Estimating Components of Variance for a Linear Model Having a Mixed
Fixed/Random Factor in a Large Data Structure

Dawn VanLeeuwen, Department of Experimental Statistics,
New Mexico State University, Las Cruces, New Mexico
and
N. Scott Urquhart, Department of Statistics, Oregon State University, Corvallis, Oregon

Abstract
Real situations can lead to random or mixed linear models with many levels. Variance component estimation with SAS® software's PROC GLM can have memory requirements exceeding the memory available on any particular platform. This paper presents a simple computation scheme using both SAS statistical procedures and a small SAS/IML® software program to obtain Type II mean squares and their expectations so that the variance components may be estimated. This scheme is applicable to classes of models possessing certain types of nested or interaction structures.

1. INTRODUCTION

This report grows out of the need, related to the EPA's Environmental Monitoring and Assessment Program (EMAP), to estimate several components of variance of a novel linear model when one of the random factors has many levels. EMAP is charged with estimating the status of and detecting trends in indicators of ecological health of all major ecosystems. Evaluation of the performance of the sampling design prescribed for EMAP involves three major components of variance (Urquhart, Overton and Birkes, 1993 and Larsen and Urquhart, in review). Estimating these important components of variance from historical data requires more complex components of variance models than those needed for the design evaluation, because the historical data were gathered without regard to sampling design specification. Data sets that are computationally unwieldy can occur due to the complexity of the components of variance models and the replication of each random factor required to obtain reliable estimates of the variance components.

When one of the random factors has enough levels, the General Linear Models (GLM) procedure will have memory requirements exceeding the memory available on any particular platform. We propose a computational scheme to obtain Type II estimates for variance components from large, cross-classified, unbalanced data sets. The example that motivated this scheme concerned several water quality variables such as chlorophyl-a, total phosphorus, and Secchi depth (a measure of water clarity). Samples were collected from many lakes over a period of about 10 years, and there was substantial local variation in the number of sampling points and sampling times.

The novel components of variance model incorporated a random factor, years, across which there could be a fixed trend. Until now, a factor having both random and fixed parts has received only limited attention in the statistical literature (Burdick and Graybill, 1992 and VanLeeuwen, 1993). We explain this model variation and its analysis using PROC GLM in section 2. Section 3 outlines a computational approach using PROC ANOVA, PROC GLM, and SAS/IML. Section 4 presents an example. Section 5 describes the general computational procedure.

2. THE MODEL

A response variable, chlorophyl-a, was collected on various weeks, from 86 lakes, over a period of 11 years. Data only from the months of August and September were used (8 weeks). This time window matched that to which inferences were to apply. A linear trend over years was believed to be present, an intended consequence of the Clear Water Act. In addition, year also was considered a random classification factor. Let yearc be the variable name for year as a continuous variable, rather than as a class variable. We wished to obtain estimates of the variance components for the following model, represented in the notation of PROC GLM:

\begin{verbatim}
class week year site;
model response = yearc week year site*year 
week(site*year);
\end{verbatim}

Week, site, and year all are random factors, whereas yearc is fixed.

It is important to realize that PROC GLM can be used to obtain the ANOVA variance component estimators when the data set is small enough. The
estimators obtained using PROC GLM output will be correct for this model.

3. COMPUTATIONAL APPROACH

To obtain reliable estimates of variance components, it is necessary to replicate at the level of the source of variation. For the model presented in Section 2, we need replication of all three factors - week, year, and site. In the presence of the interaction and nesting in the model, all this replication can lead to a model with excessive (computer) memory requirements. We now propose a computational scheme for obtaining the Type II ANOVA estimators for the variance components. This scheme pieces together output from PROC GLM and PROC ANOVA with some SAS/IML computations to obtain the Type II ANOVA and expected mean squares for the full model. Obtaining the Type II estimators for the variance components is then a straightforward process. This scheme may be used to obtain the Type II ANOVA because the Type II sums of squares do not take into account the terms that "contain" the effect being tested, but do take into account all other effects.

