RECENT DEVELOPMENTS IN VARIANCE COMPONENT ESTIMATION

R. R. Hocking

Follow this and additional works at: http://newprairiepress.org/agstatconference

Part of the Agriculture Commons, and the Applied Statistics Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-No Derivative Works 4.0 License.

Recommended Citation


This is brought to you for free and open access by the Conferences at New Prairie Press. It has been accepted for inclusion in Conference on Applied Statistics in Agriculture by an authorized administrator of New Prairie Press. For more information, please contact cads@k-state.edu.
RECENT DEVELOPMENTS IN
VARIANCE COMPONENT ESTIMATION

R. R. Hocking
Texas A&M University

ABSTRACT

The purpose of this paper is to introduce some recent developments in variance component estimation with emphasis on techniques which provide diagnostic information on the data and the model assumptions. This paper concentrates on the balanced data situation, but suggests a natural extension to the case of unbalanced data. The basic ideas are illustrated by several numerical examples.

Key Words and Phrases: Mixed Models, Diagnostics

1. INTRODUCTION

The concept of linear models with non-scalar covariance structure has a long history dating back to the early writings of Fisher (1918) and (1925). The idea of equating mean squares to expected mean squares in an analysis of variance table to obtain estimates appeared in Fisher (1925). Henderson (1953) proposed a generalization of that concept for the unbalanced case and his methods still are the basis for many estimation programs. The problem of negative estimates motivated Hartley and Rao (1967) to develop a maximum likelihood (ML) algorithm in which the estimates could be constrained to be non-negative at the cost of a significant increase in computing. Patterson and Thompson (1971) proposed a partition of the likelihood function so that one factor depended only on the variance components and suggested a maximization of this function for estimation. This idea, subsequently called restricted maximum likelihood (REML) generalized a procedure used by Anderson and Bancroft (1952). The REML estimates agree with the AOV estimates in the case of balanced data and are generally more acceptable than the biased ML estimators. They do not resolve the problem of negative estimates. The MINQUE estimates proposed by Rao (1971) can be viewed as a non-iterative version of REML.

In this paper, we shall revisit the balanced data case and examine an alternative formulation of the model. The interpretation of this model leads to an examination of the AOV estimators which reveals an interesting structure in terms of sample covariances. This suggests a means of examining the data for unusual observations and also allows a check on certain model assumptions. The notation and basic concepts will be developed in the next section. In Section 3, we provide several examples to illustrate the ideas. The development of the details for these examples is given by Hocking et al (1990). The form of these estimators suggests a new approach for the unbalanced data situation. The procedure is described by Hocking et al (1989).

2. NOTATION AND BASIC CONCEPTS

The general mixed linear model can be conveniently described in a notation which was used by Hartley and Rao (1967). With Y denoting the N-vector of observations, X the full column rank design matrix for the fixed effects, with associated parameter vector \( \theta \), \( U_i \) the indicator matrices associated with the random effects, and \( \alpha_i \) the...
random vectors, assumed to be independent, \( N(0, \sigma_i^2 I) \), we have,

\[
Y = X\theta + \sum_{i=0}^{t} U_i \alpha_i
\]  

(1)

To illustrate this notation, consider the two-factor model with factor one assumed to be fixed and factor two random. In algebraic form, we write this model as

\[
y_{ijk} = \mu + \tau_i + \beta_j + (\tau\beta)_{ij} + \epsilon_{ijk}
\]  

(2)

Here, \( \mu + \tau_i \) denotes the mean of the \( i^{th} \) treatment, and \( \beta_j \), \( (\tau\beta)_{ij} \) and \( \epsilon_{ijk} \) are independent normal variables with zero means and variances \( \sigma_\beta^2, \sigma_{\tau\beta}^2, \) and \( \sigma_\epsilon^2 \).

