# MIXED MODELS AND UNBALANCED DATA: WHEREFROM, WHEREAT AND WHERETO?

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, N.Y.

BU-507 October 1987
BU-927-M* March 1987

---

*An invited paper for the "Conference on the Analysis of the Unbalanced Mixed Model" at Gainesville, Florida, April 6-10, 1987.

# MIXED MODELS AND UNBALANCED DATA: WHEREFROM, WHEREAT AND WHERETO?

Shayle R. Searle

Biometrics Unit, Cornell University, Ithaca, N.Y.

*Key Words: ANOVA, history, MINQUE, ML, prediction, REML, simulation, variance components*

## ABSTRACT

A brief history of the early years (1820-1947) of random effects models and the estimation of variance components is followed by a personal evaluation of ML, REML and MINQUE estimation. A method is suggested for combining ML estimators obtained from subsets of a large data set, and comments are made on the need for simulation studies to assess the degree of approximation in using asymptotic properties of ML-type estimators as if they were exact for finite-sized unbalanced data sets.

## I. WHEREFROM

### 1.1. From 1820 to 1947

Clear specification of mixed models, as involving a mixture of fixed effects and random effects, began with Eisenhart (1947). But the concepts of fixed effects and of random effects originated more than a hundred years earlier than that. We begin with a brief account of that early work, drawing heavily on Scheffé (1956) and Anderson (1978) to do so.

Estimation of fixed effects essentially began with Legendre (1806) and Gauss (1809), the well-known independent fathers of the method of least squares. [Plackett (1972) has an intriguing discussion of their relative rights to priority.] As noted by

Scheffé, (1956), an interesting aspect of those two early nineteenth century papers is that they both appear in books on astronomy. What is even more interesting is that the first appearance of variance components is also in astronomy books, Airy (1861) and Chauvenet (1863). Scheffé (1956) refers to Airy (1861, especially Part IV) as being a "very explicit use of a variance components model for one-way layout ... with all the subscript notation necessary for clarity." It is noteworthy (as remarked upon by Anderson, 1978) that in this earliest known use of a variance component model there is provision for unbalanced data - unequal numbers of telescopic observations from night to night on the same phenomenon of interest. Despite Airy's now-accepted originality he did not see himself in this light for, in the preface of his book, quoted by Anderson (1978), he writes "No novelty, I believe, of fundamental character, will be found in these pages."; and "... the work has been written without reference to or distinct recollection of any other treatise (excepting only Laplace's *Théorie des Probabilitiés*) ... ." As Anderson (1978) says, this, insofar as endeavors to establish the exact origin of the components of variance concept are concerned, is an unfortunate style of writing.

The second use of a random effects model appears, according to Scheffé, to be Chauvenet (1863, Vol. II, Articles 163 and 164). Although he did not write model equations, he certainly implied a 1-way classification random model in which, using today's notation, he derived the variance of $\bar{y}_{..} = \sum_{i=1}^{a} \sum_{j=1}^{n} y_{ij}/an$ as

$$v(\bar{y}_{..}) = (\sigma_a^2 + \sigma_e^2/n)/a \quad .$$

Chauvenet suggests that there is little practical advantage in having n greater than 5, and refers to Bessel (1820) for this idea; but Scheffé says that that reference is wrong, although it "does contain a formula for the probable error of a sum of independent random variables which could be the basis for such a

conclusion. Probably Bessel made the remark elsewhere." If so, the question is "Where?" and might that other reference be an early germ of an idea about optimal design? Preitschopf (1987) has searched the 1820-1826 and 1828 yearbooks containing Bessel (1920) and finds not even a hint about not having "n geater than 5"; the only pertinent remark is on page 166 of the 1823 yearbook which has with $x_i$ being the "random error of part i, l=1,···,n, total error is $y = \sqrt{x_1^2+\cdots+x_n^2}$ .

More modern beginnings of variance components are in Fisher's (1918) paper on quantitative genetics wherein he made [adapting freely from Anderson (1978)]

(i) Inceptive use of the terms "variance" and "analysis of the variance."

(ii) Implicit, but unmistakable, use of variance components models.

(iii) Definitive ascription of percentages of a total variance to constituent causes; e.g., that dominance deviations accounted for 21% of the total variance in human stature.

Following that genetics paper, Fisher's book (1925; Sec. 40) made a major contribution to variance component models through initiating what has come to be known as the analysis of variance (ANOVA) method of estimation: equate sums of squares from an analysis of variance to their expected values and thereby obtain a set of equations that are linear in the variance components to be estimated. This idea arose from using an analysis of variance to estimate an intra-class correlation, which he wrote as $\rho = A/(A+B)$ and described as

> "... merely the fraction of the total variance due to that cause which observations in the same class have in common. The value of B may be estimated directly, for variation within each class is due to this cause alone, ... ."

From this he was led to expressions which, in to-day's notation for the 1-way classification random model with balanced data, are

$$E(SSE) = E \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 = a(n-1)\sigma_e^2 \qquad (1)$$

and

$$E(SSA) = E \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2 = (a-1)(n\sigma_\alpha^2 + \sigma_e^2) \ . \qquad (2)$$

From these the estimation equations are taken as

$$SSE = a(n-1)\hat{\sigma}_e^2 \quad \text{and} \quad SSA = (a-1)(n\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2) \qquad (3)$$

and so

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

These are, for a 1-way classification random model, the ANOVA estimators of the variance components from balanced data.

Had Fisher foreseen even a small part of the methodology for estimating variance components that he thus heralded he might have given more attention to this topic. But he did not. Section 40 of Fisher (1925) remains quite unchanged in subsequent editions, even after variance component principles were well established. Furthermore, even when he extended the analysis of variance to a 1-way classification model with unbalanced data, to a 2-factor model with interaction and to more complex settings, he did not address the estimation of variance components in those settings.

Following Fisher's work of 1918 and 1925 came Tippett (1931) who clarified and extended the ANOVA method of estimation and in his second edition (1937) displayed some explicit estimators. He also addressed (1931, Sec. 10.11) the problem of considering "the best way of distributing the observations between and within groups" for a 1-way model, as had Chauvenet (1863) and perhaps Bessel (1820). This was followed by Yates and Zacopanay's (1935) comprehensive study on sampling for yield in cereal experiments,

which dealt with designs corresponding to higher-order models. In the same vein, Neyman *et al.* (1935) considered the efficiency of randomized blocks and Latin square designs, and in doing so made extensive use of linear models (including mixed models). Maybe this is the first recognizable appearance of a mixed model.

Although Fisher (1935) used the term "components of variation" in an acrimonious review of Neyman *et al.* (1935), who themselves had used the phrase "error components", the first apparent use of "components of variance" is Daniels (1939):

> "... it is natural to use the analysis of variance ... to arrive at estimates of the components of total variance assignable to each factor. The components of variance can then be used to establish an efficient sampling scheme ... ."

It seems that the papers by Daniels (1939) and by Winsor and Clark (1940) can well be considered as the solid beginnings of the work on variance components of the last fifty years or so. Each paper, independently, derives (1) and (2) that Fisher (1925) has, using the "expected value" concept that is implicit in Fisher (1925). Daniels mentions Tippett (1931) but not Fisher, whereas Winsor and Clark describe their derivation as being "a straightforward extension of the suggestions of R. A. Fisher in his *Statistical Methods for Research Workers* [Sec. 40]." Presumably this is the seventh edition, published in 1938, in which Sec. 40 deals with intraclass correlation, exactly as it does, unchanged, in both the first edition of 1925 and the twelfth edition of 1954. Yet, although Fisher (1925) has the idea of taking expected values he had not there specifically formulated it using the E operator as do Daniels, and Winsor and Clark. Their papers were soon followed by Snedecor (1940), his third edition, which has virtually no reference to variance components at all. Page 205 contains discussion of estimating the intraclass correlation as A/(A + B), just as does Fisher (1938). The nearest thing to characterizing A as a variance component is

the description that "A is the same for all ··· samples – it is the common element, analogous to covariance." And that is, of course, the case: the covariance between two observations in the same class is $\sigma^2_\alpha$.

