THE ANALYSIS OF COVARIANCE FOR SPLIT-UNIT AND
REPEATED MEASURES EXPERIMENTS USING THE GLM PROCEDURE (II)

M.P. Meredith, Biometrics Unit, Cornell University
N. J. Miles-McDermott, SAS Institute Inc.
W.T. Federer, Biometrics Unit, Cornell University

ABSTRACT

This is the second of two papers on the analysis of covariance in split-unit and repeated measures studies. The first paper (Miles-McDermott, et al., 1988) documented analysis difficulties using SAS® PROC GLM and illustrated a two-step approach to conduct correct analyses for several examples. A one-step approach is presented in this paper and is illustrated with the same examples. The first example has whole plot experimental units (EUs) in a RCBD and a covariate measured on each subplot. The second example has whole plot EUs arranged in a CRD with a covariate measured on each whole plot. Both Type I and III sums-of-squares are necessary to construct the proper ANOVA table when using the one-step approach which requires a single procedural call to PROC GLM.

Although the correct ANOVA table may be easily constructed in a single call to GLM, additional work is required for estimation. It is shown how terms required for calculating adjusted means and associated standard errors are easily gleaned from an inexpensive call to PROC ANOVA using the MANOVA option. Formulas for computing standard errors of differences between adjusted whole plot and subplot means are reported.

Extensions for studies involving two or more splits and/or covariates are immediate. This work is easily adapted to the analysis of two-period crossover trials where baseline and/or washout responses are used as covariates.

INTRODUCTION

The purpose of this paper is to demonstrate the relative ease with which a SAS/STAT™ user may perform an analysis of covariance (ANCOVA) on data derived from a split-unit experiment. By "split-unit" is meant a study designed with several different sizes of experimental units (EUs). A consequence of these designs is that there are now several error terms to contend with. "Relative ease" means that the correct ANOVA table is easy to obtain via a single procedural call to the GLM procedure by gleaning the appropriate Type I and Type III sums-of-squares. If adjusted treatment means and standard errors of differences amongst adjusted treatment means are required then additional work is necessary. However, it is shown that the cost of this additional work is minimal and subsequent calculations may be performed on a hand calculator. Cost is an important factor because the analysis of split-unit experiments in GLM can quickly become prohibitively expensive by requiring large amounts of CPU time and memory.

The "classical" (i.e., typical textbook) split-plot design randomly allocates levels of a whole plot treatment factor (such as irrigation method) in complete blocks (such as rows in fields). Then within each of these blocks (rows) levels of a subplot treatment factor (such as pesticides) are randomly allocated to plots within the rows. Thus, each EU for a whole plot treatment (a row) is a complete block for the subplot treatment factor.

Suppose now that EUs have been randomly allocated to levels of a treatment factor via some experimental design, say a completely randomized design (CRD). A common practice is to consider measurements made in time sequence on these EUs as subplot EUs (with time period being the subplot treatment factor). Such an experiment shall be called a repeated measures study in this paper. It is well known that the univariate split-plot analysis is justified only if the assumption of spherical symmetry (Huyhn and Feldt, 1970) is valid. Various conservative tests (Box, 1954) and tests with estimated adjustments in the within time degrees-of-freedom due to Greenhouse and Geisser (1959, and Geisser and Greenhouse, 1958) and Huyhn and Feldt (1976) are now standard fare for analysis of repeated measures in major statistical software packages. These adjusted tests may also be applied when a covariate is included in the model.
For both split-plot experiments and repeated measures studies it is not uncommon that some ancillary measurements are taken on either the subplot or whole plot units with the intention of removing a possible source of bias and increasing the precision of the experiment. Such ancillary data are often employed as covariates. However, an analysis is now complicated because the covariate may be thought of as having different slopes associated with the different sizes of experimental units (while still assuming homogeneity of slopes between levels a treatment factor). For example, the soil fertility may be recorded on each plot within a row. Likewise, the ambient room temperature or relative humidity may be recorded each time an EU is measured. It is for this reason that the problem at hand is an important one.

In this paper, as in the first (Miles-McDermott, et al.) two data sets are considered in detail. First, a hypothetical data set is examined where whole plot units are in a RCBD and a covariate is measured on each subplot. A second example has whole plot units arranged in a CRD and a covariate measured on each whole plot.

