Lecture 8
Mixed Models, BLUP Breeding Values

Guilherme J. M. Rosa

University of Wisconsin-Madison

Introduction to Quantitative Genetics
SISG, Seattle
16 – 18 July 2018

OUTLINE

- The General Linear Model
- Linear Mixed Models
- The 'Animal Model'
- EBV and Prediction Accuracy
- Multiple Random Effects

General Linear Model

(Fixed Effects Model)

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

- $\mathbf{y}$: responses
- $\boldsymbol{\varepsilon}$: residuals
- $\mathbf{X}$: design/incidence matrix (known)
- $\boldsymbol{\beta}$: overall mean + fixed effects parameters

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, I_n\sigma^2) \rightarrow \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

⇨ Fixed effect: levels included in the study represent all levels about which inference is to be made. Fixed effects models: models containing only fixed effects

Example 1

Experiment to compare growth performance of pigs under two experimental groups (Control and Treatment), with three replications each.

Control	Treatment
53	61
46	66
58	57

Model:

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

- y_{ij}: weight gain of pig j of group i
- μ: constant; general mean
- δ_i: effect of group i
- e_{ij}: residual term

Matrix Notation

Control	Treatment
53	61
46	66
58	57

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{bmatrix} = \begin{bmatrix} 53 \\ 46 \\ 58 \\ 61 \\ 66 \\ 57 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} \mu \\ \delta_1 \\ \delta_2 \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{21} \\ e_{23} \end{bmatrix}$$

Example 2

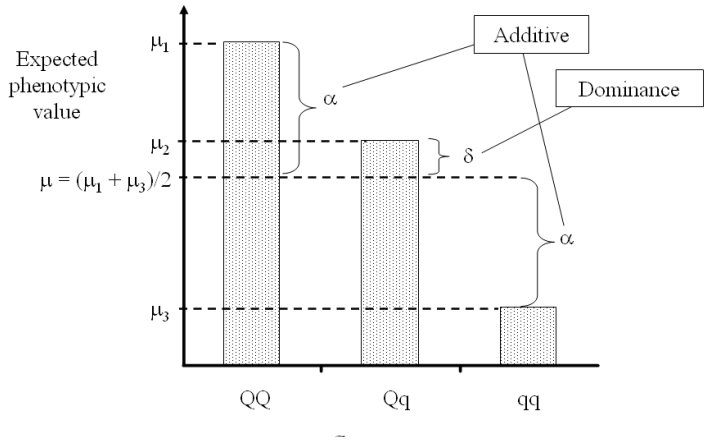
Flowering time (days, log scale) of *Brassica napus* according to genotype in specific locus, such as a candidate gene

Genotype		
qq	Qq	QQ
3.4	2.9	3.1
3.7	2.5	2.6
3.2		

Model: $y_{ij} = \mu_i + e_{ij}$

- y_{ij}: flowering time of replication j (j = 1,..., n_i) of genotype i (i = qq, Qq and QQ)
- μ_i: expected flowering time of plants of genotype i
- e_{ij}: residual (environment and polygenic effects)

⇨ The expected phenotypic values μ_i, however, can be expressed as a function of the additive and dominant effects



Expected phenotypic value according to the genotype on a specific locus.

The model can be written then as: $y_{ij} = \mu + x_{ij}\alpha + (1 - |x_{ij}|)\delta + e_{ij}$

- μ: constant (mid-point flowering time between homozygous genotypes)
- x_{ij}: indicator variable (genotype), coded as -1, 0 and 1 for genotypes qq, Qq and QQ
- α and β: additive and dominance effects

In matrix notation:

$$\begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{bmatrix} = \begin{bmatrix} 3.4 \\ 3.7 \\ 3.2 \\ 2.9 \\ 2.5 \\ 3.1 \\ 2.6 \end{bmatrix} = \begin{bmatrix} 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & -1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mu \\ \alpha \\ \delta \end{bmatrix} + \begin{bmatrix} e_{11} \\ e_{12} \\ e_{13} \\ e_{21} \\ e_{22} \\ e_{31} \\ e_{32} \end{bmatrix}$$

Least-Squares Estimation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\varepsilon} \sim (\mathbf{0}, \mathbf{I}_n\sigma^2) \quad \rightarrow \quad \varepsilon_i \overset{iid}{\sim} (0, \sigma^2)$$

