USING SAS FOR MIXED MODEL CONSTRAINED SOLUTIONS FOR VARIANCE COMPONENTS

BU-1495 -M **July, 2000**

Walter T. Federer
Dept. of Biometrics
Cornell University

Keywords: ANOVA solutions, unconstrained solutions, harmonic means, expected mean squares.

Abstract:

Computer software such as SAS makes use of unconstrained solutions for variance components when using a MIXED procedure. It appears advisable in genetic and breeding studies that the use of constrained solutions are more appropriate and realistic. However, experimenters are unlikely to use the more appropriate procedure unless computer code is available for doing this. Such a code is presented herein. Analysis of variance (ANOVA) solutions for the variance components are obtained. These are then used to recover the information from the random effects in a MIXED model analysis. Note that ANOVA solutions do not depend upon the requirement of normality as do maximum likelihood and restricted maximum likelihood (REML) solutions for the variance components.

USING SAS FOR MIXED MODEL CONSTRAINED SOLUTIONS FOR VARIANCE COMPONENTS

by

Walter T. Federer

ABSTRACT

Computer software such as SAS makes use of unconstrained solutions for variance components when using a MIXED procedure. It appears advisable in genetic and breeding studies that the use of constrained solutions are more appropriate and realistic. However, experimenters are unlikely to use the more appropriate procedure unless computer code is available for doing this. Such a code is presented herein. Analysis of variance (ANOVA) solutions for the variance components are obtained. These are then used to recover the information from the random effects in a MIXED model analysis. Note that ANOVA solutions do not depend upon the requirement of normality as do maximum likelihood and restricted maximum likelihood (REML) solutions for the variance components.

Key Words: ANOVA solutions, unconstrained solutions, harmonic means, expected mean squares.

BU-1496-M July 2000
In the Technical Report Series of the Department of Biometrics, Cornell University, Ithaca, New York 14853

INTRODUCTION

The default option in the SAS/MIXED procedure is to use *unconstrained* solutions for variance components. The question arises as to how to utilize SAS procedures to use *constrained* solutions for the variance components. In order for researchers to do this, it is desirable to present a SAS code for doing the procedure for unequal numbers of observations per cell in a two-way table. This is the object of this report.

It should be noted that the constraints on the parameters are the result of the definition of main effects and interactions. They are not the result of using an over-parameterized model. A discussion of the appropriateness of using constrained or unconstrained solutions for variance component is given by Basford *et al.* (2000). These authors denote the constrained solutions as CP solutions and the unconstrained solutions as UP solutions. Basford *et al.* (2000) recommend that the CP solutions be used in genetic and breeding investigations. A mixed model situation arises when genotypes are considered to be random effects and environments as fixed effects, or *vice versa*. Note that when a genotype is selected, it is compared in all the environments of interest, meaning that the entire set of its interactions are present. From the definition then, these interaction effects sum to zero. This results in constrained solutions for the variance components.

SAS FOR CP SOLUTIONS

To obtain CP solutions for the variance components, SAS/GLM is utilized. The following code is used. The data file is called test.dat, y is yield, e is environment, and g is genotype.

```
Data test;
   Infile 'test.dat';
   Input y g e;
SAS GLM data = test;
   Class e g;
   Model y = e g e*g;
   Random g e*g;
Run;
```

The SAS output for e fixed and g and e*g random will include the following:

Source	Type III Expected Mean Square
E	Var(Error) + a Var(E*G) + Q(E)
G	Var(Error) + b Var(E*G) + c Var(G)
E*G	Var(Error) + d Var(E*G)

The coefficients a, b, c, and d will be numerical. For CP solutions, delete the term b Var(E*G) from the G line. Also, instead of using the coefficient d, one should use sd/(s - 1) where s is the number of environments. In obtaining the expected value of the E*G interaction mean square, The sum of the interaction terms over environments sum to zero. With these modifications, analysis of variance (ANOVA) solutions for the variance components are obtained as

Var(G) = mean square for G minus mean square for Error divided by c

Var(E*G) = mean square for E*G minus mean square for Error divided by the coefficient sd/(s - 1)

These ANOVA solutions for the variance components are used in the SAS/MIXED procedure as follows:

```
SAS MIXED data = test;
   Class e g;
   Model y = e;
   Random g e*g/solution;
   PARMS (VAR(G)) (VAR(E*G))  (VAR(ERROR))/noiter;
   Lsmeans e;
Run;
```

The PARMS statement allows use of ANOVA solutions for the variance components. The term "noiter" indicates that no iteration is to be performed to obtain variance component solutions as is the case with REML solutions.

