An Overview of Variance Component Estimation

by

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, N.Y., U.S.A., 14853

BU-1231-M April 1994

AN OVERVIEW OF VARIANCE COMPONENT ESTIMATION

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, N.Y., U.S.A.

BU-1231-M April 1994

ABSTRACT

Variance components estimation originated with estimating error variance in analysis of variance by equating error mean square to its expected value. This equating procedure was then extended to random effects models, first for balanced data (for which minimum variance properties were subsequently established) and later for unbalanced data. Unfortunately, this ANOVA methodology yields no optimum properties (other than unbiasedness) for estimation from unbalanced data. Today it is being replaced by maximum likelihood (ML) and restricted maximum likelihood (REML) based on normality assumptions and involving nonlinear equations that have to be solved numerically. There is also minimum norm quadratic unbiased estimation (MINQUE) which is closely related to REML but with fewer advantages.

ORIGINS

The analysis of variance table, as developed by R.A. Fisher in the 1920s, is a well-established summary of the arithmetic for testing hypotheses using ratios of mean squares which, under normality assumptions, have F-distributions (so named by Snedecor in honor of Fisher). This arithmetic was initially designed for what are now called fixed effects models, for which the F-statistics are suited to testing hypotheses that the effects of levels of a factor are all equal: e.g., in a randomized complete block experiment involving t treatments for testing if all t treatment effects (on the response variable) are equal.

An inherent part of Fisher's procedures was that of estimating the error variance. This was (and still is) done by equating the error mean square to its expected value, which is the error variance. In this way the error variance is estimated by the mean square for error. Thus with MSE being the mean square for error and E(MSE) its expected value, we have

$$E(MSE) = \sigma_e^2$$

yielding (1)

$$\hat{\sigma}_e^2 = MSE\ ,$$

where σ_e^2 is the error variance and $\hat{\sigma}_e^2$ is its estimate.

Traditional analysis of variance was developed without the use of models and model equations such as

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + e_{ijk}$$

which we use today. And in those models, in analysis of variance situations, we think of the e_{ijk} as being random error terms, and symbols like μ, α_i, β_j and γ_{ij} are unknown fixed constants, i.e., fixed effects; the model is called a fixed effects model. It has only one variance, the error variance. But in what are called variance components models some or all of the symbols like α_i, β_j and γ_{ij} represent random variables, with variances. Thus variance components models involve more than one variance. Moreover, since (on assuming, which we do, that all covariances among those random variables are zero) those variances add to the variance of the response variable, e.g., $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\beta^2 + \sigma_\gamma^2 + \sigma_e^2$, they are called variance components.

The history of estimating these variance components begins with extending the estimation idea involved in $E(MSE) = \sigma_e^2$ yielding $MSE = \hat{\sigma}_e^2$, as in (1). That idea is simply one of equating a mean square to its expected value and calling the result an estimate. The extension is no more than applying this to other mean squares of the analysis of variance.

BALANCED DATA: ANOVA ESTIMATION

Consider, for example, the completely randomized design (or 1-way classification) of a groups and n observations in each. The usual model equation for y_{ij}, the j'th observation in the i'th group, is

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad (2)$$

for i = 1, 2, $\cdots$, a and j = 1, 2, $\cdots$, n. With μ representing an overall mean, α_i the effect of the observation being in the i'th group, and e_{ij} being the random error term, the usual conditions for the variance components form of this model, i.e., when α_i is a random variable, are as follows:

$$
\begin{aligned}
&E(\alpha_i) = 0 && E(e_{ij}) = 0 && \forall \text{ i and j},\\
&\text{var}(\alpha_i) = \sigma_\alpha^2 && \text{var}(e_{ij}) = \sigma_e^2 && \forall \text{ i and j},\\
&\text{cov}(\alpha_i, \alpha_{i'}) = 0 \ \forall\, i \neq i' && \text{cov}(\alpha_i, e_{i'j'}) = 0 && \forall \text{ i, i' and j},
\end{aligned}
\tag{3}
$$

and

$$\text{cov}(e_{ij}, e_{i'j'}) = 0 \text{ except when } i = i' \text{ and } j = j'.$$

Notice two things. First, that (2) is the model equation; but it is (2) *and* (3) that is the model. Equation (3) is where assumed properties of elements of the model equation are specified. Second, that there is no assumption of normality in the model. That can be introduced later when needed for considering properties of estimators or for maximum likelihood estimation.

Now, with this model we have the simple between- and within-groups sums of squares summarized in Table 1, where $\bar{y}_{i\cdot} = \sum_{j=1}^{n} y_{ij}/n$ and $\bar{y}_{\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{n} y_{ij}/an$.

