MIXED MODELS
(BLUP)
MIXED EFFECTS MODELS

→ Up to now we discussed models including fixed effects only.
→ Frequently, however, linear models contain also factors whose levels represent a random sample of a population of all possible factor levels.
→ Models containing both fixed and random effects are called mixed effects models.
→ Linear mixed effects models have been widely used in analysis of data where responses are clustered around some random effects, such that there is a natural dependence between observations in the same cluster.
→ For example, consider repeated measurements taken on each subject in longitudinal data, or observations taken on members of the same family in a genetic study.
Inferences regarding mixed effects models refer to the estimation of fixed effects, the prediction of random effects, and the estimation of variance and covariance components, which are briefly discussed next.

**LINEAR MIXED EFFECTS MODEL**

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

- responses
- incidence matrices
- fixed effects
- random effects
- residuals

⇒ Generally, it is assumed that \( u \) and \( e \) are independent from each other and normally distributed with zero-mean vectors and variance-covariance matrices \( G \) and \( \Sigma \), respectively, i.e.:

\[
\begin{bmatrix}
\mathbf{u} \\
\mathbf{e}
\end{bmatrix}
\sim \text{MVN}
\left(
\begin{bmatrix}
0 \\
0
\end{bmatrix},
\begin{bmatrix}
G & 0 \\
0 & \Sigma
\end{bmatrix}
\right)
\]

⇒ Inferences regarding mixed effects models refer to the estimation of fixed effects, the prediction of random effects, and the estimation of variance and covariance components, which are briefly discussed next.
Let \( y = X\beta + \varepsilon \), where \( \varepsilon = Zu + e \)

\[
\begin{align*}
E[\varepsilon] &= E[Zu + e] = ZE[u] + E[e] = 0 \\
\text{Var}[\varepsilon] &= \text{Var}[Zu + e] = Z\text{Var}[u]Z^T + \text{Var}[e] = ZGZ^T + \Sigma
\end{align*}
\]

Such that \( y \sim \text{MVN}(X\beta, V) \), where \( V = ZGZ^T + \Sigma \)

Under these circumstances, the MLE for \( \beta \) is:

\[
\hat{\beta} = (X^TV^{-1}X)^{-1}X^TV^{-1}y \sim \text{MVN}(\beta, (X^TV^{-1}X)^{-1})
\]
As $G$ and $\Sigma$ are generally unknown, an estimate of $V$ is used instead such that the estimator becomes: 
$$\hat{\beta} = (X^T \hat{V}^{-1} X)^{-1} X^T \hat{V}^{-1} y$$

The variance-covariance matrix of $\hat{\beta}$ is now approximated by $(X^T \hat{V}^{-1} X)^{-1}$

**Note:** $(X^T \hat{V}^{-1} X)^{-1}$ is biased downwards as a consequence of ignoring the variability introduced by working with estimates of (co)variance components instead of their true (unknown) parameter values.

Approximated confidence regions and test statistics for estimable functions of the type $K^T \beta$ can be obtained by using the result:

$$\left( K^T \beta^0 \right)^T \left( K^T (X^T V^{-1} X) K \right)^{-1} \left( K^T \beta^0 \right) \approx F_{[\phi_N, \phi_D]}$$

where $F_{[\phi_N, \phi_D]}$ refers to an F-distribution with $\phi_N = \text{rank}(K)$ degrees of freedom for the numerator, and $\phi_D$ degrees of freedom for the denominator, which is generally calculated from the data using, for example, the Satterthwaite’s approach.
In addition to the estimation of fixed effects, very often in genetics interest is also on prediction of random effects.

In linear (Gaussian) models such predictions are given by the conditional expectation of \( u \) given the data, i.e. \( \text{E}[u | y] \).

Given the model specifications, the joint distribution of \( y \) and \( u \) is:

\[
\begin{bmatrix} y \\ u \end{bmatrix} \sim \text{MVN} \left( \begin{bmatrix} X\beta \\ 0 \end{bmatrix}, \begin{bmatrix} V & ZG \\ GZ^T & G \end{bmatrix} \right)
\]

From the properties of multivariate normal distribution, we have that:

\[
\text{E}[u | y] = \text{E}[u] + \text{Cov}[u, y^T] \text{Var}^{-1}[y](y - \text{E}[y]) = GZ^T V^{-1} (y - X\beta) = GZ^T (ZGZ^T + \Sigma)^{-1} (y - X\beta)
\]

The fixed effects \( \beta \) are typically replaced by their estimates, so that predictions are made based on the following expression:

\[
\hat{u} = GZ^T (ZGZ^T + \Sigma)^{-1} (y - X\hat{\beta})
\]
The solutions \( \hat{\beta} \) and \( \hat{u} \) discussed before require \( V^{-1} \). As \( V \) can be of huge dimensions, especially in animal breeding applications, its inverse is generally computationally demanding if not unfeasible.