Three steps are involved in calculating the Type II mean squares and their coefficients on the components of variance in their expected mean squares. First we use PROC GLM on the model response = year*week*site*year with a random statement specifying that week, year, site, and site*year are random factors. Second, we apply PROC ANOVA to the model response = week*site*year. These two runs provide all sums of squares necessary to complete the Type II ANOVA for the original full model. Most of the quantities involved in the expected mean squares also are provided by these two procedures. Only those terms involving the variance component for week(site*year) or week*site*year are missing from the expected mean squares given in the PROC GLM step. The coefficients for this variance component may be computed using SAS/IML.

4. EXAMPLE

We now apply the computational scheme suggested in Section 3 to the model presented in Section 2. The data set had a total of 4542 observations collected over 8 weeks, 86 sites (lakes) and 11 years.

In the following, we denote the error, week, site, year, site*year, and week(site*year) components of variance by \( \sigma^2, \sigma^2_w, \sigma^2_s, \sigma^2_y, \sigma^2_{sy}, \) and \( \sigma^2_{sw} \). We also denote the site*year interaction by \( S*Y \) and the week(site*year) effect by \( S*Y*W \).

STEP 1

From the General Linear Models procedure, we obtain:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>920</td>
<td>20457.945487</td>
<td>22.236897</td>
</tr>
<tr>
<td>Error</td>
<td>3621</td>
<td>2860.293483</td>
<td>0.789918</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>4541</td>
<td>23318.238970</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II SS</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE</td>
<td>17638.048339</td>
<td>207.506451</td>
</tr>
<tr>
<td>YEAR</td>
<td>190.967810</td>
<td>21.218646</td>
</tr>
<tr>
<td>S*Y</td>
<td>2450.372587</td>
<td>2.995566</td>
</tr>
<tr>
<td>WEEK</td>
<td>14.243299</td>
<td>2.034757</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Type II Expected Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Site</td>
<td>( \sigma^2_s + 5.7644\sigma^2_w + 22.375\sigma^2_y )</td>
</tr>
<tr>
<td>Year</td>
<td>( \sigma^2_y + 7.2597\sigma^2_s + 403.27\sigma^2_y )</td>
</tr>
<tr>
<td>S*Y</td>
<td>( \sigma^2_s + 4.8447\sigma^2_y )</td>
</tr>
<tr>
<td>Week</td>
<td>( \sigma^2_s + 473.51\sigma^2_y )</td>
</tr>
</tbody>
</table>

Each of the coefficients present in the expected mean squares is correct for the Type II analysis of the original full model, including the three-factor interaction. However all these expected mean squares lack the variance component for the week(site*year) interaction, a component present in all the above expected mean squares for the full model. SAS/IML calculations in Step 3 are used to obtain the coefficients for this variance component.

STEP 2

From the Analysis of Variance procedure, we obtain:

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Sum of Squares</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td>3415</td>
<td>22656.497732</td>
<td>6.634406</td>
</tr>
<tr>
<td>Error</td>
<td>1126</td>
<td>661.741238</td>
<td>0.587692</td>
</tr>
<tr>
<td>Corrected Total</td>
<td>4541</td>
<td>23318.238970</td>
<td></td>
</tr>
</tbody>
</table>

Note that the adjusted sum of squares for week(site*year) for the full model can be obtained by subtracting the residual sum of squares for the model in step two from the residual sum of squares for the model in step one: (2860.293483 - 661.741238) = 2198.552245. Because the analogous degrees of freedom are (3621 - 1126) = 2495, its mean square is 0.881183264529. We now have all the results needed to construct the complete Type II analysis of variance for the full model.

