VARIANCE COMPONENTS – SOME HISTORY AND A SUMMARY ACCOUNT OF ESTIMATION METHODS

Shayle R. Searle[1/]

Biometrics Unit, Cornell University, Ithaca, NY

BU-959-M February 1988

Abstract

A brief recounting of some of the history of estimating variance components is followed by discussion of the maximum likelihood methods that are becoming more feasible as computers get larger, faster and cheaper to use. Estimation of fixed effects is outlined, as is prediction of random variables, the latter leading to the procedure known as BLUP.

1. Fixed and Random Effects

Basic concepts of traditional analysis of variance can be described in terms of a completely randomized design experiment. Suppose we are interested in testing the efficacy of five different diets for feeding dairy cattle – maybe five different dose levels of the bovine growth hormone bovine somatropin. In a herd of 50 purebred Holstein cows, the cows are divided at random into five groups of 10 animals each, and each group is fed with one of the five diets. Let the milk yield of cow j receiving diet i be denoted y_{ij}. Then for describing an analysis of variance of these data, a suitable model equation for y_{ij} is

$$y_{ij} = \mu + \alpha_i + e_{ij} \qquad (1)$$

[1/] Part of this article was prepared while the author was Gast Professor, at the Institut für Mathematik, Universität Augsburg, Federal Republic of Germany, supported by the Alexander von Humboldt Foundation on a Senior U.S. Scientist Award, 1985-6.

for $i = 1,\cdots,5$ and $j = 1,\cdots,10$. In (1), μ represents a general mean, and α_i is the effect on milk yield of a cow's receiving diet i; and e_{ij} is a random error defined as $e_{ij} = y_{ij} - E(y_{ij})$ for $E(y_{ij}) = \mu + \alpha_i$ where E represents expectation over repeated randomization of the 50 cows in groups of 10 to the five diets. Thus e_{ij} is taken as a random variable, with $E(e_{ij}) = 0$. The further distributional assumptions attributed to the error terms are that the variance of e_{ij} is the same for all i and j and that the covariance between any two error terms is zero; i.e., that

$$v(e_{ij}) = \sigma^2 \ \forall \ i \text{ and } j \text{ and } \operatorname{cov}(e_{ij},e_{i'j'}) = 0 \ \forall \ i \neq i' \text{ and } j \neq j' \ . \qquad (2)$$

The further assumption of normality is often made also.

1.1 Fixed effects

The parameters of interest in this model are the mean μ and the α_is, the effects of the diets on milk yield; and one object of analyzing the data is that of estimating linear functions of μ and the α_is. For example, the best linear unbiased estimators of $\mu + \alpha_i$ and of $\alpha_i - \alpha_h$ are

$$\text{BLUE}(\mu + \alpha_i) = \bar{y}_{i\cdot} = \sum_{j=1}^{n} y_{ij}/n \quad \text{and} \quad \text{BLUE}(\alpha_i - \alpha_h) = \bar{y}_{i\cdot} - \bar{y}_{h\cdot} \ .$$

In the context of this model the μ and the α_is are taken as being constants, albeit unknown and unknowable, but nevertheless fixed constants. As such, they are called *fixed effects*. They are deemed to be constants representing the effects of the different diets being studied. The diets are the things of particular interest, chosen because of our interest in them. Those different diets could just as well be different fertilizers applied to a corn crop, different forage crops grown in the same region, different machines used in a manufacturing process, different drugs given for the same illness, and so on. The possibilities are legion – as are the

varieties of models and their complexities, reaching far beyond those of (1).

1.2 Random effects

Now consider the situation where we have first lactation records of ten 2-year-old cows from each of five young, untried bulls in an artificial insemination center. Equation (1) could be used to represent the milk yield of cow j sired by bull i. Then α_i in (1) would represent the effect on milk yield of a cow's having been sired by bull i. Thus α_i represents the genetic contribution of bull i to the milk yield y_{ij}. Genetically, the α_is are therefore treated as random variables, and in this context are called *random effects*. As such we attribute distributional properties of zero mean and the same variance to each α_i:

$$E(\alpha_i) = 0 \qquad \text{and} \qquad v(\alpha_i) = \sigma^2_\alpha \quad \forall \quad i \; ; \tag{3}$$

and zero covariance between each pair of α_is, and between any α_i and any e_{ij}:

$$\text{cov}(\alpha_i,\alpha_h) = 0 \;\; \forall \; i \neq h \qquad \text{and} \qquad \text{cov}(\alpha_i,e_{i'j}) = 0 \;\; \forall \; i, \; i', \; j \; . \tag{4}$$

Then $E(y_{ij}) = \mu$; and e_{ij} is defined as $e_{ij} = y_{ij} - E(y_{ij}|\alpha_i)$ where $E(y_{ij}|\alpha_i)$ is the conditional expected value of y_{ij} for given α_i.

When the α_is in (1) are considered as random effects, as in (3) and (4), and with there being no other terms in (1) except μ and e_{ij}, the model is called a *random effects model* or, more usually a *random model*. Since the α_is occurring in the data are then envisaged as being a random sample from some population of αs, they are no longer the αs of sole interest, as they are when they are fixed effects. There is therefore little or no reason for estimating either the α_is or differences between them; the parameter of interest so far as they are concerned is now σ^2_α. Since (1)

– (4) then give $\sigma^2_y = \sigma^2_\alpha + \sigma^2_e$, the variances σ^2_α and σ^2_e, being components of the variance of y, are called variance components, and the random model is then correspondingly sometimes called a *variance components model.*

1.3 Mixed models

In the random model (1) – (4), the μ is still deemed to be a fixed effect, just as in the model (1) and (2). In that case (1) – (4) is a mixture of fixed and random effects and as such could be called a *mixed effects model* or, more simply, a *mixed model.* In the particular case of (1) – (4), where μ is the only fixed effect, the term random model is, in fact, retained. But whenever a model equation has terms in it (other than μ and error terms) that are a mixture of fixed and random effects, then the model is called a mixed model. For example, the model equation

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

with the α_is being fixed effects and the β_js being random effects is the equation of a mixed model.

The decision as to whether effects are to be taken as fixed or as random is made on the basis of what they represent. A useful dichotomy is that when we want to make inferences only to the effects of a factor that occur in the data, then those effects are considered as fixed effects, but when inferences are going to be made to a set of effects that is larger than just those that occur in the data, then the effects will be taken as random effects. Extensive discussion of this point is to found in the landmark paper of Eisenhart (1947), with further comments in Searle (1971) and Kempthorne (1975).

2. A BRIEF HISTORY

The origin of analysis of variance of what we now called fixed effects models is well known to be with R.A. Fisher during his years at Rothamsted Experimental Station. To complement this it is of interest to trace the history of the development of variance components estimation. We do so by freely adapting, with kind permission of the editors, from Searle (1988), which in turn draws heavily on Scheffé (1956) and Anderson (1978). Later sections of the paper also drawn on Searle (1988).

2.1 The early years

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 Probabilités*) ...". 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

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

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 the 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 greater than 5"; the only pertinent remark is on page 166 of the 1823 yearbook which with x_i being the "random error of part i, i = 1,$\cdots$, 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 today's notation for the 1-way classification random model with balanced data, are

$$E(SSE) = E \sum_{i=1}^{a} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 = a(n-1)\sigma_e^2 \qquad (5)$$

and

$$E(SSA) = E \sum_{i=1}^{a} \sum_{j=1}^{n} (\bar{y}_{i.} - \bar{y}_{..})^2 = (a-1)(n\sigma_\alpha^2 + \sigma_e^2) \ . \qquad (6)$$

From these the estimation equations are taken as

$$SSE = a(n-1)\hat{\sigma}_e^2 \quad \text{and} \quad SSA = (a-1)(n\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2) \qquad (7)$$

and so

$$\hat{\sigma}_e^2 = MSE \qquad \text{and} \qquad \hat{\sigma}_\alpha^2 = (MSA - MSE)/n \ . \qquad (8)$$

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 σ_α^2.

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 occur-

rence 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., (5) and (6)] "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

$$E(SSA) = E\sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 = \left(n_. - \sum_{i=1}^{a} n_i^2/n_.\right)\sigma_\alpha^2 + (a-1)\sigma_e^2 \qquad (9)$$

and

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

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.

2.2 Forty years, 1947-87

After what can now be viewed as the foundation writings of the 30s and 40s, 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 because 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 detailed historical survey, if for no other reason than the availability of the excellent survey by Khuri and Sahai (1985), where the interested reader will find a full account. So just a brief and somewhat personal outline is given of the major advances made in 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 (1966, 1968), and Rohde and Tallis (1969). In every case the results are 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 judgment 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.

3. ANOVA Estimation from Balanced Data

3.1 Method

Analysis of variance relies on data being classified by levels of the different factors. Data are described as being balanced when there are the same number of observations in each of the subclasses: balanced data are equal-subclass-numbers data. Also included in this description are cases of what can more accurately be described as data that exhibit planned unbalancedness, such as those from latin squares, balanced incomplete

blocks and other such experiments where data do not occur at every possible "intersection", so to speak, of one level of every factor. For example, in a latin square of order n, only n^2 of the possible n^3 such "intersections" contain data: the other $n^3 - n^2$ are empty (e.g., Searle, 1987, p. 6).

The basic principle for estimating variance components from balanced data is that of equating analysis of variance mean squares to their expected values. Expectations in this context are, of course, taken under the properties of whatever random effects are in the model such as those, for example, in (2) - (4). This "equate mean squares to their expectations" method of estimating variance components is essentially a method of moments, although estimators obtained this way are nowadays called analysis of variance (ANOVA) estimators. Equations (8) show these estimators for the 1-way classification with balanced data.