TABLE 1. Analysis of variance for a 1-way classification of n observations in each of a groups

Source of Variation	Degrees of Freedom	Sum of Squares	Mean Square
Between groups	a − 1	$SSA = n\sum_{i=1}^{a}\left(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot}\right)^2$	$MSA = SSA/(a-1)$
Within groups	a(n − 1)	$SSE = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{i\cdot}\right)^2$	$MSE = SSE/a(n-1)$
Total (c.f.m.)	an − 1	$SST_m = \sum_{i=1}^{a}\sum_{j=1}^{n}\left(y_{ij} - \bar{y}_{\cdot\cdot}\right)^2$	$MSE = SSE/a(n-1)$

In the fixed effects form of the model equation (2), the ratio MSA/MSE is, under normality, used for testing the hypothesis H: all α_i equal, i.e., $H: \alpha_1 = \alpha_2 = \cdots = \alpha_a$. But in the random effects form of (2), with the conditions (3) applying, we use MSA and MSE to estimate σ_α^2 and σ_e^2 by finding the expected values of MSA and MSE, using those conditions (3). Then we extend to MSA and MSE the procedure of "equating mean squares to their expected values" originated in (1). Thus, using (3) gives

$$E(MSA) = n\sigma_\alpha^2 + \sigma_e^2$$

and

$$E(MSE) = \sigma_e^2$$

and so we take

$$n\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2 = MSA$$

$$\hat{\sigma}_e^2 = MSE$$

which yield

$$\hat{\sigma}_e^2 = MSE \quad \text{and} \quad \hat{\sigma}_\alpha^2 = (MSA - MSE)/n\ . \tag{4}$$

This method of estimating variance components was extended in the 1930s and '40s to numerous different experiment designs; Daniels (1939), Tippett (1931) and Anderson and Bancroft (1952) are just a few of the interesting references. It is now known as the ANOVA method of estimation and has, as shall be described, been applied to unbalanced (unequal-subclass-numbered) data as well as to balanced data from well-planned experiments. In those situations it has some advantages: it is easy to calculate; it is easy to understand; it has very few basic assumptions, e.g., those of (3); the resulting estimators are unbiased, and they have the nice property of being minimum variance quadratic unbiased (Graybill and Hultquist, 1961) and under normality assumptions on the errors and on the random effects [still under conditions like those of (3)] these ANOVA estimators are minimum variance unbiased (Graybill, 1954, and Graybill and Wortham, 1956). These are all attractive properties. But there are also some unattractive features: estimates can be negative (which is embarrassing, a negative estimate of a positive parameter); and although under normality their sampling variances and unbiased estimates thereof can be derived, analytic forms of their distribution cannot be derived. Despite this, ANOVA estimators of variance components are widely used when data are balanced; and wisely so.

UNBALANCED DATA: ANOVA ESTIMATION METHODOLOGY

Some of the earliest users of variance components were geneticists with their particular interest in heritability defined as $4\sigma_G^2 \Big/ \left(\sigma_G^2 + \sigma_E^2\right)$, in which subscripts G and E indicate genetic and environmental variance. This was the context that motivated the landmark Henderson (1953) paper in *Biometrics*, dealing with the estimation from unbalanced data of variance components in mixed

models – those having both fixed and random effects. An example of this would be estimating dose effects and clinic variance in a nationwide clinical trial of several dose rates of a drug in clinics all over the country. An example in dairy breeding would be estimating herd-year effects and sire variance in the milk production records of cows, sired by many different bulls, milked in numerous herds over a period of years. In each of these examples it would be quite usual to have a large number of cells in the data grid that had no data in them. It was just such a situation that Henderson was interested in for estimating variance components. And his paper described three ways of doing this, which have come to be well known as Henderson's Methods 1, 2 and 3. They are, in fact, three different ways of selecting a set of sums of squares (or mean squares) for using in the algorithm "equate observed mean squares to their expected values". Henderson suggested using this principle on sums of squares which are either (i) analogous to analysis-of-variance sums of squares for balanced data or (ii) are adaptations of the former or (iii) come from fitting sub-models of the model being used for the data. And he carefully described these three different classes of sums of squares and illustrated them with a small numerical example.

Involving, as they do, the principle of "equate sums of squares to their expected values", these methods are nothing more than extensions of that ANOVA estimation principle from balanced data to unbalanced data. And it was a major step forward, especially for practical application. Method 1, for instance, has always been readily computable, even before computers came on the scene. And Methods 2 and 3, which in those pre-computer days were very impractical computationally, have now also become computable, even for large data sets.