However, Henderson (1950) presented the mixed model equations (MME) to estimate \( \beta \) and \( u \) simultaneously, without the need for computing \( V^{-1} \).

The MME were derived by maximizing (for \( \beta \) and \( u \)) the joint density of \( y \) and \( u \), expressed as:

\[
p(y, u \mid \beta, G, \Sigma) \propto |\Sigma|^{-1/2} |G|^{-1/2} \exp \left\{ -\frac{1}{2} (y - X\beta - Zu)^T \Sigma^{-1} (y - X\beta - Zu) - \frac{1}{2} u^TG^{-1}u \right\}
\]

The logarithm of this function is:

\[
\ell = \log[p(y, u \mid \beta, G, \Sigma)] \propto |\Sigma| + |G| + (y - X\beta - Zu)^T \Sigma^{-1} (y - X\beta - Zu) + u^T G^{-1} u
\]

\[
= |\Sigma| + |G| + y^T \Sigma^{-1} y - 2y^T \Sigma^{-1} X\beta - 2y^T \Sigma^{-1} Zu + \beta^T X^T \Sigma^{-1} X\beta + 2\beta^T X^T \Sigma^{-1} Zu + u^T Z^T \Sigma^{-1} Zu + u^T G^{-1} u
\]
MIXED MODEL EQUATIONS

\[ \frac{\partial l}{\partial \beta} = X^T \Sigma^{-1} y - X^T \Sigma^{-1} X \hat{\beta} - X^T \Sigma^{-1} Z \hat{u} \]
\[ \frac{\partial l}{\partial u} = Z^T \Sigma^{-1} y - Z^T \Sigma^{-1} X \hat{\beta} - Z^T \Sigma^{-1} Z \hat{u} - G^{-1} \hat{u} \]

\[ \begin{bmatrix}
X' \Sigma^{-1} X \hat{\beta} + X' \Sigma^{-1} Z \hat{u} \\
Z' \Sigma^{-1} X \hat{\beta} + Z' \Sigma^{-1} Z \hat{u} + G^{-1} \hat{u}
\end{bmatrix} = \begin{bmatrix}
X' \Sigma^{-1} y \\
Z' \Sigma^{-1} y
\end{bmatrix} \]

which can be expressed as:

\[ \begin{bmatrix}
X^T \Sigma^{-1} X & X^T \Sigma^{-1} Z \\
Z^T \Sigma^{-1} X & Z^T \Sigma^{-1} Z + G^{-1}
\end{bmatrix} \begin{bmatrix}
\hat{\beta} \\
\hat{u}
\end{bmatrix} = \begin{bmatrix}
X^T \Sigma^{-1} y \\
Z^T \Sigma^{-1} y
\end{bmatrix} \]

known as the mixed model equations (MME).
Using the second part of the MME, we have that:

\[ Z^T \Sigma^{-1} X \hat{\beta} + (Z^T \Sigma^{-1} Z + G^{-1}) \hat{u} = Z^T \Sigma^{-1} y \]

so that:

\[ \hat{u} = (Z^T \Sigma^{-1} Z + G^{-1})^{-1} Z^T \Sigma^{-1} (y - X \hat{\beta}) \]

It can be shown that this expression is equivalent to \( \hat{u} = GZ^T (ZGZ^T + \Sigma)^{-1} (y - X \hat{\beta}) \) and, more importantly, that \( \hat{u} \) is the best linear unbiased predictor (BLUP) of \( u \).

Using this result into the first part of the MME, we have that:

\[ X^T \Sigma^{-1} X \hat{\beta} + X^T \Sigma^{-1} Z \hat{u} = X^T \Sigma^{-1} y \]

\[ X^T \Sigma^{-1} X \hat{\beta} + X^T \Sigma^{-1} Z(Z^T \Sigma^{-1} Z + G^{-1})^{-1} Z^T \Sigma^{-1} (y - X \hat{\beta}) = X^T \Sigma^{-1} y \]

\[ \hat{\beta} = \left( X^T [\Sigma^{-1} - \Sigma^{-1} Z(Z^T \Sigma^{-1} Z + G^{-1})^{-1} Z^T \Sigma^{-1}] X \right)^{-1} X^T [\Sigma^{-1} - \Sigma^{-1} Z(Z^T \Sigma^{-1} Z + G^{-1})^{-1} Z^T \Sigma^{-1}] y \]

Similarly, it is shown that this expression is equivalent to \( \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \), which is the best linear unbiased estimator (BLUE) of \( \beta \).
It is important to note that $\hat{\beta}$ and $\hat{\mu}$ require knowledge of $G$ and $\Sigma$. These matrices, however, are rarely known. This is a problem without an exact solution using classical methods. The practical approach is to replace $G$ and $\Sigma$ by their estimates ($\hat{G}$ and $\hat{\Sigma}$) into the MME.