The ANOVA estimators are clearly a natural extension of the practice in the analysis of variance of fixed effects models whereby the residual variance had always been (and still is) estimated by the error mean square. This is the equation $\hat{\sigma}^2_e$ = MSE in (8), based on (5); and (6) as the extension of (5) leads to $\hat{\sigma}^2_\alpha$ = (MSA - MSE)/n, also in (8). More generally, for balanced data (and for data exhibiting planned unbalancedness), the ANOVA method of estimating variance components is based on those mean squares that have expected values that involve only variance components. The method consists of equating those mean squares to their expected values (with each σ^2 replaced by $\hat{\sigma}^2$) and the resulting equations are solved for the $\hat{\sigma}^2$s. Thus for $\mathbf{m}$ being the vector of mean squares and $\boldsymbol{\sigma}^2$ the vector of variance components with $\mathbf{T}$ being determined by

$$E(\mathbf{m}) = \mathbf{T}\boldsymbol{\sigma}^2, \quad \text{then} \quad \hat{\boldsymbol{\sigma}}^2 = \mathbf{T}^{-1}\mathbf{m}, \tag{10}$$

where $\mathbf{T}^{-1}$ is presumed to exist (and usually does).

One of the first books with any extended treatment of variance component estimation is Anderson and Bancroft (1952) with its four full chapters on the subject. This book really set the ANOVA method of estimation on a firm footing with its several examples of the procedure of equating analysis of variance sums of squares to their expectations as a method of estimating variance components. The book deals very thoroughly with estimation from balanced data for both mixed and random models; it also discusses unbalanced data for nested classifications and, after considering incomplete blocks designs, it poses a number of pertinent research problems, many of which have still not been answered satisfactorily. In all, the book is a milestone in variance components estimation.

3.2 Properties

Some of the important properties of ANOVA estimators of variance components from balanced data are as follows.

a. Sampling distributions

Even under normality assumptions the only estimator having a sampling distribution in closed form is $\hat{\sigma}^2_e$ = MSE which is distributed as $\sigma^2_e\chi^2$ with degrees of freedom being those of MSE. Other than this, every ANOVA estimator is a linear combination of mean squares that involves some of the mean squares negatively, e.g., $\hat{\sigma}^2_\alpha$ = (MSA - MSE)/n of (8). Therefore even though, under normality, those mean squares are stochastically independent and each is distributed proportionately as a χ^2, the ensuing $\hat{\sigma}^2$ does not have a χ^2 density.

Satterthwaite (1946) is a landmark paper that deals with approximate distributions for these estimators. And Robinson (1965) shows that they are infinite sums of weighted χ^2s with the weights being functions of the unknown variance components. Unfortunately this is not very helpful for practical applications.

b. Unbiasedness

Despite not having closed form density functions for ANOVA estimators, they are (of course) unbiased: (10) gives

$$E(\hat{\boldsymbol{\sigma}}^2) = \mathbf{T}^{-1}E(\mathbf{m}) = \mathbf{T}^{-1}\mathbf{T}\boldsymbol{\sigma}^2 = \boldsymbol{\sigma}^2 \ .$$

c. Minimum variance

Graybill and Wortham (1956) show that ANOVA estimators from balanced data are minimum variance unbiased on assuming normality of the random effects and error terms; and Graybill and Hultquist (1961) extend this, showing that even without normality they are minimum variance, quadratic, unbiased. This means that among all quadratic functions of the observations that are unbiased for the variance components, the functions obtained as ANOVA estimators have minimum variance.

d. Sampling variances

In addition to having minimal variance properties, explicit expressions for the sampling variances (and unbiased estimators thereof) of these estimators can be derived quite straightforwardly – when data are balanced and under normality assumptions. This is so because analysis of variance mean squares are then independent with distributions proportional to χ^2s. As a result, each mean square, M, having f degrees of freedom, say, has variance

$$v(M) = 2E(M)]^2/f \ . \qquad (11)$$

Because, by definition of variance, v(M) also has the form

$$v(M) = E(M^2) - [E(M)]^2 \ , \tag{12}$$

equating (11) and (12) yields $M^2/(f+2)$ as an unbiased estimator of $[E(M)]^2/f$. Therefore an unbiased estimator of (11) is

$$\hat{v}(M) = 2M^2/(f+2) \ . \tag{13}$$

Hence if

$$\hat{\sigma}^2 = \Sigma k_i M_i$$

is a variance component estimator obtained by this method its variance is

$$v(\hat{\sigma}^2) = 2\Sigma_i k_i^2 [E(M_i)]^2/f_i \quad \text{with} \quad \hat{v}(\hat{\sigma}^2) = 2\Sigma_i k_i^2 M_i^2/(f+2) \tag{14}$$

being an unbiased estimator of that variance. For example, applying (14) to (8) and using the independence of MSA and MSE, the variance of $\hat{\sigma}^2_\alpha$ is

$$v(\hat{\sigma}^2_\alpha) = \frac{2}{n^2}\left[\frac{(n\sigma^2_\alpha + \sigma^2_e)^2}{a-1} + \frac{\sigma^4_e}{a(n-1)}\right] \ , \tag{15}$$

an unbiased estimator of which is

$$\hat{v}(\hat{\sigma}^2_\alpha) = \frac{2}{n^2}\left[\frac{MSA^2}{a+1} + \frac{MSE^2}{a(n-1)+2}\right] = \frac{2}{n^2}\left[\frac{(n\hat{\sigma}^2_\alpha + \hat{\sigma}^2_e)^2}{a+1} + \frac{\hat{\sigma}^4_e}{a(n-1)+2}\right] \ .$$

e. Negative estimates

These minimum variance unbiasedness properties of ANOVA estimators (of variance components from balanced data) are sufficiently attractive that ANOVA estimators are almost always what are used when available data are balanced. Despite the attractiveness of these properties there is at least one big difficulty.

There is nothing inherent in the estimation method that necessarily prevents estimators (other than $\hat{\sigma}^2_e$) from being negative. In other words, although $\hat{\sigma}^2_e$ is always positive, other estimators can (and sometimes do)

yield negative estimates. Thus, any data for the 1-way classification random model that are such that MSA < MSE will, by (8), yield a negative estimate of $\hat{\sigma}^2_\alpha$. For example, for two observations on each of two classes, 10, 30 in one class and 14, 30 in the other, it is easily verified that $\hat{\sigma}^2_\alpha$ from (8) is $\hat{\sigma}^2_\alpha = -80$. Clearly, this is an embarrassment – having a negative estimate $\hat{\sigma}^2_\alpha$, of a parameter σ^2_α which by definition is positive. Nevertheless it can happen and, indeed, the probability of its happening can, under certain circumstances be appreciably large. For example, under normality, Searle (1971, p. 415) shows that for (9)

$$\Pr\{\hat{\sigma}^2_\alpha < 0\} = \Pr\{F_{a-1,a(n-1)} < (1 + n\sigma^2_\alpha/\sigma^2_e)^{-1}\}$$

where $F_{a-1,a(n-1)}$ is a random variable having an F-distribution with degrees of freedom (a - 1) and a(n - 1). Leone *et al.* (1968) have other examples of this kind of probability statement.

To the question "what does one do with a negative estimate?" there appears to be no satisfactory answer. Four possible courses of action are (i) to use zero in place of the negative value, (ii) to do that and eliminate the corresponding factor from the model and then re-estimate, (iii) to collect more data in the hope of then not getting any negative estimate, or (iv) to use a different estimation method. None of these is particularly attractive, in the face of which maybe the only course of action would be the statistician's last hope: collect more data. Whatever one does, it seems essential to always report the negative estimate even if it is not used subsequently.

4. Generalizing ANOVA Estimation to Unbalanced Data

4.1 Basic ideas

Unbalanced data are data that have unequal numbers of observations in the subclasses, including the possibility of some subclasses containing no observations at all. Indeed, in animal breeding data the empty subclasses are sometimes more numerous than the filled ones.

Applying the ANOVA method of estimation to the 1-way classification model of (1) – (14) when data are unbalanced is easy. Instead of n observations in every class we now have n_i observations in the i'th class. Then with

$$SSA = \sum_{i=1}^{a} n_i(\bar{y}_{i.} - \bar{y}_{..})^2 \quad \text{and} \quad SSE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 \qquad (16)$$

used in

$$MSA = SSA/(a - 1) \quad \text{and} \quad MSE = SSE/(N - a) ,$$

the expected mean squares are

$$E(MSA) = \lambda\sigma_\alpha^2 + \sigma_e^2 \quad \text{and} \quad E(MSE) = \sigma_e^2 \qquad (17)$$

for

$$\lambda = (\Sigma_i n_i - \Sigma_i n_i^2/\Sigma_i n_i)/(a - 1) . \qquad (18)$$

Hence applying the "equate mean squares to their expectations" principle of ANOVA estimation from balanced data gives

$$MSA = \lambda\hat{\sigma}_\alpha^2 + \hat{\sigma}_e^2 \quad \text{and} \quad MSE = \hat{\sigma}_e^2 . \qquad (19)$$

Thus

$$\hat{\sigma}_\alpha^2 = (MSA - MSE)/\lambda \quad \text{and} \quad \hat{\sigma}_e^2 = MSE \qquad (20)$$

are the ANOVA estimators. Comparison of (20) with (8) reveals that the only differences, notationally, are that λ in (20) is used in place of n in (8); and from (18) we can see that when every n_i is n, as with balanced data, then λ reduces to being $\lambda = n$. At first sight this difference of having λ in (20) in place of n in (8) appears to be the only difference of (20) from (8); but, in fact, there are other differences that lead to complications entailed with (20) that do not occur with (8). For example, although under normality of the random effects and error terms, MSA and MSE of (16) are independent, as are MSA and MSE of (8), the sampling distribution of MSA of (16) is not proportional to a χ^2, whereas it is for MSA of (8), the balanced data case. This complicates properties of the estimators (20) for the unbalanced case. It means that, although the distribution of $\hat{\sigma}_e^2$ is proportional to that of a χ^2, nothing is known about the distribution of $\hat{\sigma}_\alpha^2$ of (20). Nevertheless, its variance is known and is, for $N = n_. = \Sigma_i n_i$,

$$v(\hat{\sigma}_\alpha^2) = \frac{2\sigma_e^4 N^2(N-1)(a-1)}{(N-a)(N^2-\Sigma_i n_i^2)^2} + \frac{4\sigma_e^2\sigma_\alpha^2 N}{N^2-\Sigma_i n_i^2} + \frac{2\sigma_\alpha^4[N^2\Sigma n_i^2 + (\Sigma_i n_i^2)^2 - 2N\Sigma_i n_i^3]}{(N^2-\Sigma_i n_i^2)^2}. \quad (21)$$

This is clearly a more complicated function of a and the n_is than is its counterpart, (15), of a and n for balanced data. Of course (21) reduces to (15) when every $n_i = n$.

