Lecture 5: BLUP (Best Linear Unbiased Predictors) of genetic values

Bruce Walsh lecture notes
Tucson Winter Institute
9 - 11 Jan 2013

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

The classic designs (ANOVA, P-O regression) for variance components are simple, involving only a single type of relative comparison. Further, they assume balanced designs, with the number of offspring the same in each family.

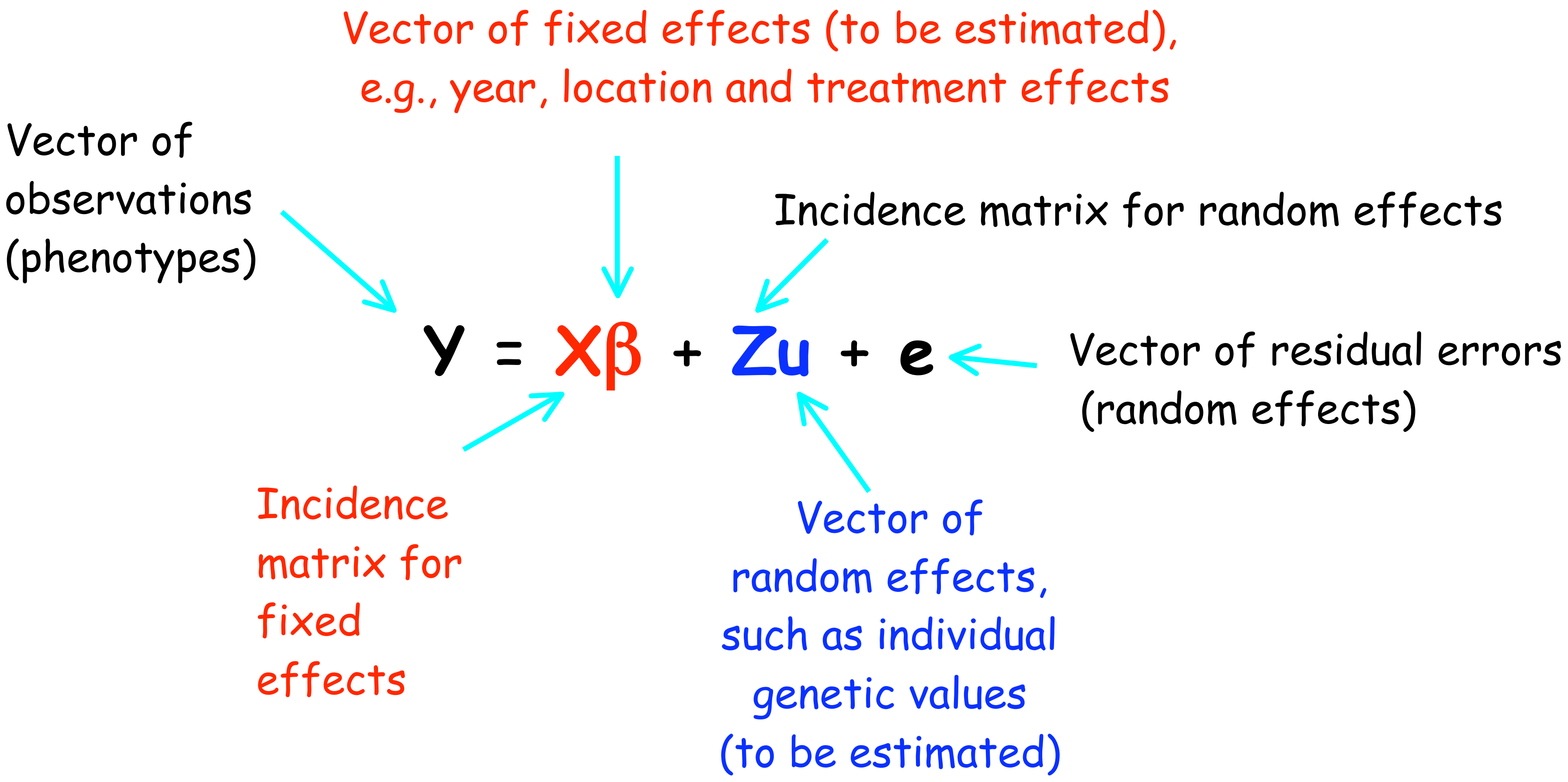
In the real world, we often have a pedigree of relatives, with a very unbalanced design. Fortunately, the general mixed model (so called because it includes both fixed and random effects), offers an ideal platform for both estimating genetic variances as well a predicting the breeding values of individuals.

