BU-1625-M

CONSTRAINED AND UNCONSTRAINED FORMULATIONS OF THE MIXED MODEL

by

Walter T. Federer

ABSTRACT

When one of the categories has randomly selected levels from a distribution and the other category of a two-way classification (factorial) has a fixed number of levels, this is known as the mixed model. Two formulations of this model have been discussed in the literature since the 1950's. One formulation, constrained, uses the constraint that the sum of the interaction effect parameters over the levels of the fixed category adds to zero. The second formulation, unconstrained, does not uses this constraint on the interaction parameters. Estimates of variance components are different for the two formulations. This has practical implications in genetic and other studies. Estimates of genetic correlation and heritability, but not genetic advance, are different and this has caused difficulties for the geneticist and breeder. Using accepted definitions of main effects and interactions and the design of the sampling procedure of the random category, it is shown that the unconstrained formulation assumptions are invalid.

Keywords: Population parameter; population structure; definition of main effect and interaction; variance component estimation; sample design; fixed effects model; random effects model; unequal numbers of observations.

BU-1625-M in the Technical Report Series of the Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, March.

INTRODUCTION

For a two-way classification (factorial), a mixed model arises when one of the classes has a fixed number of levels, fixed effect, and the other class has a randomly selected number of levels from a distribution, random effect. When both classes are fixed, this is the fixed effects case. When both classes are random this is called the random effects case. When one of classes has a fixed number of effects and the second has an infinite number, this is called the mixed effects case. Since the 1950's, many papers have been written on how to estimate variance components and make tests of hypotheses for the mixed effects or model situation. One formulation, constrained, uses the constraint that the sum of the interaction parameters adds to zero when summed over

the fixed effect category or class. The other formulation, unconstrained, does not use this constraint in developing the expected values of mean squares in an analysis of variance.

The fixed, random, and mixed effects cases are discussed below. A definition of fixed effects and interaction effects in terms of the population parameters is discussed. This definition is the commonly accepted one. Then, the random effects case is considered and the expected values of mean squares from an analysis of variance are given. In the section on mixed effects, the sampling procedure in terms of the population structure is described. The assumptions for both the constrained and unconstrained formulations are given and discussed. The validity of the assumptions is investigated. It is shown that the assumptions for the unconstrained model are not consistent with the fixed effects model, does not consider the sampling plan, and uses invalid assumptions to obtain the expected value for mean squares. The problem of unequal numbers of observations is also discussed.

FIXED EFFECTS

Consider a two-way classification (factorial) such as g genotypes and s sites (environments). Suppose there are n observations for each of the gs combinations of genotypes and sites. The population means for this situation are:

			Site			
Genotype	1	2	3	...	s	mean
1	μ_{11}	μ_{12}	μ_{13}	...	μ_{1s}	$\mu_{1.}$
2	μ_{21}	μ_{22}	μ_{23}	...	μ_{2s}	$\mu_{2.}$
3	μ_{31}	μ_{32}	μ_{33}	...	μ_{3s}	$\mu_{3.}$
...	...	...	...	...	...	...
g	μ_{g1}	μ_{g2}	μ_{g3}	...	μ_{gs}	$\mu_{g.}$
mean	$\mu_{.1}$	$\mu_{.2}$	$\mu_{.3}$	...	$\mu_{.s}$	$\mu_{..}$

where μ_{ij} is the population mean for genotype i = 1, 2, ..., g at site j = 1, 2, ..., s, $\mu_{i.}$ is the population mean of genotype i over the s sites, $\mu_{.j}$ is the population mean of site j over the g genotypes, $\mu_{..}$ is the overall mean of the gs combinations. The ith genotype effect is defined as $\alpha_i = \mu_{i.} - \mu_{..}$, the jth site effect is defined as $\beta_j = \mu_{.j} - \mu_{..}$, and the ijth interaction effect is defined as $\delta_{ij} = \mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}$.

By definition then

$$\sum_{i=1}^{g}\alpha_i = \sum_{j=1}^{s}\beta_j = \sum_{i=1}^{g}\delta_{ij} = \sum_{j=1}^{s}\delta_{ij} = 0.$$