4.2 Generalization

As we have seen, the estimators $\hat{\sigma}_\alpha^2$ and $\hat{\sigma}_e^2$ in (20) were first available in 1939-40 (Cochran, 1939, and Winsor and Clark, 1940), and were extended to nested models in general by Ganguli 1941). These results aside, the statistical literature shows little evidence of interest in the problem of estimating variance components until well into the fifties [save for Crump's (1951) derivation of (21)]. Yet there was (and still is) at

least one large class of scientists, the plant and animal breeders whose interests (e.g., selection for increased yields) demand the estimation of variance components, often, from large data sets that are almost always unbalanced, and often seriously so; e.g., data with 70% empty class cells and only an average of 1.6 observations per cell for those cells that do have data. Appreciable contributions to the development of methods of estimating variance components from unbalanced data are therefore to be found coming from animal breeders, the landmark example being Henderson (1953). This is where what are now known as Henderson's three methods were first expounded. To describe them we can benefit from hindsight by first considering a simple generalization of ANOVA estimation from balanced data; for then the Henderson methods can be seen as simply special cases of that generalization.

The ANOVA method for balanced data equates analysis of variance mean squares to their expected values. The latter are linear combinations of the variance components of whatever model is being used. A simple generalization of this is to use in place of mean squares a set of quadratic forms of the observations. Suppose $\mathbf{y'Ay}$ is one such quadratic form. Then, when $\mathbf{y}$ has expected value $\boldsymbol{\mu}$ and variance covariance matrix $\mathbf{V}$, it is well known that the expected value of $\mathbf{y'Ay}$ is

$$E(\mathbf{y'Ay}) = \text{tr}(\mathbf{AV}) + \boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} \qquad (22)$$

where $\text{tr}(\mathbf{AV})$ is the trace of $\mathbf{AV}$, the sum of its diagonal elements. Since, in variance components models, elements of $\mathbf{V}$ are either zero, individual variance components or sums thereof, (22) will be a linear combination of variance components whenever $\boldsymbol{\mu}'\mathbf{A}\boldsymbol{\mu} = 0$. Therefore, provided $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} = 0$ for

as many symmetric matrices $\mathbf{A}_i$ as we have variance components in the model being used, the use of (22) for each $\mathbf{y}'\mathbf{A}_i\mathbf{y}$ (where $\mathbf{y}$ is the data vector) provides generalization of the ANOVA method of estimation.

Suppose $\mathbf{q}$ is a vector of quadratic forms $\mathbf{y}'\mathbf{A}_i\mathbf{y}$, each having $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} = 0$. Then with $\boldsymbol{\sigma}^2$ being the vector of variance components to be estimated the generalization is that, for some matrix $\mathbf{C}$,

$$E(\mathbf{q}) = \mathbf{C}\boldsymbol{\sigma}^2 \qquad \text{and} \qquad \hat{\boldsymbol{\sigma}}^2 = \mathbf{C}^{-1}\mathbf{q} \ . \tag{23}$$

Clearly, there is an infinity of ways of using (23). Any quadratic forms $\mathbf{y}'\mathbf{A}_i\mathbf{y}$ can be used as elements of $\mathbf{q}$ just so long as $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} = 0$ for each of them and so long as $\mathbf{C}^{-1}$ exists. Other than these requirements, the method of estimation summarized by (23) places no demands on the quadratic forms to be used as elements of $\mathbf{q}$, and it gives no indication whatever as to how to choose what quadratic forms to use. For example, in a 1-way classification involving 100 classes and 5000 observations, $(y_{11} - y_{12})^2$ and $(y_{11} - y_{21})^2$ are, so far as (23) is concerned, perfectly feasible as quadratic forms for estimating σ^2_α and σ^2_e:

$$E(y_{11} - y_{12})^2 = 2\sigma^2_e \quad \text{and} \quad E(y_{11} - y_{21})^2 = 2(\sigma^2_\alpha + \sigma^2_e)$$

giving

$$\hat{\sigma}^2_e = \tfrac{1}{2}(y_{11} - y_{12})^2 \quad \text{and} \quad \hat{\sigma}^2_\alpha = \tfrac{1}{2}[(y_{11} - y_{21})^2 - (y_{11} - y_{12})^2] \ . \tag{24}$$

Nevertheless, no one with 5000 observations would be satisfied with using only three of them for estimation of the two variance components. Yet this generalized ANOVA method of estimation does not preclude doing so. Moreover, although in (23) $\mathbf{q}$ implicitly has as many elements as there are variance components to be estimated, there is no real need for this. There

could be more. Thus if $\mathbf{C}$ is rectangular, with more rows than columns and of full column rank, then a reasonable estimation method based on $E(\mathbf{q}) = \mathbf{C}\sigma^2$ would be

$$\hat{\boldsymbol{\sigma}}^2 = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{q} \ . \tag{25}$$

This is, of course, the same as (23) when $\mathbf{C}$ is non-singular.

4.3 Properties

This generalization of the ANOVA method has the following features.

a. Sampling distributions

Many quadratic forms used as elements of $\mathbf{q}$ do not, under normality, have χ^2 distributions. And even when they do as with balanced data, estimated components involve some of those quadratic forms negatively. Hence distributions of estimators are unknown.

b. Unbiasedness

Resulting estimators are unbiased: (25) gives

$$E(\hat{\boldsymbol{\sigma}}^2) = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'E(\mathbf{q}) = (\mathbf{C}'\mathbf{C})^{-1}\mathbf{C}'\mathbf{C}\sigma^2 = \boldsymbol{\sigma}^2 \ .$$

c. Sampling variances

Under normality assumptions, with $\mathbf{y} \sim N(\boldsymbol{\mu},\mathbf{V})$, it is well known with $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} = 0 = \boldsymbol{\mu}'\mathbf{A}_k\boldsymbol{\mu} = 0$ that

$$\operatorname{cov}(\mathbf{y}'\mathbf{A}_i\mathbf{y}, \ \mathbf{y}'\mathbf{A}_k\mathbf{y}) = 2\operatorname{tr}(\mathbf{A}_i\mathbf{V}\mathbf{A}_k\mathbf{V}) \ . \tag{26}$$

Therefore the variance-covariance matrix of $\mathbf{q}$ is

$$\operatorname{var}(\mathbf{q}) = 2\{\operatorname{tr}(\mathbf{A}_i\mathbf{V}\mathbf{A}_k\mathbf{V})\}$$

and so

$$\operatorname{var}(\hat{\boldsymbol{\sigma}}^2) = 2\mathbf{C}^{-1}\{\operatorname{tr}(\mathbf{A}_i\mathbf{V}\mathbf{A}_k\mathbf{V})\}\mathbf{C}^{-1'} \ . \tag{27}$$

However, since elements of $\mathbf{V}$ are (usually) variance components or sums thereof, or zero, the elements of (27) are quadratic functions of those components – and these are the very components being estimated. Equation (21) is an example. Thus estimating $\text{var}(\hat{\sigma}^2)$ in any optimal way is not necessarily straightforward. Obviously, one possibility is to replace σ^2 in $\mathbf{V}$ by $\hat{\sigma}^2$ to give $\hat{\mathbf{V}}$, and so calculate

$$\hat{\text{var}}(\hat{\sigma}^2) = 2\mathbf{C}^{-1}\{\text{tr}(\mathbf{A}_i\hat{\mathbf{V}}\mathbf{A}_k\hat{\mathbf{V}})\}\mathbf{C}^{-1'} \quad . \tag{28}$$

An unbiased estimator of $\text{var}(\hat{\sigma}^2)$ is obtainable as follows. Define

$\mathbf{v}$ = vector of variances and covariances of all $\hat{\sigma}^2$s

and

$\boldsymbol{\gamma}$ = vector of all squares and products of all σ^2s .

Then (27) implies that there is some matrix $\mathbf{M}$ such that

$$\mathbf{v} = \mathbf{M}\boldsymbol{\gamma} \quad .$$

It can then be shown (Ahrens, 1965; Searle, 1971, p. 437), with $\hat{\boldsymbol{\gamma}}$ being the vector of all squares and products of $\hat{\sigma}^2$s, that

$$\hat{\mathbf{v}} = (\mathbf{I} + \mathbf{M})^{-1}\mathbf{M}\hat{\boldsymbol{\gamma}}$$

is an unbiased estimator of $\mathbf{v}$. Its elements thus provide an unbiased estimator of $\text{var}(\hat{\sigma}^2)$.

d. Effects of unbalancedness

It is the particular forms used as elements of $\mathbf{q}$ in (23) that determine the elements of $\mathbf{C}$ in (23). Those elements are usually complicated functions of the n_{ij}s, the numbers of observations in the different classes and subclasses of the data. The same is often true of the $\mathbf{A}_i$-matrices used in the quadratic forms $\mathbf{y}'\mathbf{A}_i\mathbf{y}$. Thus elements of (27), as well as being quadratic functions of the variance components, are also usually very complicated functions of the n_{ij}s. Their complicatedness

precludes, in fact, being able to study the manner in which different values of the n_{ij}s affects elements of (27), i.e., affects the sampling variances and covariances of the estimated components. Equation (21) illustrates this. It can be shown that the first two terms of (21) increase as n-values become more disparate (e.g., 1,1,1,1,21 versus 3,4,5,6,7) but the third term can decrease; for 1,1,1,11,11 that third term is $1.106\sigma_\alpha^4$ whereas for 1,1,1,1,21 it is $.788\sigma_\alpha^4$.

Thus the effects of different degrees of unbalancedness on sampling variances of ANOVA estimators will be very difficult to study. Simulation experiments may be thought to be viable for this purpose. Certainly with today's supercomputers the great mass of computation is no longer an impediment. But planning the experiments so as to have some hope of yielding trends is probably, I think, an insuperable task.

e. Negative estimates

Just as with balanced data, so with unbalanced: there is nothing inherent in ANOVA estimation that prevents negative estimates. Again, with two classes, one with data 10,30 and the other with data 14,30,31, it will be found that $\hat{\sigma}_\alpha^2$ from (20) is negative: $\hat{\sigma}_\alpha^2 = -61.08$.