1065
Full Model Analysis of Variance

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>Type II SS</th>
<th>Mean Square</th>
</tr>
</thead>
<tbody>
<tr>
<td>SITE</td>
<td>85</td>
<td>17638.048393</td>
<td>207.506451</td>
</tr>
<tr>
<td>YEAR</td>
<td>9</td>
<td>190.967810</td>
<td>21.218646</td>
</tr>
<tr>
<td>S*Y</td>
<td>818</td>
<td>2450.372587</td>
<td>2.995566</td>
</tr>
<tr>
<td>WEEK</td>
<td>7</td>
<td>14.243299</td>
<td>2.034757</td>
</tr>
<tr>
<td>S<em>Y</em>W</td>
<td>2495</td>
<td>2198.522245</td>
<td>0.881183</td>
</tr>
</tbody>
</table>

STEP 3

For this example, let B denote the matrix associated with the week(site*year) effect. Then SAS/IML can be used to calculate terms of the form \( \text{tr}(B'PB) \), where P is an orthogonal projection operator. Such terms are involved in the coefficients in the expected mean squares. Here, there are five such terms in the following projection operators:

- \( P_1 \): The orthogonal projection operator on the range of the matrix corresponding to the site, year, and week main effects;
- \( P_2 \): The orthogonal projection operator on the range of the matrix corresponding to the year and week main effects;
- \( P_3 \): The orthogonal projection operator on the range of the matrix corresponding to the site and week main effects and year, that is, year as a continuous variable;
- \( P_4 \): The orthogonal projection operator on the range of the matrix corresponding to the site*year interaction and the week main effect;
- \( P_5 \): The orthogonal projection operator on the range of the matrix corresponding to the site*year interaction.

Let \( t_r, \ldots, t_s \) be the corresponding traces.

Then we get

\[
\begin{align*}
\text{SITE} & \quad \sigma^2 + 1.4341 \hat{\sigma}_{\text{syw}}^2 + 5.7644 \sigma_{\text{y}}^2 + 52.375 \hat{\sigma}_{\text{y}}^2 \\
\text{YEAR} & \quad \sigma^2 + 1.8325 \hat{\sigma}_{\text{syw}}^2 + 7.2597 \sigma_{\text{yr}}^2 + 403.27 \sigma_{\text{y}}^2 \\
\text{S*Y} & \quad \sigma^2 + 1.3284 \hat{\sigma}_{\text{syw}}^2 + 4.8447 \sigma_{\text{y}}^2 \\
\text{WEEK} & \quad \sigma^2 + 1.6514 \hat{\sigma}_{\text{syw}}^2 + 473.51 \sigma_{\text{y}}^2 \\
\text{S*Y*W} & \quad \sigma^2 + 1.3239 \hat{\sigma}_{\text{syw}}^2 \\
\end{align*}
\]

Likewise,

\[
\begin{align*}
\text{YEAR} & \quad \sigma^2 + 1.8325 \sigma_{\text{y}}^2 + 7.2597 \sigma_{\text{yr}}^2 + 403.27 \sigma_{\text{y}}^2 \\
\text{S*Y} & \quad \sigma^2 + 1.3284 \sigma_{\text{y}}^2 + 4.8447 \sigma_{\text{y}}^2 \\
\text{WEEK} & \quad \sigma^2 + 1.6514 \sigma_{\text{y}}^2 + 473.51 \sigma_{\text{y}}^2 \\
\text{S*Y*W} & \quad \sigma^2 + 1.3239 \sigma_{\text{y}}^2 \\
\end{align*}
\]

where additional coefficients in these expressions, respectively, are:

\[
\begin{align*}
& (t_r - t_s)/\text{df} \quad \text{year} = (152.3221 - 135.8299)/9 = 1.8325; \\
& (t_r - t_s)/\text{df site*year} = (1238.9895 - 152.3221)/818 = 1.3284; \text{ and} \\
& (t_r - t_s)/\text{df week} = (1238.9895 - 1227.4297)/7 = 1.6514.
\end{align*}
\]

Note that \( \text{tr}(B'PB) \) does not have to be calculated for \( \text{S*Y*W} \), because this trace is \( \text{tr}(B'B) = n \) = the total number of observations in the data set.

\[
(n - t_r)/\text{df site*year*week} = (4542 - 1238.9895)/2495 = 1.3239.
\]