Nevertheless, there are some difficulties with these methods. First, Method 1 should not be used with mixed models and Method 2 cannot be used with models that have interactions between fixed effects factors and random effects factors. Second, even when used outside these limitations, the Henderson Methods yield estimators that have no useful properties other than unbiasedness. Sampling variances (under normality) have been derived for a number of special cases; they are quadratic forms in the unknown variance components, with coefficients that are horribly complicated functions of the values of n_{ij}, the number of observations in the (i, j) cell. For example, in the 1-way classification

with unbalanced data the variance of the estimator of σ_α^2 is

$$v\left(\hat{\sigma}_\alpha^2\right) = \frac{2\Sigma n_i}{(\Sigma n_i)^2 - \Sigma n_i^2}\left[\frac{\Sigma n_i(\Sigma n_i - 1)(a-1)}{(\Sigma n_i - a)\left[(\Sigma n_i)^2 - \Sigma n_i^2\right]}\sigma_e^4 + 2\sigma_e^2\sigma_\alpha^2 + \frac{(\Sigma n_i)^2\Sigma n_i^2 + (\Sigma n_i^2)^2 - 2\Sigma n_i \Sigma n_i^3}{\Sigma n_i\left[(\Sigma n_i)^2 - \Sigma n_i^2\right]}\sigma_\alpha^4\right].$$

The complicated nature of the sampling variances (and that just displayed is by far the simplest) only aggravates the further result that each variance component estimator has no known analytic function for its probability density function (except for $\hat{\sigma}_e^2$, which is based on χ^2). As a result of all this there is no analytical way of comparing estimators from the three different Henderson methods. And, even though the suggestion is sometimes made that Method 3 is the best, there is truly no foundation for such a statement. Indeed, it begs the question "Which form of Method 3?", because Method 3 has the awkward feature that for many models there is no unique way of using Method 3. There can be several ways, and there are no analytical results that permit one to settle on an optimal use of Method 3.

All of these difficulties spring from the fact that the Henderson methods are just a collection of ways in which the ANOVA method of estimating variance components can be applied to unbalanced data (to any data, in fact). All that need be done, when there are r variance components to be estimated, is to choose r quadratic functions of the data that have expected values that are linear in the variance components. Denote those quadratics by a vector q, and the variance components by σ^2. Then the ANOVA method is no more than using

$$E(q) = M\sigma^2$$

to give

$$q = M\hat{\sigma}^2$$

and thus

$$\hat{\sigma}^2 = M^{-1}q\,.$$

This always gives unbiased estimates, because

$$E(\hat{\sigma}^2) = E(M^{-1}q) = M^{-1}E(q) = M^{-1}M\sigma^2 = \sigma^2\,.$$

But what other properties does it give? None. Other than demanding that the elements of E(q) be

linear in the elements of σ^2 (which leads to $\hat{\sigma}^2$ being unbiased), the method is based upon no other conditions. This means that other than unbiasedness the method has no built-in conditions – let alone any that lead to optimal properties for the resulting estimators. Indeed, the method is not even confined to using mean squares: any quadratic forms of the data will do (provided only that their expected values are linearly independent linear combinations of variance components). For example, in a 1-way classification something as wildly unorthodox as

$$\mathbf{q} = \begin{bmatrix} (y_{1,2} - y_{4,3})^2 \\ (y_{2,7} - 2y_{2,9} + y_{3,6})^2 \end{bmatrix}$$

could be used. It has

$$\mathrm{E}(\mathbf{q}) = \begin{bmatrix} 2(\sigma_\alpha^2 + \sigma_e^2) \\ 2\sigma_\alpha^2 + 6\sigma_e^2 \end{bmatrix}$$

and so estimators are given by

$$\begin{bmatrix} 2\hat{\sigma}_\alpha^2 + 2\hat{\sigma}_e^2 \\ 2\hat{\sigma}_\alpha^2 + 6\hat{\sigma}_e^2 \end{bmatrix} = \mathbf{q}\,,$$

namely

$$\hat{\sigma}_\alpha^2 = \tfrac{1}{4}\left[3(y_{1,2} - y_{4,3})^2 - (y_{2,7} - 2y_{2,9} + y_{3,6})^2\right]$$

and

$$\hat{\sigma}_e^2 = \tfrac{1}{4}\left[(y_{2,7} - 2y_{2,9} + y_{3,6})^2 - (y_{1,2} - y_{4,3})^2\right].$$

Clearly, from a commonsense point of view, these estimators are ridiculous. Nevertheless, as an application of the ANOVA method of estimation they are perfectly legitimate. But, as is obvious, unbiasedness is the only known property of the estimators. And, as we have discussed in our book (Searle, Casella and McCulloch, 1962, Sec. 5.2c), this is not necessarily an important or attractive feature for variance components estimation to have.