There is no way of universally avoiding this embarrassment with ANOVA estimation. It is a consequence of data and the method. And remedies for it are just as unsatisfactory as they are for balanced data, as spelled out in Section 3.2e.

f. Choice of quadratic forms

The method of estimation invoked by (23) gives absolutely no guidance whatever as to what quadratic forms to use as elements of **q**. Any quadratic forms can be used, even those as ridiculous as suggested in (24). The method, of itself, contains no criteria for preferring some quadratic

forms over others with a view to imposing optimal properties on the resulting estimators. The *only* implicit property is unbiasedness and that arises no matter what quadratic forms are used.

5. The Henderson Methods

The three Henderson methods (Henderson, 1953) of estimating variance components are methods for use on unbalanced data from mixed models of more than one factor. Each of the methods is simply an application of the ANOVA methodology embodied in (23), albeit a judicious and ingenious application. From this point of view, therefore, the Henderson methods are easily described and understood. All that has to be done is to describe what quadratic forms are used as elements of **q** in (23). In contrast, details needed for carrying out the methods can be lengthy and tedious. Attention is therefore confined to describing the methods on the basis of their being just particular cases of (23). They differ only through the different sets of quadratic form that are used as elements of **q**. All those methods reduce, for balanced data, to the standard ANOVA method for balanced data.

The description of the methods which follows is therefore confined solely to describing the quadratic forms (mostly sums of squares) used in each of the methods, with few attendant details. Selected aspects of the methods are illustrated using the following example.

5.1 Example: the 2-way crossed classification

Let y_{ijk} be the k'th observation in the cell defined by the i'th level of a factor to be called rows and the j'th level of a factor to be called columns. Suppose there are $n_{ij} \geq 0$ observations in that cell: then $k = 0$ for an empty cell and $k = 1,2,\cdots,n_{ij}$ for a cell containing data. And let $i = 1,\cdots,a$ and $j = 1,\cdots,b$. Then, in the usual way, define totals and means as follows:

$$n_{i\cdot} = \sum_{j=1}^{b} n_{ij}, \qquad n_{\cdot j} = \sum_{i=1}^{a} n_{ij}, \qquad n_{\cdot\cdot} = N = \sum_{i=1}^{a}\sum_{j=1}^{b} n_{ij} ;$$

$$y_{ij\cdot} = \sum_{k=1}^{n_{ij}} y_{ijk} \quad \text{and} \quad \bar{y}_{ij\cdot} = y_{ij\cdot}/n_{ij} ; \tag{29}$$

$$y_{i\cdot\cdot} = \sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \quad \text{and} \quad \bar{y}_{i\cdot\cdot} = y_{i\cdot}/n_{i\cdot} ;$$

$$y_{\cdot j\cdot} = \sum_{i=1}^{a}\sum_{k=1}^{n_{ij}} y_{ijk} \quad \text{and} \quad \bar{y}_{\cdot j\cdot} = y_{\cdot j\cdot}/n_{\cdot j} ;$$

and

$$y_{\cdot\cdot\cdot} = \sum_{i=1}^{a}\sum_{j=1}^{b}\sum_{k=1}^{n_{ij}} y_{ijk} \quad \text{and} \quad \bar{y}_{\cdot\cdot\cdot} = y_{\cdot\cdot\cdot}/n_{\cdot\cdot} .$$

5.2 Method I

This uses quadratic forms that are adaptations of sums of squares used with balanced data. In many cases those quadratic forms are also sums of squares for unbalanced data, but in some cases they are not.

Example (continued) For balanced data (every n_{ij} equal to n) the sum of squares due to rows is

$$bn\Sigma_i(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = \Sigma_i bn\bar{y}^2_{i\cdot\cdot} - abn\bar{y}^2_{\cdot\cdot\cdot} , \tag{30}$$

where Σ_i represents summation over i from i=1 to i=a; Σ_j and Σ_k are used similarly. For Method I the adaptation of (30) for unbalanced data is to rewrite its right-hand side as $\Sigma_i n_{i\cdot}\bar{y}^2_{i\cdot\cdot} - n_{\cdot\cdot}\bar{y}^2_{\cdot\cdot\cdot}$. This is a sum of squares:

$$\sum_{i=1}^{a} n_{i\cdot}\bar{y}^2_{i\cdot\cdot} - n_{\cdot\cdot}\bar{y}^2_{\cdot\cdot\cdot} = \sum_{i=1}^{a} n_{i\cdot}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot\cdot\cdot})^2 = R(\boldsymbol{\alpha}|\mu) . \tag{31}$$

It is one of the terms used in Method I for the 2-way crossed classification. In similar fashion consider the interaction sum of squares for balanced data:

$$\Sigma_i\Sigma_j n(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2 = \Sigma_i\Sigma_j n\bar{y}^2_{ij\cdot} - \Sigma_i bn\bar{y}^2_{i\cdot\cdot} - \Sigma_j an\bar{y}^2_{\cdot j\cdot} + abn\bar{y}^2_{\cdots}. \quad (32)$$

In the same manner as the right-hand side of (30) was adapted for unbalanced data to be the left-hand side of (31), so also is the right-hand side of (32) adapted to be

$$\Sigma_i\Sigma_j n_{ij}\bar{y}^2_{ij\cdot} - \Sigma_i n_{i\cdot}\bar{y}^2_{i\cdot\cdot} - \Sigma_j n_{\cdot j}\bar{y}^2_{\cdot j\cdot} + n_{\cdot\cdot}\bar{y}^2_{\cdots} . \quad (33)$$

But now, in contrast to (31), where the adaptation *is* a sum of squares, (33) is not. This is so because

$$\Sigma_i\Sigma_j n_{ij}\bar{y}_{ij\cdot} - \Sigma_i n_{i\cdot}\bar{y}^2_{i\cdot\cdot} - \Sigma_j n_{\cdot j}\bar{y}^2_{\cdot j\cdot} + n_{\cdot\cdot}\bar{y}^2_{\cdots}$$

$$= \Sigma_i\Sigma_j n_{ij}(\bar{y}_{ij\cdot} - \bar{y}_{i\cdot\cdot} - \bar{y}_{\cdot j\cdot} + \bar{y}_{\cdots})^2 - 2\Sigma_i\Sigma_j n_{ij}(\bar{y}_{i\cdot\cdot} - \bar{y}_{\cdots})(\bar{y}_{\cdot j\cdot} - \bar{y}_{\cdots}), \quad (34)$$

the last term of which is not necessarily zero. It is always zero when $n_{ij} = n$ (balanced data). Thus (33) is not a sum of squares; indeed, it can be negative. For example, with the following five observations in two rows and two columns

6	4
6,27	12

(33) is easily found to be -14.5.

It is to be noted that the possible negativity of (33) in no way precludes its being used as an element of **q** in (23). And, in Henderson's Method I it *is* used.

Merits and demerits

Method I, like all three Henderson Methods, has the properties of the generalized ANOVA's estimation method detailed in Section 4.3. The estimators have unknown sampling distributions, they *are* unbiased, their sampling variances are quadratic functions of the variance components being estimated and involve very complicated functions of the n_{ij}s [e.g., equa-

tion (21); and see Searle, 1971, Chapter for details of numerous other models]. And, of course, estimates can be negative. In addition, Method I has the following characteristics.

-i. It cannot be used for mixed models. The reason is that the quadratic forms used in Method I are such that for mixed models $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} \neq 0$. But for random models $\boldsymbol{\mu} = \mu\mathbf{1}$ and each $\mathbf{A}_i$ is such that $\mathbf{1}'\mathbf{A}_i\mathbf{1} = 0$ (all elements of $\mathbf{A}$ sum to zero) and so $\boldsymbol{\mu}'\mathbf{A}_i\boldsymbol{\mu} = 0$. Thus Method I is suitable for random, but not mixed, models. If it is used for mixed models (either by assuming the fixed effects are random or by ignoring the fixed effects) it yields biased estimators of the variance components.

-ii. It is easy to calculate. For the particular quadratic forms used in $\mathbf{q}$, derivation of coefficients of the σ^2s in $E(\mathbf{q})$ is well known, is not difficult and requires no matrix manipulations. This was an especially important feature of the method in pre-computer days. But obviously it is getting to be less so nowadays – except in third world situations where modern computers may be less available. Prior to high-speed computers, however, at the time when Henderson (1953) first published his methods, ease of computability was of paramount importance – and in this context Method I was a noteworthy development and in certain situations could still be a viable procedure.

5.3 Method II

Method II is a variation of Method I designed to embody the straightforward arithmetic of Method I but at the same time overcome the deficiency that Method I cannot be used for mixed models. Thus Method II can be used for mixed models. But those models must not contain interactions between fixed effects factors and random effects factors, whether such interactions are treated as fixed or as random. [Proof is given in Searle (1968).]

The details of Method II are very messy, see Searle (1968). The method involves adjusting the data in a particular manner that produces adjusted observations for which the linear model is a completely random model that is exactly the same as the random part of the original mixed model – except for changes to the error terms. Thus, while taking into account those changes to the error terms, the random part of the mixed model can be used for the adjusted data and on that basis Method I can be applied. An elementary description of this is as follows.

