SYLLABUS
MIXED MODELS IN QUANTITATIVE GENETICS

INSTRUCTORS:

William (Bill) Muir, Department of Animal Sciences, Purdue University
bmuir@purdue.edu

Bruce Walsh, Department of Ecology & Evolutionary Biology, University of Arizona
jbwalsh@u.arizona.edu

LW = Lynch & Walsh: Genetics and Analysis of Quantitative Traits (book)

WL = Walsh & Lynch: Evolution and Selection of Quantitative Traits (website)
http://nitro.biosci.arizona.edu/zbook/NewVolume_2/newvol2.html

LECTURE SCHEDULE

Wednesday, 20 July

1:30 3:00 pm 1. Introduction to matrix algebra and calculus (Walsh)
Background reading: LW, Chapter 8
Additional reading: LW Appendix 3; WL Appendix 5

3:00 3:30 pm Break
3:30 5:00 pm 2. The General Linear Model (Walsh)
Background reading: LW Chapter 8
Additional reading: LW Appendices 3, 4; WL Appendices 2, 3

Thursday, 21 July

8:30 10:00 am 3. Overview and Derivation of the mixed model (Muir)
Additional reading: LW Chapters 26, 27
10:00 10:30 am Break
10:30 12:00 4. Application: BLUP breeding values (Muir)
Additional reading WL Chapter 19
12:00 1:30 pm Lunch
1:30 3:00 pm 5. Application: Genomic selection (Muir)
6. Application: Correlated residuals (Muir)
3:00 3:30 pm Break
3:30 5:00 pm 7. Application: Models with multiple random effects (Walsh)
Evening Open session (review, R, etc)

Friday, 22 July

8:30	10:00 am	8. Application: Indirect Genetics (Associative) effects (Muir)
		Additional reading: WL Chapter 20
10:00	10:30 am	Break
10:30	12:00	9. Application: QTL/association mapping (Walsh)
		Additional reading: LW Chapters 14, 16
12:00	1:30 pm	Lunch
1:30	3:00 pm	10. Application: G x E (Walsh)
		Additional reading: WL Chapters 43, 44
3:00	3:30 pm	Break
3:30	5:00 pm	11. Random Regressions (Walsh)

Lecture 1: Intro/refresher in Matrix Algebra

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

Matrix/linear algebra

- Compact way for treating the algebra of systems of linear equations
- Most common statistical methods can be written in matrix form
  - $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$ is the general linear model
    - OLS solution: $\beta = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
  - $\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}\mathbf{a} + \mathbf{e}$ is the general mixed model

Topics

- Definitions, dimensionality, addition, subtraction
- Matrix multiplication
- Inverses, solving systems of equations
- Quadratic products and covariances
- The multivariate normal distribution
- Eigenstructure
- Basic matrix calculations in R

Matrices: An array of elements

Vectors: A matrix with either one row or one column.

Usually written in bold lowercase, e.g. **a, b, c**

$$\mathbf{a} = \begin{pmatrix} 12 \\ 13 \\ 47 \end{pmatrix} \quad \mathbf{b} = \begin{pmatrix} 2 & 0 & 5 & 21 \end{pmatrix}$$

Column vector Row vector

(3 x 1) (1 x 4)

Dimensionality of a matrix: r x c (rows x columns)
think of Railroad Car

General Matrices

Usually written in bold uppercase, e.g. **A, C, D**

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \quad \mathbf{D} = \begin{pmatrix} 0 & 1 \\ 3 & 4 \\ 2 & 9 \end{pmatrix}$$

(3 x 3)

Square matrix (3 x 2)

Dimensionality of a matrix: r x c (rows x columns)
think of Railroad Car

A matrix is defined by a list of its elements.
B has ij-th element B_{ij} -- the element in row i and column j

Addition and Subtraction of Matrices

If two matrices have the same dimension (both are r x c), then matrix addition and subtraction simply follows by adding (or subtracting) on an element by element basis

Matrix addition: $(A+B)_{ij} = A_{ij} + B_{ij}$

Matrix subtraction: $(A-B)_{ij} = A_{ij} - B_{ij}$

Examples:

$$\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} 4 & 2 \\ 3 & 3 \end{pmatrix} \text{ and } \mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \end{pmatrix}$$

Partitioned Matrices

It will often prove useful to divide (or partition) the elements of a matrix into a matrix whose elements are itself matrices.