THE EXAMPLE

To illustrate the above, consider the following example with s = 4 environments, v = 3 genotypes and N = 15 responses (yields):

	G					interaction effects		
E	1	2	3	mean	effect=e_i	1	2	3
1	4,4	3	2	2	-11/6	25/12	-8/12	-17/12
2	5	6	7	6	7/6	1/12	-8/12	7/12
3	4	8	9	7	13/6	-23/12	4/12	19/12
4	3,1	5	2, 4	10/3	-9/6	-3/12	12/12	-9/12

mean 15/4 22/4 21/4 29/6 0
eff=g_j -12/12 8/12 5/12 0

The means are means of means. Following the method described by Federer and Zelen (1966) for multi-way classifications with unequal numbers, we compute the harmonic means of numbers of observations for the two main effects as:

he(i)	hg(j)
he(1)=3/(1/2+1+1)=12/10	hg(1)=4/(1/2+1+1+1/2)=28/21
he(2)=3/(1+1+1)=10/10	hg(2)=4/(1+1+1+1)=21/21
he(3)=3/(1+1+1)=10/10	hg(3)=4/(1+1+1+1/2)=24/21
he(4)=3/(1/2+1+1/2)=15/10	sum = hg(.) = 73/21
sum = he(.) = 47/10	

The type III sum of squares for environments is computed as:

$SS(e)= s\{\sum_i he(i)e_i^2 - [\sum_i he(i)e_i]^2/he(.)\} = 3\{(6/5)(-11/6)^2 + 1(7/6)^2 + 1(13/6)^2 + (3/2)(-9/6)^2 - [(6/5)(-11/6) + 1(7/6) + 1(13/6) + (3/2)(-9/6)]^2/(47/10)\}$
$= 39.595744.$

The Type III sum of squares for genotypes is computed as:

$SS(g) = v\{\sum_j hg(j)g_j^2 - [\sum_j hg(i)g_j]^2/hg(.)\} = 4\{(28/21)(-13/12)^2 + (21/21)(8/12)^2 + (24/21)(5/12)^2 - [28/21)(-13/12) + (21/21)(8/12) + (24/21)(5/12)]^2/(73/21)\}$
$= 8.7260244.$

The sums of squares agree with those given in the output below. The e*g interaction sum of squares may be computed as described by Federer and Zelen (1966), and is 19.75862069 with 6 degees of freedom. The residual or error sum of squares is 4 with 3 degrees of freedom. Therefore the estimated variance components are

$$Var(G)= (4.36301370 - 4/3)/4.6027 = 0.6582,$$

where 4.6027 is the coefficient obtained in the SAS/GLM output. From Federer and Zelen (1966), this coefficient is computed as

$$s\{\sum hg(j) - \sum hg(j)^2/\sum hg(j)\}/(v - 1) = 4\{73/21 - [(28/21)^2 + (21/21)^2 + (24/21)^2]/(73/21)\}/2 = 3\{73/21 - 1.17482\}/2 = 4.60274.$$

A COMPUTER CODE AND OUTPUT FOR THE EXAMPLE

The computer code for obtaining BLUPs using CP solutions for variance components is given below:

```
data test; input y e g; datalines;
4  1  1
4  1  1
3  1  2
2  1  3
5  2  1
6  2  2
7  2  3
4  3  1
8  3  2
```

```
9  3  3
3  4  1
1  4  1
5  4  2
4  4  3
2  4  3
; run;
proc glm data = test;
  class e g;
  model y = e g e*g;
  random g e*g;
run;
proc mixed data = test;
  class e g;
  model y = e/solution;
  random g e*g/solution;
  parms (0.6582) (1.1318) (1.3333)/noiter;
  lsmeans e;
run;
```

The output of this program is:

General Linear Models Procedure
Class Level Information

Class	Levels	Values
E	4	1 2 3 4
G	3	1 2 3

Number of observations in data set = 15

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	11	67.73333333	6.15757576	4.62	0.1171
Error	3	4.00000000	1.33333333		
Corrected Total	14	71.73333333			

R-Square	C.V.	Root MSE	Y Mean
0.944238	25.85150	1.154701	4.466667

Dependent Variable: Y

Source	DF	Type I SS	Mean Square	F Value	Pr > F
E	3	42.98333333	14.32777778	10.75	0.0411
G	2	5.99137931	2.99568966	2.25	0.2533
E*G	6	18.75862069	3.12643678	2.34	0.2586

Source	DF	Type III SS	Mean Square	F Value	Pr > F
E	3	39.59574468	13.19858156	9.90	0.0459
G	2	8.72602740	4.36301370	3.27	0.1762
E*G	6	18.75862069	3.12643678	2.34	0.2586

Source	Type III Expected Mean Square
E	Var(Error) + 1.1631 Var(E*G) + Q(E)
G	Var(Error) + 1.1507 Var(E*G) + 4.6027 Var(G)
E*G	Var(Error) + 1.1882 Var(E*G)

The MIXED Procedure

Parameter Search

COVP1	COVP2	COVP3	Variance	RLL	-2RLL	Objective
0.6582	1.1318	1.3333	1.3483	-23.4127	46.8254	26.6088

Covariance Parameter Estimates (Parms)

Cov Parm	Estimate
G	0.66560332
E*G	1.14453028
Residual	1.34829672

Model Fitting Information for Y

Description	Value
Observations	15.0000
Res Log Likelihood	-23.4127
Akaike's Information Criterion	-26.4127

Model Fitting Information for Y

Description	Value
Schwarz's Bayesian Criterion	-27.0096
-2 Res Log Likelihood	46.8254

Solution for Fixed Effects

Effect	E	Estimate	Std Error	DF	t	Pr > |t|
INTERCEPT		3.20184601	0.94351582	2	3.39	0.0769
E	1	-0.03346479	1.18663424	6	-0.03	0.9784
E	2	2.79815399	1.22445770	6	2.29	0.0624
E	3	3.79815399	1.22445770	6	3.10	0.0211
E	4	0.00000000	.	.	.	.

Solution for Random Effects

Effect	E	G	Estimate	SE Pred	DF	t	Pr > |t|
G		1	-0.53112866	0.65031824	3	-0.82	0.4739
G		2	0.33953519	0.65748107	3	0.52	0.6412
G		3	0.19159348	0.65353297	3	0.29	0.7885
E*G	1	1	0.85760381	0.88330441	3	0.97	0.4032
E*G	1	2	-0.23319938	0.90806973	3	-0.26	0.8139
E*G	1	3	-0.62440443	0.90733651	3	-0.69	0.5408
E*G	2	1	-0.21527264	0.91465060	3	-0.24	0.8291
E*G	2	2	-0.15589060	0.91572944	3	-0.17	0.8757
E*G	2	3	0.37116324	0.91513349	3	0.41	0.7122
E*G	3	1	-1.13353153	0.91465060	3	-1.24	0.3033
E*G	3	2	0.30323884	0.91572944	3	0.33	0.7623
E*G	3	3	0.83029269	0.91513349	3	0.91	0.4311
E*G	4	1	-0.42209563	0.87205686	3	-0.48	0.6615
E*G	4	2	0.66969484	0.90170725	3	0.74	0.5115
E*G	4	3	-0.24759921	0.87281971	3	-0.28	0.7951

Tests of Fixed Effects

Source	NDF	DDF	Type III F	Pr > F
E	3	6	4.97	0.0458

Least Squares Means

Effect	E	LSMEAN	Std Error	DF	t	Pr > |t|
E	1	3.16838121	0.98176431	6	3.23	0.0180
E	2	6.00000000	1.02606535	6	5.85	0.0011
E	3	7.00000000	1.02606535	6	6.82	0.0005
E	4	3.20184601	0.94351582	6	3.39	0.0146

LITERATURE CITED

Basford, K. E., W. T. Federer, and I. DeLacy (2000).

Federer, W. T. and M. Zelen (1966). Analysis of multifactor classifications with unequal numbers of observations. Biometrics 22(3):525-552.