These constraints are imposed upon the parameters by definition, a point apparently disagreed on by Nelder (1998). He did not explain what he means by main effects and interactions in terms of population means. These constraints are not arbitrary but arise as a result of the definition of main effects and interactions. An estimate of μ_{ij} is $\overline{y}_{ij}$. Then

the ith genotype effect is estimated by $\bar{y}_{i.} - \bar{y}_{..}$, the jth site effect is estimated by $\bar{y}_{.j} - \bar{y}_{..}$, and the ijth interaction effect is estimated by $\bar{y}_{ij} - \bar{y}_{i.} - \bar{y}_{.j} + \bar{y}_{..}$. The sums of these estimated effects are zero. This also is a fact and not an arbitray set of conditions. The linear model for the above two-way classification is:

$$\begin{aligned} Y_{ijh} &= \mu_{ij} + \epsilon_{ijh} \\ &= \mu_{..} + (\mu_{i.} - \mu_{..}) + (\mu_{.j} - \mu_{..}) + (\mu_{ij} - \mu_{i.} - \mu_{.j} + \mu_{..}) + \epsilon_{ijh} \\ &= \mu_{..} + \alpha_i + \beta_j + \delta_{ij} + \epsilon_{ijh} \end{aligned}$$

where h = 1, 2, 3, ..., n_{ij} and the other symbols are defined above and ϵ_{ijh} are independently and identically distributed as IID(0, σ^2_ϵ). Normality is required for testing but not for estimation of effects.

There appears to be agreement among statisticians of the above definition of effects. If so, then the rest follows by definition. However some statisticians have contended that a factorial arrangement *must* have equal numbers of observation for each of the gs categories. This not a requirement as demonstrated by Zelen and Federer (1965) and Federer and Zelen (1966). Unequal numbers of observation is a consequence of the sampling procedure and not of the population structure. Parameters being estimated and hypotheses being tested remain unchanged whether or not n_{ij} = n, a constant.

The expected values of the mean squares for the fixed effects case is:

Source of variation	Degrees of freedom	Mean square expected value
Genotypes	g - 1	$\sigma^2_\epsilon + f(\alpha_i)$
Sites	s - 1	$\sigma^2_\epsilon + f(\beta_j)$
Genotype × site	(g - 1)(s - 1)	$\sigma^2_\epsilon + f(\delta_{ij})$
Residual or error	gs(n - 1)	σ^2_ϵ

where f(x) means a function of x. There is general agreement among statisticians on the above. When testing hypotheses about the effects, the residual or error mean square is used. When n_{ij} is not equal to n, the computations and algebra become complicated but estimation of effects and hypotheses being tested remain unchanged.

RANDOM EFFECTS

Letting the number of genotypes and of sites go to infinity (or some large number) and obtaining a simple random sample of g genotypes and of s sites, we have the random effects situation. The definitions of population structure and of effects and interaction does not change. However the estimation procedure does change and BLUP solutions for the random effects, interactions, and means would ordinarily be obtained. Often the variance components are the quantities of interest. The linear model of the previous section is used here. The expected values of the mean squares in an analysis of variance are:

Source of variation	Degrees of freedom	Mean square expected value
Genotypes	g - 1	$\sigma^2_\epsilon + n\,\sigma^2_{\alpha\beta} + ns\,\sigma^2_\alpha$
Sites	s - 1	$\sigma^2_\epsilon + n\,\sigma^2_{\alpha\beta} + ng\,\sigma^2_\beta$
Genotype × site	(g - 1)(s - 1)	$\sigma^2_\epsilon + n\,\sigma^2_{\alpha\beta}$
Residual or error	gs(n - 1)	σ^2_ϵ

To obtain the expected values of the mean squares, it is assumed that α_i are distributed as IID(0, σ^2_α), β_j are distributed as IID(0, σ^2_β), δ_{ij} are distributed as IID(0, σ^2_δ), and ϵ_{ijh} are distributed as IID(0, σ^2_ϵ), where IID means identically and independently distributed. Note that normality is not required for estimation of the variance components but only for testing of hypotheses. The requirement that a simple random sample of genotypes and sites is made validates the assumptions. There is general agreement on the above. Here the interaction means square would be used to test hypotheses for zero variance components for genotypes and for sites.

MIXED EFFECTS

As indicated for the above two situations, there is general agreement of understanding. However, when it comes to the mixed model where one of the categories, say sites (environments) is fixed and the other category is random, say genotypes, there has been considerable discusion in the literature (e. g., Federer, 1955; Cornfield and Tukey, 1956; Nelder, 1998; Voss, 1999, Basford *et al.*, 2002, to name a few) about how to proceed in estimating variance components and tests of hypotheses.

Letting the number of one of the factors, say genotypes, go to infinity while retaining the s sites does not change the population structure described for the fixed effects case. For every one of the genotypes, there are s interaction terms. That is, for a given genotype, there is not a distribution of random interaction effects but only the s terms. Hence, to remain consistent with the definition of main effects and interactions for the fixed effects case, the sum of the interaction parameters for any randomly selected genotype *must* add to zero.