An estimate ($\hat{\boldsymbol{\beta}}$) of the vector β can be obtained by the method of least-squares, which aims to minimize the residual sum of squares, given (in matrix notation) by:

$$\text{RSS} = \sum_{i=1}^{n} (\hat{\varepsilon}_i)^2 = \hat{\boldsymbol{\varepsilon}}^{\mathrm{T}}\hat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Taking the derivatives and equating to zero, it can be shown that the least-squares estimator of β is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$$

- It is shown that $\mathrm{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$ and $\mathrm{Var}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\sigma^2$

Least-Squares Estimation

The estimator $\hat{\boldsymbol{\beta}}_{\mathrm{OLS}} = \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{y}$ is called ordinary least squares (OLS) estimator, and it is indicated only in situations with homoscedastic and uncorrelated residuals

If the residual variance is heterogeneous (i.e., $\mathrm{Var}(\varepsilon_i) = \sigma_i^2 = w_i\sigma^2$), the residual variance matrix can be expressed as $\mathrm{Var}(\boldsymbol{\varepsilon}) = \mathbf{W}\sigma^2$, where **W** is a diagonal matrix with the elements w_i, a better estimator of β is given by: $\hat{\boldsymbol{\beta}}_{\mathrm{WLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{W}^{-1}\mathbf{y}$

which is generally referred to as weighted least squares (WLS) estimator.

Furthermore, in situations with a general residual variance-covariance matrix **V**, including correlated residuals, a generalized least squares (GLS) estimator $\hat{\boldsymbol{\beta}}_{\mathrm{GLS}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$ is obtained by minimizing the generalized sum of squares, given by: $\text{GSS} = \boldsymbol{\varepsilon}^{\mathrm{T}}\boldsymbol{\varepsilon} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$

Maximum Likelihood Estimation

Likelihood Function: any function of the model parameters that is proportional to the density function of the data

Hence, to use a likelihood-based approach for estimating model parameters, some extra assumptions must be made regarding the distribution of the data

In the case of the linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, if the residuals are assumed normally distributed with mean vector zero and variance-covariance matrix **V**, i.e. $\boldsymbol{\varepsilon} \sim \text{MVN}(\mathbf{0}, \mathbf{V})$, the response vector **y** is also normally distributed, with expectation $\text{E}[\mathbf{y}] = \mathbf{X}\boldsymbol{\beta}$ and variance $\text{Var}[\mathbf{y}] = \mathbf{V}$

Maximum Likelihood Estimation

The distribution of **y** has a density function given by:

$$p(\mathbf{y} \mid \boldsymbol{\beta}, V) = (2\pi)^{-n/2} \mid \mathbf{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{T} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

so that the likelihood and the log-likelihood functions can be expressed respectively as:

$$L(\boldsymbol{\beta}, V) \propto \mid \mathbf{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{T} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right\}$$

and

$$l(\boldsymbol{\beta}, V) = \log[L(\boldsymbol{\beta}, V)] \propto -\frac{1}{2}\log \mid \mathbf{V} \mid -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^{T} \mathbf{V}^{-1} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Maximum Likelihood Estimation

Assuming **V** known, the likelihood equations for **β** are given by taking the first derivatives of l(**β**,**V**) with respect to **β** and equating it to zero

The maximum likelihood estimator (MLE) for **β** is then shown to be:

$$\text{MLE}(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{y}$$

Note: Under normality the MLE coincides with the GLS estimator discussed previously

In addition, it is shown that: $\hat{\boldsymbol{\beta}} \sim \text{MVN}(\boldsymbol{\beta}, (\mathbf{X}^{\mathrm{T}}\mathbf{V}^{-1}\mathbf{X})^{-1})$

Two-stage Analysis of Longitudinal Data

Step 1

Supposed a series of longitudinal data (e.g., repeated measurements on time) on n individuals. Let y_{ij} represent the observation j (j = 1,2,…,n_i) on individual i (i = 1,2,…,n), and the following quadratic regression of measurements on time (z_{ij}) for each individual:

$$y_{ij} = \beta_{0i} + \beta_{1i} z_{ij} + \beta_{2i} z_{ij}^2 + \varepsilon_{ij}$$

where β_{0i}, β_{1i} and β_{2i} are subject-specific regression parameters, and ε_{ij} are residual terms, assumed normally distributed with mean zero and variance σ_ε^2

