The purpose of this paper is to discuss some methods which increase the computational efficiency of analyzing repeated measures designs - i.e. designs in which the same subject is observed under each of the different levels of an experimental manipulation. More specifically, subjects are crossed with one or more of the independent variables. For example trials is a repeated measures factor since each subject is measured at each trial of training.

In repeated measures designs, the effect of subjects is included in the analysis of variance model. The focus of this paper will be on discussing different methods for coding subjects. There are two different approaches to coding subjects that will be considered. These two methods can easily be illustrated by considering the simplest case - single factor repeated measures design. The traditional method for coding this design can be represented by equation 1 shown below.

\[ Y_{ij} = b_0 + S_i + T_j + ST_{ij} \]  

(1)

In this equation \( Y_{ij} \) is the score obtained from subject \( i \) on trial \( j \). The parameter \( S_i \) is a constant associated with subject \( i \), where the index \( i \) ranges from \( 1 \) to \( n \) subjects. The parameter \( T_j \) is the effect of trial \( j \), where \( j \) is the index ranging from \( 1 \) to \( m \) trials. The parameter \( ST_{ij} \) represents the subject by trial interaction effect. The parameter \( b_0 \) is simply a constant.

The coding method represented by equation 1 is the method that is used by the SAS 76.6 programs for analyzing repeated measures designs.

A different method for coding this design is shown as equation 2 below.

\[ Y_{ij} = b_0 + b_1 X_i + T_j + ST_{ij} \]  

(2)

where

\[ j = m \]

\[ X_i = \sum_{j=1}^{m} Y_{ij} \]  

In this equation, \( X_i \) is the sum of all the scores obtained from each subject. This equation is similar to equation 1 except that the observed score \( X_i \) is substituted for the unknown parameter \( S_i \). The value of \( X_i \) is linearly related to the least squares estimate of \( S_i \) so that the between subjects sums of squares (that is the sum of squares due to \( S_i \)) will be the same for both equations.

The advantage of equation 2 is that the number of unknown parameters has been reduced by \( n-2 \), which results in a reduction of \( (n-2)^2 \) elements in the sums of squares and crossproducts matrix. Thus this method becomes extremely useful as the number of subjects becomes large, which is often the case in social research in order to obtain sufficient statistical power. For example, if the experiment employed 50 subjects then the reduction in the number of elements in the sums of squares and crossproducts matrix will equal 2400.

One may ask at this point why there is no concern about the number of parameters needed to estimate the subject x training interaction. Estimating this effect is never any problem because it can simply be obtained from the residual sum of squares of a model which includes only the main effects of subjects and training.

Obviously, one could also recode the parameter \( T_j \) in the same way as \( S_i \) and further increase the computational efficiency. However, since the number of levels of \( T_j \) is typically small (for example, a pre-post design has only 2 levels), the gain in efficiency may not be worth the effort of recoding.

The coding method represented by equation 1 will be referred to as the "standard method", while the coding method represented by equation 2 will be referred to as "Pedhazur's method". This is due to the fact that Pedhazur's 1977 Psychological Bulletin article is responsible for making this alternative method popular.

At this point it will be useful to demonstrate how Pedhazur's method can be extended to a split plot design.

Suppose you are analyzing a study consisting of an experimental group and a control group, and each subject was measured at three points in training - before, during, and after treatment. Suppose that you were also concerned with how sex interacted with these effects. Finally, assume that the cell frequencies were unequal so that the design was unbalanced.

The standard method for coding this design is shown below in equation 3.

\[ Y_{ijk} = b_0 + A_i + B_j + AB_{ij} + S_k(A B) + T_l + AT_{il} + BT_{jl} + ABT_{ijkl} + ST_{kl}(A B). \]  

(3)

In this equation, \( Y_{ijk} \) is the score where \( i \) indexes the treatment, \( j \) indexes the sex, \( k \) indexes the subject, and \( l \) indexes the trial. The parameter \( A_i \) represents the effect of treatment, the parameter \( B_j \) represents the effect of sex, and the parameter \( T_l \) again represents the effect of training. The parameter \( S_k(A B) \) represents the effect of subjects nested within treatment and sex.

