Paper 252-27

A SAS/IML® Macro for Computation of Confidence Intervals for Variance Components of Mixed Models

Ann Hess, Department of Statistics, Colorado State University, Fort Collins, CO
Hari Iyer, Department of Statistics, Colorado State University, Fort Collins, CO

ABSTRACT

The mixed linear model is used in a variety of disciplines. Point and interval estimates for variance components from these models are often required. Confidence intervals for variance components with superior coverage properties than those afforded by the application of the Satterthwaite method may be computed using the methods discussed in the book "Confidence Intervals for Variance Components" by Burdick and Graybill [1].

To facilitate these calculations, we have written a SAS/IML® macro that will compute these confidence intervals for any balanced mixed effects saturated ANOVA model involving five or fewer factors. User input to the SAS/IML® macro is the actual data set along with a matrix indicating whether each factor is fixed or random and the nesting/crossing configuration among the factors. Such a macro is expected to be extremely useful to practitioners in view of the fact that SAS® or any other commonly available statistical software package does not have built in commands for obtaining confidence intervals for variance components discussed in [1]. A copy of this macro is available at www.stat.colostate.edu/~hess/MixedModels.htm.

INTRODUCTION

The mixed model is a special case of the general linear model, where both the mean function and covariance matrix for the data have a linear structure. Specifically, the vector of observations **y** is assumed to have moments

$$E[\mathbf{y}] = \mathbf{X}\alpha,$$

$$Var[\mathbf{y}] = \sum_t \phi_t \mathbf{V}_t \,,$$

where $\alpha = (\alpha_1,...,\alpha_p)^t$ and $\phi = (\phi_1,...,\phi_q)^t$ are unknown parameters and $\mathbf{X}, \mathbf{V}_1,...\mathbf{V}_q$ are known matrices. Note that any mean vector and covariance matrix can be written in this form, but we are interested in models in which the parameters $\alpha_1,...,\alpha_p,\phi_1,...,\phi_q$ are functionally independent (except perhaps some linear estimability conditions among the α_i) and relatively few parameters are needed to obtain this form. Such general covariance structures arise by assuming that some of the effects are random.

In the discussion that follows, we will often refer to "balanced" data and "saturated" models. For balanced data, at each factor level combination, the same number of responses are available. A saturated model includes all possible main effect and interaction terms of all orders. We will restrict our discussion to models with *n* ≤ 5 factors. In addition, we will generally consider only balanced mixed models.

Both PROC GLM and PROC MIXED are capable of fitting mixed models. In addition, both procedures can provide estimates of the variance components and PROC MIXED can provide confidence intervals for the variance components using the CL option.

In the following, we discuss methods used to construct confidence intervals on variance components and present a SAS/IML® macro used to implement these methods.

GENERATING MIXED MODELS WITH UP TO FIVE FACTORS

Hess [2] created an algorithm to generate all mixed models with up to five factors (not including the error term). The algorithm uses matrices to represent mixed models.

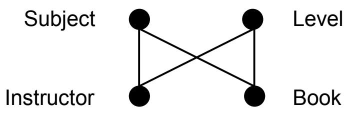
If a mixed model has *n* factors, the matrix representing this model is the *n* x *n* matrix whose (*i,j*) entry is equal to 1 if factor *j* is nested in factor *i* and 0 otherwise. Note that if *j* is nested in *i* the (*i,j*) entry of the matrix is 1, but the (*j,i*) entry is 0. In addition, if factor *j* is random then the (*j,j*) entry is 1 and if factor *j* is fixed then the (*j,j*) entry is 0.

EXAMPLE 1:

Consider a study designed to test the effectiveness of two methods for teaching math and statistics. For each subject (math and statistics) at each of two levels (100 and 200) there are two (randomly selected) instructors. For each subject there are two different books using different teaching methods. Note that the books are different for each subject x level combination. For this study there are four factors: subject (fixed), level (fixed), instructor (random) and book (fixed). Instructor and book are crossed, but both are nested within subject and level. Subject is factor 1; level is factor 2; instructor is factor 3; and book is factor 4. The matrix representing this model is given below.

$$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

This matrix is an algebraic representation that is equivalent to a graphical representation known as the Hasse diagram for the model (see [2]). For this example, the Hasse diagram is shown below.