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

BLUP in plant breeding

- BLUP has migrated from animal breeding into plant breeding.
- Advantages:
  - Handles unbalanced designs
  - Uses information for all relatives measured to improve estimates
- BLUP can be used to estimate a variety of genetic values
  - GCA, SCA, line values (i.e., genotypic values of pure lines)
  - One can also use BLUP machinery to estimate environmental effects

The general mixed model



Vector of fixed effects (to be estimated), e.g., year, location and treatment effects

Vector of observations (phenotypes)

Incidence matrix for random effects

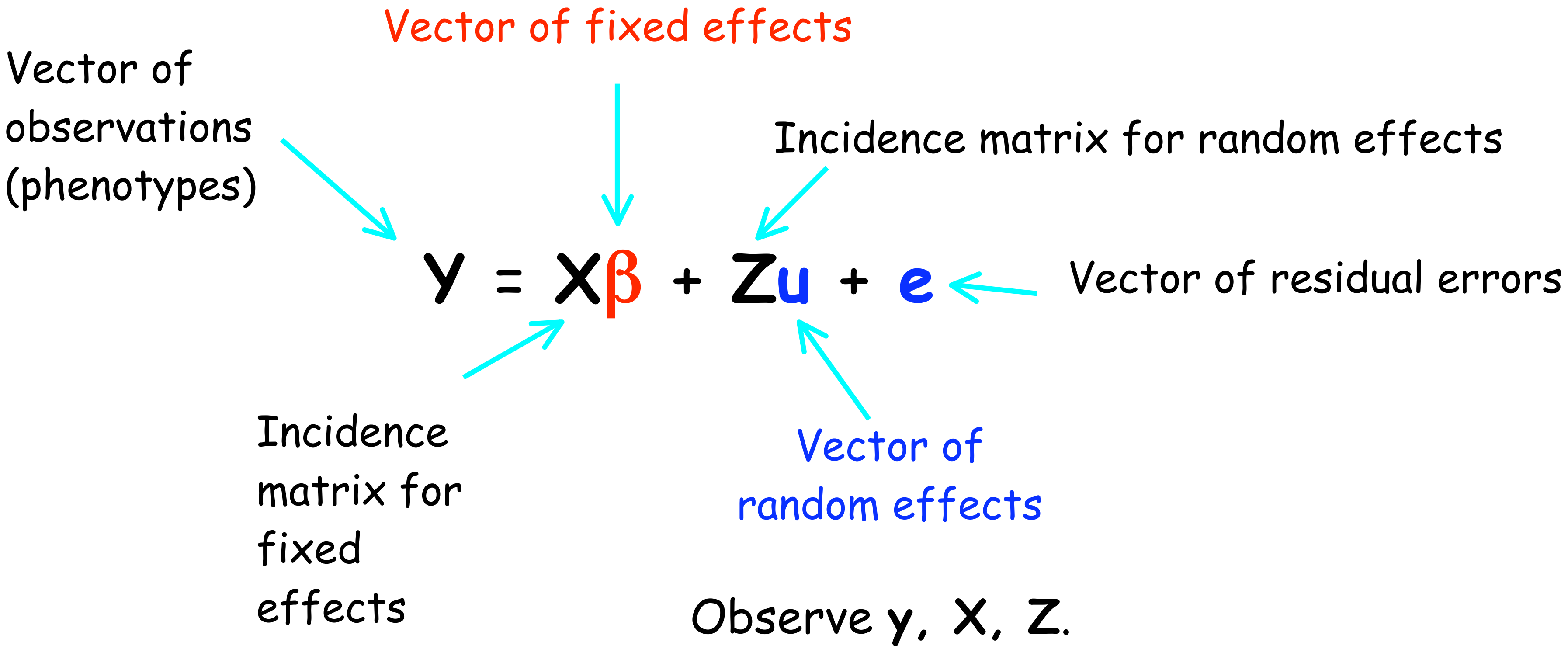
$$y = X\beta + Zu + e$$

Vector of residual errors (random effects)