Write the model equation for $\mathbf{y}$ as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \ , \tag{35}$$

where $\mathbf{1}$ is a vector of N unities, $\boldsymbol{\beta}$ is the vector of fixed effects, $\mathbf{u}$ is the vector of random effects and $\mathbf{e}$ is the vector of error terms. $\mathbf{X}$ and $\mathbf{Z}$ are known incidence matrices corresponding to the fixed and random effects, respectively; and μ is a general mean. Then the general procedure of Method II is that it uses

$$\mathbf{z} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} \tag{36}$$

for a $\hat{\boldsymbol{\beta}}$ such that the model equation for $\mathbf{z}$ is

$$\mathbf{z} = \mu^{*}\mathbf{1} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon} \tag{37}$$

with $\boldsymbol{\varepsilon} = \mathbf{W}\mathbf{e}$ for $\mathbf{W}$ that depends on the way in which $\hat{\boldsymbol{\beta}}$ is obtained, as does μ^{*} also, a scalar that is different from μ. So long as μ^{*} is scalar, its form is of no interest. Then, since $\mathbf{Z}\mathbf{u}$ is the same in the model equation (37) for $\mathbf{z}$, as it is in (35) for $\mathbf{y}$, Method I on $\mathbf{z}$ is the same as on $\mathbf{y}$, apart from taking account of $\boldsymbol{\varepsilon}$ being $\mathbf{W}\mathbf{e}$ rather than $\mathbf{e}$. The question is "how is $\hat{\boldsymbol{\beta}}$ derived in order to achieve this?" Henderson (1953) shows this largely by means of an example; Searle (1968) gives a general description. Henderson *et al.* (1974) show how to calculate $\mathbf{W}$ for $\boldsymbol{\varepsilon} = \mathbf{W}\mathbf{e}$, for this affects the coefficients of σ^2_e in E(q) applied to (37).

The precise way in which $\hat{\boldsymbol{\beta}}$ of (36) is to be calculated so that (37) applies is what makes the details of Method II difficult to describe. Moreover, for a number of years this was thought to render Method II non-unique for any set of data; but Henderson *et al.* (1974) showed otherwise. Thus Method II is a perfectly viable special case of the generalized ANOVA method for unbalanced data.

Merits and demerits

Method II does, of course, suffer the defects already described for all ANOVA estimators from unbalanced data. Additionally,

-i. it can be used for mixed models provided they involve no interactions between fixed and random factors;

-ii. computation of $\mathbf{z} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$ requires care, but after that the computation is as easy as is that of Method I, save for coefficients of $\hat{\sigma}^2_e$ in the estimation equations;

-iii. no analytical expressions are available for sampling variances of estimators.

5.4 Method III

Method III borrows from the analysis of fixed effects models. As elements of $\mathbf{q}$ it uses the sums of squares due to fitting one's model and various submodels of it. Unfortunately, with many models there are several possible sets of sums of squares (for a given data set) that could be used; and usually no method for deciding between one set and another.

As an example, consider fitting the over-parameterized model

$$E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij}$$

for the 2-way classification of Section 5.1. Suppose we are analyzing first lactation records of daughters of dairy sires in an artificial

insemination stud that has been used in a large number of herds. Then y_{ijk} could be the record of daughter k by sire j in herd i: μ would be a general mean, α_i a herd effect, β_j a sire effect and γ_{ij} an interaction effect. Reductions in sums of squares are denoted

$$\begin{aligned} R(\mu,\alpha) \quad &\text{for fitting} \quad E(y_{ijk}) = \mu + \alpha_i \\ R(\mu,\beta) \quad &\text{for fitting} \quad E(y_{ijk}) = \mu + \beta_j \\ R(\mu,\alpha,\beta) \quad &\text{for fitting} \quad E(y_{ijk}) = \mu + \alpha_i + \beta_j \end{aligned}$$

and

$$R(\mu,\alpha,\beta,\gamma) \quad \text{for fitting} \quad E(y_{ijk}) = \mu + \alpha_i + \beta_j + \gamma_{ij} \ .$$

Differences between these sums of squares are defined as

$$R(\alpha|\mu) = R(\mu,\alpha) - R(\mu) \qquad \text{and} \qquad R(\beta|\mu) = R(\mu,\beta) - R(\mu)$$

$$R(\alpha|\mu,\beta) = R(\mu,\alpha,\beta) - R(\mu,\beta) \quad \text{and} \quad R(\beta|\mu,\alpha) = R(\mu,\alpha,\beta) - R(\mu,\alpha) \ ;$$

and (40)

$$R(\gamma|\mu,\alpha,\beta) = R(\mu,\alpha,\beta,\gamma) - R(\mu,\alpha,\beta)$$

with

$$SSE = \Sigma_i\Sigma_j\Sigma_k y^2_{ijk} - R(\mu,\alpha,\beta,\gamma) \ .$$

Detailed calculation formulae for these sums of squares are available in many places, e.g., Searle (1971, p. 298 and 1987, p. 352).

Method III uses these sums of squares as elements of $\mathbf{q}$ in $E(\mathbf{q}) = \mathbf{C}\boldsymbol{\alpha}^2$. The difficulty is that there are six of these sums of squares in the random model form of (38) and there are only four variance components: σ^2_α, σ^2_β, σ^2_γ and σ^2_e. The question immediately arises: which sums of squares shall be used? At least three sets of sums of squares have been suggested in the literature:

Set 1	Set 2	Set 3
$R(\alpha\|\mu)$	$R(\beta\|\mu)$	$R(\alpha\|\mu,\beta)$
$R(\beta\|\mu,\alpha)$	$R(\alpha\|\mu,\beta)$	$R(\beta\|\mu,\alpha)$
$R(\gamma\|\mu,\alpha,\beta)$	$R(\gamma\|\mu,\alpha,\beta)$	$R(\gamma\|\mu,\alpha,\beta)$
SSE	SSE	SSE
SST_m	SST_m	

In each of Sets 1 and 2, the sums of squares add to the total sum of squares corrected for the mean, $SST_m = \Sigma_i\Sigma_j\Sigma_k(y_{ijk} - \bar{y}_{...})^2$. Those in Set 3 do not have this "adding up" property. The problem is that there is no theoretical basis whatever for deciding which of Sets 1 and 2 is to be preferred, nor for deciding if "adding up" is actually a useful property, nor for deciding if Set 3 is preferable to Sets 1 and 2. Moreover, there are obviously other sets that could be suggested: e.g., $R(\alpha|\mu)$, $R(\beta|\mu)$, $R(\gamma|\mu,\alpha,\beta)$ and SSE; and so on. Sets 1 and 2 have traditionally been given consideration because they do have that "adding up" property. But then in any p-way classification model, for $p > 2$, so do many more different sets. Even without interactions of any sort, a p-way classification random model has p! different partitionings of its SST_m akin to Sets 1 and 2 in Table 1; and each could be used in Method III to estimate the relevant variance components. Thus in unbalanced data classified by 3, 4 or 5 factors, even without any interactions, there would be 6, 24 and 120 different sets of sums of squares, respectively, that could each constitute a Method III estimation procedure. Clearly, then, for many data sets Method III is not a uniquely defined procedure. And (aside from simulation, with all its attendant difficulties) there is no way of deciding which of the many alternative Method III procedures that are available for a data set are to be preferred. This lack of uniqueness of Method III is a distinctly unsatisfactory characteristic.

Of course when several forms of Method III are available, as in Table 1, some of the resulting estimators will be the same from one form to another. In Table 1, $E(SSE) = (N - s)\sigma_e^2$ when there are s cells of the 2-way classification that have data in them, and

$$E\ R(\gamma|\mu,\alpha,\beta) = t\sigma_\gamma^2 + (s-1)\sigma_e^2 \qquad (41)$$

for some t. Therefore in all of Sets 1, 2 and 3 in Table 1 the estimators of σ^2_e and of σ^2_γ are the same. Moreover, the estimators of σ^2_β in Set 1 and of σ^2_α in Set 2 are the same as their counterparts in Set 3.

Merits and demerits

As with Methods I and II, so with III: it has the same general weaknesses as have been described for all ANOVA estimation methods. Its further features are as follows.

-i. It is a viable method for any mixed model – unlike Method I which cannot be used for mixed models at all, or Method II which is suited to only those mixed models that have no interactions between fixed and random factors.

-ii. Computing the $R(\cdot|\cdot)$ terms can be lengthy. All such terms except those like (31) involve inverting matrices which, when one has voluminous amounts of data as animal breeders often do, then those matrices can have very large dimensions, so that inverting them demands very up-to-date computing facilities.

-iii. Sampling variances can be calculated arithmetically, through a series of matrix operations and with using estimated values for the variance components, but specific analytical expressions are not available. General methodology for this is given in Rohde and Tallis (1969).

5.5 Relationships between methods

As has been said, all three Henderson methods reduce, for balanced data, to the standard ANOVA method for such data. One well might ask if other equivalences exist. Essentially there are none, other than like those just discussed for the model (38), or for trivially simple models. This is so for the kind of reason that although for the 2-way classification the Method I quadratic

$$\Sigma_i n_{i.}(\bar{y}_{i..} - \bar{y}_{...})^2 = R(\alpha|\mu)$$

of Method III, and similarly for $R(\beta|\mu)$, and SSE is the same in both methods, the inequalities

$$R(\alpha|\mu) \neq R(\alpha|\mu,\beta)$$

and

$$\Sigma_i\Sigma_j n_{ij}\bar{y}^2_{ij\cdot} - \Sigma_i n_{i\cdot}\bar{y}^2_{i\cdot\cdot} - \Sigma n_{\cdot j}\bar{y}^2_{\cdot j\cdot} + n_{\cdot\cdot\cdot}\bar{y}^2_{\cdot\cdot\cdot} \neq R(\gamma|\mu,\alpha,\beta)$$

ensure that for unbalanced data Method I and Method III estimators are not generally the same. And Method II, because its estimation is based on $\mathbf{z} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}$, yields estimators that are different again from those of Methods I and II.

5.6 Computing package output

The only widely-used statistical computing package that yields some ANOVA estimates is the VARCOMP routine in the SAS package from the SAS Institute, Cary, North Carolina. Those estimates come from a version of Method III. However, the RANDOM routine in the same package yields expected mean squares for all four of the Types 1-4 sums of squares calculated by the SAS GLM routine. Judicious use of the Types 1 and 2 sums of squares and of their expected values can readily produce Method I and Method III estimates of variance components.