Using the constraint that a fixed factor may not be nested within a random factor, Hess [2] enumerated nonisomorphic mixed models with up to 5 factors. These results are shown in Table 1.

Table 1: Number of Mixed Models with *n* Factors

n	Number of Mixed Models
1	2
2	6
3	22
4	101
5	576

CONFIDENCE INTERVALS FOR FUNCTIONS OF VARIANCE COMPONENTS

In many cases, we are interested in estimating and obtaining confidence intervals for functions of variance components. Only the mean square terms for random effects are used to estimate variance components. We will use s_t to indicate the sum of squares for factorial effect *t* and r_t to indicate the corresponding degrees of freedom. Equating the mean squares to their expected mean squares gives a set of equations, which can be solved to estimate variance components. This is described as the AOV method or Method of Moments (MOM) for estimating variance components. For balanced data situations they are also the unconstrained REML estimates. In the following discussion, the notation from [3] is used.

If we let λ and ϕ represent the vectors of expected mean squares and variance components, then the relation between these vectors is

$$\lambda = L\phi$$

for a suitable matrix ***L***. By ordering the elements of ϕ appropriately, we note that ***L*** can be written as an upper triangular matrix. Thus we see that there is a one-to-one mapping between variance components and the expected mean squares corresponding to random effects.

The estimate of the vector of variance components is then given by

$$\hat{\phi} = L^{-1}\hat{\lambda}$$

where $\hat{\lambda}$ is the vector of mean squares. Note that the estimates of the expected mean squares are nonnegative, but the estimates of the variance components may sometimes be negative.

In addition to estimating single variance components, one may be interested in estimating a linear function of variance components. Either way, the quantity of interest may be expressed as a linear function of expected mean square terms

$$\gamma = \sum_q c_q \lambda_q \ .$$

An unbiased estimate of γ is given by

$$\hat{\gamma} = \sum_q c_q \frac{s_q}{r_q}$$

Viewed in this way, we can classify the estimation of linear functions of variance components as follows:

1. $\gamma = c_q\lambda_q$ for some *q*,
2. $\gamma = \sum_q c_q\lambda_q, \quad c_q > 0, q > 0,$
3. $\gamma = \sum_{q=1}^{P} c_q\lambda_q - \sum_{r=1}^{N} d_r\lambda_r$, where $c_q, d_q > 0, P \ge 1, N \ge 1$.

Note that in the following discussion, it is assumed that we have a balanced design. Although the methods discussed here can be used when the design is not balanced, the performance of these methods is untested.

CASE 1. ($\gamma = c_q\lambda_q$ **)**

Since $\dfrac{s_q}{\lambda_q}$ has a chi-squared distribution with r_q degrees of freedom, an exact 1-2α confidence interval on γ is

$$\left[c_q \frac{s_q}{\chi^2_{1-\alpha,r_q}}, c_q \frac{s_q}{\chi^2_{\alpha,r_q}} \right]$$

where χ^2_{α,r_q} is the (1-α)th percentile of a χ^2 random variable with r_q degrees of freedom (i.e., right-tail area=α), [1].

CASE 2. ($\gamma = \sum_q c_q\lambda_q, \quad c_q > 0, q > 0$ **)**

In this case, there are multiple approximate methods for computing confidence intervals for γ. These include Satterthwaite's Procedure, Welch's Procedure, and the Modified Large-Sample (MLS) Procedure. Among these, the MLS Procedure, proposed by Graybill and Wang, has been found to have superior coverage probabilities. See [1] for details.

The MLS method is a modification of Welch's Procedure which is exact for special cases. This method is exact when all but one of the λ_q are zero or when all but one of the r_q approach infinity. The Graybill-Wang two-sided 1-2α interval on γ is

$$\left[\hat{\gamma} - \sqrt{\sum_q \mathrm{G}_q^2 \mathrm{c}_q^2 \left(\frac{\mathrm{s}_q}{\mathrm{r}_q} \right)^2} \ , \ \hat{\gamma} + \sqrt{\sum_q \mathrm{H}_q^2 \mathrm{c}_q^2 \left(\frac{\mathrm{s}_q}{\mathrm{r}_q} \right)^2} \right]$$

where

$$G_i = 1 - \frac{\mathrm{r}_q}{\chi^2_{1-\alpha,r_q}} \quad \text{and} \quad H_i = \frac{\mathrm{r}_q}{\chi^2_{\alpha,r_q}} - 1.$$

