AGR 6322 Advanced Plant Breeding

Fall 2018

Use of BLUP in Plant Breeding

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

Goals for today

Introduction of BLUP in Plant Breeding

Linear Mixed Model

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

where:

- $\mathbf{y}$: response vector; observations
- $\boldsymbol{\beta}$: vector of fixed effects
- $\mathbf{u}$: vector of random effects; $\mathbf{u} \sim N(\mathbf{0}, \mathbf{G})$
- $\mathbf{X}$ and $\mathbf{Z}$: (known) incidence matrices
- $\mathbf{e}$: residual vector; $\mathbf{e} \sim N(\mathbf{0}, \boldsymbol{\Sigma})$

Mixed Models in 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.

Data from natural populations and agricultural species are generally highly unbalanced and fragmented by numerous kinds of relationships.

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, andnatural or artificial selection.

Linear Mixed Model

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

- **Mixed models** extend the linear model by allowing a more flexible specification of the errors (and other random factors). Hence, it allows for a different type of inference and also allows to incorporate *correlation* and *heterogeneous variances* between the observations.
- **Fixed effects:** are those factors whose levels are selected by a nonrandom process or whose levels consist of the entire population of possible levels. Inferences are made *only* to those levels included in the study. Hint: all levels of interest are in your data set.
- **Random effects:** a factor where its levels consist of a random sample of levels from a population of possible levels. The inference is about the population of levels, not just the subset of levels included in the study.
- Mixed linear models contain both random and fixed effects.

Linear Mixed Model

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

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

α_i fixed effect of the i^{th} block

g_j random effect of the j^{th} variety, $E(g_j) = 0$, $V(g_j) = \sigma_g^2$

e_{ij} random error of the ij^{th} observation, $E(e_{ij}) = 0$, $V(e_{ij}) = \sigma^2$

$i = 1, \ldots, 6$ (r blocks)

$j = 1, \ldots, 12$ (t treatments)

Linear Mixed Model

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

Hypothesis of interest

Fixed effects: $H_0: \mu_1 = \mu_2 = \ldots = \mu_t$

$H_1: \mu_i \neq \mu_j$ for some i, j in the set $1 \ldots t$

(i.e. is there a significant treatment effect)

Test statistic: F or t

Random effects: $H_0: \sigma_g^2 = 0$

$H_1: \sigma_g^2 > 0$

(i.e. is there a significant variation due to the random effects)

Test statistic: Chi-square (likelihood ratio test)

Linear Mixed Model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{g} + \mathbf{e} \qquad E\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix} \qquad Var\begin{bmatrix}\mathbf{g}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{bmatrix}$$

Assumptions

- Random effects: $E(\mathbf{g}) = \mathbf{0}$, $\mathbf{V}(\mathbf{g}) = \mathbf{G} = \mathbf{G}(\boldsymbol{\theta})$
- Deviations: $E(\mathbf{e}) = \mathbf{0}$, $\mathbf{V}(\mathbf{e}) = \mathbf{R} = \mathbf{R}(\boldsymbol{\theta})$
- $\mathbf{g}$ and $\mathbf{e}$ independent.

hence, $E(\mathbf{y}) = \mathbf{X}\boldsymbol{\beta}$

$Var(\mathbf{y}) = \mathbf{V} = \mathbf{V}(\boldsymbol{\theta}) = \mathbf{V}(\mathbf{y}) = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}$

Note: normality assumptions can be made about $\mathbf{g}$ and $\mathbf{e}$.

$\mathbf{g} \sim \text{MVN}(0, \mathbf{G})$ and $\mathbf{e} \sim \text{MVN}(0, \mathbf{R})$

Linear Mixed Model

- Variance components need to be estimated before obtaining estimates of fixed/random effects and performing any type of inference.

$$\hat{\theta} \Rightarrow \begin{matrix} \hat{\mathbf{G}} = \mathbf{G}(\hat{\theta}) \\ \hat{\mathbf{R}} = \mathbf{R}(\hat{\theta}) \end{matrix} \Rightarrow \hat{\mathbf{V}} = \mathbf{V}(\hat{\theta}) = \mathbf{Z}\hat{\mathbf{G}}\mathbf{Z}' + \hat{\mathbf{R}}$$

- Restricted/residual maximum likelihood (REML) is a likelihood-based method used to estimate these variance components and is based assuming that both **g** and **e** follow a multivariate normal distribution.
- The REML variance component estimates are later used to estimate the solutions of fixed and random effects.
- Henderson (1950) derived the Mixed Model Equations (MME) to obtain the solutions of all effects:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{y} \qquad \text{BLUE} \rightarrow \text{EBLUE}$$

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \qquad \text{BLUP} \rightarrow \text{EBLUP}$$

Linear Mixed Model

Process:

- Requires to estimate treatment effects and variance components.
- Based in maximization of a non-linear function (*maximum likelihood*).