Note that if $G$ and $\Sigma$ are known, the variance covariance matrix of the BLUE and BLUP is:

$$\text{Var} \begin{bmatrix} \hat{\beta} \\ \hat{\mu} \end{bmatrix} = \begin{bmatrix} X^T \Sigma^{-1} X & X^T \Sigma^{-1} Z \\ Z^T \Sigma^{-1} X & Z^T \Sigma^{-1} Z + G^{-1} \end{bmatrix}$$

If $G$ and $\Sigma$ are unknown and their values are replaced in the MME by some sort of point estimates $\hat{G}$ and $\hat{\Sigma}$, the new solutions $\tilde{\beta}$ and $\tilde{\mu}$ of the system:

$$\begin{bmatrix} X' \hat{\Sigma}^{-1} X & X' \hat{\Sigma}^{-1} Z \\ Z' \hat{\Sigma}^{-1} X & Z' \hat{\Sigma}^{-1} Z + \hat{G}^{-1} \end{bmatrix} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mu} \end{bmatrix} = \begin{bmatrix} X' \hat{\Sigma}^{-1} y \\ Z' \hat{\Sigma}^{-1} y \end{bmatrix}$$

are no longer BLUE and BLUP solutions, as they are not even linear functions of the data $y$. It is shown also that generally:

$$\text{Var} \begin{bmatrix} \tilde{\beta} \\ \tilde{\mu} \end{bmatrix} > \begin{bmatrix} X^T \hat{\Sigma}^{-1} X & X^T \hat{\Sigma}^{-1} Z \\ Z^T \hat{\Sigma}^{-1} X & Z^T \hat{\Sigma}^{-1} Z + \hat{G}^{-1} \end{bmatrix}$$
Consider the data set below, related to observations of half-sib families of k unrelated sires. The following model can be used to represent these data:

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

where \( y_{ij} \) represents the phenotypic trait observation of progeny \( j \) (\( j = 1, 2, \ldots, n_i \)) in family \( i \), \( \mu \) is a mean, \( s_i \) is an effect common to all animals having sire \( i \), and \( e_{ij} \) is a residual term.

The sire effect \( s_i \) is equivalent to the transmitting ability (which is equal to one-half additive genetic value) of sire \( i \), as one-half of its genes are (randomly) transmitted to each of its \( n_i \) progeny. The residual terms \( e_{ij} \) refer to additional genetics effects (such as the effect of dams) and environmental components.

It is assumed that \( s_i \sim (0, \sigma^2_s) \) and \( e_{ij} \sim (0, \sigma^2_e) \).
ESTIMATION OF VARIANCE COMPONENTS

ANOVA Estimation

⇒ From the model settings discussed before we have that:

\[ E[y_{ij}] = \mu \quad \text{and} \quad \text{Var}[y_{ij}] = \sigma_s^2 + \sigma_e^2. \]

The overall sample mean is given by:

\[ \bar{y}_{..} = \frac{1}{N} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{N} \sum_{i=1}^{k} y_{i.} \]

where \( N = \sum_{i=1}^{k} n_i \) and \( \bar{y}_{i.} = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij} \) are the sire-specific means.

The analysis of variance (ANOVA) approach consists of an orthogonal decomposition of the total sum of squares (TSS) into between classes (or, in our case, sires) and within classes (or residual) components. The corrected (in terms of the general mean) TSS is given by:

\[ \text{TSS} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{..})^2 \]
ESTIMATION OF VARIANCE COMPONENTS

ANOVA Estimation

By adding and subtracting $\bar{y}_{i.}$ within the parentheses, the TSS can be expressed as:

$$
TSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} [(y_{ij} - \bar{y}_{i.}) + (\bar{y}_{i.} - \bar{y}_{..})]^2
$$

$$
= \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2 + \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 + 2 \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})(\bar{y}_{i.} - \bar{y}_{..})
$$