CASE 3. ($\gamma = \sum_{q=1}^{P} c_q\lambda_q - \sum_{r=1}^{N} d_r\lambda_r, \quad c_q, d_q > 0, P \ge 1, N \ge 1$ **)**

In this case, the method proposed by Ting, Burdick, Graybill, Jeyaratnam, and Lu is recommended. We will refer to this as the TBGJL method. The 1-α lower confidence bound on γ is given by

$$L = \hat{\gamma} - \sqrt{V_L}$$

where

$$V_L = \sum_{\mathrm{q}=1}^{\mathrm{P}} \mathrm{G}_\mathrm{q}^2 \mathrm{c}_\mathrm{q}^2 \left(\frac{s_q}{r_q} \right)^2 + \sum_{\mathrm{r}=1}^{\mathrm{N}} \mathrm{H}_\mathrm{r}^2 \mathrm{d}_\mathrm{r}^2 \left(\frac{s_r}{r_r} \right)^2$$

$$+ \sum_{\mathrm{q}=1}^{\mathrm{P}} \sum_{\mathrm{r}=1}^{\mathrm{N}} \mathrm{G}_{\mathrm{qr}} \mathrm{c}_\mathrm{q} \mathrm{d}_\mathrm{r} \frac{s_q}{r_q} \frac{s_r}{r_r} + \sum_{\mathrm{q}=1}^{\mathrm{P}-1} \sum_{\mathrm{t}>\mathrm{q}}^{\mathrm{P}} \mathrm{G}_{\mathrm{qt}}^* \mathrm{c}_\mathrm{q} \mathrm{c}_\mathrm{t} \frac{s_q}{r_q} \frac{s_t}{r_t}$$

$$G_q = 1 - \frac{r_q}{\chi^2_{1-\alpha,r_q}}$$

$$H_q = \frac{r_q}{\chi^2_{\alpha,r_q}} - 1$$

$$\mathrm{G}_{\mathrm{qr}} = \frac{(\mathrm{F}_{1-\alpha,r_q,r_r} - 1)^2 - \mathrm{G}_\mathrm{q}^2 \mathrm{F}_{1-\alpha,r_q,r_r}^2 - \mathrm{H}_\mathrm{r}^2}{\mathrm{F}_{1-\alpha,r_q,r_r}}$$

$$\mathrm{G}_{\mathrm{qt}}^* = \left[\left(1 - \frac{\mathrm{r}_q + \mathrm{r}_t}{\chi^2_{1-\alpha,r_q+r_t}} \right)^2 \frac{(\mathrm{r}_q + \mathrm{r}_t)^2}{\mathrm{r}_q r_t} - \frac{\mathrm{G}_\mathrm{q}^2 \mathrm{r}_q}{\mathrm{r}_\mathrm{t}} - \frac{\mathrm{G}_\mathrm{t}^2 \mathrm{r}_t}{\mathrm{r}_\mathrm{q}} \right] / (\mathrm{P} - 1).$$

If P=1 then G_{qt}^* is defined to be zero.

The upper 1-α confidence bound for γ is given by

$$U = \hat{\gamma} + \sqrt{V_U}$$

where

$$V_U = \sum_{\mathrm{q}=1}^{\mathrm{P}} \mathrm{H}_\mathrm{q}^2 \mathrm{c}_\mathrm{q}^2 \left(\frac{s_q}{r_q} \right)^2 + \sum_{\mathrm{r}=1}^{\mathrm{N}} \mathrm{G}_\mathrm{r}^2 \mathrm{d}_\mathrm{r}^2 \left(\frac{s_r}{r_r} \right)^2$$

$$+ \sum_{\mathrm{q}=1}^{\mathrm{P}} \sum_{\mathrm{r}=1}^{\mathrm{N}} \mathrm{H}_{\mathrm{qr}} \mathrm{c}_\mathrm{q} \mathrm{d}_\mathrm{r} \frac{s_q}{r_q} \frac{s_r}{r_r} + \sum_{\mathrm{r}=1}^{\mathrm{N}-1} \sum_{u>r}^{\mathrm{N}} \mathrm{H}_{\mathrm{ru}}^* \mathrm{c}_\mathrm{r} \mathrm{c}_\mathrm{u} \frac{s_r}{r_r} \frac{s_u}{r_u}$$

