A brief outline is given of the mixed model, of difficulties associated with it, and of attempts made to overcome those difficulties.

Mention of the mixed model of analysis of variance usually evokes three questions:

(i) What is the mixed model?
(ii) What are the difficulties in using it?
(iii) How are those difficulties overcome?

Skeleton answers to these questions are offered here.

1. FIXED EFFECTS MODELS

We are all familiar with traditional analysis of variance. It is one of the oldest statistical methods and still one of the most frequently used. Although its originator, R. A. Fisher, saw no need for having a linear model associated with an analysis of variance, there is now widespread use of such models. For example, the analysis of data from a randomized complete block experiment is often accompanied by the following model:
where
\[ y_{ijk} = \mu + t_i + b_j + (tb)_{ij} + e_{ijk} \]

and
- \( y_{ijk} \) = yield on the \( k \)'th observation of the \( i \)'th treatment in the \( j \)'th block, for \( k = 1, \ldots, n \), \( i = 1, \ldots, a \) and \( j = 1, \ldots, b \);
- \( \mu \) = a general mean;
- \( t_i \) = effect on yield of the \( i \)'th treatment;
- \( b_j \) = effect on yield of the \( j \)'th block;
- \( (tb)_{ij} \) = interaction effect on yield of the \( i \)'th treatment and the \( j \)'th block;
- \( e_{ijk} \) = residual, random, error term, with \( \mathbb{E}(e_{ijk}) = 0 \), for all \( i, j \) and \( k \)
  and \( \mathbb{E}(e_{ijk}e_{i'j'k'}) = \sigma^2 \) for \( i = i' \), \( j = j' \) and \( k = k' \) and zero otherwise, \( \mathbb{E} \) being expectation over repeated sampling.

The thing of interest in a model like this is estimating treatment effects or, more particularly, contrasts among treatment effects such as
\[ t_i - t_i' = \bar{y}_{i..} - \bar{y}_{i'..} \]

and
\[ t_i - \frac{1}{2} (t_i' + t_i'') = \bar{y}_{i..} - \frac{1}{2} (\bar{y}_{i'..} - \bar{y}_{i''..}) \]

The \( t_i \)'s in this model represent items of particular, pre-planned, interest: treatments, be they different fertilizers, in agriculture, diets in a nutrition experiment, levels of a pesticide in pest management, drugs in a clinical trial, different temperatures in firing china, or bricks; in general, different levels of whatever factor is of interest in an experiment. The important aspect of the \( t_i \)'s is that they represent, in our concept of the real world as mimicked by the
model, the effect on yield of the different treatments — and they are treatments of particular interest. Our attention is fixed upon just the treatments in the experiment, upon these and no others, and so the model is called a fixed effects model. There is no thought for any other treatments. The $t_i$'s are deemed to be fixed constants (albeit unknown and unknowable) representing the effects on yield of the different treatments, and we are interested in estimating them, or estimating contrasts among them.

2. RANDOM AND MIXED MODELS

Suppose a dairy farmer is going to buy a new cow for his herd. He goes to an auction sale — or maybe visits neighboring farmers to see what cows they have for sale. Before leaving home he knows what is wanted: a young cow, of the Holstein breed, say. But that is all that is known: the intended purchase can be specified only as female, young and Holstein. It cannot be specified any more precisely and certainly not with the specificity that an entomologist, for example, can describe levels of concentration of an insecticide as his treatments in an experiment for efficacy of that insecticide.

Suppose the farmer buys three cows and subsequently has milk yields from them for two seasons. A linear model equation that could be used for these records is

$$y_{ij} = \mu + c_i + s_j + e_{ij}$$

where $c_i$ represents the effect of cow $i$ on yield for the three cows purchased, $i = 1, 2$ and 3. The farmer, in making the purchases knows only that they are cows, young and Holstein. In no way can he specify them as three pre-planned levels of a factor called cow. They are three levels of a factor called cow, but in terms of interpreting their effect on yield, there is no specific, pre-planning of the levels in the manner that the entomologist uses three levels of concentration of
an insecticide. So far as their effects on yield are concerned, the cows are just a random trio of cows; each is a random selection of genes for milk production. The cow effects are therefore random; i.e., the \( c_i \)'s in the model are random variables. As such, means and variances can be attributed to them, usually \( E(c_i) = 0, \sigma^2(c_i) = \sigma^2 \) and zero covariances, and distributional properties can also be assumed when deemed appropriate.

In contrast, \( \mu \) represents some general mean and clearly is not a random effect. Neither is each \( s_j \), the effect due to season \( j \); it is a fixed effect. Thus the model is a mixture of fixed effects and random effects, and so is called a **mixed model**. If, for some reason the \( s_j \)-terms were to be assumed random effects rather than fixed effects, then all the effects in the model save \( \mu \) would be random, and the model would be a **random model**.