It is seen that the last part of this expression is equal to zero, so that TSS can be written as two components, given by:

$$
SSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\bar{y}_{i.} - \bar{y}_{..})^2 \quad \text{and} \quad RSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2
$$

which are the sire and the residual sum of squares, respectively. The SSS term measures the variation of each progeny family around the overall mean, while the RSS term measures the extra variation related to each observation around its sire average.
It can be shown that the expectation of these sums of squares terms are:

\[
E[SSS] = \left( N - \frac{1}{N} \sum_{i=1}^{n_i} n_i^2 \right) \sigma_s^2 + (k-1)\sigma_e^2 \quad \text{and} \quad E[RSS] = (N - k)\sigma_e^2
\]

so that the ANOVA estimators of the sire and residual variance components are given by:

\[
\hat{\sigma}_s^2 = \left( N - \frac{1}{N} \sum_{i=1}^{n_i} n_i^2 \right)^{-1} [SSS - (k-1)\hat{\sigma}_e^2] \quad \text{and} \quad \hat{\sigma}_e^2 = \frac{1}{(N - k)} RSS
\]

respectively.

In the specific case of balanced data, i.e. the same progeny size for all sires, \( n_i = n = N / k \) and the ANOVA estimators become:

\[
\hat{\sigma}_s^2 = \frac{1}{n} \left[ \frac{1}{(k-1)} SSS - \hat{\sigma}_e^2 \right] \quad \text{and} \quad \hat{\sigma}_e^2 = \frac{1}{k(n-1)} RSS
\]
In general, the ANOVA approach works well for simple models (such as a one-way structure) or balanced data (such as data from designed experiments with no missing data), but they are not indicated for more complex models and data structures such as those generally found in animal breeding.

A number of methods have been proposed for estimating variance components in more complex scenarios, such as the expected mean squares approach of Henderson (1953), and the minimum norm quadratic unbiased estimation (Rao 1971a, 1971b), but maximum likelihood based methods are currently the most popular ones, especially the restricted (or residual) maximum likelihood (REML) approach, which attempts to correct for the well-known bias in the classical maximum likelihood (ML) estimation of variance components. These two methods are briefly described next.
Maximum likelihood estimates of the variance components can be obtained by maximizing the log-likelihood \( L(\beta, G, \Sigma) \) with respect to each element of \( G \) and \( \Sigma \), after replacing \( \beta \) by \( \hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \).

Alternatively, \( G, \Sigma, \) and \( \beta \) can be estimated simultaneously by maximizing their joint log-likelihood with respect to the variance components and the fixed effects. Standard errors can then be obtained by the inverse of the estimated Fisher information matrix. This approach provides an estimator for the variance-covariance matrix of \( \hat{\beta} \) which takes into account the extra variability related to the estimation of the variance components.
As a simple example of maximum likelihood estimation of variance components, consider the balanced case (i.e., constant progeny sizes) half-sib families data set discussed previously, and the linear model:

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

with the same definitions as before, but with the additional assumption of normality of both the sire and the residual effects, i.e.:

\[ s_i \sim \text{N}(0, \sigma_s^2) \quad \text{and} \quad e_{ij} \sim \text{N}(0, \sigma_e^2) \]
On a matrix notation, this model can be expressed as:

\[
\begin{pmatrix}
\mathbf{y}_1 \\
\mathbf{y}_2 \\
\vdots \\
\mathbf{y}_k
\end{pmatrix} =
\begin{pmatrix}
1_n \\
1_n \\
\vdots \\
1_n
\end{pmatrix} \mu +
\begin{pmatrix}
1_n & 0_n & \ldots & 0_n \\
0_n & 1_n & \ldots & 0_n \\
\vdots & \vdots & \ddots & \vdots \\
0_n & 0_n & \ldots & 1_n
\end{pmatrix}
\begin{pmatrix}
s_1 \\
s_2 \\
\vdots \\
s_k
\end{pmatrix} +
\begin{pmatrix}
e_1 \\
e_2 \\
\vdots \\
e_k
\end{pmatrix}
\]

where \( \mathbf{y}_i = [y_{i1}, y_{i2}, \ldots, y_{ik}]^\top \) represents the vector of observations of progeny \( i \) (i.e., relative to sire \( i \)); \( 1_n \) and \( 0_n \) represent \( n \)-dimensional column vectors of 1’s and 0’s, respectively; and \( \mathbf{e}_i = [e_{i1}, e_{i2}, \ldots, e_{ik}]^\top \) is the vector of residuals associated with progeny \( i \).
\[ \text{The vector of observations } y = [y_1^T, y_2^T, \ldots, y_k^T]^T \text{ has then a multivariate normal distribution with mean vector } \mu = 1_N \mu \text{ and variance-covariance matrix given by } I_s \otimes (1_n \sigma_s^2 1_n^T) + I_N \sigma_e^2, \text{ and its density function (from which the likelihood function obtained) can be written as:} \]