$$\mathrm{H}_{\mathrm{qr}} = \frac{(1 - \mathrm{F}_{\alpha,r_q,r_r)})^2 - \mathrm{H}_\mathrm{q}^2 \mathrm{F}_{\alpha,r_q,r_r}^2 - \mathrm{G}_\mathrm{r}^2}{\mathrm{F}_{\alpha,r_q,r_r}}$$

$$\mathrm{H}^*_{\mathrm{ru}} = \left[\left(1 - \frac{\mathrm{r}_r + \mathrm{r}_u}{\chi^2_{1-\alpha, r_r + r_u}}\right)^2 \frac{(r_r + r_u)^2}{\mathrm{r}_r \mathrm{r}_u} - \frac{\mathrm{G}^2_{\mathrm{r}} r_r}{r_u} - \frac{\mathrm{G}^2_{\mathrm{u}} r_u}{r_r}\right] / (\mathrm{N} - 1).$$

If N=1 then H^*_{ru} is defined to be zero.

Ting, Burdick, Graybill, Jeyaratnam, and Lu used computer simulation to demonstrate that this interval provides confidence coefficients close to the stated levels over a wide range of conditions; see [6].

There are alternatives to the TBGJL method. Let

$$\gamma = \sum_{\mathrm{q}=1}^{\mathrm{P}} \mathrm{c}_{\mathrm{q}} \lambda_q - \sum_{\mathrm{r}=1}^{\mathrm{N}} \mathrm{d}_{\mathrm{r}} \lambda_r = \theta_P - \theta_N .$$

One alternative to the TBGJL method is to apply chi-squared approximations to $\hat{\theta}_P$ and to $\hat{\theta}_N$, then use Howe's method to obtain a confidence interval for $\theta_P - \theta_N$ Refer to [4] for a discussion of Howe's method.

A SAS/IML® MACRO FOR COMPUTING CONFIDENCE INTERVALS ON LINEAR FUNCTIONS OF VARIANCE COMPONENTS

At the time of this work, SAS® PROC MIXED output generally includes only asymptotic confidence intervals on variance components. Specifically, for parameters that have a lower bound constrained by zero, a Satterthwaite approximation is used to construct a confidence interval. The 1-2α confidence interval for σ^2 is

$$\left[\frac{v\hat{\sigma}^2}{\chi^2_{1-\alpha,v}}, \frac{v\hat{\sigma}^2}{\chi^2_{\alpha,v}}\right]$$

where $v = 2Z^2$, and *Z* is the Wald statistic $\dfrac{\hat{\sigma}^2}{SE(\hat{\sigma}^2)}$. For all other parameters, Wald Z-scores and normal quantiles are used to construct the limits. See [5] for details.

Better methods exist for computing confidence intervals for variance components. Hence, it is believed that a macro that would provide superior intervals would be useful.

Specifically, the goal of the following macro is that the user need only input a data set, the factor names, and an adjacency matrix describing the model. The macro would fit a saturated model to the data and provide an ANOVA table as well as estimates and confidence intervals for the variance components. Additional estimates and confidence intervals for linear functions of variance components can be provided. The following algorithm was used to accomplish these goals.

1. Start with the list of matrices describing all mixed models with up to five factors (not including the error term). We chose to include only those mixed models which obey the constraint that a fixed factor may not be nested within a random factor.
2. The user input matrix (or a permutation of this matrix) is matched to one of the matrices in the list from Step 1.
3. A (saturated) model statement corresponding to the user input matrix is run through PROC MIXED, generating an ANOVA table with Expected Mean Squares.
4. A matrix (***L***) defining the linear relationships between EMS and variance components is generated.
5. Using the matrix, ***L***, from above, the appropriate method (Exact, MLS, TBGJL) is used to obtain a confidence interval for each of the variance components.

EXAMPLE 2:

Suppose an experiment is conducted to examine the causes of variation in milk production. At each of two randomly selected farms, three milking machines were selected. Five cows were assigned to each machine, and the amount of milk produced was recorded for each of three days. All the factors are random, and cow is nested within machine which is nested within farm.