Incidence matrix for fixed effects

Vector of random effects, such as individual genetic 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 (fathers), each of which is mated to a random female (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

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

Breeding value of sire i

Effect of environment j

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

The mixed model vectors and matrices become

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

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

$$V = ZGZ^T + R$$

Estimating fixed Effects & Predicting Random Effects

For a mixed model, we observe **y**, **X**, and **Z**

β, u, R, and G are generally unknown

Two complementary estimation issues

(i) Estimation of β and **u**

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

Estimation of fixed effects

BLUE = Best Linear Unbiased Estimator

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

Prediction of random effects

BLUP = Best Linear Unbiased Predictor

Recall $V = ZGZ^T + R$

Let's return to our example

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

Suppose σ^2_e = 6, giving R = 6* I

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

Suppose σ^2_A = 8, giving G = 8/4*I

Solving, recalling that $V = ZGZ^T + R$

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

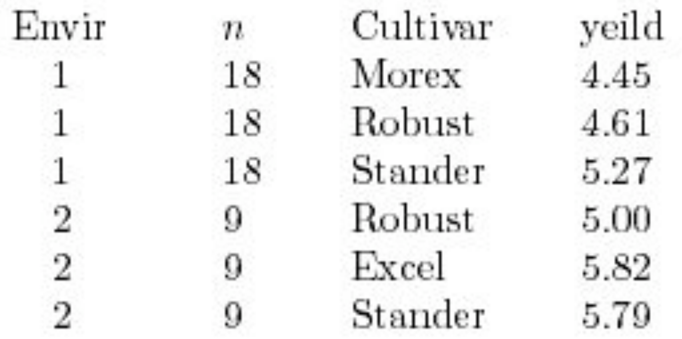
$$= \begin{pmatrix} 8 & 2 & 0 & 0 & 0 & 0 \\ 2 & 8 & 0 & 0 & 0 & 0 \\ 0 & 0 & 8 & 2 & 0 & 0 \\ 0 & 0 & 2 & 8 & 0 & 0 \\ 0 & 0 & 0 & 0 & 8 & 2 \\ 0 & 0 & 0 & 0 & 2 & 8 \end{pmatrix} \quad \text{giving} \quad \mathbf{V}^{-1} = \frac{1}{30}\cdot\begin{pmatrix} 4 & -1 & 0 & 0 & 0 & 0 \\ -1 & 4 & 0 & 0 &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

$$\widehat{\boldsymbol{\beta}} = \begin{pmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{pmatrix} = \left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} = \frac{1}{18}\begin{pmatrix} 148 \\ 235 \end{pmatrix}$$

$$\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ u_3 \end{pmatrix} = \mathbf{G}\mathbf{Z}^T\mathbf{V}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \frac{1}{18}\begin{pmatrix} -1 \\ 2 \\ -1 \end{pmatrix}$$

BLUP estimates of line values

Bernardo example (11.3.1): yield in four (related) inbred lines of Barley raised over two sets of environments



Envir	n	Cultivar	yeild
1	18	Morex	4.45
1	18	Robust	4.61
1	18	Stander	5.27
2	9	Robust	5.00
2	9	Excel	5.82
2	9	Stander	5.79

Environment type (fixed)

Line value (random)

$$y_{ij} = \mu + t_i + u_j + e_{ij}$$

Model $\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \quad \mathbf{u} \sim MVN(\mathbf{0}, \mathbf{G}), \quad \mathbf{e} \sim MVN(\mathbf{0}, \mathbf{R})$

Envir	n	Cultivar	yeild
1	18	Morex	4.45
1	18	Robust	4.61
1	18	Stander	5.27
2	9	Robust	5.00
2	9	Excel	5.82
2	9	Stander	5.79