\[
p(y | \mu, \sigma_s^2, \sigma_e^2) = \frac{1}{(2\pi)^{N/2} | I_s \otimes J_n \sigma_s^2 + I_N \sigma_e^2 |^{1/2}} \times \exp \left\{ -\frac{1}{2} (y - 1_N \mu)^T (J_n \sigma_s^2 + I_n \sigma_e^2)^{-1} (y - 1_N \mu) \right\}
\]

\[
= (2\pi)^{-\frac{N}{2}} \left( \sigma_e^2 \right)^{-\frac{(N-k)}{2}} \left( \sigma_e^2 + n \sigma_s^2 \right)^{-\frac{k}{2}} \exp \left\{ -\frac{1}{2} (y - 1_N \mu)^T \left[ I_s \otimes J_n \left( \frac{1}{n} \left( \frac{1}{\sigma_e^2 + n \sigma_s^2} - \frac{1}{\sigma_e^2} \right) \right) \right] (y - 1_N \mu) \right\}
\]

where \( J_n = 1_n 1_n^T \) is an \((n \times n)\) matrix of 1’s, and \( \otimes \) is the Kronecker product.
The log-likelihood function can be written then as:

\[ l(\mu, \sigma^2, \sigma^2_e) \propto -\frac{(N-k)}{2} \log(\sigma^2_e) - \frac{k}{2} \log(\sigma^2_e + n\sigma^2_s) - \frac{1}{2\sigma^2_e} \sum_i \sum_j (y_{ij} - \bar{y}_{ij\cdot})^2 - \frac{1}{2} \frac{n(\bar{y}_{ij\cdot} - \mu)^2}{\sigma^2_e + n\sigma^2_s} \]

By taking the derivatives and setting them to 0, the following solutions are obtained:

\[ \hat{\mu} = \bar{y}_{\cdot\cdot}, \quad \hat{\sigma}^2_e = \frac{1}{k(n-1)} \text{RSS} \quad \text{and} \quad \hat{\sigma}^2_s = \frac{1}{n} \left[ \frac{\text{SSS}}{k} - \hat{\sigma}^2_e \right] \]

from which maximum likelihood estimates of the variance components are obtained, except if \( \hat{\sigma}^2_s < 0 \), in which case the estimate is set to zero.

Note the difference between the maximum likelihood and the ANOVA estimators of \( \sigma^2_s \). It is well known that maximum likelihood estimates of variance components are biased downwards as they do not take into account the degrees of freedom used for estimating the fixed effects.
Another alternative likelihood-based method for inferring variance components in mixed models is the restricted (or residual) maximum likelihood approach (REML), which corrects the bias associated with maximum likelihood estimates by taking into account the degrees of freedom used for estimating the fixed effects.

The REML approach for estimation of variance components maximizes the likelihood function of a set of error contrasts \( d = L^T y \), where \( L \) is a \([n \times (n-p)]\) full-rank matrix with columns orthogonal to the columns of the incidence matrix \( X \). The vector \( d \) then follows a multivariate normal distribution with null mean vector and variance-covariance matrix \( L^T V L = L^T (Z G Z^T + \Sigma) L \). Note that the distribution of \( d \) does not depend on \( \beta \).
The residual likelihood function for the variance components is then:

\[ L(G, \Sigma | y) = (2\pi)^{-(n-p)/2} \left| L^T V L \right|^{-1/2} \exp \left\{ -\frac{1}{2} d^T (L^T V L)^{-1} d \right\} \]

Another approach for obtaining the residual likelihood function for the variance components is by integrating the fixed effects out of the ‘full’ likelihood function, i.e.:

\[ L(G, \Sigma | y) = \int L(\beta, G, \Sigma | y) d\beta \]

as illustrated in the following example.
Recall the balanced half-sib families data set, and its associated likelihood function:

\[
L(\mu, \sigma_s^2, \sigma_e^2) = (2\pi)^{-\frac{N}{2}} \left( \sigma_e^2 \right)^{-\frac{N-k}{2}} \left( \sigma_e^2 + n\sigma_s^2 \right)^{-\frac{k}{2}}
\times \exp\left\{- \frac{1}{2\sigma_e^2} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 - \frac{1}{2} \sum_{i=1}^{k} \frac{n(\bar{y}_{i.} - \mu)^2}{\sigma_e^2 + n\sigma_s^2}\right\}
\]