In matrix notation such subject-specific regressions can be expressed as:

$$\mathbf{y}_i = \mathbf{Z}_i\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i \qquad (1)$$

where $\mathbf{y}_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})^T$, $\boldsymbol{\beta}_i = (\beta_{0i}, \beta_{1i}, \beta_{2i})^T$, $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \ldots, \varepsilon_{in_i})^T \sim N(\mathbf{0}, \mathbf{I}\sigma^2_{\varepsilon})$ and

$$\mathbf{Z}_i = \begin{bmatrix} 1 & z_{i1} & z_{i1}^2 \\ 1 & z_{i2} & z_{i2}^2 \\ \vdots & \vdots & \vdots \\ 1 & z_{in_i} & z_{in_i}^2 \end{bmatrix}$$

Under these specifications, it is shown that the least-squares estimate of β_i is:

$$\hat{\boldsymbol{\beta}}_i = (\mathbf{Z}_i^T\mathbf{Z}_i)^{-1}\mathbf{Z}_i^T\mathbf{y}_i$$

Note that this is also the maximum likelihood estimate of β_i

Such estimates can be viewed as summary statistics for the longitudinal data, the same way one could use area under the curve (AUC), or peak (maximum value of y_{ij}), or mean response.

Two-stage Analysis of Longitudinal Data
Step 2

Supposed now we are interested on the effect of some other variables (such as gender, treatment, year, etc.) on the values of β_i

Such effects could be studied using a model as:

$$\hat{\boldsymbol{\beta}}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i$$

where $\mathbf{u}_i \sim N(\mathbf{0},\mathbf{D})$, which is an approximation for the model:

$$\boldsymbol{\beta}_i = \mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i \qquad (2)$$

Single-stage Analysis of Longitudinal Data

The two step-analysis described here can be merged into a single stage approach by substituting (2) in (1):

$$\mathbf{y}_i = \mathbf{Z}_i[\mathbf{W}_i\boldsymbol{\beta} + \mathbf{u}_i] + \boldsymbol{\varepsilon}_i$$

which can be expressed as:

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{u}_i + \boldsymbol{\varepsilon}_i$$

where $\mathbf{X}_i = \mathbf{Z}_i\mathbf{W}_i$. By concatenating observations from multiple individuals, we have the following mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}$$

Linear Mixed Effects Model

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

responses — incidence matrices — fixed effects — random effects — residuals

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

Estimation of Fixed Effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with $\boldsymbol{\varepsilon} = \mathbf{Z}\mathbf{u} + \mathbf{e}$, such that $\text{Var}[\boldsymbol{\varepsilon}] = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\text{T}} + \boldsymbol{\Sigma}$

➔ MLE of β :

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{y} \sim \text{MVN}(\boldsymbol{\beta}, (\mathbf{X}^{\text{T}}\mathbf{V}^{-1}\mathbf{X})^{-1})$$

where $\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^{\text{T}} + \boldsymbol{\Sigma}$

Prediction of Random Effects

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

$$\begin{aligned}\mathrm{E}[\mathbf{u}\,|\,\mathbf{y}] &= \mathrm{E}[\mathbf{u}] + \mathrm{Cov}[\mathbf{u}, \mathbf{y}^{\mathrm{T}}]\mathrm{Var}^{-1}[\mathbf{y}](\mathbf{y} - \mathrm{E}[\mathbf{y}])\\ &= \mathbf{GZ}^{\mathrm{T}}\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta}) = \mathbf{GZ}^{\mathrm{T}}(\mathbf{ZGZ}^{\mathrm{T}} + \mathbf{\Sigma})^{-1}(\mathbf{y} - \mathbf{X\beta})\end{aligned}$$

Replacing $\boldsymbol{\beta}$ by its estimate:

$$\hat{\mathbf{u}} = \mathbf{GZ}^{\mathrm{T}}(\mathbf{ZGZ}^{\mathrm{T}} + \mathbf{\Sigma})^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

Mixed Model Equations

$$\begin{bmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{X} & \mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{X} & \mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{X}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{y}\\ \mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{y}\end{bmatrix}$$