Mixed and random models are used in a wide variety of situations. In agriculture their use is common in statistical studies associated with genetic improvement of livestock production: milk yield in dairy cows, egg production in poultry, rate of gain in weight in broiler poultry, in beef cattle, hogs and lambs; back fat thickness in hogs, and crimp, staple length, and weight of fleece in wool production of sheep. In clinical trials in the drug industry, clinical effects are treated as random (e.g., Chakravorti and Grizzle, 1975), in psychological and educational testing the effect of subject is often considered random (e.g., Alwin, 1976); analyses of trajectory and orbital data in rocket flight testing have used the mixed model (e.g., Bush, 1971), as have Kalman filtering techniques of engineering (e.g., Duncan and Horn, 1972), and so also have economists in combining cross-section with time series data (e.g., Houthakker et al., 1974), referring to their models as error components models. Mixed models therefore continue to be important in data analysis, and also attract attention theoretically.
3. PREDICTION

3.1. An example

Consider another example, that of measuring intelligence in humans. Each of us has some degree of intelligence, usually called I.Q. It can never be measured exactly. As a substitute, we have test scores. Suppose we model \( y_{ij} \), the j'th test score for the i'th person, as

\[
y_{ij} = \mu + q_i + e_{ij}
\]

where \( q_i \) is the person's true I.Q. and \( e_{ij} \) is a residual error term. In this context an exercise in Mood's (1950) book is of great interest:

"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.)"

This exercise raises the first of three difficulties with mixed models - what is nowadays called prediction of random variables, although when 30 years ago it was referred to as estimating random variables that phrase brought forth scorn from many a mathematical statistician. Today, as prediction, it is an acceptable technique in many of the applications already alluded to. In animal improvement plans for farm livestock it has long been known as estimating genetic (or additive genetic) merit. And quite recently (e.g., Lindley and Smith, 1972) it has attracted attention as being Bayesian estimation in linear model theory.

To solve Mood's exercise note that I.Q. and score are jointly distributed with bivariate normal density:
\[
\begin{bmatrix}
\text{I.Q.} \\
\text{Score}
\end{bmatrix} = \begin{bmatrix} q_i \\
y_{ij}
\end{bmatrix} \sim N\left(\begin{bmatrix} 100 \\
100 \end{bmatrix}, \begin{bmatrix} 15^2 & 15^2 \\
15^2 & 15^2 + 5^2 \end{bmatrix}\right).
\]

From this, the maximum likelihood estimate of the conditional mean \(E(q_i | y_{ij} = 130)\) is

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

This is what today is called the predicted value of \(q_i\) (given that \(y_{ij} = 130\)).

### 3.2. General formulation

Write the equation of a general mixed model as

\[
y = X\beta + Zu + e
\]

where \(y\) represents the data, \(\beta\) is the vector of fixed effects, \(u\) is the vector of random effects, \(X\) and \(Z\) are the corresponding incidence matrices (with \(X\) including columns of covariates, if present) and \(e\) is a vector of random errors. For the random \(u\) and \(e\) assume the following mean and covariance structure:

\[
E\left[\begin{bmatrix} u \\
e
\end{bmatrix}\right] = 0 \quad \text{and} \quad \text{var}\left[\begin{bmatrix} u \\
e
\end{bmatrix}\right] = \begin{bmatrix} D & 0 \\
0 & R \end{bmatrix}.
\]

Then

\[
\text{var}(y) = V = ZDZ' + R
\]

and

\[
\begin{bmatrix}
\sim u \\
\sim y
\end{bmatrix} \sim \begin{bmatrix} 0 \\
X\beta \end{bmatrix}, \begin{bmatrix} D & DZ' \\
ZD & ZDZ' + R \end{bmatrix}.
\]

(1)
No matter what joint distribution is assumed for \( u \) and \( y \), the predictor \( \tilde{u} \) of \( u \) which is best is the conditional mean:

\[
(\text{best predictor of } u) = E(u|y). \tag{2}
\]

This means that, given the observations \( y \), the best predictor is \( E(u|y) \). Best, in this context, means the predictor which minimizes \( E(\tilde{u} - u)'A(\tilde{u} - u) \) for positive definite symmetric \( A \), a generalization of minimum mean squared error of prediction.