In describing a 2-factor no-interaction situation Jackson (1939) writes that one factor is "a measure of the trial effect," and the other is "a measure of the individual effect." This seems to be the first occurrence of the word "effect" in what is now its customary usage in linear models. Jackson also described his model as having one factor random and one non-random – a crystal-clear specification of a mixed model, although not called such at that time. In this connection it is surprising that it was eight more years before someone, Eisenhart (1947), saw the need for carefully describing and distinguishing between what we now know as fixed, random and mixed models.

Although unbalanced data were provided for in that very early description of a random model in Airy (1861), they received little attention for another eighty years. Tippett (1931, Sec. 6.5) makes a passing comment that for unbalanced data certain relations [e.g., (1) and (2)] "do not hold, for in summing the squares of the deviations of the group means from the grand mean, each group has been given a different weight", the number of observations in the group. Nevertheless, in Section 9.6, he provides an approximation for calculating an intraclass correlation coefficient from such data. In contrast, Snedecor (1934, Sec. 31) simply states: "The direct relation between analysis of variance and intraclass correlation disappears if there are unequal frequencies in the classes." Even in his third edition (Snedecor, 1940, Example 12.21, p. 205), in referring to the unbalanced data of Table 10.8, he asks "Why can't you calculate intraclass correlation accurately" for such data? Needless to say, that example does not appear in the sixth edition, Snedecor and Cochran (1967). The reason is, of course, that the now well-known results

$$E(SSA) = E\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \left( n_. - \sum_{i=1}^{a} n_i^2/n_. \right)\sigma_\alpha^2 + (a-1)\sigma_e^2$$

and (5)

$$E(SSE) = E\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_{i.})^2 = (n_. - a)\sigma_e^2 ,$$

had been derived, independently, by Cochran (1939) and Winsor and Clark (1940). Soon after, Ganguli (1941) specified the details of ANOVA estimation of variance components from unbalanced data in fully nested models, no matter how many nestings there are.

## 1.2 Forty years, 1947-87

After what can now be viewed as the foundation writings of the 30's and 40's, interest in variance component estimation expanded at an ever-increasing rate. Much of the activity continued to be motivated, as had the early publications, by practical problems. Statisticians with minimal concern for data showed no interest whatever. Geneticists, particularly (perhaps fired by Fisher's 1918 paper), quickly became users of variance components models in applications to humans, dairy cows, wheat, beef cattle, corn, pigs and poultry - to name but a few. Almost all these applications involved unbalanced data.

This is no place for a historical survey, if for no other reason than most of those attending this conference are familiar with the details of the last forty years. Moreover, the excellent survey by Khuri and Sahai (1985) is where the interested reader will find a full account. So just a brief and somewhat personal outline will be given of the major advances made in the matter of methods of estimation.

For estimating variance components from unbalanced data the landmark paper is undoubtedly Henderson (1953). In that paper the ANOVA method of estimation, based on equating analysis of variance sums of squares to their expected values, was extended for unbalanced data to equating a wide variety of quadratic forms

(not all of them sums of squares) to their expected values. Then followed a period of trying to evaluate those methods mostly through deriving expressions, under normality assumptions, for sampling variances of the resulting estimates, e.g., Crump (1951), Searle (1956, 1958, 1961), Mahamunulu (1963), Low (1964), Hirotsu (1966), Blischke (1968), and Rohde and Tallis (1969). In every case the results are, of course, quadratic functions of the unknown variance components; but the coefficients of the squares and products of those components are such hopelessly intractable functions of the numbers of observations in the cells of the data (see Searle, 1971, Chapter 11) that it is impossible to make analytic comparisons either of different estimation methods, or of the effects of different degrees of data unbalancedness on any one method of estimation. This absence of tractable criteria on which judgement can be made as to which application of the ANOVA method has any optimal features thus became very frustrating. For balanced data this frustration does not exist: Graybill and his co-workers (e.g., with Wortham, 1956, and with Hultquist, 1961) had established minimum variance properties.

But for unbalanced data that frustration persisted. Distinctions between the three Henderson methods could be made on the basis of computing requirements, and, after proofs given in Searle (1968), on the basis of the nature of the model being used: for mixed models, Method I is not suitable and neither is Method II if the model is to have interactions between fixed and random effects. But, no matter what form of the ANOVA method is used, the resulting estimators are unbiased - and no other optimal properties have been established. Of course, the extensive work of R. L. Anderson and colleagues (e.g., Anderson, 1975, Anderson and Crump, 1967, Bainbridge, 1963) gives some indication of which of some applications of the ANOVA methods may be better than others for quite a variety of special designs planned for estimating variance components. But it can be difficult to extrapolate from those designs to situations often

found with survey-style data; for example, to breeding data from farm livestock, where there may be several hundreds of levels of a random factor, and some thousands of cells in the data but with only 20-30% of them actually containing data.

This unavailability of methods for estimating variance components from unbalanced data that have optimality criteria was radically changed during the 1967-72 years when three different (but related) methods were developed that came with built-in optimality criteria. The first was Hartley and Rao's (1967) paper presenting maximum likelihood (ML) estimation, based on normality assumptions being made of the data. The second was restricted maximum likelihood (REML) estimation initiated for balanced data by Anderson and Bancroft (1952) and Thompson (1962), and extended by Patterson and Thompson (1971) to block designs and thence to unbalanced data generally. The third was minimum norm quadratic unbiased estimation (MINQUE) coming from both LaMotte (1970, 1973) and Rao (1970, 1971a, b, 1972). And currently there is in development a method designed by Hocking and colleagues; it is based on treating variance components as covariances and estimating them from utilizing all the available cross-product covariance estimates that are appropriate. For balanced data this method is shown in Hocking *et al.* (1986) to be equivalent to ANOVA estimation; and for unbalanced data I anticipate hearing more about this method later to-day.

## 2. WHERAT?

It is convenient to have before us the matrix formulation of a mixed model, a formulation which today is considered quite standard but which when first proposed in Hartley and Rao (1967) was deemed to be (and still is) very ingenious.

### 2.1 A general mixed model

Let $\mathbf{y}$ of order $N\times 1$ be the vector of data. The model equation for $\mathbf{y}$ is taken as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} \qquad (6)$$

where $\boldsymbol{\beta}$ is a k×1 vector of fixed effects, including coefficients of covariables whenever covariables are to be part of the model. $\mathbf{X}$ is a known N×k matrix corresponding to the occurrence in the data of the elements in $\boldsymbol{\beta}$. When covariates are part of the model $\mathbf{X}$ will contain columns of observed values of the covariates; otherwise $\mathbf{X}$ is usually an incidence matrix. $\mathbf{Z}$ is usually an incidence matrix, but does not have to be.