Thus it is that the ANOVA method of estimating variance components contains no procedures at all for indicating what quadratic forms to use as elements of $\mathbf{q}$ in $\mathrm{E}(\mathbf{q}) = \mathbf{M}\sigma^2$; indeed not even an indication of how many to use. It has been assumed that $\mathbf{M}$ is square and nonsingular, in order to

have $\hat{\sigma}^2 = \mathbf{M}^{-1}\mathbf{q}$. But even that is not dictated by the method. It is perfectly possible to choose more quadratic forms as elements of q than there are variance components as elements of σ^2. Then M would be rectangular, the equations $\mathbf{M}\hat{\sigma}^2 = \mathbf{q}$ would not have to be consistent, but provided M had full column rank a "least squares" solution of $\mathbf{M}\hat{\sigma}^2 = \mathbf{q}$ could be used; i.e., $\hat{\sigma}^2 = (\mathbf{M}'\mathbf{M})^{-1}\mathbf{M}'\mathbf{q}$; or if M did not have full column rank then $(\mathbf{M}'\mathbf{M})^{-}\mathbf{M}'\mathbf{q}$ would be a solution, with $(\mathbf{M}'\mathbf{M})^{-}$ being any generalized inverse of M'M.

In addition to this lack of uniqueness of the ANOVA method, it has other serious weaknesses: first, it can yield the embarrassment of negative estimates. Second, even under normality assumptions, ANOVA estimators have distributions that are unknown; and sampling variances (which are quadratic functions of the unknown components) involve very complicated functions of the numbers of observations in the subclasses of the data. Analytic comparison of different applications of the ANOVA method is therefore impossible. And arithmetic comparison cannot be satisfactorily designed to produce informative results.

It is within this environment of deficiencies of the ANOVA method that maximum likelihood estimation is coming to be the preferred method, certainly for unbalanced data and for many (if not all) cases of balanced data.

MAXIMUM LIKELIHOOD

The method of maximum likelihood estimation, developed by R.A. Fisher in the 1920s (Fisher, 1925) seems to have been first applied to the estimation of variance components by Crump (1947, 1951). In this and almost all subsequent presentation of this topic, normality is assumed for the error terms and all the random effects, normality with zero means, homogeneous variance of all random effects pertaining to each factor, and all covariances zero. Within this framework, Herbach (1959) gave careful attention, for balanced data, to the need for maximum likelihood estimators (MLEs) to be non-negative, this being essential because ML theory demands that maximization be over the parameter space. In describing ML for variance components it is therefore essential to distinguish between solutions of the ML equations and estimators. They are not necessarily the same. Nor are they always the same as ANOVA estimators. For example, in the 1-way classification with balanced data, using

the mean squares MSA and MSE of Table 1, the ANOVA estimators are, as in (4),

$$\hat{\sigma}^2_\alpha = \tfrac{1}{n}(\text{MSA} - \text{MSE}) \quad \text{and} \quad \hat{\sigma}^2_e = \text{MSE}\,.$$

The solutions to the maximum likelihood (under normality) equations are

$$\dot{\sigma}^2_\alpha = \tfrac{1}{n}\left[(1 - \tfrac{1}{a})\text{MSA} - \text{MSE}\right] \quad \text{and} \quad \dot{\sigma}^2_e = \text{MSE}\,.$$

Because $\dot{\sigma}^2_\alpha$ can be negative, it cannot be an MLE. ML theory indicates that whenever $\dot{\sigma}^2_\alpha$ is negative, the MLE of σ^2_α is 0; and this changes the MLE of σ^2_e to be not $\dot{\sigma}^2_e = \text{MSE}$ but SST_m/an. Thus the MLEs are as follows:

$$\tilde{\sigma}^2_\alpha = \dot{\sigma}^2_\alpha \quad \text{and} \quad \tilde{\sigma}^2_e = \dot{\sigma}^2_e \qquad \text{if} \quad \dot{\sigma}^2_\alpha \geq 0$$

$$\tilde{\sigma}^2_\alpha = 0 \quad \text{and} \quad \tilde{\sigma}^2_e = \text{SST}_m/\text{an} \qquad \text{if} \quad \dot{\sigma}^2_\alpha < 0\,.$$

With balanced data, there are at least three other cases (2-way nested, random model and the 2-way crossed, mixed model, with and without interaction) where this avoidance of negative estimators is relatively easy (see Searle *et al.*, 1992, Section 4.7b). But with a model no more complicated than the 2-way crossed classification, random model, the ML equations for balanced data are easily written down (Miller, 1977), but have no closed-form solution (*loc. cit.* Section 4.7d). They have to be solved numerically.