$b_i = \mu + t_i$

Line values

$$\begin{pmatrix} 4.45 \\ 4.61 \\ 5.27 \\ 5.00 \\ 5.82 \\ 5.79 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix} + \begin{pmatrix} e_{11} \\ e_{12} \\ e_{14} \\ e_{22} \\ e_{23} \\ e_{24} \end{pmatrix}$$

$$\mathrm{Var}(\mathbf{u}) = \mathbf{G} = 2\sigma_A^2 \begin{pmatrix} 1 & 0.5 & 0.44 & 0.34 \\ 0.5 & 1 & 0.84 & 0.67 \\ 0.44 & 0.84 & 1 & 0.71 \\ 0.34 & 0.67 & 0.71 & 1 \end{pmatrix}, \quad \mathrm{Var}(\mathbf{e}) = \mathbf{G} = \mathbf{R} = \frac{\sigma_e^2}{18} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 2 \end{pmatrix}$$

Relationship values (from pedigree data on how lines are related)

Observations differ in their residual error due to sample size differences

Henderson's Mixed Model Equations

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{u} + \mathbf{e}, \quad \mathbf{u} \sim (0, G),\ \mathbf{e} \sim (0,\ \mathbf{R}),\ \mathrm{cov}(\mathbf{u},\mathbf{e}) = 0,$$

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

$$\begin{array}{c} \begin{array}{cc} p \times p & p \times q \end{array} \\ \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \\ \begin{array}{cc} q \times pq & q \times q \end{array} \end{array} \begin{pmatrix} \widehat{\beta} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

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

Easier to numerically work with than BLUP/BLUE equations

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

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

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}^T + \mathbf{R}$$

$\uparrow$

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 example on slide 6 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

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

As found above

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 $R = \sigma^2_e * I$, so that $R^{-1} = 1/(\sigma^2_e) * I$.

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

The "animal" model estimates the breeding value for each individual, even for a plant or tree! Same approach also works to estimate line (genotypic) values for inbreds.

Returning to the animal model

Henderson's mixed model equations

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

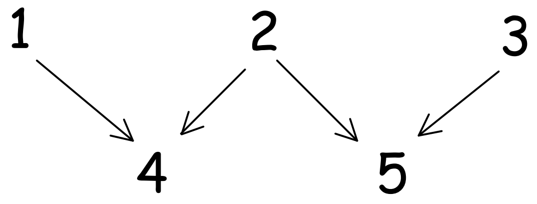
here $\lambda = \sigma^2_e / \sigma^2_A = (1-h^2)/h^2$

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



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

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
  - $\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{e}$
  - $\mathbf{a} \sim (0, \sigma_A^2 \mathbf{A})$, $\mathbf{e} \sim (0, \sigma_e^2 \mathbf{I})$
  - $\text{BLUP}(\mathbf{a}) = \sigma_A^2 \mathbf{AZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta)$
  - Where $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R} = \sigma_A^2 \mathbf{ZAZ}^T + \sigma_e^2 \mathbf{I}$
- 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 $\mathbf{Z} = \mathbf{A} = \mathbf{I} = (1)$,
  - $\mathbf{V} = \sigma_A^2 + \sigma_e^2$
  - $\sigma_A^2 \mathbf{AZ}^T\mathbf{V}^{-1} = \sigma_A^2 / (\sigma_A^2 + \sigma_e^2) = h^2$
  - $\text{BLUP}(a) = h^2(y - \mu)$

- More generally, with single observations on n unrelated individuals,
  - $A = Z = I_{n \times n}$
  - $\mathbf{V} = \sigma_A^2\mathbf{ZAZ}^T + \sigma_e^2\mathbf{I} = (\sigma_A^2 + \sigma_e^2)\ \mathbf{I}$
  - $\sigma_A^2\ \mathbf{AZ}^T\mathbf{V}^{-1} = \ h^2\,\mathbf{I}$
  - $\text{BLUP}(\mathbf{a}) = \sigma_A^2\mathbf{AZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\beta) = h^2(\mathbf{y} - \mu)$
- 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

BLUP is a shrinkage estimator

- For a single observation on one individual, BLUP(a) = $h^2(y-\mu)$
  - The difference between the observed value (y) and the mean (μ) is shrunk by the factor h^2 --- shrinks the estimate back towards the mean (zero in the case of BVs)
- More generally, BLUP(**a**) = $\mathbf{GZ}^T\mathbf{V}^{-1}(\mathbf{y}- \mathbf{X}\beta)$
  - First adjusts observations (**y**) for fixed effects (**X**β) and then regresses this difference back towards zero (the mean BV), as **Cov*Var**$^{-1}$ is a generalized regression coefficient