Its residual likelihood is then:

\[
L(\sigma_s^2, \sigma_e^2) = \int L(\mu, \sigma_s^2, \sigma_e^2) d\mu
\]

\[
= (2\pi)^{-\frac{N}{2}} \left( \sigma_e^2 \right)^{-\frac{N-k}{2}} \left( \sigma_e^2 + n\sigma_s^2 \right)^{-\frac{k}{2}}
\times \exp\left\{- \frac{1}{2\sigma_e^2} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 \right\} \int \exp\left\{- \frac{1}{2} \sum_{i=1}^{k} \frac{n(\bar{y}_{i.} - \mu)^2}{\sigma_e^2 + n\sigma_s^2}\right\} d\mu
\]

which is equal to:

\[
L(\sigma_s^2, \sigma_e^2) = (2\pi)^{-\frac{N}{2}} \left( \sigma_e^2 \right)^{-\frac{N-k}{2}} \lambda^{-\frac{k}{2}}
\times \exp\left\{- \frac{1}{2\sigma_e^2} \sum_{i=1}^{k} \sum_{j=1}^{n} (y_{ij} - \bar{y}_{i.})^2 \right\} \exp\left\{- \frac{n}{2\lambda} \sum_{i=1}^{k} (\bar{y}_{i.} - \mu)^2 \right\} \sqrt{\frac{2\pi}{kn}}
\]

where \( \lambda = \sigma_e^2 + n\sigma_s^2 \).
By taking the derivatives with respect to $\lambda$ and $\sigma^2_c$, and by using the invariance property of maximum likelihood estimators, the following solutions are obtained:

$$\hat{\sigma}_e^2 = \frac{1}{k(n-1)} \text{RSS} \quad \text{and} \quad \hat{\sigma}_s^2 = \frac{1}{n} \left[ \frac{1}{(k-1)} \text{SSS} - \hat{\sigma}_e^2 \right]$$

which are the REML estimates of the variance components, except if $\hat{\sigma}_s^2 < 0$, i.e. if $\text{SSS} < \frac{(k-1)}{k(n-1)} \text{RSS}$.

As explicit forms of ML and REML estimators are often not available for more complex mixed effects models, ML and REML estimates are generally obtained by iterative approaches such as the expectation-maximization (EM) algorithm and Newton-Raphson-based procedures.
Animal breeding programs are based on the principle that phenotypic observations on related individuals can provide information about their underlying genotypic values.

The additive component of genetic variation is the primary determinant of the degree to which offspring resemble their parents, and therefore this is usually the component of interest in artificial selection programs.

Many statistical methods for analysis of genetic data are specific (or more appropriate) for phenotypic measurements obtained from planned experimental designs and with balanced data sets.

While such situations may be possible within laboratory or greenhouse experimental settings, data from natural populations and agricultural species are generally highly unbalanced and fragmented by numerous kinds of relationships.
Culling of data to accommodate conventional statistical techniques (e.g. ANOVA) may introduce bias and/or lead to a substantial loss of information.

The mixed model methodology allows efficient estimation of genetic parameters (such as variance components and heritability) and animal breeding values while accommodating extended pedigrees, unequal family sizes, overlapping generations, sex-limited traits, assortative mating, and natural or artificial selection.

To illustrate such application of mixed models in animal breeding, we consider here the so-called Animal Model in situations with a single trait and a single observation (including missing values) per animal.
The animal model can be described as:

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

where:

- \( y \) is an \((n \times 1)\) vector of observations (phenotypic scores)
- \( \beta \) is a \((p \times 1)\) vector of fixed effects (e.g. herd-year-season effects)
- \( u \sim N(0, G) \) is a \((q \times 1)\) vector of breeding values (relative to all animals with record or in the pedigree file, such that \( q \) is in general bigger than \( n \))
- \( e \) represents residual effects, assumed \( e \sim N(0, R = I_n \sigma_e^2) \), where \( \sigma_e^2 \) is the residual variance.
The matrix $G$ describing the covariances among the random effects (here the breeding values) follows from standard results for the covariances between relatives.

It is seen that the additive genetic covariance between two relatives $i$ and $i'$ is given by $2\theta_{ii'}\sigma_a^2$, where $\theta_{ii'}$ is the coefficient of coancestry between individuals $i$ and $i'$, and $\sigma_a^2$ is the additive genetic variance in the base population.

Hence, under the animal model, $G = A\sigma_a^2$, where $A$ is the additive genetic (or numerator) relationship matrix, having elements given by $a_{ii'} = 2\theta_{ii'}$. 
In general in animal breeding interest is on predicting breeding values (for selection of superior individuals), and on estimation of variance components and functions thereof, such as heritability. The fixed effects are, in some sense, nuisance factors with no central interest in terms of inferences, but which need to be taken into account (i.e., they need to be corrected for when inferring breeding values).