The form of $\mathbf{u}$ is important. It is partitionable as

$$\mathbf{u} = [\mathbf{u}_o' \; \mathbf{u}_1' \; \ldots \; \mathbf{u}_r']' \qquad (7)$$

where

$$\mathbf{u}_o = \mathbf{e}$$

is the N×1 vector of error terms in the model; and for the r random factors in the model, $\mathbf{u}_i$ of order $q_i$ for $i = 1, \ldots r$, is the vector of $q_i$ effects occurring in the data, corresponding to the i'th random factor in the model, be it a main effects factor, a nested factor or an interaction factor. The distributional properties attributed to $\mathbf{u}$ that are customary for random effects are

$$E(\mathbf{u}_i) = \mathbf{0} \quad \forall \quad i, \text{ which includes } E(\mathbf{u}_o) = E(\mathbf{e}) = \mathbf{0} \qquad (8)$$

and

$$\text{var}(\mathbf{u}_i) = \sigma_i^2 \mathbf{I}_{q_i} \text{ and } \text{cov}(\mathbf{u}_i, \mathbf{u}_{i'}') = \mathbf{0} \quad \forall \quad i \neq i'. \qquad (9)$$

In the case of $\mathbf{u}_o = \mathbf{e}$ we have $q_o = N$ and

$$\text{var}(\mathbf{u}_o) = \text{var}(\mathbf{e}) = \sigma_o^2 \mathbf{I}_N = \sigma_e^2 \mathbf{I}_N \text{ and } \text{cov}(\mathbf{u}_i, \mathbf{e}') = \mathbf{0} \qquad (10)$$

for $i = 1, 2, \ldots, r$.

In (6) the coefficient matrix $\mathbf{Z}$ is partitioned conformably with the product $\mathbf{Zu}$ for $\mathbf{u}$ partitioned as in (7). Thus

$$\mathbf{Z} = [\mathbf{Z}_o, \mathbf{Z}_1, \ldots, \mathbf{Z}_r] \text{ with } \mathbf{Z}_o = \mathbf{I}_N \qquad (11)$$

corresponding to $\mathbf{u}_o = \mathbf{e}$. And thus (6) is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{u}_i \quad . \qquad (12)$$

The variance-covariance features of $\mathbf{y}$ are determined from (12). First

$$\mathbf{D} = \mathrm{var}(\mathbf{u}) = \begin{bmatrix} \sigma_0^2 \mathbf{I}_{q_0} & \cdot & \cdots & \cdot \\ \cdot & \sigma_1^2 \mathbf{I}_{q_1} & & \cdot \\ \vdots & & \ddots & \vdots \\ \cdot & \cdot & \cdots & \sigma_r^2 \mathbf{I}_{q_r} \end{bmatrix} \tag{13}$$

is block diagonal, with the blocks being the scalar matrices $\sigma_i^2 \mathbf{I}_{q_i}$ for $i = 0,1,\ldots,r$. Thus $\mathbf{D}$ is diagonal. Then from (12) and (13)

$$\mathbf{V} = \mathrm{var}(\mathbf{y}) = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{D}_i \mathbf{Z}_i' = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2 \tag{14}$$

and

$$\mathbf{C} = \mathrm{cov}(\mathbf{y},\mathbf{u}') = \sum_{i=1}^{r} \mathbf{Z}_i \sigma_i^2 \quad . \tag{15}$$

## 2.2. The mixed model problem

One often hears the phrase "the mixed model problem" - frequently voiced in awesome, sometimes even fearsome, tones, almost as if the problem was impossibly difficult to describe and even more difficult to solve. Surely the problem is not difficult to state - and we already have some answers that go a long way towards being satisfactory.

Confining attention to point and/or interval estimation, the three aspects of a mixed model that demand attention are estimation of the fixed effects, of the random effects and of the variance components. There are therefore three parts to "the mixed model problem." We consider each in turn.

### a. Estimation of fixed effects

To estimate estimable functions of elements of $\boldsymbol{\beta}$ in (6) we consider estimation of $\mathbf{X}\boldsymbol{\beta}$. Every element, and any linear combination of elements, of $\mathbf{X}\boldsymbol{\beta}$ is estimable. The ordinary least squares estimator of $\mathbf{X}\boldsymbol{\beta}$ is $\mathbf{X}(\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{y}$; but this takes no account

whatever of the random effects in the model, and so it is of little interest as an estimator. Limiting ourselves to cases when $\mathbf{V}$ is non-singular, we then have the best linear unbiased estimator of $\mathbf{X\beta}$ as

$$\mathbf{X\beta}^{o} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \quad . \tag{16}$$

Although this is confined to non-singular $\mathbf{V}$, that is not very restrictive in actual practice. In most applications $\mathbf{V}$ will not be singular: e.g., whenever $\mathbf{Z}_o = \mathbf{I}$ and $\sigma_o^2 > 0$.

As an estimator of $\mathbf{X\beta}$, the $\mathbf{X\beta}^{o}$ in (16) has many good properties: it is not only best linear unbiased, but it is also ML under normality. However, it has an obvious deficiency: $\mathbf{V}$ is usually unknown. The "obvious" thing to do is to estimate

$$\boldsymbol{\sigma}^2 = [\sigma_o^2 \quad \sigma_1^2 \quad \ldots \quad \sigma_r^2]'$$

in some way, call the estimator $\hat{\boldsymbol{\sigma}}^2$, use it in place of $\boldsymbol{\sigma}^2$ in $\mathbf{V} = \Sigma_i \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2$ of (14) to get

$$\hat{\mathbf{V}} = \Sigma_i \mathbf{Z}_i \mathbf{Z}_i' \hat{\sigma}_i^2 \tag{17}$$

and then calculate

$$\mathbf{X\beta}^{o}_{\hat{\mathbf{V}}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \quad . \tag{18}$$

(18) is not a best linear unbiased estimator. But Kackar and Harville (1981, 1984) have shown that $\mathbf{X\beta}^{o}_{\hat{\mathbf{V}}}$ is unbiased, and that its sampling variance can be calculated, "provided the data vector is symmetrically distributed about its expected value and provided the variance component estimators are translation-invariant and are even fucntions of the data vector. The standard procedures for estimating the variance components yield even, translation-invariant estimators."

An even nicer result concerns maximum likelihood. Denote the ML estimator of $\boldsymbol{\sigma}^2$ by $\tilde{\boldsymbol{\sigma}}^2$ [See (26), which follows.] Now replace $\boldsymbol{\sigma}^2$ by $\tilde{\boldsymbol{\sigma}}^2$ in (14) and (16) to yield $\tilde{\mathbf{V}}$ and

$$\mathbf{X\beta}^{o}_{\tilde{\mathbf{V}}} = \mathbf{X}(\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y} \quad . \tag{19}$$

Then (19) is the ML estimator of $\mathbf{X\beta}$.

Thus for this first part of the mixed model problem, we have two useful possibilities: (i) for $\hat{\sigma}^2$ being in the wide range of estimators described by Kackar and Harville use $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\hat{\mathbf{V}}}$ of (18); it is unbiased and its sampling dispersion matrix can be calculated. Unfortunately its distribution is unknown, so that exact interval estimates are unavailable (ii) use the ML estimator [see (26)] $\tilde{\sigma}^2$ to calculate $\tilde{\mathbf{V}}$ and hence $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\tilde{\mathbf{V}}}$ of (19); it is the ML estimator of $\mathbf{X}\boldsymbol{\beta}$ with all the usual attendant properties. And a third possibility is to (iii) use the REML estimator [see (27)] $\dot{\sigma}^2$ of $\sigma^2$ to calculate $\dot{\mathbf{V}}$ and then $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\dot{\mathbf{V}}}$. Properties of $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\dot{\mathbf{V}}}$ are unknown, but they are, hopefully, quite similar to those of $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\tilde{\mathbf{V}}}$.