The Relationship Matrix *A*

- Typically given from a pedigree, but increasingly being estimated from marker data
- The diagonal elements indicate the amount of inbreeding
  - A_{ii} = 1 + F_i, where F_i is inbreeding coefficent for individual i.
  - For a fully-inbred, A_{ii} = 2

Marker-based relationship matrices

- There are two reasons for using a marker-estimated relationship matrix
  - Pedigree either unknown or poorly known
  - With very dense markers, provides a better estimate than a known pedigree. Why?
    - Consider two (non-inbred) full-sibs. The expectation under a pedigree is that they share exactly half their genes.
    - However, there is a sampling variance about this expected value, so that some pair of sibs may share more than 50%, while another may share less. Using markers to detect such pairs improves the estimated values
    - This is called G-BLUP (in animal breeding) and is a form of genomic selection

Marker-based relationship matrix

Simplest case is to consider a very large number (L) of SNPs, and treat alike in state as IBD, and then compute the probability f_{xy} that x and y share a randomly-drawn allele for each SNP marker. Twice the average over all markers is the entry for x and y in the relationship matrix (as $A_{xy} = 2f_{xy}$)

SNP genotype for x

SNP genotype for y

	00	01	11
00	1	0.5	0
01	0.5	0.5	0.5
11	0	0.5	1

Values for f_{xy} given the SNP genotypes

Estimation of R and G

A second estimation issue concerns 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

Multiple random effects

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

y is a n x 1 vector of observations

β is a q x 1 vector of fixed effects

a is a p x 1 vector of random effects

u is a m x 1 vector of random effects

X is n x q, **Z** is n x p, **W** is n x m

y, X, Z, **W** observed. β, **a**, **u**, **e** to be estimated

Covariance structure

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{Wu} + \mathbf{e}$$

Defining the covariance structure key in any mixed-model

Suppose **e** ~ $(0,\sigma_e^2\ I)$, **u** ~ $(0,\sigma_u^2\ I)$, **a** ~ $(0,\sigma_A^2\ A)$,
as with breeding values

These covariances matrices are still not sufficient, as we have yet to give describe the relationship between **e**, **a**, and **u**. If they are independent:

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{Wu} + \mathbf{e}$

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{u} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_u^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

Covariance matrix for the vector of observations **y**

$$\mathrm{Var}(\mathbf{y}) = \mathbf{V} = \mathbf{ZAZ}^T \sigma_A^2 + \mathbf{WW}^T \sigma_u^2 + \mathbf{I}\,\sigma_e^2$$

Note that if we ignored the second vector **u** of random effects, and assumed $\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{e}^*$, then $\mathbf{e}^* = \mathbf{Wu} + \mathbf{e}$, with $\mathrm{Var}(\mathbf{e}^*) = \sigma_e^2\, \mathrm{I} + \sigma_u^2\, \mathbf{WW}^T$

Consequence of ignoring random effects is that these are incorporated into the residuals, potentially compromising its covariance structure

Mixed-model Equations

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z} & \mathbf{X}^T\mathbf{W} \\ \mathbf{Z}^T\mathbf{X} & \mathbf{Z}^T\mathbf{Z}+\lambda_A\mathbf{A}^{-1} & \mathbf{Z}^T\mathbf{W} \\ \mathbf{W}^T\mathbf{X} & \mathbf{W}^T\mathbf{Z} & \mathbf{W}^T\mathbf{W}+\lambda_u\mathbf{I} \end{pmatrix} \begin{pmatrix} \widehat{\beta} \\ \widehat{\mathbf{a}} \\ \widehat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}^T\mathbf{y} \\ \mathbf{W}^T\mathbf{y} \end{pmatrix}$$

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_u^2}$$

The repeatability model

- Often, multiple measurements (aka “records”) are collected on the same individual
- Such a record for individual k has three components
  - Breeding value a_k
  - Common (permanent) environmental value p_k
  - Residual value for ith observation e_{ki}
- Resulting observation is thus
  - $z_{ki} = \mu + a_k + p_k + e_{ki}$
- The repeatability of a trait is $r = (\sigma_A^2+\sigma_p^2)/\sigma_z^2$
- Resulting variance of the residuals is $\sigma_e^2 = (1-r)\,\sigma_z^2$