Since under the animal model \( G^{-1} = A^{-1}\sigma_a^{-2} \) and \( R^{-1} = I_n\sigma_e^{-2} \), the mixed model equations can be expressed as:

\[
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + \lambda A^{-1}
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{u}
\end{bmatrix} =
\begin{bmatrix}
X'y \\
Z'y
\end{bmatrix}
\]

where \( \lambda = \frac{\sigma_e^2}{\sigma_a^2} = 1 - h^2 \), such that:

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{u}
\end{bmatrix} =
\begin{bmatrix}
X'X & X'Z \\
Z'X & Z'Z + \lambda A^{-1}
\end{bmatrix}^{-1}
\begin{bmatrix}
X'y \\
Z'y
\end{bmatrix}
\]
Conditional on the variance components ratio \( \lambda \), the BLUP of the breeding values are given by
\[
\mathbf{u} = (\mathbf{Z}'\mathbf{Z} + \lambda \mathbf{A}^{-1})^{-1} \mathbf{Z}'(\mathbf{y} - \mathbf{X}\hat{\beta})
\]. These are generally referred to as estimated breeding values (EBV).

Alternatively, some breeders associations express their results as predicted transmitting abilities (PTA) or estimated transmitting abilities (ETA), which are equal to half the EBV, representing the portion of an animal’s breeding values that is passed to its offspring.

The amount of information contained in an animal’s genetic evaluation depends on the availability of its own record, as well as how many (and how close) relatives it has with phenotypic information. As a measure of amount of information in livestock genetic evaluations, EBVs are typically reported with its associated accuracies.

Accuracy of predictions is defined as the correlation between true and estimated breeding values, i.e., \( r_i = \rho(\hat{u}_i, u_i) \). Instead of accuracy, some livestock species genetic evaluations use reliability, which is the squared correlation of accuracy \( (r_i^2) \).
The calculation of $\rho(\hat{u}_i, u_i)$ requires the diagonal elements of the inverse of the MME coefficient matrix, represented as:

$$C = \begin{bmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{bmatrix}^{-1} = \begin{bmatrix} C^{\beta\beta} & C^{\beta u} \\ C^{u\beta} & C^{uu} \end{bmatrix}$$

It is shown that the prediction error variance of EBV $\hat{u}_i$ is given by:

$$\text{PEV} = \text{Var}(\hat{u}_i - u_i) = c_{ii}^{uu} \sigma_e^2$$

where $c_{ii}^{uu}$ is the $i$-th diagonal element of $C^{uu}$, relative to animal $i$. The PEV can be interpreted as the fraction of additive genetic variance not accounted for by the prediction. Therefore, PEV can be expressed also as:

$$\text{PEV} = (1 - r_i^2)\sigma_a^2$$

such that $c_{ii}^{uu} \sigma_e^2 = (1 - r_i^2)\sigma_a^2$, from which the reliability is obtained as:

$$r_i^2 = 1 - c_{ii}^{uu} \sigma_e^2 / \sigma_a^2 = 1 - \lambda c_{ii}^{uu}$$
ANIMAL MODEL

\[
\begin{bmatrix}
310 \\
270 \\
350
\end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_3 \\ e_4 \end{bmatrix}
\]

\[
y = X \beta + Z u + e
\]
Breeding values: \( u \sim \mathcal{N}(0, A\sigma_u^2) \), with

\[
A = \begin{bmatrix}
1 & 0 & 0.5 & 0.5 & 0.25 \\
0 & 1 & 0 & 0.5 & 0 \\
0.5 & 0 & 1 & 0.25 & 0.5 \\
0.5 & 0.5 & 0.25 & 1 & 0.125 \\
0.25 & 0 & 0.5 & 0.125 & 1
\end{bmatrix}
\]

\[
\begin{bmatrix}
\hat{\beta} \\
\hat{u}
\end{bmatrix} = \left[ X'X \quad X'Z \right]^{-1} \left[ X'y \right] - \left[ Z'X \quad Z'Z + \lambda A^{-1} \right] \left[ Z'y \right]
\]

\[
\alpha = \frac{\sigma_e^2}{\sigma_u^2} = \frac{1 - h^2}{h^2}
\]

\[
h^2 = \frac{1}{3} \rightarrow \alpha = 2 \quad \Rightarrow \quad \begin{cases}
\hat{h}_1 = 290 \\
\hat{h}_2 = 348 \\
\hat{u}_1 = 4.0 \\
\hat{u}_2 = 0.0 \\
\hat{u}_3 = -4.0 \\
\hat{u}_4 = 2.0 \\
\hat{u}_5 = -2.0
\end{cases}
\]
The animal model can be extended to model multiple (correlated) traits, multiple random effects (such as maternal effects and common environmental effects), repeated records (e.g. test day models), and so on.

