Mixed-Model Estimation of genetic variances

Bruce Walsh lecture notes
Uppsala EQG 2012 course
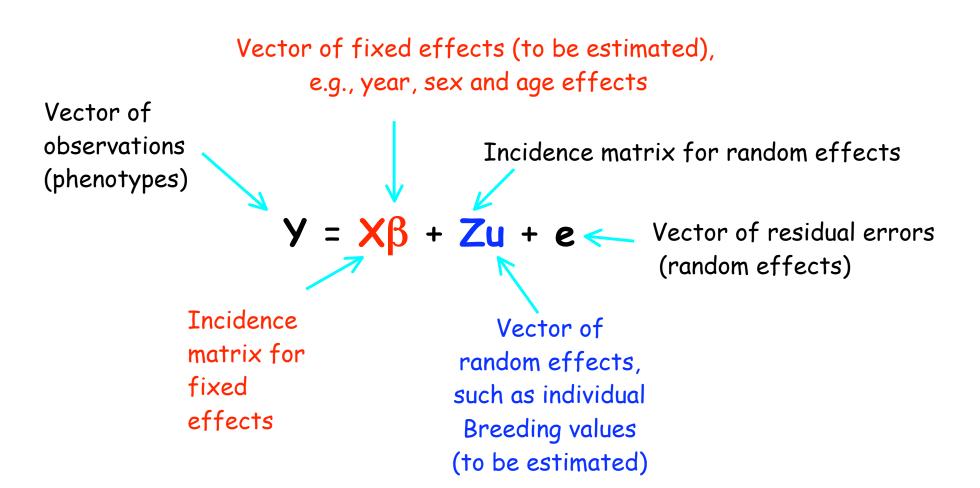version 28 Jan 2012

Estimation of Var(A) and Breeding Values in General Pedigrees

The above designs (ANOVA, P-O regression) are simple, involving only a single type of relative comparison. Further, we assumed balanced designs, with the number of offspring the same in each family.
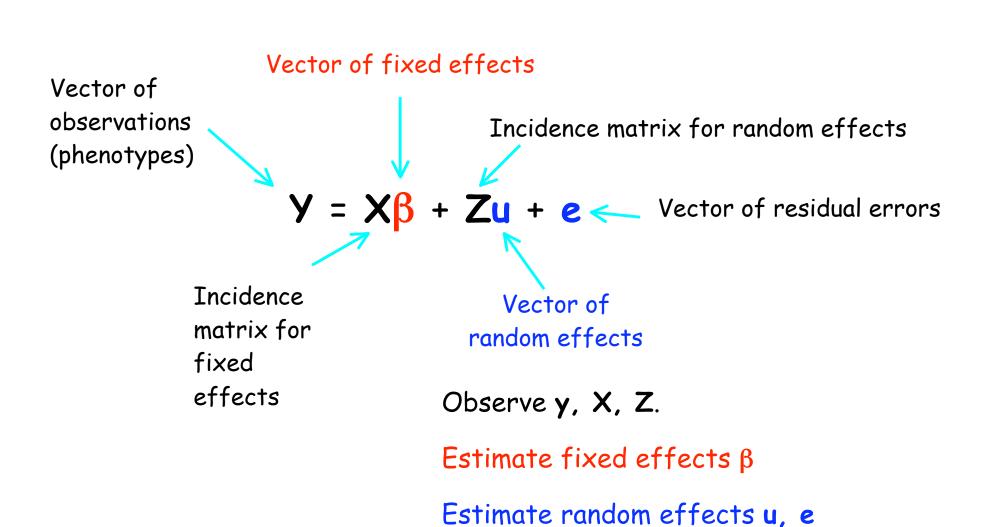
In the real world, we often have a pedigree of relatives, with a very unbalanced design. Fortunately, the general mixed model (so called because it includes both fixed and random effects, offers an ideal platform for both estimating genetic variances as well a predicting the breeding values of individuals.

Almost all animal breeding is based on such models, with REML (restricted max likelihood) used to estimated variances and BLUP (best linear unbiased predictors) used to predict BV

The general mixed model

Vector of fixed effects (to be estimated),
e.g., year, sex and age effects

Vector of
observations
(phenotypes)

Incidence matrix for random effects

$$Y = X\beta + Zu + e$$

Vector of residual errors
(random effects)

Incidence
matrix for
fixed
effects

Vector of
random effects,
such as individual
Breeding values
(to be estimated)

The general mixed model

Vector of fixed effects

Vector of
observations
(phenotypes)

Incidence matrix for random effects

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \mathbf{e}$$

Vector of residual errors

Incidence
matrix for
fixed
effects

Vector of
random effects

Observe **y, X, Z**.