Using the constrained parameter model formulation for the linear model desribed above, the following assumptions are made:

(1) α_i are distributed as IID(0, σ^2_α)

(2) $E[\beta_j] = 0 = \sum_{j=1}^{s} \beta_j / s$

(3) $\delta_{ij}|j$ (|j means given j) are distributed as IID(0, σ^2_δ)

(4) $E[\delta_{ij} \mid i] = 0$ which is equivalent to $\sum_{j=1}^{s} \delta_{ij} / s = 0$

(5) ϵ_{ijh} are distributed as IID(0, σ^2_ϵ)

The expected values of the mean squares in an analysis of variance are:

Source of variation	Degrees of freedom	Mean square expected value
Genotypes	g - 1	$\sigma_\epsilon^2 + \text{ns}\,\sigma_\alpha^2$
Sites	s - 1	$\sigma_\epsilon^2 + \text{ns}\,\sigma_{\alpha\beta}^2/(\text{s} - 1) + \text{f}(\beta_j)$
Genotype × site	(g - 1)(s - 1)	$\sigma_\epsilon^2 + \text{ns}\,\sigma_{\alpha\beta}^2/(\text{s} - 1)$
Residual or error	gs(n - 1)	σ_ϵ^2

Using the unconstrained parameter model formulation for the above linear model, the following assumptions are made:

(a) α_i are IID(0, σ_α^2)
(b) δ_{ij} are IID(0, σ_δ^2)
(c) ϵ_{ijh} are IID(0, σ_ϵ^2)

These are the same conditions as for the infinite model formulation. The expected values of the mean squares in an analysis of variance are:

Source of variation	Degrees of freedom	Mean square expected value
Genotypes	g - 1	$\sigma_\epsilon^2 + \text{n}\,\sigma_{\alpha\beta}^2 + \text{ns}\,\sigma_\alpha^2$
Sites	s - 1	$\sigma_\epsilon^2 + \text{n}\,\sigma_{\alpha\beta}^2 + \text{f}(\beta_j)$
Genotype × site	(g - 1)(s - 1)	$\sigma_\epsilon^2 + \text{n}\,\sigma_{\alpha\beta}^2$
Residual or error	gs(n - 1)	σ_ϵ^2

Assumption (b) is inconsistent with the fixed model case. This inconsistency plus the failure to consider the sampling procedure that when a random genotype is selected *all* interaction terms associated with this genotype are also selected invalidates this mixed model formulation. This model also assumes that $E[\delta_{ij}|i] = E[\delta_{ij}|j] = 0$. The use of $E[\delta_{ij}|i]$ implies that the sum of the interaction effects over sites is zero as all of them are present. The definition of effects also appears to have been changed from the fixed model formulation.

COMMENTS

It appears that the unconstrained parameter model formulation is based upon invalid assumptions. Of course, if one is willing to accept the assumptions, then the results in the literature (e. g., Nelder and Lane, 1995; Nelder, 1998) are correct. Failure to consider the definition of effects, the sampling plan, a consistent approach for all three formulations, and the use of invalid assumptions has led to the development of the unconstrained parameter model formulation. The question as to why the mixed model assumption (b) should be the same as for the infinite model case has not been answered

by the proponents of the unconstrained formulation when the population structure and sampling plan are considered.

Quite different estimates of variance components are possible using the two different parameter model formulations. Such differences can greatly affect the estimates of genetic correlations and heritability estimates as shown by Basford *et al.* (2002). They also show that either formulation results in the same estimate of genetic advance. A data set is included to demonstrate the effects of the two formulations on genetic correlations, heritability, and gentic advance.

lITERATURE CITED

Basford, K. E., W. T. Federer, and I. H. Delacy (2002). Mixed model formulations for multi-environment trials. BU-1602-M in the Technical Report Series of the Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York (submitted for publication).

Cornfield, J. and J. W. Tukey (1956). Average values of mean squares in factorials. Annals of Mathematical Statistics 27:907-949.

Federer, W. T. (1955). *Experimental Design--Theory and Application.* Macmillan, New York.

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

Nelder, J. A. (1998). The great mixed model muddle is alive and flourishig, alas! Food Quality and Preference 9(3):157-159.

Nelder, J. A. and P. W. Lane (1995). The computer analysis of factorial experiments: In Memorium--Frank Yates. The American Statistician 49:383-385.

Voss, D. T. (1999). Resolving the mixed models controversy. The American Statistician 53:352-356.

Zelen, M. and W. T. Federer (1965). Applications of the calculus for factorial arrangeemts: III. Analysis of factorials with unequal numbers of observations. Sankhya, Series A 27:383-400.