The models and notation used in the examples are given in the following section. The calculations and appropriate ANOVA tables are given in the first paper. Thus, this paper will report appropriate formulae for calculation of standard errors for differences amongst the adjusted means. These are necessary to draw conclusions from comparisons amongst the treatments in the experiment.

Annotated SAS/STAT™ output showing the proper model statements and portions of annotated analyses for the two examples are given in the Appendix.

**MODELS & STANDARD ERRORS**

The model for SP-2 may be written as a cell means model with several error components. The population cell means may then be thought of as being overparameterized as indicated below.

Model: $Y_{ik} = \mu_{ik} + \rho_j + \tau_i + \alpha_k + (\alpha \tau)_{ik} + \epsilon_{ijk}$

for $\mu_{ik} = \mu + \delta_i + \alpha_k + (\alpha \tau)_{ik}$

i.e., $Y_{ik} = \mu + \rho_j + \delta_i + \alpha_k + (\alpha \tau)_{ik} + \epsilon_{ijk}$

where:
- $\mu$ = overall mean
- $\delta_{ij}$ = error(a)
- $\rho_j$ = random effect of block j
- $\epsilon_{ijk}$ = error(b)
- $\tau_i$ = effect of whole plot i
- $\alpha_k$ = effect of split plot k
- $(\alpha \tau)_{ik}$ = effect of interaction of whole plot level i and split plot level k,

and,
- $\rho_j \sim N(0, \sigma^2_\rho)$, $\delta_{ij} \sim N(0, \sigma^2_\delta)$, and $\epsilon_{ijk} \sim N(0, \sigma^2_\epsilon)$,

and $\rho_j$, $\delta_{ij}$, and $\epsilon_{ijk}$ are mutually independent random variables. $i=1,2,...,a$, $j=1,2,...,r$, and $k=1,2,...,s$.

The model including the covariate measured on each subplot EU with separate slopes for the whole plot and subplot responses is given by the following:

$Y_{ijk} = \mu_{ik} + \beta_1(Z_{ji} - \bar{Z}_{..}) + \beta_2(Z_{jk} - \bar{Z}_{..}) + \rho_j + \delta_{ij} + \epsilon_{ijk}$

where $\beta_1$ and $\beta_2$ are the whole plot and subplot slope coefficients, respectively, and $\mu_{ik}$, $\rho_j$, $\delta_{ij}$, and $\epsilon_{ijk}$ are defined as above.

The variances of differences amongst various whole and subplot means are given by:

$$\operatorname{Var}(\bar{Y}_{..j} - \bar{Y}_{..i}) = \frac{\sigma^2_\epsilon + \sigma^2_\delta}{ar} + \frac{\bar{Z}_{..j} - \bar{Z}_{..i}}{W\times B_{ZZ}}$$

$$\operatorname{Var}(\bar{Y}_{..k} - \bar{Y}_{..i}) = \frac{\sigma^2_\epsilon}{ar} + \frac{\bar{Z}_{..k} - \bar{Z}_{..i}}{S\times B\times W_{ZZ}}$$

$$\operatorname{Var}(\bar{Y}_{..k} - \bar{Y}_{..i}) = \frac{\sigma^2_\epsilon}{ar} + \frac{\bar{Z}_{..k} - \bar{Z}_{..i}}{S\times B\times W_{ZZ}}$$

and, for $i\neq i'$,

$$\operatorname{Var}(\bar{Y}_{i..} - \bar{Y}_{i'..}) = \frac{\sigma^2_\epsilon + \sigma^2_\delta}{ar}$$

$$\bar{Z}_{i..} - \bar{Z}_{i'..} = \frac{(Z_{i..} - Z_{i'..})^2}{\sigma^2_\epsilon + \sigma^2_\delta} + \frac{(Z_{i..} - Z_{i'..})^2}{S\times B\times W_{ZZ}}$$