Maximum likelihood estimators, in general, have a number of nice properties, most of which are, unfortunately, asymptotic in nature. In the case of a simple sample of n observation those asymptotics are quite straightforward: they are based on limits as $n \to \infty$. However, when data have multiple classifications, defining the limiting conditions is not so easy as just $n \to \infty$. For example, in a 2-way layout of rows and columns, limiting conditions have to involve not just the total number of observations but also the numbers of rows and columns and numbers of observations therein and in the cells. Just exactly what is meant by "in the limit" in such situations is discussed in Hartley and Rao (1967) and Miller (1977).

Accepting that "in the limit" can be prescribed, the two most useful properties of MLEs are that in the limit such estimators are normally distributed and have sampling variances given by the inverse of the information matrix. These are two exceedingly useful properties, compared to those of ANOVA estimators – even though they are sustainable only in the limit. This begs the question, of course:

What does "in the limit" mean in practical terms? It certainly does not mean estimating variance components from a Latin Square of order 5! (I was once asked how to do that when the Latin Square had some missing cells!) But I think it does mean that a data set of 500,000 observations with 400 rows and 1000 columns is probably close enough to being at the limit for one to be content with using the asymptotic results. And data sets as large as this are not uncommon in some of the genetic uses of variance components. Even for data sets more modest in size I would prefer ML to ANOVA, for the prime reason of knowing what is involved, of not being bedeviled by wondering if some ANOVA application would be better than the one I've used, and of having no doubt – save for the "in the limit" meaning – about how to calculate sampling variances.

The preceding description of MLE and its properties apply quite generally to any mixed model for which the model equation can be written as

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}\,,$$

where $\mathbf{y}$ is the vector of data, β is a vector of fixed effects, $\mathbf{u}$ is a vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are the model matrices (usually incidence matrices of zeros and ones) corresponding to β and $\mathbf{u}$, respectively, and $\mathbf{e}$ is a vector of error terms. The vector $\mathbf{u}$ can be thought of as being partitioned into r subvectors:

$$\mathbf{u}' = \left[\mathbf{u}_1' \quad \mathbf{u}_2' \quad \cdots \quad \mathbf{u}_i' \quad \cdots \quad \mathbf{u}_r'\right].$$

Each $\mathbf{u}_i$ will have elements that are all the q_i random effects of some random factor (main effects or interaction factor) that is in the data. Then, under the usual assumption of homogeneity of variances and zero covariances, the variance-covariance matrix of $\mathbf{u}$, denoted by $\mathbf{D}$, is the block diagonal matrix

$$\text{var}(\mathbf{u}) = \mathbf{D} = \begin{bmatrix} \sigma_1^2\mathbf{I}_{q_1} & 0 & \cdots & & & 0 \\ 0 & \sigma_2^2\mathbf{I}_{q_2} & & & & 0 \\ \vdots & & \ddots & & & \vdots \\ 0 & & & \sigma_i^2\mathbf{I}_{q_i} & & 0 \\ \vdots & & & & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & \cdots & \sigma_r^2\mathbf{I}_{q_r} \end{bmatrix},$$

which can be written more compactly in various forms as

$$\mathbf{D} = \bigoplus_{i=1}^{r} \sigma_i^2\mathbf{I}_{q_i} = \mathrm{D}\left[\sigma_1^2\mathbf{I}_{q_1} \ \cdots \ \sigma_r^2\mathbf{I}_{q_r}\right] = \left\{ \underset{d}{\ } \sigma_i^2\mathbf{I}_{q_i} \right\}_{i=1}^{r} .$$

Then, with $\mathbf{Z}$ of $\mathbf{Zu}$ partitioned conformably with $\mathbf{u}$,

$$\text{var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZDZ}' + \sigma_e^2\mathbf{I}_N = \sum_{i=1}^{r} \mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 + \sigma_e\mathbf{I}_N .$$

Furthermore, $\mathbf{e}$ can be absorbed into the $\mathbf{Zu}$ notation by defining

$$\mathbf{e} = \mathbf{u}_0, \quad \mathbf{I}_N = \mathbf{Z}_0 \quad \text{and} \quad \mathrm{N} = q_0 ,$$

to give

$$\mathbf{V} = \sum_{i=0}^{r} \mathbf{Z}_i\mathbf{Z}_i'\sigma_i^2 . \tag{5}$$

It was this notational formulation that contributed, in my opinion, to Hartley and J.N.K. Rao (1967) being so successful in deriving ML equations – those obtained by differentiating the likelihood (under normality) with respect to the parameters, the σ^2s and the elements of β, and equating the resulting expressions to zero. The ML solutions, $\dot{\beta}$ and $\dot{\mathbf{V}}$ [being $\mathbf{V}$ of (5) with each σ^2 replaced by its $\dot{\sigma}^2$] are given by the equations