To describe the model, it is sufficient to specify the mean and covariance structure. In matrix form, the general model (1) is given by

\[
E(Y) = X\theta
\]
\[
VAR(Y) = \sum_{i=0}^{t} \sigma_i^2 U_i U_i^T
\]  

(3)

\[
= \sum_{i=0}^{t} \sigma_i^2 V_i = V
\]

While this result is general, it is more informative to consider the two-factor model in (2). That is,

\[
E(Y_{ijk}) = \mu + \tau_i = \mu_i
\]
\[
VAR(Y_{ijk}) = \sigma_\beta^2 + \sigma_{\tau\beta}^2 + \sigma_\epsilon^2
\]
\[
COV(y_{ijk}, y_{i'j'k'}) = \sigma_\beta^2 + \sigma_{\tau\beta}^2 \quad i = i' \quad j = j' \quad k \neq k'
\]  

(4)
\[
= \sigma_\beta^2 \quad i \neq i' \quad j = j'
\]
\[
= 0 \quad j \neq j'
\]

In this form, note that the variance components may be given a different interpretation. Thus, \( \sigma_\beta^2 \), the variance of the random variable \( \beta_j \) in the linear model (2), may be viewed as the covariance between observations on different treatments but at the same level of the second factor. Similarly, \( \sigma_\beta^2 + \sigma_{\tau\beta}^2 \) denotes the covariance between observations in the same cell of the two-way table. This is interesting from a modeling point of view since, under this formulation, negative values for these parameters, and hence, negative estimates might be acceptable. This formulation suggests an examination of the form of the estimators, and we shall see in the next section that these parameters, or certain functions of them are estimated by averages of sample covariance (or sample variances).

To generalize the notation, we find it convenient to use numerical subscripts for the factors and the variance components. Also, in view of the above interpretation, we shall use the symbol \( \phi \) rather than \( \sigma^2 \) to denote the components. Thus, for example, in the two-factor random model, we will write \( \phi_1, \phi_2, \phi_{12}, \) and \( \phi_0 \) rather than \( \sigma_\tau^2, \sigma_\beta^2, \sigma_{\tau\beta}^2 \).
and $\sigma^2$. The extension of this notation to higher ordered factorial and nested-factorial models should be clear from the examples.

3. EXAMPLES OF ESTIMATES AND DIAGNOSTIC ANALYSIS

Example 1. The randomized block design.

To introduce the basic ideas, consider the randomized block design, with $a_1$ treatments and $a_2$ blocks. The model is given by equation (2) in which we assume one observation per cell and delete the term $(r\beta)_{ij}$. Equivalently, the covariance assumptions are that observations in the same block are correlated with covariance $\phi_2$, observations in different blocks are uncorrelated and the variance of an observation is given by $\phi_0 + \phi_2$. The estimates of these parameters are easily obtained as follows: If we view the data as displayed in a two-way table with treatments as columns and blocks as rows, compute the sample covariance matrix assuming that this data represents $a_2$ observations on the $a_1$ column variables. The resulting covariance matrix, of dimension $a_1$, has the property that each off-diagonal element has expected value $\phi_2$ and each diagonal element has expected value $\phi_0 + \phi_2$. Further, the AOV estimate of $\phi_2$ is the average of these off-diagonal elements and the AOV estimate of $\phi_0 + \phi_2$ is the average of the diagonal elements.

The diagnostic information in this matrix is evident. If the off-diagonal elements do not vary in a reasonable manner about their mean, it suggests that the assumption of equal covariance for each pair of treatments should be questioned. That is, does a common value of $\phi_2$ persist throughout the design, or, should a more general structure be considered. Further, an individual covariance can be examined by constructing a scatter plot of the data in the two columns which are used to compute it. Individual covariances can be negative, and a negative estimate of $\phi_2$ arises when the negatives dominate. The scatter plots may indicate that the problem is caused by one or more outlying observations. Of course, negative estimates should not be the only reason to examine the plots. The analysis of the diagonal elements of the covariance matrix for consistency should also be considered and histograms or box plots of the data may be constructed. In the following example, we indicate a simple graphical summary of this information.

To illustrate these ideas, we consider an example from Johnson and Wichern (1988), page 218. In this case, there are four treatments and nineteen blocks. Actually, the blocks are different dogs and each treatment is administered to each animal. The authors treat this as a repeated measures design and we shall discuss the relation of the two analyses later. The covariance (correlation) matrix for this ‘sleeping dog’ data is given in Table 1. Here, the diagonal elements are sample variances for a given treatment, the elements below the diagonal are the sample covariances and the sample correlations are given above the diagonal.