Estimates of the variance components $\sigma^2_\epsilon$ and $\sigma^2_\delta$ are required to calculate standard errors of the above differences amongst adjusted treatment means. From the expected mean squares of the ANOVA table it is known that error(a) and error(b) estimate $\sigma^2_\epsilon + \sigma^2_\delta$ and $\sigma^2_\epsilon$, respectively. If error(a) and error(b) are denoted $E_a$ and $E_b$, respectively, then $E_b$ estimates $\sigma^2_\epsilon$, and $\sigma^2_\delta$ is
estimated by \((E_a - E_b)/s\). Hence, the desired standard errors are given by:

\[
\text{SE}(\bar{Y}_{ij}-\bar{Y}_{ik}) = \sqrt{E_a \left[ \frac{2}{rs} + \frac{(Z_{i} - Z_{j})^2}{W \times B_{ZZ}} \right]}
\]

\[
\text{SE}(\bar{Y}_{ik}-\bar{Y}_{ik'}) = \sqrt{E_b \left[ \frac{2}{ar} + \frac{(Z_{ik} - Z_{ik'})^2}{S \times B \cdot W_{ZZ}} \right]}
\]

\[
\text{SE}(\bar{Y}_{ik}-\bar{Y}_{ik'}) = \sqrt{E_b \left[ \frac{2}{r} + \frac{(Z_{ik} - Z_{ik'})^2}{S \times B \cdot W_{ZZ}} \right]}
\]

and, for \(i \neq i'\):

\[
\text{SE}(\bar{Y}_{ik}-\bar{Y}_{ik'}) = \sqrt{E_b \left[ \frac{2}{r} + \frac{(Z_{ik} - Z_{ik'})^2}{S \times B \cdot W_{ZZ}} \right]}
\]

For the second data set, SP3, the whole plot EUs are allocated to treatments in a CRD. The model may be written as follows:

Model: \(Y_{ijk} = \mu + \delta_{i} + \tau_{j} + \alpha_{k} + \epsilon_{ijk}\)

for where:

\(\mu\) = overall mean  
\(\delta_{i}\) = error(a)  
\(\epsilon_{ijk}\) = error(b)  
\(\tau_{j}\) = A effect (whole plot )  
\(\alpha_{k}\) = B effect (split plot)  
\(\epsilon_{ijk}\) = interaction effect of whole plot factor A at level i and split plot factor B at level k,

and, \(\delta_{i} \sim N(0, \sigma^2_{\delta})\), and \(\epsilon_{ijk} \sim N(0, \sigma^2_{\epsilon})\),

for \(\delta_{i}\), and \(\epsilon_{ijk}\) are mutually independent random variables, \(i = 1, 2, \ldots, a\), \(j = 1, 2, \ldots, r\), and \(k = 1, 2, \ldots, s\).

For this data set, SP3, the following variances for differences amongst various whole and subplot means are reported:

\[
\text{Var}(\bar{Y}_{ij}-\bar{Y}_{ik}) = \left(\sigma^2_{\delta} + \sigma^2_{\epsilon}\right) \left[ \frac{2}{rs} + \frac{(Z_{i} - Z_{j})^2}{E(a)Z} \right]
\]

\[
\text{Var}(\bar{Y}_{ik}-\bar{Y}_{ik'}) = \frac{2 \sigma^2_{\epsilon}}{r} = \text{Var}(\bar{Y}_{ik} - \bar{Y}_{ik'})
\]

and, for \(i \neq i'\):

\[
\text{Var}(\bar{Y}_{ik}-\bar{Y}_{ik'}) = \frac{2 \sigma^2_{\epsilon} + \sigma^2_{\alpha}}{E(a)Z} + \frac{(Z_{ik} - Z_{ik'})^2}{E(a)Z}
\]

Estimates of the variance components \(\sigma^2_{\delta}\) and \(\sigma^2_{\epsilon}\) are given by \(E(b)\) and \(\{E(a) - E(b)\}/s\), respectively. Hence, the desired standard errors are given by:

\[
\text{SE}(\bar{Y}_{ij}-\bar{Y}_{ik}) = \sqrt{E(a) \left[ \frac{2}{rs} + \frac{(Z_{i} - Z_{j})^2}{E(a)Z} \right]}
\]

\[
\text{SE}(\bar{Y}_{ik} - \bar{Y}_{ik'}) = \sqrt{E(a) \left[ \frac{2}{r} + \frac{(Z_{ik} - Z_{ik'})^2}{E(a)Z} \right]}
\]

