;

$H_1 = 1$ \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1$

*HA TESTS THE BETWEEN-SUBJECTS GROUPS;

$H_A = 1$ \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1 \hspace{1cm} 1$

*HT TESTS THE REPEATED MEASURES MAIN EFFECT;

$H_T = 1$ \hspace{1cm} -1 \hspace{1cm} 0 \hspace{1cm} 0 \hspace{1cm} 0 \hspace{1cm} -1 \hspace{1cm} 0 \hspace{1cm} 0$

*HAT TESTS THE INTERACTION;

$H_A T = (H_A' @ H_T')'$

$B_1 = H_1 \cdot B$

$B_A = H_A \cdot B$

$B_T = H_T \cdot B$

$B_A T = H_A T \cdot B$

$V_1 = H_1 \cdot V \cdot H_1$

$V_A = H_A \cdot V \cdot H_A$

$V_T = H_T \cdot V \cdot H_T$

$V_A T = H_A T \cdot V \cdot H_A T$

$C_H_1 = B_1' \cdot \text{inv}(V_1) \cdot B_1$

$C_H_A = B_A' \cdot \text{inv}(V_A) \cdot B_A$

$C_H_T = B_T' \cdot \text{inv}(V_T) \cdot B_T$

$C_H_A T = B_A T' \cdot \text{inv}(V_A T) \cdot B_A T$

\text{Print} \ H_1 \ B_1 \ V_1 \ \text{ch}_1 \ \text{ha} \ \text{ba} \ \text{va} \ \text{ch}_1 \ A \ \text{th} \ \text{vt} \ \text{bt} \ \text{ch}_1 \ T$

Dr. Kenneth N. Berk
Department of Mathematics
Illinois State University
Normal, IL 61761

Phone: 309/438-8781