Method:

REML (most common)

Iteratively searches over the parameters space for parameter values that maximize a function of the parameters related to the chance of observing the data collected, but that is defined as free of fixed effects.

Efficient method that allows for missing values and unbalanced designs.

Linear Mixed Model

LRT: likelihood ratio test

- Based on asymptotic derivations.
- Used to compare nested models and is valid if the fixed effects are the same (under REML).
- Examples: $H_0: \rho = 0$ against $H_0: \rho \neq 0$
  $H_0: \sigma^2_g = 0$ against $H_0: \sigma^2_g > 0$
- Test Statistic: $d = 2\,[\,\log L_2 - \log L_1] \sim \chi^2_{r2\text{-}r1}$

Hypothesis	P-value
Two-sided	$\text{Prob}(\chi^2_{r2\text{-}r1} > d)$
One-sided	$0.5(1 - \text{Prob}(\chi^2_1 \leq d))$

Breeding Value (BLUP)

Definition

- The **average effect** of the parental *alleles* passed to the offspring determine the mean genotypic value of its offspring, or
- The **genetic value** of an individual (or cross) judged by mean value of its progeny.

  - Sum of average effects across loci (theoretical, now molecular).
  - Mean value of offspring (practical).

- Not equivalent concepts if interaction between loci is present or if mating is not at random.

Estimation

- By **BLUP** (Best Linear Unbiased Predictor), i.e. the *prediction* of the random effects from linear mixed models.

Breeding Value (BLUP)

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$\hat{\mathbf{g}}$	vector of random effect predictions.
$\hat{\mathbf{G}}\mathbf{Z}' = \mathbf{C}'$	covariance matrix between observations and random (genetic) effects to be predicted.
$\hat{\mathbf{V}}$	variance-covariance matrix for the observations.
$(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$	individual observations 'corrected' by fixed effects.

$$\hat{\mathbf{g}} = \hat{\mathbf{G}}\mathbf{Z}'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

$$\hat{g}_i = [\sigma_a^2 / \sigma_p^2] \times (y_i - \bar{y})$$

$$\hat{g}_i = h^2 \times (y_i - \bar{y}) \rightarrow \Delta\text{Gain}$$

Genetic Models

Individual Models

- **Animal model.**
  - One or two parents known. Individual/parent selection.
- **Reduced animal model.**
  - One or two parents known. Individual/parent selection (only individuals with records).

Parental Models

- **Half-sib crosses / sire model.**
  - One parent known. Parent selection.
- **Full-sib crosses model.**
  - Both parents known. Parent/cross selection. Add and Dom effects estimable.
- **Family model.**
  - Both parents known. Cross selection. Add and Dom effects confounded.
- **Clonal model.**
  - Clonally replicated individuals. Parent/cross/individual selection.

Incorporating Pedigree in the Genetic Models

- Why worry about the pedigree in genetic analyses?

  Statistically, random genetic effects (i.e. BLUPs) are not independent and their matrix of correlations or co-variances (**G** or **A**) needs to be specified.

  Genetically, it is important to consider information about relatives as they will share some alleles, and therefore their response is correlated.

- How to incorporate this information?