A difficulty with (ii) and (iii) is that, for data that are not normally distributed, procedures for deriving ML and REML estimators of $\sigma^2$ have not been worked out. One possible course of action would be to try transforming the data in some manner that makes them normal or at least more nearly so than are crude data.

b. Prediction of random effects

A random effect that occurs in data is a realization of a random variable. But it is usually unobservable: e.g., the genetic make-up of the dairy cow Daisy, whose annual milk yield has been recorded; or the true intelligence of the schoolboy Tom Brown, for whom we have a score on an I.Q. test. Whilst we cannot measure these realizations, and hesitate to refer to estimating them (since estimation of random variables is counter-intuitive statistically), we can think of predicting these values – in the following sense. Consider two school children, Tom and Jane, who have taken three different versions of an I.Q. test, versions that have been touted as being equivalent. Tom scored 115, 117 and 122, with an average of 118. Jane scored 110 and 126 on the first two tests but missed taking the third. She, too, has an average of 118 (for her two tests). Thus Tom and Jane have the same average. But they do not necessarily have the same actual true, underlying, unobservable intelligence. (We may wish

to use their test average as our indicator of intelligence, but it is not necessarily the same as actual intelligence.) But what we might wish to do is to answer the question: from all school children who average 118 on the tests what is a good prediction of their intelligence? And in answering that question we will also be interested to see to what extent the answer takes into account the number of test scores that a person has.

This is the sense in which we speak of predicting a random effect. One of its early occurrences (Henderson, 1955) seems to have been as a classroom exercise used by A.M. Mood in the late 1940's at Iowa State University, and subsequently appearing in Mood (1950, p. 164, exercise 23), and in revised form in Mood and Graybill (1963, p. 195, exercise 32) and in Mood, Graybill and Boes (1974, p. 370, exercise 52). The 1950 version is as follows.

> "23. Suppose intelligence quotients for students in a particular age group are normally distributed about a mean of 100 with standard deviation 15. The I.Q., say $x_1$, of a particular student is to be estimated by a test on which he scores 130. It is further given that test scores are normally distributed about the true I.Q. as a mean with standard deviation 5. What is the maximum-likelihood estimate of the student's I.Q.? (The answer is not 130.)"

The final sentence is tantalizing. Overcoming its implied temptation can be achieved by modeling $y_{ij}$, the j'th test score for some i'th person, as

$$y_{ij} = \mu + u_i + e_{ij}$$

where $u_i$ is the person's true I.Q. and $e_{ij}$ is a residual error term. At first we think of $u_i$ as certainly being a fixed effect insofar as the particular person who has been labeled as the i'th person is concerned. But in thinking about people in general, that particular person is really just a random person: and $u_i$ is, accordingly, simply a realized (but unobservable) value of a random effect — the effect on test score of the intelligence

level of the i'th randomly chosen person. Therefore, we treat $u_i$ as random and have I.Q. and score, namely $u_i$ and $y_{ij}$, jointly distributed with bivariate normal density:

$$\begin{bmatrix} \text{I.Q.} \\ \text{Score} \end{bmatrix} = \begin{bmatrix} u_i \\ y_{ij} \end{bmatrix} \sim N \left[ \begin{pmatrix} 100 \\ 100 \end{pmatrix}, \begin{pmatrix} 15^2 & 5^2 \\ 15^2 & 15^2 + 5^2 \end{pmatrix} \right] .$$

Then, what we want from this is the maximum likelihood estimate of the conditional mean of the variable "$u_i$, given $y_{ij} = 130$", i.e., we want $E(u_i | y_{ij} = 130)$, which is

$$E(u_i | y_{ij}) = 130) = 100 + \frac{15^2}{15^2 + 5^2} (130 - 100) = 127 \neq 130 .$$

This is what is called the predicted value of $u_i$ (given that $y_{ij} = 130$).

There are many situations of this nature in the real world: of having a vector (or scalar value) of observations on some random variables from which we wish to predict the value of some other random variable or variables that cannot be observed. Biological examples include predicting the genetic merit of a dairy bull from the milk yields of his daughters' records, a practice which, as a basis for ranking bulls and selecting those of high genetic merit for widespread use in artificial insemination programs, has contributed greatly to the spectacular increase of per-cow milk production that has occurred over the last thirty years in many countries around the world. Other examples are those of predicting instrument bias in a micrometer selected randomly out of a munufacturer's lot, using measurements made on a number of objects; and of predicting peoples' intelligence from scores on a battery of tests.

So, in terms of the general model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu}$ of (6), the feature of interest is prediction of $\mathbf{u}$, where $\mathbf{y}$ and $\mathbf{u}$ are jointly distributed:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \sim \left[ \begin{pmatrix} \mathbf{X\beta} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{V} & \mathbf{C} \\ \mathbf{C}' & \mathbf{D} \end{pmatrix} \right] \qquad (20)$$

for $\mathbf{D}$, $\mathbf{V}$ and $\mathbf{C}$ of (13) - (15). Cochran (1951), Rao (1965, pages 79 and 220-222) and Searle (1974) all describe how, for (20) the best (i.e., minimum mean square) predictor for $\mathbf{u}$ is

$$\tilde{\mathbf{u}} = E(\mathbf{u}|\mathbf{y}) \quad . \tag{21}$$

And the best linear predictor (linear in elements of $\mathbf{y}$) is

$$\tilde{\mathbf{u}}_L = E(\mathbf{u}) + \mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{22}$$

for $\mathbf{X}\boldsymbol{\beta}$ of (6). These results, $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{u}}_L$, hold for any distribution (satisfying the usual regularity conditions) having finite first and second moments. Furthermore, when that distribution is the normal distribution we have (21) and (22) being equal; i.e., under normality

$$\tilde{\mathbf{u}} = \tilde{\mathbf{u}}_L = E(\mathbf{u}|\mathbf{y}) = E(\mathbf{u}) + \mathbf{C}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \quad . \tag{23}$$

Moreover, this expression can also be viewed as a Bayes estimator, with the prior on $\mathbf{u}$ being the normal density having mean $E(\mathbf{u})$ and variance-covariance matrix $\mathbf{D}$. The posterior density of $\mathbf{u}$ is then the density of $\mathbf{u}|\mathbf{y}$, which is normal with mean $E(\mathbf{u}|\mathbf{y})$ shown in (23). That mean can therefore be taken as a Bayes estimator, the mean of the posterior density of $\mathbf{u}$.

So here, for the second part of the "mixed model problem", we also have procedures that have a high degree of satisfaction - except that again they depend on knowing $\sigma^2$ and thus $\mathbf{V}$ and $\mathbf{C}$. Just as with using $\mathbf{X}\boldsymbol{\beta}\overset{o}{\hat{\mathbf{V}}}$, so here: replacing $\sigma^2$ in $\tilde{\mathbf{u}}_L$ by $\hat{\sigma}^2$ seems the "obvious" thing to do. Properties of the resulting expressions have been considered by Jeske and Harville (1986) -and we will no doubt hear more about this tomorrow. Although (23) can be calculated using $\hat{\mathbf{V}}$ as can their sampling variances, their distributions are not known, so giving rise to difficulties in interval estimation.

An extension of (22) given by Henderson (1963) is to the mixed model. He shows that the best linear predictor of $\mathbf{w} = \mathbf{T}'\mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ is

$$\tilde{w} = T'X\beta^{o} + E(u) + C'V^{-1}(y - X\beta^{o})$$

where $X\beta^{o} = X(X'V^{-1}X)X'V^{-1}y$ of (16). This is what animal breeders refer to as BLUP - best linear unbiased prediction. Again, it requires values of $\sigma^2$ for practical application.

c. Estimating variance components

It is clear from the preceding two subsections that estimation of fixed effects and prediction of random effects could be achieved with some degree of satisfaction if $\sigma^2$ could be estimated satisfactorily. To do this we have, at the present time, four main available options as a method of estimation: ANOVA, ML, REML and MINQUE; and along with the latter, two variants of it: MIVQUE, minimum variance quadratic unbiased and MIMSQE, minimum mean square quadratic unbiased. We confine attention to the four main methods.

ANOVA has already been discussed. Its lack of optimality criteria on which to pass judgement on the various forms of ANOVA is a serious deficiency. In many computing environments Henderson's Method I may be the only feasible form - or possibly Method II. And even if Method III is computationally feasible there is no unique application of it to models of two or more crossed factors. Therefore, except when limited computational facilities demand using Henderson's Method I, my recommendation is for abandonment of the ANOVA method of estimating variance components from unbalanced data.