BLUP and BLUE:

$$\begin{cases}\hat{\mathbf{u}} = (\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\\ \hat{\boldsymbol{\beta}} = \{\mathbf{X}^{\mathrm{T}}[\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}]\mathbf{X}\}^{-1}\\ \qquad \times \mathbf{X}^{\mathrm{T}}[\mathbf{\Sigma}^{-1} - \mathbf{\Sigma}^{-1}\mathbf{Z}(\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}^{\mathrm{T}}\mathbf{\Sigma}^{-1}]\mathbf{y}\end{cases}$$

Estimation of Variance Components

BLUE and BLUP require knowledge of $\mathbf{G}$ and $\boldsymbol{\Sigma}$

These matrices, however, are rarely known and must be estimated

Variance and covariance components estimation:

- Analysis of Variance (ANOVA)
- Maximum Likelihood
- Restricted Maximum Likelihood (REML)
- Bayesian Inference

Mixed Models in Animal and Plant Breeding

Animal/plant 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

Mixed Models in Animal and Plant Breeding

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

Animal Model

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 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 breeding programs, we consider here the so-called Animal Model in situations with a single trait and a single observation (including missing values) per individual

Animal Model

The animal model can be described as:

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

- $\mathbf{y}$ is an $(n \times 1)$ vector of observations (phenotypic scores)
- $\boldsymbol{\beta}$ is a $(p \times 1)$ vector of fixed effects (e.g. herd-year-season effects)
- $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$ is a $(q \times 1)$ vector of breeding values (relative to all individuals with record or in the pedigree file, such that q is in general bigger than n)
- $\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}_n\sigma_e^2)$ represents residual effects, where σ_e^2 is the residual variance

The Matrix *A*

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 σ_a^2 is the additive genetic variance in the base population

Hence, under the animal model, $\mathbf{G} = \mathbf{A}\sigma_a^2$, where ***A*** is the additive genetic (or numerator) relationship matrix, having elements given by $a_{ii'} = 2\theta_{ii'}$

The Matrix *A*

For each animal i in the pedigree (i = 1, 2,…,n), going from older to younger animals, compute a_{ii} and a_{ij} (j = 1, 2,…,i-1) as follows:

If both parents (s and d) of animal i are known:

$a_{ij} = a_{ji} = (a_{js} + a_{jd})/2$ and $a_{ii} = 1 + a_{sd}/2$

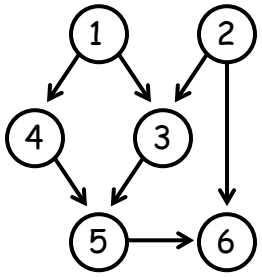
If only one parent (e.g. d) of animal i is known:

$a_{ij} = a_{ji} = a_{jd}/2$ and $a_{ii} = 1$

If parents unknown:

$a_{ij} = a_{ji} = 0$ and $a_{ii} = 1$

Example



Animal	Sire	Dam
1	-	-
2	-	-
3	1	2
4	1	-
5	4	3
6	5	2

$$A = \begin{bmatrix} 1 & 0 & .5 & .5 & .5 & .25 \\ 0 & 1 & .5 & 0 & .25 & .625 \\ .5 & .5 & 1 & .25 & .625 & .563 \\ .5 & 0 & .25 & 1 & .625 & .313 \\ .5 & .25 & .625 & .625 & 1.125 & .688 \\ .25 & .625 & .563 & .313 & .688 & 1.125 \end{bmatrix}$$

pedigree matrix A

Animal Model

In general, in animal/plant breeding interest is on prediction of 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)

Animal Model

Since under the animal model $\mathbf{G}^{-1} = \mathbf{A}^{-1}\sigma_a^{-2}$ and $\mathbf{R}^{-1} = \mathbf{I}_n\sigma_e^{-2}$, the mixed model equations can be expressed as:

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

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

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

Conditional on the variance components ratio λ, the BLUP of the breeding values are given then by:

$$\hat{\mathbf{u}} = (\mathbf{Z}^T\mathbf{Z} + \lambda\mathbf{A}^{-1})^{-1}\mathbf{Z}^T(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

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) or Expected Progeny Difference (EPD)), 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)

Prediction Accuracy

The calculation of $\rho(\hat{u}_i, u_i)$ requires the diagonal elements of the inverse of the MME coefficient matrix, represented as:

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

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

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

where c_i^{uu} is the i-th diagonal element of $\mathbf{C}^{uu}$, relative to animal i.