In this example, apart from the fact that the sample variance for treatment one is a little small, there is not strong evidence to reject the model assumptions. In Figure 1, we provide this same information in graphical form. Here, the off-diagonal scatter plots are constructed for each pair of treatments and the diagonal box plots are constructed for each treatment. The advantage of such a plot should be clear. The role of an observation can be viewed throughout the design. In this case, we see one observation in treatment four which is slightly outlying but appears to have little influence on the results. (This figure was obtained from a sample program called GEX2.SAS in SAS-GRAPH, SAS-GRAPH is a product of the SAS Institute, Inc. Box 8000, Cary, N.C. 27511.)

To relate to the multivariate analysis of this data in Johnson and Wichern (1988), we see that they assume a multivariate model in which the the rows of the data ma-
matrix are assumed to be independent with the same mean vector and covariance matrix. Indeed, their estimate of this covariance matrix is given by Table 1. Our diagnostic analysis of this matrix in the randomized block design is just the check for the assumption of compound symmetry which appears to be reasonable in this case. The issue here is not to question their analysis, but rather to point out that the multivariate analysis of a repeated measures model recognizes the possibility of a general covariance structure while the randomized block model does not.

Example 2. A three-factor mixed model.

To see how these concepts extend to more complex models, we consider a three-factor mixed model in which factors one and two are fixed and factor three is random. In linear model form, we would write

$$y_{ijk} = \mu_{ij} + (\alpha_3)_{ik} + (\alpha_{13})_{jk} + (\alpha_{23})_{ij} + (\alpha_{123})_{ijk} + \epsilon_{ijk}$$

(5)

Here, the random factors are assumed to have variances $\phi_3$, $\phi_{13}$, $\phi_{23}$, $\phi_{123}$ and $\phi_0$. The number of levels of the factors are $a_1$, $a_2$, and $a_3$ with $n$ observations per cell. Using the row, column and level terminology to refer to the three factors, we see that the covariance structure identifies $\phi_3$ as the covariance between observations in the same level but different rows and columns. Observations in the same row and level but different columns have covariance $\phi_3 + \phi_{13}$ and for the same row and level but different rows, the covariance is $\phi_3 + \phi_{23}$. The variance of a cell mean is $\phi_0/n + \phi_3 + \phi_{13} + \phi_{23} + \phi_{123}$.

To estimate these parameters, consider the two-way table with $a_3$ rows and $a_1 \times a_2$ columns containing the cell means. The columns are labeled by all pairs of indices for factors one and two. Computing the sample covariance matrix treating the columns of this table as variables enables us to determine the estimates of these parameter functions as follows:

The off-diagonal elements for which $i \neq i'$ and $j \neq j'$ are unbiased estimates of $\phi_3$ and their average is the AOV estimate of that parameter. The covariances for which $i = i'$ but $j \neq j'$ each estimate $\phi_3 + \phi_{13}$ and there average is the estimate of that parameter function. Similarly, the remaining off-diagonal elements are used to estimate $\phi_3 + \phi_{23}$ and the average of the diagonal elements yields the estimate of the variance of a cell mean. Using the within cell variance to estimate $\phi_0$, we may then obtain estimates of the individual variance components. The diagnostic analysis of these evaluations follows as in Example 1 and is visually aided by a graph as in Figure 1.

To illustrate, we use the milk data taken from Brownlee (1960), page 419. In this case, we have $n = 1$ observation on two levels of factor one and three levels of factor two which are fixed and twelve levels of factor three which is random. The covariance matrix for this data is shown in Table 2 and is interpreted as follows:

- The six off-diagonal elements in the off-diagonal block of this matrix are averaged to estimate $\phi_3$.
- The three diagonal elements of that same block are averaged to estimate $\phi_3 + \phi_{23}$.
- The six off-diagonal elements in the two diagonal blocks are averaged to estimate $\phi_3 + \phi_{13}$.
- The average of the six diagonal elements is the estimate of $\phi_3 + \phi_{13} + \phi_{23}$ which is the variance of an observation. (With $n = 1$ we set $\phi_{123}$ to zero.)