ML, REML and MINQUE are all to be preferred over ANOVA because they have built-in optimality properties. To succinctly display these methods we use the notation of

$$\left\{ {}_{m}a_{ij} \right\}_{i,j=0}^{i,j=t} \quad \text{and} \quad \left\{ {}_{c}v_{i} \right\}_{i=0}^{i=t} \tag{24}$$

respectively, for a square matrix of order t + 1 and a column vector of order t + 1. Then, along with the model formulation given earlier, together with

$$P = V^{-1} - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}, \qquad (25)$$

these methods can be summarized as follows.

ML: Solve by iteration for $\tilde{\sigma}^2$:

$$\left\{ \operatorname{tr}\left(Z_i Z_i' \tilde{V}^{-1} Z_j Z_j' \tilde{V}^{-1}\right) \right\}_{i,j=0}^{i,j=r} \tilde{\sigma}^2 = \left\{ y'\tilde{P}Z_i Z_i'\tilde{P}y \right\}_{i=0}^{i=r} . \qquad (26)$$

REML: Solve by iteration for $\dot{\sigma}^2$:

$$\left\{ \operatorname{tr}\left(Z_i Z_i' \dot{P} Z_j Z_j' \dot{P}\right) \right\}_{i,j=0}^{i,j=r} \dot{\sigma}^2 = \left\{ y'\dot{P}Z_i Z_i'\dot{P}y \right\}_{i=0}^{i=r} . \qquad (27)$$

MINQUE: For a set of r+1 pre-assigned numerical values $w$ used in place of $\sigma^2$ in $V$ and then $P$, to yield $P_w$, solve (with no iteration needed)

$$\left\{ \operatorname{tr}\left(Z_i Z_i' P_w Z_j Z_j' P_w\right) \right\}_{i,j=0}^{i,j=r} \dot{\sigma}^2_{(w)} = \left\{ y'P_w Z_i Z_i' P_w y \right\}_{i=0}^{i=r} . \qquad (28)$$

A few comments about the equations are in order.

(i) Each set of equations has order r + 1, the number of variance components.

(ii) Each set of equations has on its left-hand side a matrix of elements that are each the trace of the product of six matrices: but each such trace can be expressed as tr($TT'$) for some $T$ and so can be calculated as the sum of squares of elements of $T$.

(iii) Each set of equations has a right-hand side that is a vector of quadratic forms in the observations.

(iv) The REML equations differ from the ML equations only in having the $P$-matrix where the ML equations have the $V^{-1}$-matrix.

(v) Both the ML and the REML equations are very non-linear in the sought-after estimator - $\hat{\sigma}^2$ in ML and $\dot{\sigma}^2$ in REML. These equations therefore have to be solved iteratively; and this raises a number of questions that are in the realm of numerical

analysis. Does the choice of starting value affect the attained value at convergence? Does that attained value always correspond to a global maximum of the likelihood that is being maximized – or does it sometimes correspond to a local maximum? And if so, when? Since at each successive round of the iteration a numerical matrix is being used for $\mathbf{V}$ how does one ensure that it is always positive-definite? And if it is not, what are the consequences? If, after some iteration, the numerical value $\hat{\sigma}^2_t$ to be given to $\sigma^2_t$ is negative, what action is to be taken? Were that $\hat{\sigma}^2_t$ to be at the last round of iteration then it would, in accord with the principles of maximum likelihood estimation of a variance (e.g., Herbach, 1959), be changed to zero. The model would then be altered correspondingly, and the remaining variance components re-estimated. But suppose $\hat{\sigma}^2_t < 0$ occurs before convergence; and suppose it is changed to zero and the model altered, and iteration continues using that altered model. If, as a result of some numerical quirk of those data, continuing with that unchanged negative $\hat{\sigma}^2_t$ had, at a subsequent round of iteration, led to a positive $\hat{\sigma}^2_t$, then changing it to zero and altering the model is presumably the wrong thing to do. How is this situation provided for in solving the ML and REML equations? Does any present computing package do this?

(vi) Solving the MINQUE equations requires no iteration. Once the pre-assigned numerical values that are to be elements of $\mathbf{w}$ have been decided on as replacements for elements of $\boldsymbol{\sigma}^2$ in $\mathbf{P}$ to yield $\mathbf{P}_w$ – once this has been done, the MINQUE equations are just a simple set of linear equations in the unknown variance components estimates.

(vii) The MINQUE equations are exactly the same as the REML equations except with the $\dot{\mathbf{P}}$-matrix in REML replaced by $\mathbf{P}_w$ for MINQUE. Thus, as first observed by Hocking and Kutner (1975),

$$\text{a MINQUE} = \text{a first iterate of REML.} \qquad (28)$$

(viii) Solutions to the MINQUE equations depend on the pre-assigned $\mathbf{w}$.

## 3. WHERETO?

### 3.1 ML, REML or MINQUE?

The over-riding question is: which of the ML, REML and MINQUE methods of estimation should be used in analyzing unbalanced data? This is of particular importance when analyzing the very large data sets that often arise in situations where mixed models are appropriate. Certainly, my first conclusion is to not use MINQUE. Reasons for this are three-fold. First, and foremost is that for different values of the pre-assigned vector $\mathbf{w}$, it gives different values of the estimated $\sigma^2$. This means that for N people all having the same data, but each person using a different $\mathbf{w}$, there will be N different estimated $\sigma^2$-vectors. Somehow I do not see this as being an acceptable feature of an estimation procedure to investigators who have large data sets. Any kind of argument about making use of prior knowledge in some manner, in the form of pre-assigned weights that play a role akin to the unknown variances, is unlikely to sit well with someone who has 50,000 observations from which to estimate two or three variance components. Second, MINQUE can produce negative estimates – which are not attractive. Third, having obtained a MINQUE estimator of $\sigma^2$ it would be very natural for any investigator to contemplate using it as a new $\mathbf{w}$ – and in this way be led to iterating on MINQUE. This is known as the I-MINQUE method of estimation. Under normality assumptions it is the same as REML (Hocking and Kutner, 1975) - and without those assumptions Brown (1976) has shown that I-MINQUE has a limiting distribution that is normal. Hence my conclusion is to favour REML over MINQUE.

Then comes the question "ML or REML?" This is difficult to answer. One favoured characteristic of REML is that with balanced data the REML equations reduce to the same equations as

are used in ANOVA estimation - and ANOVA estimators are known to have the attractive minimum variance properties established by Graybill and colleagues. But, of course, whereas ANOVA estimators may well be the same as REML solutions with balanced data, REML solutions are not necessarily REML estimators; they are, only if they are positive. For example, in the 1-way classification, random model, with balanced data, of a classes and n observations in each, suppose the mean squares between and within groups are denoted by MSA and MSE respectively. Then the REML solutions are $\dot{\sigma}^2_\alpha = (MSA - MSE)/n$ and $\dot{\sigma}^2_e = MSE$. But only when $\dot{\sigma}^2_\alpha > 0$ are $\dot{\sigma}^2_\alpha$ and $\dot{\sigma}^2_e$ the REML estimators. When $\dot{\sigma}^2_\alpha \leq 0$ the REML estimator of $\sigma^2_\alpha$ is zero and that of $\sigma^2_e$ is $SST/(an - 1)$ for $SST = SSE + SSA$ of (1) and (2) – see Thompson, (1962).

It is sometimes said that REML gives unbiased estimators. This is not so. It is true that the expected value of the right-hand side of the REML equations in (27) can be written in the same form as the left-hand side of those equations. But this does not imply unbiasedness. The non-negativity of any form of maximum likelihood estimators (as distinct from solutions of maximum likelihood equations) has to be taken into account. For example, with balanced data from a 1-way classification random model, the solutions of the REML equations are unbiased, but the two-pronged procedure just described for adapting those solutions so as to get REML estimators gives an estimator of $\sigma^2_\alpha$ that is clearly upwardly biased, as can also be shown for the estimator of $\sigma^2_e$.