Estimate fixed effects β

Estimate random effects **u, e**

Example

Suppose we wish to estimate the breeding values of three sires, each of which is mated to a random dam, producing two offspring, some reared in environment one, others in environment two. The data are

Observation	Value	Sire	environment
Y_{111}	9	1	1
Y_{121}	12	1	2
Y_{211}	11	2	1
Y_{212}	6	2	1
Y_{311}	7	3	1
Y_{321}	14	3	2

Here the basic model is

$$Y_{ijk} = \beta_j + u_i + e_{ijk}$$

Effect of environment j

Breeding value of sire i

The mixed model vectors and matrices become

$$\mathbf{y} = \begin{pmatrix} y_{1,1,1} \\ y_{1,2,1} \\ y_{2,1,1} \\ y_{2,1,2} \\ y_{3,1,1} \\ y_{3,2,1} \end{pmatrix} = \begin{pmatrix} 9 \\ 12 \\ 11 \\ 6 \\ 7 \\ 14 \end{pmatrix}$$

$$\mathbf{X} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix}, \quad \mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$$

Means & Variances for $y = X\beta + Zu + e$

Means: $E(u) = E(e) = 0$, $E(y) = X\beta$

Variances:

Let R be the covariance matrix for the residuals. We typically assume $R = \sigma^2_e * I$

Let G be the covariance matrix for the breeding values (the vector **u**)

The covariance matrix for y becomes
$$V = ZGZ^T + R$$

Effects of model misspecification

Suppose we simply used a General Linear model (only fixed effects) for this example?

Here $y = X\beta + e^*$, where $e^* \sim MVN(0, V)$, implying $Y \sim MVN(X\beta, V)$

The effect of using a mixed model is that it partitions the residual e^* as
$e^* = Zu + e$

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe $\mathbf{y}$, $\mathbf{X}$, and $\mathbf{Z}$

β, $\mathbf{u}$, $\mathbf{R}$, and $\mathbf{G}$ are generally unknown

Two complementary estimation issues

(i) Estimation of β and $\mathbf{u}$

$$\widehat{\beta} = \left(\mathbf{X}^T \mathbf{V}^{-1} \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{V}^{-1} \mathbf{y}$$ Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

$$\widehat{\mathbf{u}} = \mathbf{G} \mathbf{Z}^T \mathbf{V}^{-1} \left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)$$ Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$

Let's return to our example

Assume residuals uncorrelated & homoscedastic,
$R = \sigma^2_e * I$. Hence, need σ^2_e to solve BLUE/BLUP equations.

Suppose $\sigma^2_e = 6$, giving R = 6* I

Now consider G, the covariance matrix for u (the vector of the three sire breeding values). Assume sires are unrelated, so G is diagonal with element σ^2_G =sire variance, where $\sigma^2_G = \sigma^2_A /4$.

Suppose $\sigma^2_A = 8$, giving   G = 8/4*I

Solving, recalling that $V = ZGZ^{\mathsf{T}} + R$

$$\mathbf{V} = \frac{8}{4}\begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix} + 6\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} 8 & 2 & 0 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & 0 & 2 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{pmatrix} \quad \text{giving} \quad \mathbf{V}^{-1} = \frac{1}{30} \cdot \begin{pmatrix} 4 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 4 & -1 & 0 & 0 \\ 0 & 0 & -1 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 4 & -1 \\ 0 & 0 & 0 & 0 & -1 & 4 \end{pmatrix}$$

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta_1} \\ \widehat{\beta_2} \end{pmatrix} = \left(\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^{T}\mathbf{V}^{-1}\mathbf{y} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \end{pmatrix}$$

$$\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{u_1} \\ \widehat{u_2} \\ u_3 \end{pmatrix} = \mathbf{G}\mathbf{Z}^{T}\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \frac{1}{18}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

Henderson's Mixed Model Equations

$$y = X\beta + Zu + e, \ u \sim (0, G), \ e \sim (0, R), \ \text{cov}(u, e) = 0,$$

If X is n x p and Z is n x q

$$
\begin{array}{cc}
p \times p & p \times q
\end{array}
$$

$$
\begin{pmatrix}
\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\
\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}
\end{pmatrix}
\begin{pmatrix}
\widehat{\beta} \\
\widehat{u}
\end{pmatrix}
=
\begin{pmatrix}
\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\
\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y}
\end{pmatrix}
$$

$$q \times q$$

The whole matrix is (p+q) x (p+q)

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$$

$$V = ZGZ^T + R$$

$$\widehat{u} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\beta}\right)$$