In this example, we obtain a negative estimate of $\phi_{23}$. Examination of the off-diagonal block in Table 2 reveals the source of this problem. Inspection of the scatter plot matrix indicates some potential outlying observations that should be examined but we shall not pursue this example further as the essential points have been made.
Example 3. A nested factorial example.

To illustrate another class of models, consider a three-factor situation in which factors one and two are crossed and factor three is nested in factor one. For simplicity, we assume that the first two factors are fixed and factor three is random. The linear model is written as

$$y_{ijkt} = \mu_{ij} + (\alpha_3(1))_{ik} + (\alpha_{23}(1))_{ijk} + \epsilon_{ijkt}$$

(6)

Here, $\mu_{ij}$ denotes the mean response for the $i,j^{th}$ combination of the first two factors and the remaining terms are random variables with variances $\phi_3(1)$, $\phi_{23}(1)$ and $\phi_0$. The meaning of these components is best understood by examining the covariance structure. It follows that $\phi_3(1)$ is the covariance between observation at the same level of factors one and three but different levels of factor two. The covariance between different replicates at the same levels of all three factors is $\phi_3(1) + \phi_{23}(1)$. To estimate these parameters, we first array the cell means in $a_1$ matrices with $a_3$ rows and $a_2$ columns. Sample covariance matrices are computed for each of these matrices assuming the columns are the variables. The estimate of $\phi_3(1)$ is given by the average of the off-diagonal elements of these matrices. The average of the diagonal elements gives the estimate of $\phi_0/n + \phi_3(1) + \phi_{23}(1)$ and the residual mean square is used to estimate $\phi_0$. The use of scatter plots to diagnose problems with the data proceeds in the usual way. The model assumptions can be assessed by an examination of these matrices. Each matrix should have constant diagonal elements and constant off-diagonal elements and the matrices should be identical. Departures from this which cannot be attributed to sampling variability should be examined.

To illustrate, we consider another example from Brownlee (1960), page 434. In this case, the first two factors are at two levels and the third has three levels. There are $n = 2$ observations at each factor combination. The two covariance matrices are shown in Table 3.

The striking feature of Table 3 in the difference between the covariance matrices for the two levels of factor one. The AOV estimate of $\phi_3(1)$ (593) is the average of the two covariances, 1160 and 26. The negative estimate of $\phi_{23}(1)$ (-26.5) is the average of 520 and -573 since $\hat{\phi}_0 = 106$. Based on this evidence, it would be difficult to accept the model assumptions in this case. Of course, we must recognize that each of the covariance matrices is based on only three observations, but this example illustrates that there is information about the estimates which is not evident from the standard analysis.

Again, we make a comparison with a repeated measures design. Consider an experiment in which each of $a_1$ drugs was administered at $a_2$ times to $a_3$ subjects. The multivariate model would assume that the observations on a given subject would be correlated and an estimate of this common covariance matrix would be obtained for each drug. Relating the drugs to factor one and the times to factor two in the nested-factorial example, we see that these are just the matrices in Table 3. The variance component model makes the added assumption that these matrices have constant diagonal and off-diagonal elements.

Example 4. A split-plot design.

As a final example, we consider a simple split-plot design. The model may be written as follows:

$$y_{ijk} = \mu_{jk} + \rho_i + \delta_{ij} + \epsilon_{ijk}$$

(7)
Here, \( \mu_{jk} \) denotes the mean response of the \( j^{th} \) whole plot and \( k^{th} \) subplot treatment combinations, and the remaining terms are random variables with variances \( \phi_1, \phi_{12} \) and \( \phi_0 \). We note that this model, apart from changes in subscripting, is a special case of the model for Example 2 as given in (5). As in the milk data, we have \( n = 1 \), but more importantly, we have deleted the term corresponding to the block by subplot interaction. Equivalently, we have assumed that \( \phi_{13} \) is zero. The purpose of this example is to examine the consequences of this modification.

Examining the covariance structure which is implicit in this model, we see that \( \phi_1 \) is the covariance between observations in the same block but different whole plots. Assuming that \( \phi_{13} \) is zero is equivalent to assuming that this covariance does not depend on the subplot label. Similarly, \( \phi_1 + \phi_{12} \) is the covariance between observations in the same block and whole plot and different subplots.