If we denote the estimators for the error, week, site, year, site*year, and week(site*year) components of variance by \( \hat{\sigma}_1, \hat{\sigma}_2, \hat{\sigma}_3, \hat{\sigma}_4, \hat{\sigma}_5, \) and \( \hat{\sigma}_{\text{syw}} \), then we obtain the following equations:

\[
\begin{align*}
207.506451 & = \hat{\sigma}^2 + 1.4341 \hat{\sigma}_{\text{syw}}^2 + 5.7644 \sigma_{\text{y}}^2 + 52.375 \hat{\sigma}_{\text{y}}^2 \\
21.218646 & = \hat{\sigma}^2 + 1.8325 \hat{\sigma}_{\text{syw}}^2 + 7.2597 \sigma_{\text{yr}}^2 + 403.27 \sigma_{\text{y}}^2 \\
2.995566 & = \hat{\sigma}^2 + 1.3284 \hat{\sigma}_{\text{syw}}^2 + 4.8447 \sigma_{\text{y}}^2 \\
2.034757 & = \hat{\sigma}^2 + 1.6514 \hat{\sigma}_{\text{syw}}^2 + 473.51 \sigma_{\text{y}}^2 \\
0.881183 & = \hat{\sigma}^2 + 1.3239 \hat{\sigma}_{\text{syw}}^2 \\
0.587692 & = \hat{\sigma}^2 \\
\end{align*}
\]

whose solutions are:

\[
\begin{align*}
\hat{\sigma}^2 & = 0.587692 \quad \hat{\sigma}_{\text{syw}}^2 = 0.22169 \\
\hat{\sigma}_1 & = 0.00228 \quad \hat{\sigma}_{\text{yr}}^2 = 0.43623 \\
\hat{\sigma}_2 & = 0.04230 \quad \hat{\sigma}_{\text{y}}^2 = 3.89663.
\end{align*}
\]
5. GENERAL COMPUTATIONAL SCHEME

The scheme demonstrated in Section 4 may be applied to other models. We now describe the computational scheme in more general terms. We may write the linear model in the following way:

\[
E(y) = X\beta, \quad \text{Cov}(Y) = \sigma^2 A A' + \cdots + \sigma^2_{r-1} A_{r-1} A'_{r-1} + \sigma^2 I
\]

where the matrices \(X, A_1, \ldots, A_r\) are known. In particular, \(A_1, \ldots, A_r\) correspond to classification factors and are composed of zero's and one's. In addition, \(\beta\) is an unknown \(p \times 1\) vector of fixed parameters and the variance components \(\sigma^2_1, \ldots, \sigma^2_{r-1}, \sigma^2\) also are unknown. We confine our attention to either nested models or models that contain both main effects and interactions. In particular, the following scheme applies when \(R(A_r) \supseteq R(X, A_1, \ldots, A_{r-1})\), so that interactions of the highest order must be present in interaction models.

As suggested by the example, models of this type may become far too large computationally for SAS software to handle on most platforms. If, however, SAS software can do the Type II computations for the model that remains after dropping the \(r\)-th random effect, and this random effect is a higher order interaction containing all other effects, a simple scheme can be employed to estimate the variance components. The following scheme is given in three steps; not listed as a separate step are hand calculations that could be included in the SAS/IML program:

1. Use PROC GLM to obtain the Type II analysis of the model:

\[
Y = X\beta + A_1\alpha_1 + \cdots + A_{r-1}\alpha_{r-1} + e.
\]

Include a RANDOM statement, so that expected mean squares for this model are calculated.

The mean squares provided by this analysis are the correct mean squares for the Type II analysis of the full model, but the expected mean squares each lack the component involving the parameter \(\alpha_r\). Denote by \(c_i\) the coefficient missing in the \(i\)-th expected mean square, where \(i = 1, \ldots, r-1\).

2. Use PROC ANOVA to obtain the analysis of the following completely random model:

\[
Y = \mu + A_1\alpha_1 + e.
\]

The mean square error for this model provides an estimate of \(\sigma^2\) for the full model.