Another favoured feature of REML is that it takes account of degrees of freedom used for estimating fixed effects; e.g., in a simple sample of $x_i \sim$ i.i.d.$N(\mu,\sigma^2)$ the REML estimate of $\sigma^2$ is $\Sigma_i(x_i - \bar{x})^2/(n - 1)$ whereas the ML estimator is $\Sigma_i(x_i - \bar{x})^2/n$. In this simple case REML is unbiased – but that is not the general rule. And, of course, nothing is unbiased after iteration, neither in ML or REML.

One of the merits of ML over REML is that the ML procedure includes providing an ML estimator for the fixed effects, namely $\mathbf{X}\underset{\sim}{\beta}^{\circ}_{\mathbf{V}}$ of (19), with all the attendant properties of ML. The REML method provides no such estimator, although intuitively one would be inclined to use $\mathbf{X}\beta^{\circ}_{\dot{\mathbf{V}}}$ based on (18), where $\dot{\sigma}^2$, the REML estimator, leads to $\dot{\mathbf{V}}$, which in turn is used as $\hat{\mathbf{V}}$ in (18).

In contrast to ANOVA estimation, both ML and REML are methods of estimating variance components from unbalanced data that can be used with any mixed or random model. They accommodate crossed and/or nested classifications, with or without covariates, and they are based on the maximum likelihood principle of estimation that has a long history of well-established, large-sample properties. The applications (26) and (27) do, of course, depend on normality assumptions for the data. Nevertheless, there are also more difficult situations that are not so easily accommodated, where $\mathbf{V}$ may have a special structure that cannot be expressed as $\Sigma \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2$. Jennrich and Schlucter (1986) and Berk (1987) discuss such possiblities as pertaining to unbalanced data arising from repeated measures situations. In those cases, just as with (26) and (27), there are the computational difficulties associated with solving nonlinear equations by iteration. But these, I believe, are difficulties that are progressively being overcome. For example, these difficulties used to also include problems of sheer size; e.g., the enormous amount of time and money needed for the inverting of matrices of large dimension, of order 5,000, say. The advent of supercomputers will see this problem of size becoming of less and less an impediment to calculating ML and REML estimates.

It is difficult to be anything but inconclusive about which of ML and REML is the preferred method. ML has the merit of simultaneously providing estimators of both the fixed effects and the variance components - and that is appealing. On the other hand, REML has the attraction of providing variance components

estimators that are unaffected by the fixed effects. The dependence of both ML and REML on normality assumptions may, for some data, be bothersome; and if that were to be felt overpowering then using the REML procedure and calling it I-MINQUE would be acceptable. That requires no normality assumptions on the data, but nevertheless yields estimators that have asymptotic normality properties.

## 3.2 Dividing data into subsets

When local computing facilities are too limited to feasibly cope with doing the ML or REML calculations for a very large data set, one possible way of obtaining ML estimates from the whole data set is to divide the data into portions, from each of which ML estimates can be calculated, and then combine those estimates. Since estimators from one data subset are not necessarily independent of those from another, simple averaging of the subset estimates can be improved upon. A method for doing this has been developed in Babb (1986). It is as equally applicable to REML as it is to ML, so we describe it in terms of just ML.

Begin by dividing the data into t sets, the p'th of which has model equations

$$\mathbf{y}_p = \mathbf{X}_p\boldsymbol{\beta}_p + \sum_{i=0}^{r} \mathbf{Z}_{ip}\mathbf{u}_{ip} \tag{29}$$

$$= \mathbf{X}_p\boldsymbol{\beta}_p + \sum_{i=0}^{r} \dot{\mathbf{Z}}_{ip}\mathbf{u}_i \quad . \tag{30}$$

Model (29) is the same as (6) but with subscript p for the p'th data set and with $\mathbf{u}_{ip}$ (possibly) different from $\mathbf{u}_i$ because $\mathbf{u}_{ip}$ has as elements only those of $\mathbf{u}_i$ that are actually in data set p. Thus $\dot{\mathbf{Z}}_{ip}$ of (30) has null columns corresponding to those elements of $\mathbf{u}_i$ that are not in $\mathbf{u}_{ip}$; and the non-null columns are columns of $\mathbf{Z}_{ip}$. From either (29) or (30, $\sigma^2$ can be estimated by ML using (26) [or by REML using (27)] to yield $\tilde{\sigma}^2_p$ for $p = 1,2,\cdots,t$ based on

$$\mathbf{V}_{pp} = \text{var}(\mathbf{y}_p) = \sum_{i=0}^{r} \dot{\mathbf{Z}}_{ip}\dot{\mathbf{Z}}'_{ip}\sigma_i^2 = \sum_{i=0}^{r} \mathbf{Z}_{ip}\mathbf{Z}'_{ip}\sigma_i^2 \quad . \tag{31}$$

Estimates from the t data sets are combined by a weighted least squares procedure using an approximate variance-covariance matrix of $[\sigma_1^2,\cdots,\sigma_t^2]$, developed from

$$\mathbf{V}_{pq} = \mathrm{cov}(\mathbf{y}_p,\mathbf{y}_q') = \sum_{i=1}^{r} \dot{\mathbf{Z}}_{ip}\dot{\mathbf{Z}}_{iq}'\sigma_i^2 \quad , \text{ for } p \neq q \ . \tag{32}$$

(32) contains no $\sigma_o^2 = \sigma_e^2$ because no two data sets have error terms in common. Also, whereas (31) can be expressed equivalently in terms of either $\dot{\mathbf{Z}}_{ip}$ or $\mathbf{Z}_{ip}$, (32) cannot; it is available only through using the $\dot{\mathbf{Z}}$-matrices of (30). Furthermore, in any particular case, several terms in (32) may be zero and, indeed, for many (p,q) pairs, (32) itself may be zero. For example, when the random factors are a 2-way crossed classification, the data can (and probably will) be divided into sets so that each cell of the 2-way classification is wholly in a single data set. Then (32) will not involve the interaction variance component, for the same kind of reason that it never contains $\sigma_e^2$; and only for data sets that have levels of the $\alpha$-factor in common can (30) contain $\sigma_\alpha^2$; and so on.

Define

$$\mathbf{A} = \left\{ \underset{m}{\mathrm{tr}}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1}) \right\}_{i,j=0}^{r} \quad \text{and} \quad \mathbf{f} = \left\{ \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} \right\}_{i=1}^{r} \ , \tag{33}$$

similar to the matrix and vector on the left and right sides, respectively, of (26). Denote $\mathbf{A}$ and $\mathbf{f}$ for data set p by $\mathbf{A}_p$ and $\mathbf{f}_p$. Then the covariance matrix of $\mathbf{f}_p$ and $\mathbf{f}_q$ is

$$\mathbf{G}_{pq} = \mathrm{cov}(\mathbf{f}_p,\mathbf{f}_q') = \left\{ \underset{m}{2\mathrm{tr}}(\mathbf{P}_p\mathbf{Z}_{ip}\mathbf{Z}_{ip}'\mathbf{P}_p\mathbf{V}_{pq}\mathbf{P}_q\mathbf{Z}_{jq}\mathbf{Z}_{jq}'\mathbf{P}_q\mathbf{V}_{qp}) \right\}_{i,j=0}^{r} . \tag{34}$$

Assemble these in a matrix

$$\mathbf{G} = \left\{ \underset{m}{\mathbf{G}_{pq}} \right\}_{p,q=1}^{t} . \tag{35}$$