A package commonly known as HARVEY also generates what purports to be Method III estimates. It is based on the general result that comes from rewriting the model equation (38) as

$$\mathbf{y} = \mu\mathbf{1} + \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e} \ ,$$

where $\mathbf{u}_1$ represents the random effects for just a single factor in the model, their variance being σ^2_1, say; and $\mathbf{u}_2$ represents all the other random effects. Then

$$E\ R(\mathbf{u}_1|\mu,\boldsymbol{\beta},\mathbf{u}_2) = p_1\sigma^2_1 + p_2\sigma^2_e$$

for some values p_1 and p_2. This result is similar to (41), and is described more generally in Searle [1971, p. 445, equation (79)]. Its use in the form

$$\hat{\sigma}_1^2 = [R(\mathbf{u}_1|\mu,\boldsymbol{\beta},\mathbf{u}_2) - p_2\hat{\sigma}_e^2]/p_1$$

with $\hat{\sigma}^2$ = MSE, does demand, however, that $R(\mathbf{u}_1|\mu,\boldsymbol{\beta},\mathbf{u}_2)$ be calculated correctly. A method of calculation that can lead to wrong results is that of inappropriate use of Σ-restricted models as illustrated, for example, in Searle and Henderson (1980, 1983).

6. OTHER SPECIAL CASES OF ANOVA ESTIMATION

As has been emphasized, there is an infinite number of ways of using the generalized ANOVA method of estimation summarized in (23). Two that are sometimes touted for all-cells-filled data are based on the sum of squares of the weighted squares of means analysis, and of the analysis of unweighted means.

For the 2-way crossed classification, the weighted squares of means utilizes for $E(\mathbf{q}) = \mathbf{C}\boldsymbol{\sigma}^2$ sums of squares such as

$$E\ \Sigma_i w_i(\tilde{y}_i - \Sigma_i w_i\tilde{y}_i/\Sigma w_i)^2 = (\Sigma w_i - \Sigma_i w_i^2/\Sigma w_i^2)(\sigma_\alpha^2 + \sigma_\gamma^2/b) + (a-1)\sigma_e^2 \quad (42)$$

where

$$1/w_i = (\Sigma_j 1/n_{ij})/b^2 \qquad \text{and} \qquad \tilde{y}_i = \Sigma_j \bar{y}_{ij\cdot}/b\ .$$

The left-hand side of (42) is the sum of squares originally suggested (for fixed effects models) by Yates (1934). It is calculated as a Type 3 sum of squares by the SAS GLM routine: but *only* for all-cells-filled data.

The unweighted means analysis uses terms such as

$$E\ b\Sigma_i(\tilde{y}_i - \Sigma_i\tilde{y}_i/a)^2 = (a-1)(b\ \sigma_\alpha^2 + \sigma_\gamma^2 + n_h\sigma_e^2)$$

$$n_h = \Sigma_i\Sigma_j(1/n_{ij})/ab\ .$$

Full details are available, for example, in Searle (1971, p. 452); and further extensions are given in Gosslee and Lucas (1965).

Just these two variants alone illustrate the limitless possibilities for applying the generalized ANOVA method; and all of them suffer the shortcomings detailed in Section 3.2.

7. A GENERAL MIXED MODEL

For describing further methods of estimating variance components, it is necessary to describe some alternative notations for a general linear model. We begin with the vector of error terms $\mathbf{e}$ in the model equation (35). It can be expressed as $\mathbf{e} = \mathbf{Z}_0\mathbf{u}_0$ with $\mathbf{u}_0 = \mathbf{e}$ and $\mathbf{Z}_0 = \mathbf{I}$. Then (35) can be rewritten as

$$\mathbf{y} = [\mathbf{1} \quad \mathbf{X}]\begin{bmatrix}\mu \\ \boldsymbol{\beta}\end{bmatrix} + [\mathbf{Z}_0 \quad \mathbf{Z}]\begin{bmatrix}\mathbf{u}_0 \\ \mathbf{u}\end{bmatrix} . \tag{43}$$

On redefining the symbols $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}$ and $\mathbf{u}$ to respectively represent the partitioned matrices and vectors on the right-hand side of (43), that equation becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} . \tag{44}$$

The meanings of $\mathbf{X}$, $\boldsymbol{\beta}$, $\mathbf{Z}$ and $\mathbf{u}$ shall, for the remainder of this paper, except for Section 10, be as they are in (44); i.e., $\mathbf{X}$ of (44) is $[\mathbf{1} \quad \mathbf{X}]$ of (43), and $\mathbf{Z}$ of (44) is $[\mathbf{Z}_0 \quad \mathbf{Z}]$ of (43).

$\mathbf{u}$ is a vector of random effects, its first N elements being the error terms, $\mathbf{e}$. Other elements of $\mathbf{u}$ are the random effects corresponding to the levels of the random factors that occur in the data. Let $\mathbf{u}_i$ for $i = 1,\cdots,r$ be the vector of those effects for factor i, be it a main effect, nested or interaction, random factor. Then $\mathbf{u}$ is partitioned into r subvectors $\mathbf{u}_0, \mathbf{u}_1,\cdots,\mathbf{u}_r$. The distributional properties attributed to

these sub-vectors are matrix forms of those that are customary for random effects, as in (3) and (4), namely $E(\mathbf{u}_i) = \mathbf{0}$ and $var(\mathbf{u}_i) = \sigma_i^2 \mathbf{I}_{t_i}$ where t_i is the number of elements in $\mathbf{u}_i$ (the number of levels of the i'th random factor occurring in the data, with $t_0 = N$); and for each pair of sub-vectors, $cov(\mathbf{u}_i, \mathbf{u}_j) = \mathbf{0}$ for $i \neq j$.

$\mathbf{Z}$ in (44) is now partitioned conformably for the product $\mathbf{Zu}$: thus $\mathbf{Z} = [\mathbf{Z}_0 \ \mathbf{Z}_1 \ \cdots \ \mathbf{Z}_r]$ with $\mathbf{Z}_0 = \mathbf{I}_N$. Then (44) becomes

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{u}_i \ , \tag{45}$$

with variance-covariance matrix

$$var(\mathbf{y}) = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' \sigma_i^2 = \mathbf{V}, \text{ say} \ . \tag{46}$$

Succinct description of further methods of estimation involves the $\mathbf{Z}_i$-matrices and also the matrix

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \ , \tag{47}$$

where $(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}$ is a generalized inverse of $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$, satisfying $\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$.

8. ML, REML, AND MINQUE ESTIMATION

The method of maximum likelihood estimation is well established (Fisher, 1922) in the statistical literature and has many good properties. It does, however, demand assuming some known form of distribution function for the data vector. And this, be it noticed, is not required for any form of ANOVA estimation already considered. We here summarize two applications of maximum likelihood estimation (ML and REML), based on normality; and a third method of estimation, MINQUE, which demands no distribution assump-

tions but which is nevertheless closely connected to REML methodology. Following a brief description (since details are voluminous) of each of the three methods is a discussion of their comparative merits.

8.1 Maximum Likelihood (ML)

The model equation (44) is essentially the same as that first formulated in Hartley and Rao (1967). That paper, assuming normality of $\mathbf{y}$, namely $\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$ for $\mathbf{X}\boldsymbol{\beta}$ of (44) and $\mathbf{V}$ of (46), derived equations for maximum likelihood estimation of σ^2. The derivation is lengthy, but on using the notation

$$\left\{ \underset{m}{a_{ij}} \right\}_{i,j=0}^{r} \quad \text{for a square matrix of order } r+1 \text{ of elements } a_{ij}$$

and

$$\left\{ \underset{c}{b_i} \right\}_{i=0}^{r} \quad \text{for a column vector of order } r+1 \text{ of elements } b_i \ ,$$

the equations can ultimately be stated as

$$\left\{ \underset{m}{\mathrm{tr}(\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{V}}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\tilde{\mathbf{V}}^{-1})} \right\}_{i,j=0}^{r} \tilde{\sigma}^2 = \left\{ \underset{c}{\mathbf{y}'\tilde{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\tilde{\mathbf{P}}\mathbf{y}} \right\}_{i=0}^{r} . \qquad (48)$$

These equations are very non-linear in the variance components estimators $\tilde{\boldsymbol{\sigma}}^{2\prime} = [\tilde{\sigma}_0^2 \ \tilde{\sigma}_1^2 \ \cdots \ \tilde{\sigma}_r^2]$. This is so because they are involved in $\mathbf{V} = \Sigma_i \mathbf{Z}_i\mathbf{Z}_i'\tilde{\sigma}_i^2$ which occurs as $\tilde{\mathbf{V}}^{-1}$ in (48), both explicitly and in $\tilde{\mathbf{P}}$ which is $\mathbf{P}$ of (47) with $\mathbf{V}^{-1}$ replaced by $\tilde{\mathbf{V}}^{-1}$. Even in the 1-way classification, equations (48) do not simplify to anything tractable (e.g., Searle, 1971, p. 463). Neither do they even for the balanced data case of the 2-way crossed classification, random model with interaction (see Miller, 1973).

So equations (48) have to be solved numerically, using iteration. This is discussed in Section 8.4.

8.2 Restricted maximum likelihood (REML)

Variance components are characteristics only of $\mathbf{u}$ in the model equation $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$. This suggests estimating those components from functions of $\mathbf{y}$, say $\mathbf{K}'\mathbf{y}$, that do not involve $\boldsymbol{\beta}$. To do this, $\mathbf{K}'$ must be chosen so that $\mathbf{K}'\mathbf{X} = \mathbf{0}$. Elements of $\mathbf{K}'\mathbf{y}$ have been called error contrasts, by Harville (1977); and maximum likelihood applied to $\mathbf{K}'\mathbf{y}$, first suggested by Patterson and Thompson (1971), is what is now called restricted maximum likelihood (REML). Again, considerable detail is involved in deriving the estimation equations, although once obtained they bear close resemblance to those in (48) for ML. Using $\dot{\sigma}^2$ to represent the REML estimations, the REML equations are

$$\left\{ \operatorname*{m}_{} \operatorname{tr}(\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{Z}_j\mathbf{Z}_j'\dot{\mathbf{P}}) \right\}_{i,j=0}^{r} \dot{\sigma}^2 = \left\{ \operatorname*{c}_{} \mathbf{y}'\dot{\mathbf{P}}\mathbf{Z}_i\mathbf{Z}_i'\dot{\mathbf{P}}\mathbf{y} \right\}_{i=0}^{r} . \qquad (49)$$

It is easily seen that these equations have the same right-hand side as the ML equation in (48); and the left-hand side of (49) has $\mathbf{P}$ where (48) has $\mathbf{V}^{-1}$. As with (48), so with (49): they have to be solved numerically, usually by iteration.