3. Use SAS/IML to obtain terms to be used in computing the quantities \(c_i, i = 1, \ldots, r-1\). Note that the degrees of freedom to be used in computing the \(c_i\) need not be computed by SAS/IML, since the SAS runs will already have provided this information.

All of the terms \(c_i\) are of the form \(tr[(P-Q)BB']\), where \(B=A_r\) and \(P\) and \(Q\) are orthogonal projection operators of the form \(A(A'A)^{-1}A\), where

\[
A = (A_{j_1}, \ldots, A_{j_k})\quad \text{for} \quad (A_{j_1}, \ldots, A_{j_k})
\]

some subset of the set \(\{X, A_1, \ldots, A_{r-1}\}\), and where \((A'A)^{-1}\) is a g-inverse of \(A'A\). Since \(tr[(P-Q)BB']\) can be rewritten as \(tr[PBB'] - tr[QBB']\), we need only consider these terms individually. These terms may be written \(tr[PBB'] = tr[B'B] = \sum_k b_k b_k\), where \(b_k\) is the \(k\)-th column of \(B\). This is essentially what Hartley (1967) dubbed synthesis for finding expectations of mean squares. Note that \(B\) does not need to be stored explicitly; rather its columns may be generated one at a time, used in the computation of the trace, and then discarded.

If \(A\) can be stored completely, that should be done so the program executes faster. Recall that the matrix \(A\) is made up mostly of classification matrices. Each classification matrix may have several columns but all the information contained in each classification matrix is contained in a single data variable. Thus the matrix \(A\) may be built column by column using logical operators. For interactions, a single column can be created that contains the information necessary to build the corresponding classification matrix. A rough SAS/IML algorithm follows:

- Create the matrix \(A\).
- Compute \(A'A\). Call subroutine \(ginv\) to compute \((A'A)^{-1}\).
- Initialize the variable \(sum\) to zero.
- For \(k = 1 \text{ to } m\) (where \(m = \# \text{ columns in } B\))
  - Create \(b_k\) (the \(k\)-th column of \(B\)).
  - \(sum = sum + b_k(A'A)^{-1}A'b_k\).
- End for loop.

1067
In large data sets where A has many columns, storing both A and (A'A)\(^{-1}\) may be prohibitive. In this case, the matrix A'A may be calculated element by element. That is, loops may be written that compute the \(ij\)-th element of A'A by creating the \(i\)-th and the \(j\)-th columns of A, and taking their dot product. This way only two columns of A must be in memory at any one time. The same sort of modification could be used to calculate the terms \(A'b_k\). That is, the loop to calculate the sum can be rewritten as a nested loop so the vector (A'B)\(_k\) = A'b\(_k\) is built one element at a time. In this case, the last line of the loop is simply changed to \(\text{sum} = \text{sum} + (A'B)'(A'A)^{-1}(A'B)\). These changes are not difficult; however, many columns of A may be created several times, making these computations time consuming.

Note that if A consists of a single classification matrix, A'A is diagonal. This may be used to write code that stores only the diagonal, rather than the entire matrix A'A, saving memory and implementation time. In this case, the inversion routine can be replaced by taking reciprocals of the diagonal elements.

It is important to note that this computational scheme really does not require the A; to correspond to classification factors; it is more likely to be useful in instances where some of them do. Also, the fixed effects need not be nested in the manner indicated in the example or in the manner of the Burdick-Graybill model. That is, they need not be nested in a random main effect. Rather, it is required that all of the effects be nested in one of the random effects.

6. CONCLUSION

As illustrated, even when many levels of one of several random factors have been sampled, we may be able to obtain Type II estimators for the variance components of a fairly complex model. The scheme employed in this example can be generalized to other models, and we see that for many mixed interaction or nested models we may be able to obtain estimates of the variance components for quite complex models using extremely large data sets. Although this scheme restricts us to using the Type II estimators, it may be a very good alternative in many situations where it otherwise would not be possible to obtain estimates using all of the data.

SAS, and SAS/IML are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. * indicates USA registration.