Then, on assuming that $\sigma^2$ is known,

$$\mathbf{C}_{pq} = \mathrm{cov}(\bar{\boldsymbol{\sigma}}_p^2,\ \bar{\boldsymbol{\sigma}}_q^{2\,\prime}) = \mathbf{A}_p^{-1}\mathbf{G}_{pq}\mathbf{A}_q^{-1}$$

gives

$$\mathbf{C} = \mathrm{var}\left(\left\{ {}_{c}\tilde{\sigma}^2_{p} \right\}_{p=1}^{t}\right)$$

$$= \left\{ {}_{m}C_{pq} \right\}_{p,q=1}^{t} = \left\{ {}_{d}A_{p}^{-1} \right\} \left\{ {}_{m}G_{pq} \right\} \left\{ {}_{d}A_{p}^{-1} \right) \quad . \qquad (36)$$

Now $\left\{ {}_{c}\tilde{\sigma}^2_{p} \right\}_{p=1}^{t}$ is an estimator of

$$1_t * \sigma^2 = (1_t * I_{r+1})\sigma^2 \quad ,$$

where * represents the direct (or Kronecker) product operator. Hence a weighted least squares combination of the estimates $\tilde{\sigma}_p$ for $p = 1,\cdots,t$ is

$$\tilde{\sigma}^2 = \left[(1 * I)'\mathbf{C}^{-1}(1 * I)\right]^{-1}(1 * I)'\mathbf{C}^{-1} \left\{ {}_{c}\tilde{\sigma}^2_{p} \right\}$$

$$= (\Sigma_p\Sigma_q C^{p,q})^{-1}\Sigma_p(\Sigma_q C^{p,q})\tilde{\sigma}^2_p$$

$$= (\Sigma_p\Sigma_q A_p G^{p,q} A_q)^{-1}\Sigma_p A_p(\Sigma_q G^{p,q} A_q)\tilde{\sigma}^2_p \qquad (37)$$

where $C^{p,q}$ and $G^{p,q}$ are the (p,q)'th sub-matrices of $\mathbf{C}$ and $\mathbf{G}$ respectively.

One now iterates on (37), starting with $\sigma^2_o = \Sigma_p \tilde{\sigma}^2_p/t$ for $\sigma^2$. At least two iteration schemes are available. One would not involve $A_p$, using for that at each iteration its value determined at the time $\tilde{\sigma}^2_p$ was determined from data set p. Another scheme would recalculate $A_p$ with each iterated $\tilde{\sigma}^2$ derived from (37), just as (31), (32), (34) and the inverse of (35) will be calculated after each iteration of (37). In either case, each $\tilde{\sigma}^2_p$ would be the same at each iteration: those are the values that are being combined in (37). Simulation studies are needed to evaluate this procedure.

## 3.3 Confidence intervals

With ANOVA estimation we do not have exact, analytical forms for the sampling distributions of variance components estimators (other than, in some cases, the error component). This is so even in the simplest case, the 1-way classification with balanced data. Derivation of exact confidence intervals is therefore not feasible. On the other hand, with ML and REML estimation, the estimators have several asymptotic properties that are not only of theoretical interest but also of great practical importance: consistency, and a sampling distribution that is normal, with a variance-covariance matrix that comes from the inverted information matrix. In the limit, therefore, the establishment of confidence intervals from ML and REML estimators presents no great difficulty. The problem is that data sets may be extensive, but they are not of infinite size, and so using large-sample properties for a data set involves some degree of approximation. The question is, therefore,: what is the degree of this approximation? This brings us to what, I believe, is a vital research problem.

## 3.4 Planning evaluation studies

A detraction from ANOVA estimation is that we do not have any feasible ways of comparing its various forms when using unbalanced data. Analytic comparisons are impossible, and numerical comparisons using simulations can entail two big difficulties: tremendous amounts of computing, and the planning of simulation experiments in a manner that has some hope of yielding useful conclusions. The same is true of ML methods but with two ameliorating conditions. First, the time and money needed to do voluminous computing has declined enormously from even as little as ten years ago; and in the era of supercomputers will continue to decline. Second, with ML methodology there are only two methods of interest, ML and REML. With each of them we know their large-sample characteristics and the questions of

interest that needs to be assessed numerically are "How valid are those large-sample characteristics when used with finite samples?" and "How is that validity affected by different degrees and patterns of unbalancedness?" The literature does contain some studies of these questions but, naturally, in the light of the computing effort required, they have been predominantly studies of small-scale situations; e.g, 1-way classifications of 10-15 groups with a total of 100 observations, or 2-way classifications with 5-8 levels of each factor. These give but fragile information to the investigator who has 200 levels of one factor, 1,000 of the other and only 87,000 records located in but 29% of the 200,000 cells of the 2-way classification -thus 71% of the cells are empty and in the filled cells there is only an average of $1\frac{1}{2}$ observations per cell. And people *do* have data like this. Animal breeders have been estimating variance components from this kind of data (and even larger examples) for many years, e.g. Lush (1931). Furthermore, with both computerized data collection and statistical computing packages being so readily available nowadays, the need for knowing how the large-sample properties of ML and REML estimation apply to very large highly unbalanced data sets is becoming ever more pressing.

Of course, one major problem still pertains: planning simulation experiments so as to have some hope of their yielding useful conclusions; in the 2-way layout, how many levels of each factor, how many observations in each level, and in each cell, how many filled cells, and how will they be spread throughout the total grid of available cells? Also, what sets of values shall be used for the variance components when the data are generated? All of these questions and more must be addressed when planning simulation studies. It is not going to be easy to do that planning so that the studies yield conclusions that can be helpful to investigators who have large, unbalanced data sets. But this is, I believe, one of the most important research problems currently needing attention in the estimation of

variance components from unbalanced data. We need to know just how good or bad are the large-sample properties (e.g., normality, consistency and sampling variances from the inverted information matrix) of ML and REML estimation when used on large and oft-times highly unbalanced data sets.

This is paper BU-507 in the Biometrics Unit, Cornell University, Ithaca, New York 14853.

## BIBLIOGRAPHY


Airy, G. B. (1861). *On the Algebraical and Numerical Theory of Errors of Observations and the Combinations of Observations.* London: MacMillan.

Anderson, R. D. (1978). Studies on the estimation of variance components. Ph.D. Thesis, Cornell University, Ithaca, New York.

Anderson, R. L. (1975). Designs and estimators for variance components. In *Statistical Design and Linear Models*, Ed. J. N. Srivastava, 1-30. Amsterdam, North Holland.

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

Anderson, R. L. and P. P. Crump (1967). Comparisons of designs and estimation procedures for estimating parameters in a two-stage nested process. *Technometrics* **9**, 499-416.

Babb, J. S. (1986). Pooling maximum likelihood estimates of variance components obtained from subsets of unbalanced data. M.S. Thesis, Biometrics Unit, Cornell University, Ithaca, New York.

Bainbridge, T. R. (1963). Staggered nested designs for estimating variance components. *American Society for Quality Control Annual Conference Transactions.* 93-103.

Berk, K. (1987). Computing for incomplete repeated measures. *Biometrics* **43**, 385-398.

Bessel, I. (1820). Beschreibung des auf der Könisgberger Sternwart aufgestellten Reichenbachschen Meridiankrieses, dessen Anwendung und Geranigkeit imgleichen der Repoldschen Uhr. *Astronomisches Jahrbuch für das Jahr.* 1823, Berlin.

Blischke, W. R. (1968). Variances of estimates of variance components in a three-way classification. *Biometrics* **22**, 553-565.

Brown, K. G. (1976). Asymptotic behavior of MINQUE-type estimators of variance components. *Annals of Statistics* **4**, 746-754.

Chauvenet, W. (1863). A *Manual of Spherical and Practical Astronomy,* **2**: *Theory and Use of Astronomical Instruments.* Philadelphia, Lippincott.