Inversion of an n x n matrix

Standard Errors

A relatively straightforward extension of Henderson's mixed-model equations provides estimates of the standard errors of the fixed and random effects. Let the inverse of the leftmost matrix in Equation 26.5 be

$$
\begin{pmatrix} \mathbf{X}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{X}^T \mathbf{R}^{-1} \mathbf{Z} \\ \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{X} & \mathbf{Z}^T \mathbf{R}^{-1} \mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{12}^T & \mathbf{C}_{22} \end{pmatrix} \tag{26.6}
$$

where $\mathbf{C}_{11}$, $\mathbf{C}_{12}$, and $\mathbf{C}_{22}$ are, respectively, $p \times p$, $p \times q$, and $q \times q$ submatrices. Using this notation, Henderson (1975) showed that the sampling covariance matrix for the BLUE of $\boldsymbol{\beta}$ is given by

$$
\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}) = \mathbf{C}_{11} \tag{26.7a}
$$

that the sampling covariance matrix of the prediction errors $(\widehat{\mathbf{u}} - \mathbf{u})$ is given by

$$
\boldsymbol{\sigma}(\widehat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{22} \tag{26.7b}
$$

and that the sampling covariance of estimated effects and prediction errors is given by

$$
\boldsymbol{\sigma}(\widehat{\boldsymbol{\beta}}, \widehat{\mathbf{u}} - \mathbf{u}) = \mathbf{C}_{12} \tag{26.7c}
$$

(We consider $\widehat{\mathbf{u}} - \mathbf{u}$ rather than $\widehat{\mathbf{u}}$ as the latter includes variance from both the prediction error and the random effects $\mathbf{u}$ themselves.)

Let's redo our previous example using Henderson's Equation

$$\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} = \frac{1}{6}\begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix}, \qquad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} = \left(\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X}\right)^T = \frac{1}{6}\begin{pmatrix} 1 & 2 & 1 \\ 1 & 0 & 1 \end{pmatrix}$$

$$\mathbf{G}^{-1}+\mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} = \frac{5}{6}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 33 \\ 26 \end{pmatrix}, \quad \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} = \frac{1}{6}\begin{pmatrix} 21 \\ 17 \\ 21 \end{pmatrix}$$

$$\begin{pmatrix} 4 & 0 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 \\ 1 & 1 & 5 & 0 & 0 \\ 2 & 0 & 0 & 5 & 0 \\ 1 & 1 & 0 & 0 & 5 \end{pmatrix}\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \begin{pmatrix} 33 \\ 26 \\ 21 \\ 17 \\ 21 \end{pmatrix}$$

Taking the inverse gives

$$\begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \widehat{u}_1 \\ \widehat{u}_2 \\ \widehat{u}_3 \end{pmatrix} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \\ -1 \\ 2 \\ -1 \end{pmatrix}$$

The Animal Model, $y_i = \mu + a_i + e_i$

Here, the individual is the unit of analysis, with y_i the phenotypic value of the individual and a_i its BV

$$\mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix}, \qquad \boldsymbol{\beta} = \mu, \qquad \mathbf{u} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_k \end{pmatrix} \qquad \mathbf{G} = \sigma_A^2 \, \mathbf{A},$$

Where the additive genetic relationship matrix A is given by
$A_{ij} = 2\theta_{ij}$ namely twice the coefficient of coancestry

Assume R = σ^2_e*I, so that $R^{-1} = 1/(\sigma^2_e)$*I.
Likewise, G = σ^2_A*A, so that $G^{-1} = 1/(\sigma^2_A)$* A^{-1}.

Henderson's mixed model equation here becomes

$$
\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\beta}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \end{pmatrix}
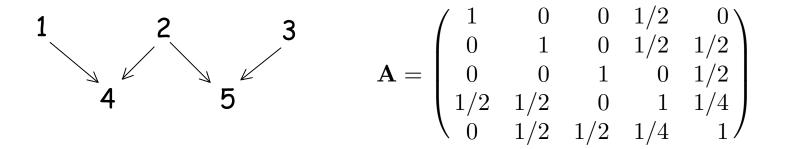$$

here $\lambda = \sigma^2_e / \sigma^2_A = (1-h^2)/h^2$

This reduces to

$$
\begin{pmatrix} n & \mathbf{1}^T \\ \mathbf{1} & \mathbf{I} + \lambda\,\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \sum^n y_i \\ \mathbf{y} \end{pmatrix}
$$

Example

Suppose our pedigree is

1 ↘ 2 ↙↘ 3 ↙
     4   5

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 & 1/2 & 0 \\ 0 & 1 & 0 & 1/2 & 1/2 \\ 0 & 0 & 1 & 0 & 1/2 \\ 1/2 & 1/2 & 0 & 1 & 1/4 \\ 0 & 1/2 & 1/2 & 1/4 & 1 \end{pmatrix}$$