8.3 Minimum norm estimation (MINQUE)

A series of papers by LaMotte (1970, 1973) and Rao (1970, 1971a, b, 1972) propounded ideas of minimum norm estimation, of which minimum norm quadratic unbiased estimation is the best known. Without making distribution assumptions about the data, it is based on seeking quadratic forms $\mathbf{y}'\mathbf{A}\mathbf{y}$ to estimate variance components in such a way that (i) $\mathbf{A}$ is symmetric, because of its occurrence in a quadratic form, (ii) $\mathbf{A}\mathbf{X} = \mathbf{0}$ so as to have $\mathbf{y}'\mathbf{A}\mathbf{y}$ free of the fixed effects, (iii) $\mathbf{y}'\mathbf{A}\mathbf{y}$ is unbiased, and (iv) a Euclidean norm that is tantamount to a generalized variance is minimized. Once more, as with deriving the ML and REML equations, so here: details

are extensive. But the result is easily stated. It demands taking some pre-assigned, numerical weights $[w_0\ w_1\ \cdots w_r] = \mathbf{w}$, say, and using them to calculate

$$\mathbf{V}_w = \sum_{i=0}^{r} \mathbf{Z}_i \mathbf{Z}_i' w_i \ ,$$

wherein $\mathbf{V}_w$ is $\mathbf{V}$ with σ^2 replaced by $\mathbf{w}$. Then the MINQUE equations are

$$\left\{ \operatorname{tr}(\mathbf{Z}_i \mathbf{Z}_i' \mathbf{P}_w \mathbf{Z}_j \mathbf{Z}_j' \mathbf{P}_w) \right\}_{i,j=0}^{r} \ddot{\sigma}^2 = \left\{ \mathbf{y}' \mathbf{P}_w \mathbf{Z}_i \mathbf{Z}_i' \mathbf{P}_w \mathbf{y} \right\}_{i=0}^{r} \ . \qquad (50)$$

These equations are similar in form to those of ML and REML, but with certain important differences as now noted.

8.4 Characteristics of the estimation equations

Several features of the three sets of equations, for ML in (48), for REML in (49) and for MINQUE in (50) are worth noting.

-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 $\operatorname{tr}(\mathbf{T}\mathbf{T}')$ for some $\mathbf{T}$ and so can be calculated as the sum of squares of elements of $\mathbf{T}$.

-iii. Each set of equations has a right-hand side that is a vector of quadratic forms in the observations; indeed, for the appropriate $\mathbf{P}$ in each case the right-hand side is a vector of sums of squares of elements of $\mathbf{Z}_i'\mathbf{P}\mathbf{y}$, for $i = 0,1,\cdots,r$.

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

-v. Each of the ML and the REML sets of 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 **V**, how does one ensure that it is always positive-definite? And if it is not, what remedial steps are to be taken, if any? If, after some round of iteration, the numerical value $\hat{\sigma}^2_t$ to be given to σ^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 theory behind maximum likelihood estimation demands that maximization be over the range of the parameters, and that the estimators be in that range. Thus negative values cannot be estimates of variance components.) 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? At least one package (REML, Scottish Agricultural Statistics Service, Edinburgh)

does something: instead of a negative $\hat{\sigma}_t^2$ being changed to zero it is changed to a small positive value.

-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 estimate.

-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),

$$\text{a MINQUE} = \text{a first iterate of REML} . \tag{51}$$

-viii. Solutions to the MINQUE equations depend on the pre-assigned $\mathbf{w}$; they are unbiased, of course, but some elements of the solution vector may be negative. There is nothing in MINQUE methodology to prevent this.

-ix. Large-sampling variances and covariance of ML estimators come from the information matrix:

$$\text{var}(\tilde{\boldsymbol{\sigma}}^2) = 2\left[\left\{ \underset{m}{\text{tr}}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{V}^{-1}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{V}^{-1})\right\}_{i,j=0}^{r}\right]^{-1} . \tag{52}$$

To use this $\mathbf{V}^{-1}$ needs to be replaced by an estimate: $\tilde{\mathbf{V}} = \Sigma_i \mathbf{Z}_i\mathbf{Z}_i'\tilde{\sigma}_i^2$ is the obvious candidate. And for REML $\text{var}(\dot{\boldsymbol{\sigma}})$ is (52) with $\mathbf{P}$ replacing $\mathbf{V}^{-1}$. With MINQUE

$$\text{var}(\ddot{\boldsymbol{\sigma}}^2) = 2\mathbf{F}_w^{-1}\left\{ \underset{m}{\text{tr}}(\mathbf{P}_w\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_w\mathbf{V}\mathbf{P}_w\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}_w\mathbf{V})\right\}_{i,j=0}^{r}\mathbf{F}_w^{-1} \tag{53}$$

for

$$\mathbf{F}_w = \left\{ \underset{m}{\text{tr}}(\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}_w\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}_w)\right\}_{i,j=0}^{r} .$$

If $\mathbf{V}_w^{-1}$ is used in (53) in place of $\mathbf{V}$, then (53) reduces to $\mathbf{F}_w$.

8.5 Assessing the three methods

ANOVA estimation has already been discussed at some length. Its lack of optimality criteria on which to pass judgment 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. But the 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 σ^2. This means that for several people all with the same data, but each person using a different $\mathbf{w}$, there can be several different estimated σ^2-vectors. Somehow this cannot be seen as 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 σ^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. It is identical (Hocking and Kutner, 1975) to iterating the REML equations but ignoring the non-negativity requirement for estimates obtained by REML. And, of course, REML estimation is based on normality assumptions. Even without those assumptions Brown (1976) has shown that I-MINQUE has a limiting distribution that is normal. All this makes for the conclusion of favouring REML over MINQUE.

Then comes the question of 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 those 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, as in (1), the REML solutions are $\dot{\sigma}^2_e = \text{MSE}$ and $\dot{\sigma}^2_\alpha = (\text{MSA} - \text{MSE})/n$, as in (8). 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 σ^2_α is zero and that of σ^2_e is $\text{SST}_m/(an - 1)$ for $\text{SST}_m = \Sigma_i \Sigma_j (y_{ij} - \bar{y}_{..})^2$; 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 (49) 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 σ^2_α that is clearly upwardly biased, as can also be shown for the estimator of σ^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 σ^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 ML estimation for fixed effects. The REML method provides no such estimator, although intuitively one would be inclined to use the REML estimates of the variance components to adapt the ML estimator of the fixed effects in an obvious way. This is discussed in Section 9.

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 have a long history of well-established, large-sample properties. True, the development of (48) and (49) rests on normality assumptions for the data; and, as already discussed, there are the computational difficulties associated with solving non-linear 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.

9. LINEAR ESTIMATION IN MIXED MODELS

Although this paper is about variance component estimation, it would be incomplete without brief mention of linear estimation in mixed models, namely estimation of the fixed effects and prediction of the random effects.

9.1 Estimating fixed effects

To estimate estimable functions of the fixed effects that are elements of $\boldsymbol{\beta}$ in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ 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 of random effects in the model, and so is of no interest. Limiting attention to $\mathbf{V}$ being non-singular (which is not very restrictive because in most applications this will be true), the best linear unbiased estimator of $\mathbf{X}\boldsymbol{\beta}$ is

$$\mathbf{X}\boldsymbol{\beta}^{o} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \ . \qquad (54)$$

As an estimator of $\mathbf{X\beta}$, this 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 known. The "obvious" thing to do is to estimate σ^2, as $\hat{\sigma}^2$, say, use it in place of σ^2 in $\mathbf{V}$ to have

$$\hat{\mathbf{V}} = \Sigma_i \mathbf{Z}_i \mathbf{Z}_i' \hat{\sigma}^2_i \tag{55}$$

and then calculate

$$\mathbf{X}\boldsymbol{\beta}^{\circ}_{\hat{\mathbf{V}}} = \mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \ . \tag{56}$$

Equation (56) is not a best linear unbiased estimator. But for a wide class of estimators of $\hat{\sigma}^2$, Kackar and Harville (1981, 1984) have shown that it is unbiased, and that its sampling variance can be calculated.

An even nicer result concerns maximum likelihood. With the ML estimator of σ^2 being $\tilde{\sigma}^2$, use $\tilde{\sigma}^2$ in (55) and (56) to yield

$$\mathbf{X}\boldsymbol{\beta}^{\circ}_{\tilde{\mathbf{V}}} = \mathbf{X}(\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{X})^{-}\mathbf{X}'\tilde{\mathbf{V}}^{-1}\mathbf{y} \ . \tag{57}$$

Then $\mathbf{X}\boldsymbol{\beta}^{\circ}_{\tilde{\mathbf{V}}}$ is the ML estimator of $\mathbf{X\beta}$, with all the usual properties attendant to ML estimation. A third possibility is to use the REML estimator $\dot{\sigma}^2$ of σ^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 $\tilde{\mathbf{V}}$ and $\dot{\mathbf{V}}$ is that, for data that are not normally distributed, procedures for deriving ML and REML estimators of σ^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.

9.2 Predicting random effects

Any random effect that occurs in the data is not actually a random variable, but it a realization of a random variable. Nevertheless, it is usually unobservable: e.g., the genetic value of the dairy cow, Daisy,

whose first lactation milk production has been recorded as 5,000 kilograms; or the true intelligence of the schoolboy, Tom Brown, who has scored 130 on an I.Q. test. Whilst we cannot measure these realizations we hesitate to speak of estimating them, since estimation of random variables is counter-intuitive statistically. But we can think of predicting these values, in the following sense: of all cows (of the same breed and age as Daisy) that had first lactation milk production in 1987 of 5,000 kg, as did Daisy, what is their average genetic value? Similarly, of all schoolboys (of the same age as Tom Brown) who scored 130 on the same I.Q. test, what is their average true intelligence?