REFERENCES


APPENDIX

Procedural call for SP-2

```plaintext
DATA ONE; INPUT EU BLOCK WHOLE SUBPLOT Z ZTOTAL Y; CARDS;
{ data }
TITLE 'SPLIT PLOT HYPOTHETICAL DATA: COVARIATE ADDED';
PROC PRINT; VAR EU BLOCK WHOLE SUBPLOT Z ZTOTAL Y;
PROC GLM; CLASS WHOLE BLOCK SUBPLOT;
MODEL Y = BLOCK WHOLE ZTOTAL BLOCK*WHOLE
       SUBPLOT SUBPLOT*WHOLE Z / SS1 SS3 P;
       RANDOM BLOCK BLOCK*WHOLE;
    TEST H=ZTOTAL E=BLOCK*WHOLE / HTYPE=1 ETYPE=1;
    TEST H=WHOLE E=BLOCK*WHOLE / HTYPE=3 ETYPE=3;
    LSMEANS SUBPLOT / STDERR DDIFF;
    ESTIMATE 'SUBPLOT SLOPE' Z 1;
```

The ordering in the MODEL statement is important. The RANDOM option prints expected mean squares to aid in correctly choosing different Types of SS's for constructing F-tests.

The LSMEANS statement gives correctly adjusted subplot means, however the reported standard errors are incorrect.

The ESTIMATE statement provides the estimate of the subplot slope coefficient $\beta_2$. Unfortunately, the whole plot coefficient $\beta_1$ may not be estimated as easily.
PROC GLM Output

SOURCE DF TYPE I SS F VALUE PR > F
BLOCK = R(p|µ;τ) 2* 48.00000000 2.71 0.1107
TOTAL = R(τ|µ) 1 24.00000000 2.71 0.1282
TOTAL*BLOCK = R(τ|µ,τ,µ,τ) 1* 16.00000000 1.00 0.3063
SUBLOT = R(µ|ρ,τ) 3 156.00000000 5.81 0.0121
TOTAL*SUBLOT = R(µ|τ,µ,τ|µ,τ) 3 84.00000000 3.16 0.0083
Z = R(ρ|µ,µ,τ,µ) 1* 14.45000000 1.63 0.2281

Note: The SS's whose DF have asterisks are those that appear in the correct ANOVA table, as verified by the expected mean squares.

TESTS OF HYPOTHESES USING THE TYPE I MS FOR WHOLE*BLOCK AS AN ERROR TERM
SOURCE DF TYPE I SS F VALUE PR > F
TOTAL 1 16.00000000

TESTS OF HYPOTHESES USING THE TYPE III MS FOR WHOLE*BLOCK AS AN ERROR TERM
SOURCE DF TYPE III SS F VALUE PR > F
TOTAL 1 3.42857143

PARAMETER ESTIMATE T FOR HO: PR > [T] STD ERROR OF ESTIMATE
SUBLOT SLOPE 0.85000000 1.28 0.2281 0.66588970 => This is the result of the ESTIMATE statement to estimate the subplot regression coefficient, β2.

Procedural call for SP-2 using ANOVA with MANOVA option
[ Same input as previous call] PROC ANOVA;
CLASS WHOLE BLOCK SUBLOT;
MODEL Y Z = BLOCK WHOLE BLOCK*WHOLE SUBLOT WHOLE*SUBLOT;
MEANS WHOLE BLOCK BLOCK*BLOCK SUBLOT SUBLOT*SUBLOT;
MANOVA H=WHOLE E=BLOCK*WHOLE / PRINTE;
MANOVA H=SUBLOT / PRINTE;
The MEANS statement gives the unadjusted means of both the response Y and the covariate Z.

E = TYPE I SS&CP MATRIX FOR: WHOLE*BLOCK
DF=2
Y 16.00000000 = BxWYY
Z 4.00000000 = BxWYZ

E = ERROR SS&CP MATRIX
DF=12
Y 112.00000000 = SxSxWYY
Z 17.00000000 = SxSxWYZ
X 20.00000000 = SxSxWZZ

These are the results from the MANOVA option.
This call to GLM produces the correct ANOVA table for SP-3. Expected mean squares are printed with the RANDOM option and the TEST statement computes F-tests of $H_0: \beta_1 = 0$ and adjusted whole plot effects.

This call to PROC ANOVA is unnecessary if only correct ANOVA table is desired. However, it does give the correct table of sums of squares and cross products among Y and Z for SP-3 by using the MANOVA statement.

These may be used to correctly estimate the whole plot slope coefficient. The MEANS option prints appropriate means for Y (unadjusted) and the covariate Z. Thus, sufficient information is given to calculate any adjusted means as well as correct standard errors for adjusted means and differences amongst adjusted means.

**Relevant PROC ANOVA Output**

$\hat{\beta}_1 = \frac{E(a)yz}{E(a)z^2} = 163.0/159.5 = 1.022$