\( E(u|y) \) is not necessarily linear in \( y \). The best predictor which is, is

\[
\tilde{u} = (\text{best linear predictor of } u) = DZ'V^{-1}(y - X\beta^0) \tag{3}
\]

with

\[
X'V^{-1}X\beta^0 = X'V^{-1}y. \tag{4}
\]

Again, like (2), this result is true for all distributional forms of \( u \) and \( y \). When that form is a normal distribution, i.e., when \( u \) and \( y \) have a joint normal distribution, then (2) and (3) are the same:

\[
(\text{best predictor of } u \text{ under normality}) = (\text{best linear predictor of } u) = \tilde{u}. \tag{5}
\]

If \( V \) is known (which it seldom is, see Section 4), then under normality \( X\beta^0 \) for \( \beta^0 \) of (4) is the maximum likelihood estimator of \( X\beta \) and \( \tilde{u} \) is correspondingly the maximum likelihood estimator of \( E(u|y) \).

3.3. Mixed model prediction

A more useful prediction than \( u \) is that of \( u \) plus linear combinations of the fixed effects,

\[
w = u + K'\beta, \tag{6}
\]
say. The best predictor of \( \tilde{w} \) (best in the sense just described) which is linear in \( y \), i.e., has the general form \( a + By \) for some \( a \) and \( B \), and which is unbiased, meaning that \( E(\tilde{w}) = w \) over repeated data vectors \( y \), is

\[
\tilde{w} = \text{best linear unbiased predictor (BLUP) of } w = \tilde{u} + K\beta^o.
\]

(6)

This is the predictor that is widely used in animal breeding programs for the improvement of farm livestock.

Details of deriving (2) through (6) are recently available in Searle (1973, 1974).

3.4. Mixed model equations

The equations which simultaneously give \( \tilde{u} \), the best linear predictor, and \( \beta^o \), a solution to the (Aitken) generalized least squares equations, are

\[
\begin{bmatrix}
    X'R^{-1}X & X'R^{-1}Z \\
    Z'R^{-1}X & Z'R^{-1}Z + D^{-1}
\end{bmatrix}
\begin{bmatrix}
    \beta^o \\
    u
\end{bmatrix}
= \begin{bmatrix}
    X'R^{-1}y \\
    Z'R^{-1}y
\end{bmatrix}.
\]

They are known as the mixed model equations. Originally conceived of as giving maximum likelihood estimators (under normality) of \( \beta \) and \( u \), as in Henderson (1950), they do in fact result from maximizing a density function – see Henderson et al. (1959). That same paper and also Searle (1971) show the equivalence of the solutions of (7) to (3) and (4).

An algorithm for developing equations (7) is first to treat \( u \) as if it were a vector of fixed effects. Then \( \text{var}(y) = R \) and the generalized least squares (maximum likelihood under normality) equations for \( \beta \) and \( u \) are
These equations with \( D^{-1} \) added to the sub-matrix \( Z'R^{-1}Z \) are then the mixed model equations (7).

4. VARIANCE COMPONENTS

An important feature of \( \tilde{\theta} \) and \( \beta^0 \) is that they are functions of the variance-covariance matrices \( D = \text{var}(u) \) and \( V = \text{var}(y) = ZDZ' + R \). In most applications these matrices are unknown, a condition that raises two questions. First, how should \( D \) and \( R \) (and hence \( V \)) be estimated? Second, in replacing \( D \) and \( V \) in \( \tilde{\theta} \) and \( \beta^0 \) by those estimates, what properties will be attributable to the then estimated values of \( \tilde{\theta} \) and \( \beta^0 \)?

The problem of estimating \( D \) and \( R \) is usually simplified (somewhat) by assuming that \( R = \sigma^2 I_N \) for \( y \) of order \( N \), and by taking \( u' = \{ u_i \} \) for \( i = 1, \ldots, c \), where \( u_i \) has as elements the effects (\( N_i \) in number, say) corresponding to the \( i \)'th factor (main effect or interaction) of the random factors. Then

\[
D = \text{var}(u) = \sum_{i=1}^{c} \sigma^2_i \mathbf{I}_{N_i},
\]

where \( \oplus \) denotes a direct sum of matrices and \( \sigma^2_i \) is the variance component of the \( i \)'th random factor. Then

\[
V = \sum_{i=1}^{c} Z_i Z'^i \sigma^2_i + \sigma^2 e_i \mathbf{I}_N
\]

\[
= \sum_{i=0}^{c} \sigma^2_i \mathbf{I}_{N_i} \text{ with } \sigma^2_0 = \sigma^2 e \text{ and } Z_0 = \mathbf{I}_N.
\]

In this way the problem of estimating \( D \) and \( V \) becomes that of estimating the variance components \( \sigma^2_0, \sigma^2_1, \ldots, \sigma^2_c \).
4.1. Estimation from balanced data

Certain optimal properties have been known for some years for variance component estimates obtained from equal-subclass-numbers (balanced) data by equating mean squares of analyses of variance to their expected values, and solving the resulting equations for the variance components. These are known as analysis of variance (ANOVA) estimates. They are minimum variance quadratic unbiased [Graybill and Wortham (1956)] and, under normality, minimum variance unbiased [Graybill and Hultquist (1961)]. Despite this, they can take on negative values (which is very embarrassing as estimates of positive parameters), and their distributions even under normality assumptions have no known closed form.