It is an interesting historical aside that an early occurrence (Henderson, 1955) of a question of this nature seems to have been as a classroom exercise used by A.M. Mood in the late 1940s at Iowa State University, and subsequently appearing in Mood (1950, p. 164, exercise 23) 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:

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

What is wanted 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

$$E(u_i|y_{ij} = 130) = 100 + \frac{15^2}{15^2 + 5^2}(130 - 100) = 127 \neq 130 . \qquad (58)$$

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

It is because genetic merit of a dairy cow and true intelligence of a human cannot be observed, and because each is a random variable throughout the populations of dairy cows and schoolboys, respectively, that we speak of predicting these values. In each case this amounts to predicting **u** in the model $\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu}$. Then, with $E(\mathbf{u}) = 0$, and defining

$$\text{var}(\mathbf{u}) = \mathbf{D} = \left\{ \underset{d}{} \sigma_i^2 \mathbf{I}_{t_i} \right\}_{i=0}^{r} \qquad (59)$$

$$= \text{block diagonal matrix of matrices } \sigma_i^2 \mathbf{I}_{t_i} \text{ for } i = 0, \cdots, r ,$$

we have

$$\text{cov}(\mathbf{y},\mathbf{u}') = \mathbf{ZD} = \mathbf{C}, \text{ say, and } \text{var}(\mathbf{y}) = \mathbf{ZDZ}' \ . \tag{60}$$

Hence the joint distribution of $\mathbf{y}$ and $\mathbf{u}$ is

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

Now if $\tilde{\mathbf{u}}$ denotes a predictor of $\mathbf{u}$, $E(\mathbf{u} - \tilde{\mathbf{u}})^2$ is its mean squared error. Choosing $\tilde{\mathbf{u}}$ so that $E(\mathbf{u} - \tilde{\mathbf{u}})^2$ is minimized yields for $\tilde{\mathbf{u}}$ what is known as the best predictor. It is

$$\tilde{\mathbf{u}} = E(\mathbf{u}|\mathbf{y}) \ . \tag{61}$$

Derivation of this result is available in many places, e.g., Cochran (1951), Rao (1965, pages 79 and 220-2) and Searle (1974). Thus we find that the best predictor, in the sense of minimum mean squared error is $E(\mathbf{u}|\mathbf{y})$, the conditional mean of $\mathbf{u}$, given $\mathbf{y}$; (58) is an example.

If, in addition to this property of being best, one also demands that the predictor be a linear function of elements of $\mathbf{y}$, i.e., of the observations, then one has the best linear predictor which is

$$\tilde{\mathbf{u}}_L = E(\mathbf{u}) + \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta}) \ . \tag{62}$$

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. Moreover, when that distribution is the normal distribution we have (61) and (62) being equal; i.e., under normality $\tilde{\mathbf{u}} = \tilde{\mathbf{u}}_L$.

Henderson (1963) extends these results by showing that the best linear unbiased predictor of a combination of the fixed and random effects, $\mathbf{w} = \mathbf{T}'\mathbf{X\beta} + \mathbf{u}$, is

$$\tilde{\mathbf{w}} = \mathbf{T}'\mathbf{X\beta}^{\circ} + E(\mathbf{u}) + \mathbf{DZ}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta}^{\circ}) \ . \tag{63}$$

where $\mathbf{X}\boldsymbol{\beta}^{\circ} = \mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$. This is what animal breeders refer to as BLUP – best linear unbiased prediction.

9.3 The mixed model equations

As formulated in (44), the $\mathbf{u}$ in $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u}$ includes the error terms $\mathbf{e} = \mathbf{Z}_0\mathbf{u}_0 = \mathbf{u}_0$ since $\mathbf{Z}_0 = \mathbf{I}_N$. We return to the definition of $\mathbf{Zu}$ that *excludes* $\mathbf{Z}_0\mathbf{u}_0$, as in (35), so that $\mathbf{u}$, $\mathbf{Z}$ and $\mathbf{D}$ are now

$$\mathbf{u} = \left\{ \begin{matrix} \\ \mathbf{u}_i \end{matrix} \right\}_{i=1}^{r}, \quad \mathbf{Z}' = \left\{ \begin{matrix} \\ \mathbf{Z}_i' \end{matrix} \right\}_{i=1}^{r}, \quad \mathbf{D} = \operatorname{diag}\left\{ \sigma_i^2 \mathbf{I}_{t_i} \right\}_{i=1}^{r} . \tag{64}$$

and with these definitions the model equation is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e} \quad \text{and} \quad \mathbf{V} = \mathbf{Z}\mathbf{D}\mathbf{Z}' + \sigma_e^2 \mathbf{I}_N . \tag{65}$$

Then, with $E(\mathbf{u}) = \mathbf{0}$ it is easily established from (62) that the corresponding predictor of $\mathbf{u}$ of (64) is

$$\hat{\mathbf{u}} = \mathbf{D}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{66}$$

for $\mathbf{u}$, $\mathbf{Z}$ and $\mathbf{D}$ as in (64). Without discussing details of derivation, it can then be shown that $\boldsymbol{\beta}^{\circ}$ for $\mathbf{X}\boldsymbol{\beta}^{\circ}$ of (54), and $\hat{\mathbf{u}}$ of (66), can be obtained as solutions to

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{Z}'\mathbf{X} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \sigma_e^2 \mathbf{D}^{-1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}^{\circ} \\ \hat{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{bmatrix} . \tag{67}$$

These are the equations that are well known as Henderson's mixed model equations (e.g., Henderson *et al.*, 1959, and Henderson, 1963).

Verification that (67) yields (54) and (66) is easily established from the identity

$$\mathbf{V}^{-1} = (\mathbf{Z}\mathbf{D}\mathbf{Z}' + \sigma_e^2 \mathbf{I})^{-1} = (1/\sigma_e^2)[\mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z} + \sigma_e^2 \mathbf{D}^{-1})^{-1}\mathbf{Z}']$$

so that

$$\mathbf{Z}'\mathbf{V}^{-1} = \mathbf{D}^{-1}(\mathbf{Z}'\mathbf{Z} + \sigma_e^2 \mathbf{D}^{-1})^{-1}\mathbf{Z}' .$$

Derivation of (66) is also available from maximizing a distribution function (Henderson *et al.*, 1959), or from a Bayesian perspective (Dempfle, 1977), or from a regression viewpoint (Gianola and Goffinet, 1982).

9.4 The need for variance components estimates

All the results of this section are useful, naturally. But to be applicable to real-life situations they demand a numerical value for σ^2, and thence for **D** and **V**. Thus are we driven to estimating variance components.

The seemingly "obvious" way of using estimated components is to use them in place of elements of σ^2 in **D** and **V** and thence in β^o and so in (67). Properties of the resulting expressions have been considered by Jeske and Harville (1986). Although $\tilde{u}_L$ and $\tilde{w}$ can then be calculated using estimated **D** and **V**, as can their sampling variances, their distributions are not known, so giving rise to difficulties in interval estimation.

10. REFERENCES

Ahrens, H. (1965). Standardfehler geschatzter Varianzkomponenten eines unbalanzie Versuchplanes in r-stufiger hierarchischer Klassifikation. *Monatsb. Deut. Akad. Wiss. Berlin*, **7**, 89-94.

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.

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

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. (1966). Variances of estimates of variance components in a three-way classification. *Biometrics* **22**, 553-565.

Blischke, W. R. (1968). Variances of moment estimators of variance components in the unbalanced r-way classification. *Biometrics* **24**, 527-540.

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.

Dempfle, L. (1977). Relation entre BLUP (best linear unbiased estimation) et estimateurs bayésiens. *Annales Genetiques Selection Animales*, **9**, 27.

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. (1922). On the mathematical foundations of theoretical statistics. *Transactions of the Royal Society of London, A,222*, 309-368.

Fisher, R. A. (1925, 1938), *Statistical Methods for Research Workers*, 1st and 7th Editions. 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.

Gianola, D. and Goffinet, B. (1982). Sire evaluation with best linear unbiased predictors. *Biometrics* **38**, 1085-6.

Gosslee, D. G. and Lucas, H. L. (1965). Analysis of variance of disproportionate data when interaction is present. *Biometrics* **21**, 115-133.

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.

Harville, D. A. (1977). Maximum-likelihood approaches to variance component estimation and to related problems. *J. Amer. Statist. Assoc.* **72**, 320-340.

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.

Henderson, C. R., Kempthorne, O., Searle, S. R., and Von Krosigk, C. N. (1959). Estimation of environmental and genetic trends from records subject to culling. *Biometrics* **15**, 192-218.

Henderson, C. R., Searle, S. R., and Schaeffer, L. R. (1974). The invariance and calculation of Method 2 for estimating variance components. *Biometrics* **30**, 583-588.

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.

Jeske, D. R. and D. A. Harville (1986). Prediction, confidence and empirical Bayes intervals 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.

Kempthorne, O. (1975). Fixed and mixed random models in the analysis of variance. *Biometrics* **31**, 473-486.

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.

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.

Miller, J. J. (1973). Asymptotic properties and computation of maximum likelihood estimates in the mixed model of the analysis of variance. Tech. Rep. No. 12, Department of Statistics, Stanford University, Stanford, California.

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.

Robinson, J. (1965). The distribution of a general quadratic form in normal variables. *Australian J. of Statistics* **7**, 110-114.

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

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin* **2**, 110-114.

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.

Searle, S. R. (1988). Mixed models and unbalanced data: wherefrom, whereat and whereto? *Communications in Statistics* (in press).

Searle, S. R. and Henderson, H. V. (1980). Annotated computer output for analyses of unbalanced data: SAS HARVEY. Paper No. BU-659-M, Biometrics Unit, Cornell University.

Searle, S. R. and Henderson, H. V. (1983). Faults in a computing algorithm for reparameterizing linear models. *Communications in Statistics — Simulation & Computation* **12**, 67-76.

Snedecor, G. W. (1934). *Analysis of Variance and Covariance*. Ames, Iowa: Collegiate Press, Inc.

Snedecor, G. W. (1940). *Statistical Methods*, 3rd Edition, Ames, Iowa: Iowa State College Press.

Snedecor, G. W. and W. G. Cochran (1967). *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. *Sears Foundation 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.