$$\mathbf{X}'\dot{\mathbf{V}}^{-1}\mathbf{X}\dot{\beta} = \mathbf{X}\dot{\mathbf{V}}^{-1}\mathbf{y} \tag{6}$$

and

$$\text{trace}(\dot{\mathbf{V}}^{-1}\mathbf{Z}_i\mathbf{Z}_i') = \mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y} \qquad \text{for} \quad \mathrm{i} = 0, \cdots, \mathrm{r} , \tag{7}$$

where

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \tag{8}$$

with $\dot{\mathbf{P}}$ being $\mathbf{P}$ with $\mathbf{V}$ replaced by $\dot{\mathbf{V}}$. Clearly, equations (6) and (7) have no closed form solution, and so they have to be solved numerically, usually by iteration. This involves some thorny questions in the

realm of numerical analysis.

(i) What method of iteration should be used?

(ii) There is at least one other form of (7): does the method of iteration depend on the form of the equations?

(iii) Which form of the equations is best?

(iv) Is convergence of the iteration always assured?

(v) If convergence is achieved, is it at a global maximum?

(vi) Do starting values affect where convergence is achieved?

(vii) If so, is there some set of starting values that always lead to convergence at the global maximum?

(viii) The matrix $\mathbf{V}$ is, by definition, non-negative definite, and usually positive definite. What is to be done numerically if at an intermediate step of the iteration the true value of $\dot{\mathbf{V}}$ is not non-negative definite?

(ix) How will non-negativity conditions be taken into account?

Thus it is that programming the numerical solution of the ML equations is no job for an amateur. Indeed, one well may wonder how satisfactorily these questions have been handled by professional programmers.

Despite these numerical difficulties, I strongly adhere to the belief that maximum likelihood methodology is the correct estimation technique to use. Twenty-five years ago computer programs could barely handle the task, except for very small (and therefore effectively useless) data sets. But today computing power is continually increasing, and numerical techniques for sparse matrices and inverting matrices are always on the improve. Thus, whilst the computing power needed for *every* kind of possible data set is probably not available (nor may ever be), the power needed for a wide range of realistic data sets certainly is available. And it should be used: e.g., SAS and BMDP, to mention two of the widest-ranging computing sources.

Those packages also calculate the asymptotic sampling variance-covariance matrices for the ML estimators, using those estimators $\tilde{\sigma}^2$ in place of the population values σ^2 in the following expressions. In the limit, these variance results are

$$\begin{aligned} \text{var}(\tilde{\beta}) &\to (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\ \text{cov}(\tilde{\beta}, \tilde{\sigma}^2) &\to 0 \\ \text{var}(\tilde{\sigma}^2) &\to 2\left[\left\{_m \ \text{trace}(\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j')\right\}_{i=0\ j=0}^{r\quad r}\right]^{-1} . \end{aligned} \tag{9}$$

$\tilde{\beta}$ and $\tilde{\sigma}^2$ are the vectors of ML estimators; and the matrix to be inverted in $\text{var}(\tilde{\sigma}^2)$ is a symmetric matrix of order $r+1$, its (i, j)'th element being the trace of the product matrix

$$\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j' , \qquad \text{for} \quad i, j = 0, 1, \cdots, r.$$

RESTRICTED MAXIMUM LIKELIHOOD

An adaptation of ML estimation of variance components is to maximize only that part of the likelihood which is location invariant. It was first proposed for unbalanced data by Patterson and Thompson (1971) and has come to be known as restricted (in Europe, marginal) maximum likelihood (REML). It can also be defined as maximizing the likelihood (under normality, of course) of $N - r_X$ linearly independent linear combinations of the data, such that those combinations contain no fixed effects. This is tantamount to maximizing the likelihood of $\mathbf{K}'\mathbf{y}$ where $\mathbf{K}'\mathbf{X} = \mathbf{0}$ and $\mathbf{K}'$ has full row rank. A Bayes description is also available (Searle *et al.*, 1992, pp. 303 and 323).