The estimate of \( \phi_1 \) is obtained by arraying the marginal means (averaged over subplots) in a table with \( a_1 \) rows and \( a_2 \) columns and computing the sample covariance matrix. Here, \( a_1 \) is the number of blocks and \( a_2 \) is the number of whole plot treatments. The average of the off-diagonal elements of this matrix is the estimate of \( \phi_1 \). The average of the diagonal elements of this matrix is the estimate of the variance of such marginal means, that is, \( \phi_1 + \phi_{12} + \phi_0 / a_3 \).

To estimate \( \phi_1 + \phi_{12} \), we array the data in \( a_2 \) tables, one for each whole plot, with \( a_1 \) rows and \( a_3 \) columns. Computing sample covariance matrices for each of these tables and averaging the off-diagonal elements yields the estimate. The average of the diagonal elements of this matrix is the estimate of the variance of an observation.

The diagnostic analysis for this model should be clear. Since no new ideas are introduced, we shall not give a numerical example.

4. SUMMARY AND EXTENSIONS

The ideas presented here by example are described in general by Hocking et al. (1990) for the class of balanced designs. The extension to unbalanced data is considered by Hocking et al. (1989) for factorial models with no missing cells. We used exactly the same procedure as in the balanced case, obtaining the estimates from appropriate covariance matrices. The only difference is that the cell means are based on different numbers of observations. The resulting estimates were seen to be highly efficient except for situations in which the components were small relative to \( \phi_0 \). The efficiencies compared favorably with those of other estimators. This approach extends naturally to the nested factorial models and Hocking (1988) noted the relation to Yates method of unweighted means as described in Searle (1971).

This approach fails with missing cells, but there are some obvious suggestions. If the number of missing cells is small, we might just compute covariances from available pairs. While losing some efficiency, we would still retain the diagnostic information. Alternatively, we might insert predictions for the missing data. These are determined quite simply from the estimated regression relation for the two variables involved. The properties of this approach are under investigation.

Assuming that the details of these extensions can be resolved, this approach has two advantages over existing methods. The ability to examine the data and the model assumptions has been emphasized. There would also appear to be a definite computational advantage. The evaluation of the required covariance matrices would be simpler than fitting linear models as in the AOV methods or inverting large matrices as required in the iterative ML and REML methods.

5. REFERENCES


Table 1. Covariance Matrix for the Sleeping Dog Data

<table>
<thead>
<tr>
<th>Treatment</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2819</td>
<td>.753</td>
<td>.670</td>
<td>.619</td>
</tr>
<tr>
<td>2</td>
<td>3568</td>
<td>7.96</td>
<td>.718</td>
<td>.652</td>
</tr>
<tr>
<td>3</td>
<td>2943</td>
<td>5.04</td>
<td>6.85</td>
<td>.778</td>
</tr>
<tr>
<td>4</td>
<td>2295</td>
<td>4.06</td>
<td>4.50</td>
<td>4.87</td>
</tr>
</tbody>
</table>

Table 2. Covariance Matrix for Milk Data

<table>
<thead>
<tr>
<th>Row/Column</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>21</th>
<th>22</th>
<th>23</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>2.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>2.20</td>
<td>2.81</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>1.27</td>
<td>0.85</td>
<td>2.24</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>1.25</td>
<td>0.98</td>
<td>0.72</td>
<td>1.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>0.55</td>
<td>-0.30</td>
<td>0.75</td>
<td>0.63</td>
<td>1.88</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>2.14</td>
<td>0.95</td>
<td>2.00</td>
<td>2.05</td>
<td>1.82</td>
<td>4.63</td>
</tr>
</tbody>
</table>

Table 3. Covariance Matrices for the Nested - Factorial Example.

<table>
<thead>
<tr>
<th>Factor 1</th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Factor 2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>1271</td>
<td>94</td>
</tr>
<tr>
<td>2</td>
<td>1160</td>
<td>1061</td>
</tr>
</tbody>
</table>
Figure 1. Scatter Plots for Sleeping Dog Data