It is convenient to break the sum of squares in a split plot design into separate components. The total between subjects sum of squares using the standard method is estimated by equation 4 shown below.

\[ Y_{ijk} = b_0 + A_i + B_j + AB_{ij} + S_k(A B) \]  

(4)
The between groups sum of squares using the standard method is estimated by equation 5 below.

\[ Y_{ijkl} = b_0 + A_i + B_j + AB_{ij} \]  

(5)

The sum of squares due to subjects nested within treatment and sex can be estimated by subtracting the sum of squares predicted by equation 5 from that predicted by equation 4. This is one method for estimating the sum of squares due to subjects nested within treatment and sex. As noted before, using the standard method to code subjects can cause severe computational problems when the number of subjects becomes large, especially when it is necessary to use PROC GLM.

Using Pedhazur's method, a computationally more efficient method is available. The total between subjects sum of squares can be estimated by equation 6 shown below.

\[ Y_{ijkl} = b_0 + b_{1}X_{ijk} \]

(6)

\[ X_{ijk} = \sum_{1=1}^{m} Y_{ijkl} \]

(6)

The variable \( X_{ijk} \) is simply the sum of all the scores for each subject. By using equation 6 rather than equation 4, it is possible to avoid estimating the parameter \( b_{1} \), which greatly reduces the number of unknown parameters as the number of subjects becomes large.

In order to illustrate the advantages of Pedhazur's coding method over the standard coding method for this split plot design, the two SAS programs shown below were run on a 370/168 IBM computer using SAS 76.6. For this example, 50 subjects were generated for each of the four cells produced by combining 2 levels of treatment with 2 levels of sex, which is approximately the size of a master's thesis.

*FIRST PROGRAM;

DATA; INPUT SUBJ SEX TREAT Y1 Y2 Y3;
TRAINING = 1; SCORE = Y1; OUTPUT;
TRAINING = 2; SCORE = Y2; OUTPUT;
TRAINING = 3; SCORE = Y3; OUTPUT;
KEEP SUBJ SEX TREAT TRAINING SCORE; CARDS;
1 1 7.1 8.2 6.3
1 2 7.1 8.2 6.3
2 1 7.1 8.2 6.3
2 2 7.1 8.2 6.3

PROC GLM; CLASSES SEX TREAT TRAINING; MODEL SCORE = SEX|TREAT|TRAINING SUBSCORE/SSL; TITLE PEDHAZUR'S METHOD;

The first program corresponds to the method suggested by the SAS communications for fall 1978, volume 4, number 2.

In the first program Y1, Y2, and Y3 are the variables representing the scores for each subject obtained before, during, and after treatment respectively. The purpose of the output statements is to create the repeated measures factor called TRAINING, and to associate the appropriate score with each stage of training. The purpose of the PROC GLM is to estimate the sum of squares due to subjects nested within treatment and sex, and the sum of squares due to the subject by training interaction. The sum of squares due to subjects nested within treatment and sex is estimated by including a variable called SUBSCORE in the classes statement and in the model statement. SAS uses this variable to code subjects according to the standard method. The sum of squares due to the subject by training interaction is estimated by the residual sum of squares.

Notice in the second program, a variable called SUBSCORE is created which is simply the sum of all the scores for each subject. Notice also that SUBSCORE is not included as class variable but rather as a covariate. The sum of squares due to subjects nested within treatment and sex is estimated by subtracting the sum of squares predicted by SUBSCORE, adjusted for the treatment, sex, and treatment by sex interaction. This is essentially the same as subtracting the between groups sum of squares from the total between subjects sum of squares. The sum of squares due to the subject by training interaction is again estimated by the residual sum of squares.

The results of this example demonstrated that the PROC GLM using Pedhazur's coding method required only .7 seconds of CPU time, and only 162 K of memory. However, the PROC GLM using the standard coding method required 14 seconds of CPU time and 606 K of memory.

In summary, how the split plot analysis of variance was a minor task when Pedhazur's method was employed, the analysis blew up into a enormous problem when the standard coding method was used.

Applications of Pedhazur's method to more complicated repeated measures designs, consisting of more than 1 repeated measures factor, are discussed in Pedhazur's article.

REFERENCES