Compared to ML estimation REML has two features that some people feel sufficiently strongly about to always favor REML over ML. I'm not convinced. One feature is that for balanced data the REML solutions are identical to ANOVA estimators. The other is that REML estimators implicitly take into account the degrees of freedom associated with the fixed effects in the model. The simplest example of this is in estimating the variance from a simple sample of multinormal (μ, σ^2) data, $x_1, x_2, \cdots, x_n$. The ML estimator is $\Sigma_i(x_i - \bar{x})^2/n$, whereas the REML estimator is $\Sigma_i(x_i - \bar{x})^2/(n-1)$. A mildly negative feature of REML compared to ML is that REML contains nothing about estimating the fixed effects whereas ML does. In practice, just as the ML estimator of β is given by (6), $\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X}\tilde{\beta} = \mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y}$, which is simply generalized least squares estimation, GLSE(β), with the ML estimator $\tilde{\mathbf{V}}$ in place of $\mathbf{V}$, one would of course after estimating $\mathbf{V}$ by REML use it in place of $\mathbf{V}$ in

GLSE(β). But that is not part of REML. And then, of course, the Bayes justification for REML will, for some people, be a strong enough argument for preferring REML to ML.

None of these reasons are strong enough for me to be vociferous over preferring REML to ML. In practice I would probably calculate both – and hope there was no huge difference between two sets of estimates.

Equations for calculating REML solutions are

$$\text{trace}(\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i') = \mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y} \qquad \text{for} \quad i = 0, \cdots, r\,. \tag{10}$$

These are simply equations of (7) of ML but with the $\dot{\mathbf{V}}^{-1}$ on the left-hand side replaced by $\dot{\mathbf{P}}$. Likewise, in the asymptotic sampling dispersion matrix of (9) for ML, the $\mathbf{V}^{-1}$ is replaced for REML by $\mathbf{P}$.

Note from (8) that $\mathbf{PVP} = \mathbf{P}$. Therefore

$$\text{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i') = \text{tr}(\mathbf{PVP}\mathbf{Z}_i\mathbf{Z}_i') = \text{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{PV}) = \textstyle\sum_j \text{tr}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\sigma_j^2\,,$$

so that the REML equations (10) can also be written as

$$\textstyle\sum_j \text{trace}(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j')\sigma_j^2 = \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \qquad \text{for} \quad i = 0, 1, \cdots, r\,. \tag{11}$$

MINIMUM NORM ESTIMATION

A series of papers starting with Rao (1970) proposed a method of estimation that has come to be known as MINQUE, minimum norm quadratic unbiased estimation. It does not require normality, and it is developed by wanting to estimate a linear function of variance components, $\mathbf{p}'\boldsymbol{\sigma}^2$ say, by a quadratic function of the data, $\mathbf{y}'\mathbf{A}\mathbf{y}$ say. The symmetric matrix $\mathbf{A}$ is to be chosen so that $\mathbf{y}'\mathbf{A}\mathbf{y}$ minimizes a certain Euclidean norm (akin to a generalized variance) and is unbiased for $\mathbf{p}'\boldsymbol{\sigma}^2$. The result of this (see, for example, Searle *et al.*, 1992, Section 11.3d) is that the equations to be solved for the MINQUE estimators are

$$\sum_{j=0}^{r} \text{trace}(\mathbf{P}_0\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_0\mathbf{Z}_j\mathbf{Z}_j')\hat{\sigma}_i^2 = \mathbf{y}'\mathbf{P}_0\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_0\mathbf{y} \qquad \text{for} \quad i = 0, 1, \cdots, r\,. \tag{12}$$

Two features of these equations are noteworthy. First is $\mathbf{P}_0$: it is $\mathbf{P}$ of (8), which involves $\mathbf{V}$, but with a set of pre-assigned values $\sigma_{i,0}^2$ used in place of σ_i^2 in $\mathbf{V}$. Denote those values by $\boldsymbol{\sigma}_0^2$. Then replacing $\boldsymbol{\sigma}^2$ by $\boldsymbol{\sigma}_0^2$ in $\mathbf{V}$ gives $\mathbf{V}_0$; and replacing $\mathbf{V}$ in $\mathbf{P}$ gives $\mathbf{P}_0$, used in the MINQUE equations (12). Thus the solution of those equations, which are the MINQUE estimators, depends upon what values are chosen as elements of $\boldsymbol{\sigma}_0^2$. Different values of $\boldsymbol{\sigma}_0^2$s will, from the same data set, yield different estimates.

Personally, I cannot live with that as an estimation procedure.

A second feature of the MINQUE equations is that they only have to be solved as they stand. That is nice, because it is easy: they are just a set of linear equations in the wanted solutions. But that means there can be negative solutions; no provision is embodied in the method for avoiding them.

REML, MINQUE and I-MINQUE

MINQUE is a non-iterative procedure: just solve (12). But (12) involves the pre-assigned value σ_0^2 used in $\mathbf{P}_0$. And the REML equations (11) are the same as the MINQUE equations (12) except REML has $\mathbf{P}$ where MINQUE has $\mathbf{P}_0$. But the REML equations are used iteratively and this has to begin with a starting value for σ^2. Hence if that starting value is σ_0^2 the first iteration of REML will yield the same solution as MINQUE. Thus, in general

a MINQUE estimator = a first iterate solution of REML .