Resulting mixed model

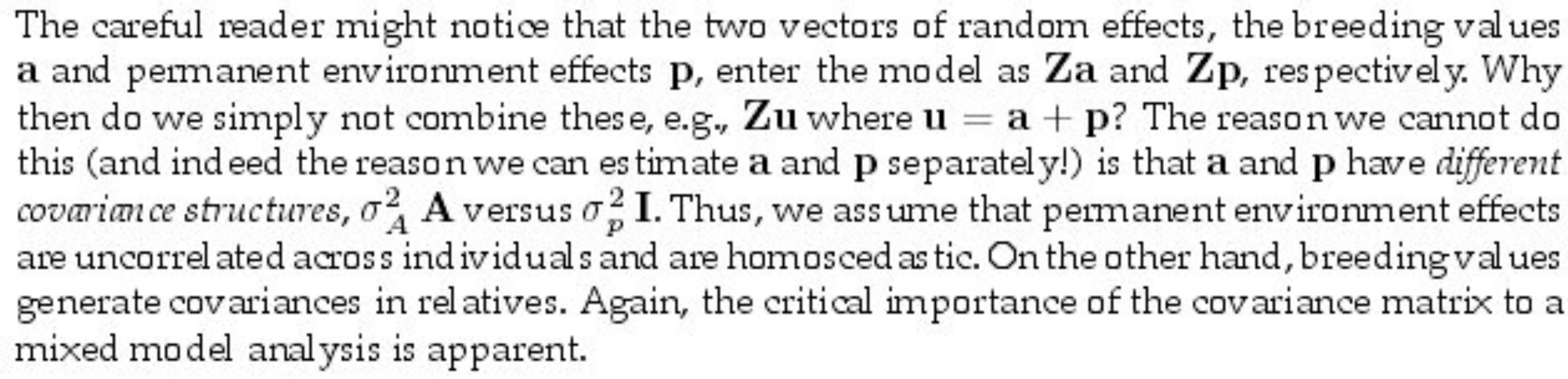
$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{Zp} + \mathbf{e}$$

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_A^2 \cdot \mathbf{A} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \sigma_p^2 \cdot \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2 \cdot \mathbf{I} \end{pmatrix}$$

Notice that we could also write this model as

$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}(\mathbf{a} + \mathbf{p}) + \mathbf{e} = \mathbf{y} = \mathbf{X}\beta + \mathbf{Zv} + \mathbf{e}$, $\mathbf{v} = \mathbf{a}+\mathbf{p}$

In class question: Why can we obtain separate estimates of **a** and **p**?



The careful reader might notice that the two vectors of random effects, the breeding values $\mathbf{a}$ and permanent environment effects $\mathbf{p}$, enter the model as $\mathbf{Za}$ and $\mathbf{Zp}$, respectively. Why then do we simply not combine these, e.g., $\mathbf{Zu}$ where $\mathbf{u} = \mathbf{a} + \mathbf{p}$? The reason we cannot do this (and indeed the reason we can estimate $\mathbf{a}$ and $\mathbf{p}$ separately!) is that $\mathbf{a}$ and $\mathbf{p}$ have *different covariance structures*, $\sigma_A^2\ \mathbf{A}$ versus $\sigma_p^2\ \mathbf{I}$. Thus, we assume that permanent environment effects are uncorrelated across individuals and are homoscedastic. On the other hand, breeding values generate covariances in relatives. Again, the critical importance of the covariance matrix to a mixed model analysis is apparent.