SAS User Input:

```
/*Step 1: Entering Variable Names and
Number of Observations */
%let V1=Farm;
%let V2=Machine;
%let V3=Cow;
%let Response=Y;
%let n=90;

/*Step 2: Entering the Balanced Data*/
data Cows; do Farm=1 to 2;
do Machine=1 to 3;
do Cow=1 to 5;
Rep=1 to 3;
input y @@;
output;
end;
end;
end;
end;
cards;
2.37  3.02  2.59
...
3.05  2.94 3.07 ;

/*Step 3: Entering the Adjacency Matrix*/
data matrix;
input Farm Machine Cow;
cards;
1 1 0
0 1 1
0 0 1 ;

/*Step 4: Requesting an CI for the sum of
Variance Components.*/
data Lfuncs;
input Farm MachineFarm CowMachineFarm
Residual;
cards;
1 1 1 1 ;
```

SAS Macro Output:

```
   Type 3 Analysis of Variance

                           Sum of
   Source            DF    Squares    Mean Square  Expected Mean Square

   Farm               1    0.645160    0.645160    Var(Residual) + 3 Var(Cow(Farm*Machine))
                                                   + 15 Var(Machine(Farm)) + 45 Var(Farm)
   Machine(Farm)      4    1.669182    0.417296    Var(Residual) + 3 Var(Cow(Farm*Machine))
                                                   + 15 Var(Machine(Farm))
   Cow(Farm*Machine) 24    2.014187    0.083924    Var(Residual) + 3 Var(Cow(Farm*Machine))
   Residual          60    5.031600    0.083860    Var(Residual)


   The following are approximate        95 % Confidence Intervals
   (except when noted as exact)

   PARAMETER          METHOD     LB      ESTIMATE    UB

   Farm               TBGJL  -0.061383 0.0050637 14.586546
   Machine(Farm)      TBGJL  0.0035236 0.0222247 0.2238027
   Cow(Farm*Machine)  TBGJL   -0.01739 0.0000215 0.0270036
   Residual           Exact  0.060405  0.08386   0.1242931


   LFCNS        PARAMETER          METHOD     LB      ESTIMATE    UB

   1  1  1  1   Farm               GWMLS  0.0867974 0.1111699 14.69615
                Machine(Farm)
                Cow(Farm*Machine)
                Residual
```

Note that this macro can be used even with unbalanced data, but methods used for computing confidence intervals for variance components are further approximations (because the sums of squares are no longer independent chi-squared random variables) of already approximate methods. The methods considered here remain largely untested for unbalanced data.

CONCLUSION

We have written a SAS/IML® macro that will compute confidence intervals for any balanced mixed effects saturated ANOVA model involving five or fewer factors. User input to the SAS/IML® macro is the actual data set along with a matrix indicating whether each factor is fixed or random and the nesting/crossing configuration among the factors. This macro is expected to be extremely useful to practitioners in view of the fact that SAS® or any other commonly available statistical software package does not have built in commands for obtaining confidence intervals for variance components discussed in [1]. A copy of this macro is available at www.stat.colostate.edu/~hess/MixedModels.htm.

REFERENCES

[1] Richard K. Burdick and Franklin A. Graybill. *Confidence Intervals on Variance Components*. Marcel Dekker, Inc., New York, NY, 1992.

[2] Ann M. Hess. Enumeration of mixed models and a SAS macro for computation of confidence intervals for variance components. Master's thesis, Colorado State University, 2001.

[3] Ronald R. Hocking. *Methods and Applications of Linear Models: Regression and the Analysis of Variance.* John Wiley and Sons, Inc., New York, NY, 1996.

[4] W.G. Howe. Approximate confidence limits on the mean of x+y where x and y are two tabled independent random variables. *Journal of the American Statistical Association*, 69:789-794, 1974.

[5] SAS Institute Inc. *SAS OnLine Doc, Version 8*, Cary,NC, 1990.

[6] Naitee Ting, Richard K. Burdick, Franklin A. Graybill, S. Jeyaratnam, and Tai-Fang C. Lu. Confidence intervals on linear combinations of variance components that are unrestricted in sign. *Journal of Statistical Computing and Simulation*, 35:135-143, 1990.

CONTACT INFORMATION

Contact the author at:

Ann Hess
Department of Statistics
Colorado State University
Fort Collins, CO 80523
(970) 491-6014
Email: hess@stat.colostate.edu
Web: www.stat.colostate.edu/~hess/MixedModels.htm