Consider the MINQUE equations (12) again. They are based on σ_0^2, and yield a solution, $\tilde{\sigma}_1^2$, say. Suppose that solution is now used in place of σ_0^2 in (12), and those equations are solved, yielding $\tilde{\sigma}_2^2$. Continuing with this iterative procedure in (12) we have what is called Iterative MINQUE, or I-MINQUE, and if continued until convergence, then providing the starting value σ_0^2 for I-MINQUE is the same as for iterating REML, we have

I-MINQUE estimates = REML solutions .

This conclusion is of some importance. MINQUE and I-MINQUE do not require normality assumptions. Yet Brown (1976) has shown that I-MINQUE estimators are asymptotically normally distributed. Therefore, from (13) so are REML solutions. And this is so even when REML calculations (which have been derived on the basis of normality assumptions) are used on data that do not necessarily satisfy the normality assumptions. This seems to be a very strong point in favor of using REML.

MINQUE(O) and MINVAR

MINQUE(O) is the special form of MINQUE wherein the preassigned value chosen for σ_0^2 is 1.0 for σ_e^2 and zero for every other σ^2. This reduces $\mathbf{P}$ to be $\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'$ and Goodnight (1978) has shown that that makes the MINQUE equations especially easy to compute. But Quaas and Bolgiano (1979) found it to be a particularly inefficient method of estimation.

The development of MINQUE stems from choosing $\mathbf{A}$ of $\mathbf{y}'\mathbf{A}\mathbf{y}$ to minimizing a Euclidean norm based on σ_0^2. If one adapts that to minimizing the variance (under normality) of $\mathbf{y}'\mathbf{A}\mathbf{y}$, without the need of any σ_0^2, one finishes up (Searle *et al.*, 1992, p. 394) with the same equations as for REML, namely (12). But since these equations can only be solved iteratively (or by some equivalent arithmetic method) the resulting solutions used as estimators are neither unbiased nor minimum variance.

References

Anderson, R. L. and Bancroft, T. A. (1952). *Statistical Theory in Research.* McGraw-Hill, New York.

Brown, K. G. (1976). Asymptotic behavior of MINQUE-like estimators of variance components. *Annals of Statistics* **73**, 141-146.

Crump, S. L. (1947). The estimation of variance in multiple classifications. Ph.D. Thesis, Iowa State University, Ames, Iowa.

Crump, S. L. (1951). The present status of variance components analysis. *Biometrics* **7**, 1-16.

Daniels, H. E. (1939). The estimation of components of variance. *J. Roy. Statist. Soc. Suppl.* **6**, 186-197.

Fisher, R. A. (1925). Theory of statistical estimation. *Proc. of the Cambridge Philosophical Soc.* **22**, 700-725.

Goodnight, J. H. (1978). Computing MIVQUEO estimates of variance components. *Technical Report R-105*, SAS Institute, Cary, North Carolina.

Graybill, F. A. (1954). On quadratic estimates of variance components. *Ann. Math. Statist.* **25**, 367-372.

Graybill, F. A. and Hultquist, R. A. (1961). Theorems concerning Eisenhart's Model II. *Ann. Math. Statist.* **32**, 261-269.

Graybill, F. A. and Wortham, A. W. (1956). A note on uniformly best unbiased estimators for variance components. *J. Amer. Statist. Assoc.* **51**, 266-268.

Hartley, H. O. and Rao, J. N. K. (1967). Maximum likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93-108.

Henderson, C. R. (1953). Estimation of variance and covariance components. *Biometrics* **9**, 226-252.

Herbach, L. H. (1959). Properties of Model II type analysis of variance tests, A: optimum nature of the F-test for Model II in the balanced case. *Ann. Math. Statist.* **30**, 939-959.

Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *Ann. Statist.* **5**, 746-762.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545-554.

Quaas, R. L. and Bolgiano, D. C. (1979). Sampling variances of the MIVQUE and Method 3 estimators of the nine components of variance. In *Variance Components and Animal Breeding*, Animal Science Department, Cornell University, pp. 99-106.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. J.A.S.A. **65**, 161-172.

Searle, S. R., Casella, G. and McCulloch, C. E. (1992). *Variance Components*, Wiley & Sons, N. Y.

Tippett, L. H. C. (1931). *The Methods of Statistics*, 1st ed. Williams and Norgate, London.

Paper number BU-677 in the Biometrics Unit, Cornell University, Ithaca, N. Y.