Prediction Accuracy

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:

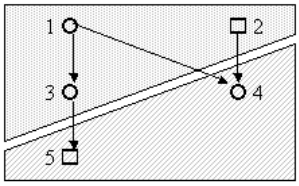
$$PEV = (1 - r_i^2)\sigma_a^2$$

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

$$r_i^2 = 1 - c_i^{uu}\sigma_e^2 / \sigma_a^2 = 1 - \lambda c_i^{uu}$$

Animal Model

herd 1



herd 2

Animal	Sire	Dam	Herd	Observation
1	–	–	h1	310
2	–	–	h1	–
3	–	1	h1	270
4	2	1	h2	350
5	–	3	h2	–

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

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

Animal Model

Breeding values: $\mathbf{u} \sim N(\mathbf{0}, \mathbf{A}\sigma_u^2)$, with

$$\mathbf{A} = \begin{bmatrix}1 & 0 & 0.5 & 0.5 & 0.25\\0 & 1 & 0 & 0.5 & 0\\0.5 & 0 & 1 & 0.25 & 0.5\\0.5 & 0.5 & 0.25 & 1 & 0.125\\0.25 & 0 & 0.5 & 0.125 & 1\end{bmatrix}$$

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

$$\lambda = \frac{\sigma_e^2}{\sigma_u^2} = \frac{1-h^2}{h^2}$$

R Code

animal model
toy example

```
y<-matrix(c(310,270,350),nrow=3)
X<-matrix(c(1,1,0,0,0,1),nrow=3)
Z<-matrix(c(1,0,0,0,0,0,0,1,0,0,0,0,0,1,0),nrow=3, byrow = TRUE)
A<-matrix(c(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),nrow=5)

h2<-1/3 # heritability
a=(1-h2)/h2

# crossproducts
XX<-crossprod(X,X)
XZ<-t(X) %*% Z
ZX<-t(Z) %*% X
ZZ<-crossprod(Z,Z)+a*solve(A)

# mixed model equations
# coefficient matrix and right hand side
C<-rbind(cbind(XX,XZ),cbind(ZX,ZZ))
rhs<-rbind(t(X) %*% y,t(Z) %*% y)

#solution
theta.hat <- solve(C) %*% rhs
```

$$h^2 = \frac{1}{3} \rightarrow \alpha = 2 \Rightarrow \begin{cases} \hat{h}_1 = 290 \\ \hat{h}_2 = 348 \\ \hat{u}_1 = \ 4.0 \\ \hat{u}_2 = \ 0.0 \\ \hat{u}_3 = -4.0 \\ \hat{u}_4 = \ 2.0 \\ \hat{u}_5 = -2.0 \end{cases}$$

Animal Model

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): Weaning weight (kg) of piglets, progeny of three sows mated to two boars:

Piglet	Sire	Dam	Sex	Weight
6	1	2	1	90
7	1	2	2	70
8	1	2	2	65
9	3	4	2	98
10	3	4	1	106
11	3	4	2	60
12	3	4	2	80
13	1	5	1	100
14	1	5	2	85
15	1	5	1	68

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**:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \mathbf{Wc} + \mathbf{e}$$

Weight — Sex — Breeding values — Common environment — Residual

Assuming that $\sigma^2_u = 20$, $\sigma^2_c = 15$ and $\sigma^2_e = 65$, the MME are as follows:

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

where $\lambda_1 = \frac{\sigma^2_e}{\sigma^2_u} = 3.25$ and $\lambda_2 = \frac{\sigma^2_e}{\sigma^2_c} = 4.\dot{3}$

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

Effects	Solutions
Sex	
1	91.493
2	75.764
Animals	
1	-1.441
2	-1.175
3	1.441
4	1.441
5	-0.266
6	-1.098
7	-1.667
8	-2.334
9	3.925
10	2.895
11	-1.141
12	1.525
13	0.448
14	0.545
15	-3.819
Environ.	
2	-1.762
4	2.161
5	-0.399

Mrode example