  Genetic relationships can be calculated using genetic theory (expected values) or molecular information (e.g. SNPs), and included into the linear mixed model by specifying a pedigree file,

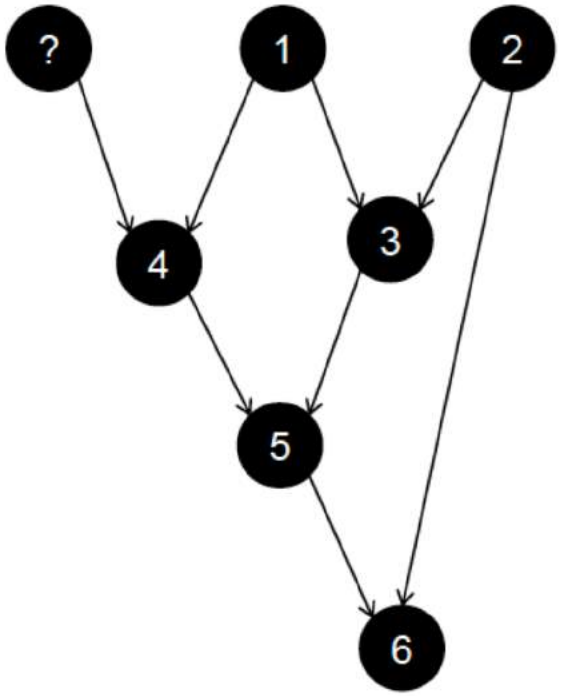
- Are there other benefits?

  Many. It is a more efficient use of the information about individuals, but also genetic values of individual not tested, but with relatives tested, can be *predicted* and selected.

Incorporating Pedigree in the Genetic Models

Example

Pedigree of a group of individuals:



Individual	Male	Female
3	1	2
4	1	Unknown
5	4	3
6	5	2

Incorporating Pedigree in the Genetic Models

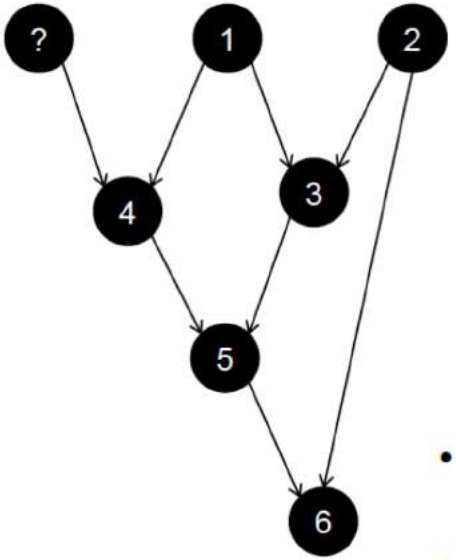
Numerator relationship matrix (A)

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{bmatrix} 1.00 & 0.00 & 0.50 & 0.50 & 0.50 & 0.25 \\ & 1.00 & 0.50 & 0.00 & 0.25 & 0.625 \\ & & 1.00 & 0.25 & 0.625 & 0.563 \\ & & & 1.00 & 0.625 & 0.313 \\ & & & & 1.125 & 0.688 \\ & & & & & 1.125 \end{bmatrix} \end{array}$$

- Linked to the concept of **identity by descent**.
- **Diagonal** $a_{ii} = 1 + F_i$ (inbreeding coefficient on individual i)

  Twice the probability that two gametes taken at random from animal i will carry identical alleles by descent.
- **Off-diagonal** a_{ij} numerator of the coefficient of relationship between animal i and j.

Incorporating Pedigree in the Genetic Models



Indiv	Male	Female
1	0	0
2	0	0
3	1	2
4	1	0
5	4	3
6	5	2

- Pedigree information is associated with proper management and validation/check of data.
- Individuals need to be ordered by generation (e.g. parents need to be defined before progeny).
- All parents need to be defined in pedigree file (the inclusion of founder parents is optional).
- All individuals present in dataset (i.e. levels associated with pedigree file) need to be defined in pedigree file.
- Individuals can be defined as male or female parents (but this should be checked if is not biologically possible).

Incorporating Pedigree in the Genetic Models

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

- $\mathbf{y}$ 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

Incorporating Pedigree in the Genetic Models

The matrix G describing the covariances among the random effects (breeding values) follows from standard results for the covariances between relatives

The additive genetic covariance between two relatives i and i' is given by $2\theta_{ii'}\sigma_a^2$

$$\mathbf{G} = \mathbf{A}\sigma_a^2$$

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

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

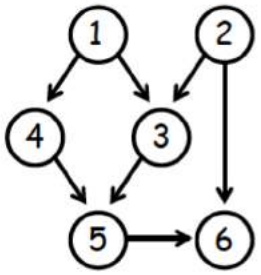
If only one parent (e.g. d) of animal i is known:

$a_{ij} = a_{ji} = a_{jd}/2$ 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

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