EXAMPLE: Mrode 1996, pp74-76

Consider the following data set on the weaning weight (kg) of piglets, which are progeny of three sows mated to two boars:

<table>
<thead>
<tr>
<th>Piglet</th>
<th>Sire</th>
<th>Dam</th>
<th>Sex</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>70</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>65</td>
</tr>
<tr>
<td>9</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>98</td>
</tr>
<tr>
<td>10</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>108</td>
</tr>
<tr>
<td>11</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>60</td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td>4</td>
<td>2</td>
<td>80</td>
</tr>
<tr>
<td>13</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>14</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td>85</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>68</td>
</tr>
</tbody>
</table>
A linear model with the (fixed) effect of sex, and the (random) effects of common environment (related to each litter) and breeding values can be expressed as $X$:

$$y = X\beta + Zu + Wc + e$$

Assuming that $\sigma_u^2 = 20$, $\sigma_c^2 = 15$ and $\sigma_e^2 = 65$, the MME are as follows:

$$
\begin{bmatrix}
X'X & X'Z & X'W \\
Z'X & Z'Z + A^{-1} \lambda_1 & Z'W \\
W'X & W'Z & W'W + I \lambda_2
\end{bmatrix}
\begin{bmatrix}
\hat{\beta} \\
\hat{u} \\
\hat{c}
\end{bmatrix} =
\begin{bmatrix}
X'y \\
Z'y \\
W'y
\end{bmatrix}
$$

where $\lambda_1 = \frac{\sigma_e^2}{\sigma_u^2} = 3.25$ and $\lambda_2 = \frac{\sigma_e^2}{\sigma_c^2} = 4.3$. 

The BLUEs and BLUPs (inverting the numerator relationship matrix) are:

<table>
<thead>
<tr>
<th>Effects</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>91.493</td>
</tr>
<tr>
<td>2</td>
<td>75.764</td>
</tr>
<tr>
<td>Animals</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>-1.441</td>
</tr>
<tr>
<td>2</td>
<td>-1.175</td>
</tr>
<tr>
<td>3</td>
<td>1.441</td>
</tr>
<tr>
<td>4</td>
<td>1.441</td>
</tr>
<tr>
<td>5</td>
<td>-0.266</td>
</tr>
<tr>
<td>6</td>
<td>-1.098</td>
</tr>
<tr>
<td>7</td>
<td>-1.667</td>
</tr>
<tr>
<td>8</td>
<td>-2.334</td>
</tr>
<tr>
<td>9</td>
<td>3.925</td>
</tr>
<tr>
<td>10</td>
<td>2.895</td>
</tr>
<tr>
<td>11</td>
<td>-1.141</td>
</tr>
<tr>
<td>12</td>
<td>1.525</td>
</tr>
<tr>
<td>13</td>
<td>0.448</td>
</tr>
<tr>
<td>14</td>
<td>0.545</td>
</tr>
<tr>
<td>15</td>
<td>-3.819</td>
</tr>
<tr>
<td>Environ.</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>-1.762</td>
</tr>
<tr>
<td>4</td>
<td>2.161</td>
</tr>
<tr>
<td>5</td>
<td>-0.399</td>
</tr>
</tbody>
</table>
Technical note: An R package for fitting generalized linear mixed models in animal breeding

A. I. Vazquez,*2 D. M. Bates,† G. J. M. Rosa,* D. Gianola,*‡ and K. A. Weigel*

©2010 American Society of Animal Science. All rights reserved.


\[ g(\mu_{Y|U}) = Zu + X\beta \]

\[ \mu_{Y|U} = E[Y|U = u] \quad u \sim N(0, A\sigma_u^2) \]

\[ u^* = L^{-1}u \rightarrow g(\mu_{Y|U}) = ZL(L^{-1}u) + X\beta = Zu^* + X\beta \]

\[ A = LL' \quad u^* \sim N(0, I\sigma_u^2) \]

(Harville and Callanan 1989)
pedigreemm: An R package for fitting generalized linear mixed models in animal breeding

(Vazquez et al. 2010)

Welcome to Pedigree-based mixed models project!

This project extends the structures of the lme4 package to fit models in which random effects can be defined with respect to a pedigree.

No content added.

The project summary page you can find here.