Suppose λ =1 (corresponds to h^2 = 0.5). In this case,

$$\mathbf{I} + \lambda \, \mathbf{A}^{-1} = \begin{pmatrix} 5/2 & 1/2 & 0 & -1 & 0 \\ 1/2 & 3 & 1/2 & -1 & -1 \\ 0 & 1/2 & 5/2 & 0 & -1 \\ -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & 0 & 3 \end{pmatrix}$$

Suppose the vector of observations is

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{pmatrix} = \begin{pmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

Here n = 5, Σ y = 41, and Henderson's equation becomes

$$\begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 5/2 & 1/2 & 0 & -1 & 0 \\ 1 & 1/2 & 3 & 1/2 & -1 & -1 \\ 1 & 0 & 1/2 & 5/2 & 0 & -1 \\ 1 & -1 & -1 & 0 & 3 & 0 \\ 1 & 0 & -1 & -1 & 0 & 3 \end{pmatrix} \begin{pmatrix} \widehat{\mu} \\ \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ \widehat{a}_5 \end{pmatrix} = \begin{pmatrix} 41 \\ 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{pmatrix}$$

Solving gives

$$\widehat{\mu} = \frac{440}{53} \simeq 8.302, \qquad \begin{pmatrix} \widehat{a}_1 \\ \widehat{a}_2 \\ \widehat{a}_3 \\ \widehat{a}_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} -662/689 \\ 4/53 \\ 610/689 \\ -732/689 \\ 381/689 \end{pmatrix} \simeq \begin{pmatrix} -0.961 \\ 0.076 \\ 0.885 \\ -1.062 \\ 0.553 \end{pmatrix}$$

More on the animal model

- Under the animal model
  - $y = X\beta + Za + e$
  - $a \sim (0, \sigma_A^2 A)$, $e \sim (0, \sigma_e^2 I)$
  - $BLUP(a) = \sigma_A^2 A Z^T V^{-1}(y - X\beta)$

  - Where $V = ZGZ^T + R = \sigma_A^2 ZAZ^T + \sigma_e^2 I$
- Consider the simplest case of a single observation on one individual, where the only fixed effect is the mean μ, which is assumed known
  - Here $Z = A = I = (1)$,
  - $V = \sigma_A^2 + \sigma_e^2$
  - $\sigma_A^2 A Z^T V^{-1} = \sigma_A^2 / (\sigma_A^2 + \sigma_e^2) = h^2$
  - $BLUP(a) = h^2(y - \mu)$

- More generally, with single observations on n unrelated individuals,
  - $A = Z = I_{n \times n}$
  - $V = \sigma_A^2 \mathbf{Z A Z}^T + \sigma_e^2 \mathbf{I} = (\sigma_A^2 + \sigma_e^2)\,\mathbf{I}$
  - $\sigma_A^2\,\mathbf{A Z}^T\mathbf{V}^{-1} = h^2\,\mathbf{I}$
  - $\text{BLUP}(\mathbf{a}) = \sigma_A^2\mathbf{A Z}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) = h^2(\mathbf{y} - \mu)$
- Hence, the predicted breeding value of individual i is just $\text{BLUP}(a_i) = h^2(y_i - \mu)$
- When at least some individuals are related and/or inbred (so that $A \neq I$) and/or missing or multiple records (so that $Z \neq I$), then the estimates of the BV differ from this simple form, but BLUP fully accounts for this

Estimation of R and G

The second estimation issue the covariance matrix for residuals R and for breeding values G

As we have seen, both matrices have the form $\sigma^2 * B$, where the variance σ^2 is unknown, but B is known

For example, for residuals, $R = \sigma^2_e * I$

For breeding values, $G = \sigma^2_A * A$, where A is given from the pedigree

REML Variance Component Estimation

REML = Restricted Maximum Likelihood.

Standard ML variance estimation assumes fixed factors are known without error. Results in downward bias in variance estimates

REML maximizes that portion of the likelihood that does not depend on fixed effects

Basic idea: Use a transformation to remove fixed effect, then perform ML on this transformed vector

Simple variance estimate under ML vs. REML

$$\text{ML} = \frac{1}{n} \sum_{i+1}^{n} (x - \overline{x})^2, \quad \text{REML} = \frac{1}{n-1} \sum_{i+1}^{n} (x - \overline{x})^2$$

REML adjusts for the
estimated fixed
effect,
in this case, the mean

With balanced design, ANOVA variance estimates are
equivalent to REML variance estimates