4.2. Estimation from unbalanced data

Estimation of variance components from unequal-subclass-numbers (unbalanced) data is far more complicated - primarily because there are several available methods and all of them have deficiencies. In essence, the most readily computed methods give estimators with few attractive properties, and the methods having broadly based foundations are difficult, and sometimes impossible, to compute.

a. ANOVA methods. Henderson (1953) gave three variations of the "equate mean squares to expectations" method of the ANOVA procedure for balanced data. Method I is easy to compute, no matter how much data one has, but it can be used only for random models and not for mixed models. Method II is somewhat harder to compute, but is usually feasible, and can be used for mixed models, but only if there are no interactions between fixed and random effects. [Objections that Method II lacked specificity - see Searle (1971, p. 443) - have since been overcome, see Henderson et al. (1974).] Method III can be used for any mixed model, but it is not uniquely defined for cross-classified random factors, and it can be quite impractical computationally - due to having to invert matrices of unprecedented order for large data sets. All three methods can give negative
estimates, even though the estimators are unbiased. No other properties are known.

b. ML methods. Equations for maximum likelihood (ML) estimation, under normality assumptions, were developed by Hartley and Rao (1967). Solution has to be by iteration, which is sometimes tricky, and can be computationally impossible because of needing matrix inverses of large order.

c. REML methods. Similar comments apply to the restricted maximum likelihood (REML) method first suggested by Patterson and Thompson (1971) and later considered by Corbeil and Searle (1976a,b) and Harville (1977). It differs from ML estimation through maximizing just that portion of the likelihood which is invariant to the fixed effects in the model. ML and REML estimators are, by definition, always non-negative.

d. MINQUE methods. A method demanding no distributional assumptions is that of minimum norm quadratic unbiased estimation, known as MINQUE (Rao, 1970, 1971a,b and 1972). It is based on minimizing a Euclidean norm, that requires assigning à priori values to the variance components. The estimates are then functions of these à priori values. The arbitrary features of introducing the Euclidean norm and of requiring à priori values are put in perspective by the observation (Hocking and Kutner, 1975) that MINQUE is the same as the first round of iteration from REML, using the à priori values required of MINQUE as starting values for REML. Iterative use of the MINQUE procedure, called I-MINQUE and studied by Brown (1976), is therefore equivalent to REML.

The MINQUE procedure is also MIVQUE (minimum variance quadratic unbiased) under normality. And a special case is when, for want of any information on the variance components, the à priori values are all taken to be zero, save for the error variance. This case is called MINQUE0 or MIVQUE(0); it has attractive computing features (e.g., Hartley et al., 1978, and Goodnight, 1978), even for very
large data sets, but otherwise seems to have little to recommend it (Quaas and Bolgiano, 1979). In general, MINQUE suffers from the same computing difficulties as does REML — namely the impracticality of inverting matrices of large order. MINQUE (and MINQUEO) estimates can also be negative.

e. The dispersion-mean model. Finally, yielding no further methodology, but providing us with a sound theoretical basis for REML and a connection with ML (and through REML with MINQUE) is the dispersion-mean, linear model \( E(y) = \sim o^2 \) where \( \sim o^2 \) is the vector of variance components to be estimated and \( \sim y \) is the Kronecker product of \( y \) with itself. Pukelsheim (1976), Drygas (1977) and Brown (1978) show details. Generalized least squares on this model yields MINQUE, ordinary least squares yields MINQUEO (Pukelsheim, 1976) and generalized least squares on a simple variation of the model yields ML (Anderson, 1978). Extensive details are available in Anderson (1978) and Searle (1979).

5. ESTIMATING FIXED EFFECTS

The generalized least squares (ML under normality) equations that provide betas-values for estimating estimable functions of the fixed effects are \( X'V^{-1}X\beta_0 = X'V^{-1}y \). When \( V \) is known, then \( \beta_0 \) has well-known optimal properties. But \( \sim V \) is seldom known, and ML is the only method where these equations are solved and variance components are estimated simultaneously, as part of the estimation method. In all other methods the variance components are estimated without estimating \( X\beta \). The usual procedure is then to solve \( X'\sim V^{-1}X\beta_0 = X'\sim V^{-1}y \) where \( \sim V \) is the same function of the estimated \( \sim o^2 \)'s (i.e., of \( \sim o^2 \)) as \( \sim V \) is of \( o^2 \). The difficulty then is to ascertain what properties \( X\beta_0 \) has, especially what its distributional properties are that might enable confidence intervals and hypothesis tests to be established for estimable functions of elements of \( \beta \).
REFERENCES