Cochran, W. G. (1939). The use of analysis of variance in enumeration by sampling. *Journal of the American Statistical Association* **34**, 492-510.

Cochran, W. G. (1951). Improvement by means of selection. *Proceedings Second Berkeley Symposium*, 449-470.

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. *Journal of the Royal Statistical Society, Supplement* **6**, 186-197.

Eisenhart, C. (1947). The assumptions underlying the analysis of variance. *Biometrics* **3**, 1-21.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc., Edinburgh* **52**, 399-433.

Fisher, R. A. (1925), *Statistical Methods for Research Workers*, 1st Edition. Edinburgh and London, Oliver and Boyd.

Fisher, R. A. (1935). Discussion of Neyman *et al.*, 1935. *Journal of the Royal Statistical Society, Series B*, **2**, 154-155.

Ganguli, M. (1941). A note on nested sampling. *Sankya* **5**, 449-452.

Gauss, K. F. (1809). *Theoria Motus Corporum Celestrium in Sectionibus Conics Solem Ambientium.* Perthes and Besser: Hamburg.

Graybill, F. A and R. A. Hultquist (1961). Theorems concerning Eisenhart's Model II. *Annals of Mathematical Statistics* **32**, 261-269.

Graybill, F. A. and A. W. Wortham (1956). A note on uniformly best unbiased estimators for variance components. *Journal of the American Statistical Association* **51**, 266-268.

Hartley, H. O. and J. N. K. Rao (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.

Henderson, C. R. (1955). Personal communication.

Henderson, C. R. (1963). Selection index and expected genetic advance. *Statistical Genetics in Plant Breeding*, National Academy of Sciences, National Research Council publication No. 982.

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. *Annals of Mathematical Statistics* **30**, 939-959.

Hirotsu, C. (1966). Estimating variance components in a two-way layout with unequal numbers of observations. *Reports of Statistical Applications*, Union of Japanese Scientists and Engineers (JUSE), **13**, No. 2, 29-34.

Hocking, R. R., J. W. Green, and R. H. Bremer (1986). Estimation of variance components in mixed factorial models including model-based diagnostics. Paper presented at American Statistical Association Meetings, Chicago, Illinois.

Hocking, R. R. and M. H. Kutner (1975). Some analytical and numerical comparisons of estimators for the mixed A.O.V. model. *Biometrics* **31**, 19-28.

Jackson, R. W. B. (1939). Reliability of mental tests. *British Journal of Psychology* **29**, 267-287.

Jennrich, R. I. and Schlucter, M. D. (1986). Unbalanced repeated-measures models with structured covariance matrices. *Biometrics* **42**, 805-820.

Jeske, D. R. and D. A. Harville (1986). Prediction, confidence and empirical Bayes interval in linear models. Paper presented at American Statistical Association Meetings, Chicago, Illinois.

Kackar, R. N. and D. A. Harville (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications In Statistics A: Theory and Methods* **10**, 1249-1261.

Kackar, R. N. and D. A. Harville (1984). Approximations for standard errors of estimators of fixed and random effects in mixed linear models. *Journal of the American Statistical Association* **79**, 853-861.

Khuri, A. I. and H. Sahai (1985). Variance component analysis: a selective literature survey. *International Statistics Review* **53**, 279-300.

LaMotte, L. R. (1970). A class of estimators of variance components. Technical Report 10, Department of Statistics, University of Kentucky, Lexington, Kentucky, 13 pages.

LaMotte, L. R. (1973). Quadratic estimation of variance components. *Biometrics* **29**, 311-330.

Legendre, A. M. (1806). *Nouvelles Méthodes pour la Determination des Orbites des Cométes; avec un Supplément Contenant Divers Perfectionnements de ces Méthodes et leur Application aux deux Cométes de 1805.* Courcier, Paris.

Low, L. Y. (1964). Sampling variances of estimates of components of variance from a non-orthogonal two-way classification. *Biometrika* **51**, 491-494.

Lush, J. L. (1931). The number of daughters necessary to prove a sire. *J. of Dairy Sciece*, **14**, 209-220.

Mahamunulu, D. M. (1963). Sampling variances of the estimates of variance components in the unbalanced three-way nested classification. *Annals of Mathematical Statistics* **34**, 521-527.

Mood, A. M. (1950). *Introduction to the Theory of Statistics.* New York: McGraw-Hill.

Mood, A. M. and Graybill, F. A. (1963). *Introduction to the Theory of Statistics.* 2nd Edition. New York: McGraw-Hill.

Mood, A. M., Graybill, F. A., and Boes, D. C. (1974). *Introduction to the Theory of Statistics.* 3rd Edition. New York: McGraw-Hill.

Neyman, J. K., Iwaszkiewicz and S. T. Kolodziejczyk (1935). Statistical problems in agricultural experimentation. *Journal of the Royal Statistical Society,* Supplement **2**, 107-154.

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

Plackett, R. L. (1972). Studies in the history of probability and statistics. XXIX. The discovery of the method of least squares. *Biometrika* **59**, 239-251.

Preitschopf, Franz (1987). Personal Communicaton.

Rao, C. R. (1965). *Linear Statistical Inference and its Applications*, Wiley, New York.

Rao, C. R. (1970). Estimation of heteroscedastic variances in linear models. *Journal of the American Statistical Association* **65**, 161-172.

Rao, C. R. (1971a). Estimation of variance and covariance components - MINQUE theory. *Journal of Multivariate Analysis* **1**, 257-275.

Rao, C. R. (1971b). Minimum variance quadratic unbiased estimation of variance components. *Journal of Multivariate Analysis* **1**, 445-456.

Rao, C. R. (1972). Estimation of variance and covariance components in linear models. *Journal of the American Statistical Association* **67**, 112-115.

Rohde, D. A. and G. M. Tallis (1969). Exact first- and second-order moments of estimates of components of covariance. *Biometrika* **56**, 517-525.

Scheffé, H. (1956). Alternative models for the analysis of variance. *Annals of Mathematical Statistics* **27**, 251-271.

Searle, S. R. (1956). Matrix methods in components of variance and covariance analysis. *Annals of Mathematical Statistics* **27**, 737-748.

Searle, S. R. (1958). Sampling variances of estimates of components of variance. *Annals of Mathematical Statistics* **29**, 167-178.

Searle, S. R. (1961). Variance components in the unbalanced two-way nested classification. *Annals of Mathematical Statistics* **32**, 1161-1166.

Searle, S. R. (1968). Another look at Henderson's methods of estimating variance components. *Biometrics* **24**, 749-778.

Searle, S. R. (1971). *Linear Models*. New York: Wiley.

Searle, S. R. (1974). Prediction, mixed models and variance components. In *Reliability and Biometry*, Ed. F. Proschan and R. J. Serfling, 229-266. Philadelphia, Society of Industrial and Applied Mathematics.

Snedecor, G. W. (1934, 1940). *Analysis of Variance and Covariance*. 1st and 3rd Editions. Ames, Iowa: Collegiate Press, Inc.

Snedecor, G. W. and W. G. Cochran (1969). *Statistical Methods.* Ames, Iowa: Iowa State College Press.

Thompson, W. A., Jr. (1962). The problem of negative estimates of variance components. *Annals of Mathematical Statistics* **33**, 273-289.

Tippett, L. H. C. (1931, 1937). *The Methods of Statistics.* 1st and 2nd Editions. London: Williams and Norgate.

Winsor, C. P. and G. L. Clarke (1940). Statistical study of variation in the catch of plankton nets. *Journal of Marine Research* **3**, 1-34.

Yates, F. and I. Zacopanay (1935). The estimation of the efficiency of sampling with special reference to sampling for yield in cereal experiments. *Journal of Agricultural Science* **25**, 545-577.