The incident matrix Z

Suppose we have a total of 7 observations/records, with 3 measures from individual 1, 2 from individual 2, and 2 from individual 3. Then:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{12} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \quad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

Why? Matrix multiplication. Consider y_{21}.

$$y_{21} = \mu + A_2 + p_2 + e_{21}$$

Consequences of ignoring p

- Suppose we ignored the permanent environment effects and assumed the model $\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{e^*}$
  - Then $\mathbf{e^*} = \mathbf{Zp} + \mathbf{e}$,
  - Var($\mathbf{e^*}$) = $\sigma_e^2\, \mathrm{I} + \sigma_p^2\, \mathbf{ZZ}^T$
- Assuming that Var($\mathbf{e^*}$) = $\sigma_e^2\, \mathrm{I}$ gives an incorrect model
- We could either
  - use $\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{e^*}$ with the correct error structure (covariance) for $e^* = \sigma_e^2\, \mathrm{I} + \sigma_p^2\, \mathbf{ZZ}^T$
  - Or use $\mathbf{y} = \mathbf{X}\beta + \mathbf{Za} + \mathbf{Zp} + \mathbf{e}$, where $e = \sigma_e^2\, \mathrm{I}$

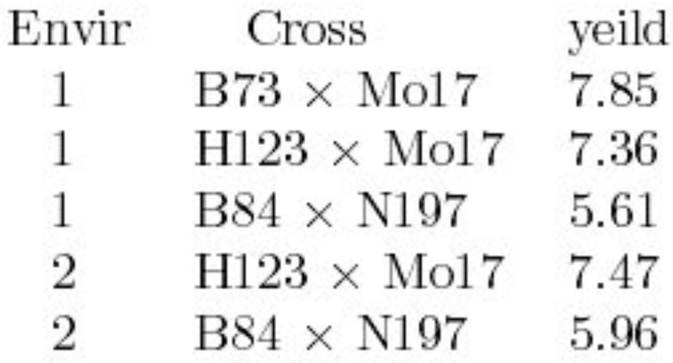
Generalizing BLUP

- Thus far, we have framed BLUP in the standard animal breeding context which estimates a vector of breeding values from the genetic relationship matrix
- More generally, we can estimate any number of vectors **g** of genetic parameters (such as CGA, SCA, line values) given some matrix of genetic relatedness
- Historically the relatedness matrix is obtained from a pedigree, but now with dense markers it can be estimated directly

BLUP for GCA, SCA

Again, Example from Bernardo (11.5.1)

B73, B84, and H123 are in one maize heteroic group (Stiff Stalk), Mo17 and N197 in another (Lancaster)



Envir	Cross	yeild
1	B73 × Mo17	7.85
1	H123 × Mo17	7.36
1	B84 × N197	5.61
2	H123 × Mo17	7.47
2	B84 × N197	5.96

Note highly unbalanced design --- not all crosses in both environments

$$y_{ijk} = \mu + t_k + GCA_i + GCA_j + SCA_{i,j} + e_{ijk}$$

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{s} + \mathbf{e}, \quad \mathrm{Var}\begin{pmatrix}\mathbf{g}\\ \mathbf{s}\\ \mathbf{e}\end{pmatrix} \sim MVN\left(\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix}, \begin{pmatrix}\mathbf{G} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{S} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{pmatrix}\right)$$

Envir	Cross	yeild
1	B73 × Mo17	7.85
1	H123 × Mo17	7.36
1	B84 × N197	5.61
2	H123 × Mo17	7.47
2	B84 × N197	5.96

$$\begin{pmatrix} 7.85 \\ 7.36 \\ 5.61 \\ 7.47 \\ 5.96 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} GCA_{B73} \\ GCA_{B84} \\ GCA_{H123} \\ GCA_{Mol7} \\ GCA_{N197} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} SCA_{B73XMol17} \\ SCA_{H123XMol17} \\ SCA_{B84XN197} \end{pmatrix} + \mathbf{e}$$

$$\mathrm{Var}(\mathbf{g}) = \mathbf{G} = \sigma^2_{GCA} \begin{pmatrix} 1 & 0.27 & 0.75 & 0 & 0 \\ 0.27 & 1 & 0.20 & 0 & 0 \\ 0.75 & 0.2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.75 \\ 0 & 0 & 0 & 0.75 & 1 \end{pmatrix}, \qquad \mathrm{Var}(\mathbf{s}) = \mathbf{S} = \sigma^2_{SCA} \begin{pmatrix} 1 & 0.75 & 0.2 \\ 0.75 & 1 & 0.15 \\ 0.20 & 0.15 & 1 \end{pmatrix}$$

Covariance matrix based on pedigree information (see Bernardo for details)