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \left(\begin{array}{cccc} 3 & \vdots & 1 & 2 \\ \cdots & \cdots & \cdots & \cdots \\ 2 & \vdots & 5 & 4 \\ 1 & \vdots & 1 & 2 \end{array}\right) = \begin{pmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{B} \end{pmatrix}$$

$$\mathbf{a} = (3), \quad \mathbf{b} = (1 \quad 2), \quad \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 5 & 4 \\ 1 & 2 \end{pmatrix}$$

One useful partition is to write the matrix as either a row vector of column vectors or a column vector of row vectors

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} \mathbf{r}_1 \\ \mathbf{r}_2 \\ \mathbf{r}_3 \end{pmatrix}$$

A column vector whose elements are row vectors

$$\begin{aligned} \mathbf{r}_1 &= (3 \quad 1 \quad 2) \\ \mathbf{r}_2 &= (2 \quad 5 \quad 4) \\ \mathbf{r}_3 &= (1 \quad 1 \quad 2) \end{aligned}$$

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = (\mathbf{c}_1 \quad \mathbf{c}_2 \quad \mathbf{c}_3)$$

A row vector whose elements are column vectors

$$\mathbf{c}_1 = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c}_2 = \begin{pmatrix} 1 \\ 5 \\ 1 \end{pmatrix}, \quad \mathbf{c}_3 = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix}$$

Towards Matrix Multiplication: dot products

The dot (or inner) product of two vectors (both of length n) is defined as follows:

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

Example:

$$\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 4 & 5 & 7 & 9 \end{pmatrix}$$

a ˙b = 1*4 + 2*5 + 3*7 + 4*9 = 60

Matrices are compact ways to write systems of equations

$$\begin{aligned} 5x_1 + 6x_2 + 4x_3 &= 6 \\ 7x_1 - 3x_2 + 5x_3 &= -9 \\ -x_1 - x_2 + 6x_3 &= 12 \end{aligned}$$

$$\begin{pmatrix} 5 & 6 & 4 \\ 7 & -3 & 5 \\ -1 & -1 & 6 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 6 \\ -9 \\ 12 \end{pmatrix}$$

$$\mathbf{Ax} = \mathbf{c}, \qquad \text{or} \quad \mathbf{x} = \mathbf{A}^{-1}\mathbf{c}$$

The least-squares solution for the linear model

$$y = \mu + \beta_1 z_1 + \cdots \beta_n z_n$$

yields the following system of equations for the β_i

$$\begin{aligned}
\sigma(y,z_1) &= \beta_1\sigma^2(z_1) + \beta_2\sigma(z_1,z_2) + \cdots + \beta_n\sigma(z_1,z_n)\\
\sigma(y,z_2) &= \beta_1\sigma(z_1,z_2) + \beta_2\sigma^2(z_2) + \cdots + \beta_n\sigma(z_2,z_n)\\
\vdots \quad & \quad\ \vdots \qquad\qquad\quad \vdots \qquad\qquad \ddots \qquad \vdots\\
\sigma(y,z_n) &= \beta_1\sigma(z_1,z_n) + \beta_2\sigma(z_2,z_n) + \cdots + \beta_n\sigma^2(z_n)
\end{aligned}$$

This can be more compactly written in matrix form as

$$\underbrace{\begin{pmatrix} \sigma^2(z_1) & \sigma(z_1,z_2) & \dots & \sigma(z_1,z_n)\\ \sigma(z_1,z_2) & \sigma^2(z_2) & \dots & \sigma(z_2,z_n)\\ \vdots & \vdots & \ddots & \vdots\\ \sigma(z_1,z_n) & \sigma(z_2,z_n) & \dots & \sigma^2(z_n) \end{pmatrix}}_{X^TX} \underbrace{\begin{pmatrix}\beta_1\\ \beta_2\\ \vdots\\ \beta_n\end{pmatrix}}_{\beta} = \underbrace{\begin{pmatrix}\sigma(y,z_1)\\ \sigma(y,z_2)\\ \vdots\\ \sigma(y,z_n)\end{pmatrix}}_{X^Ty}$$

T = transpose, discussed in slide 17-9

or, $\beta = (X^TX)^{-1} X^Ty$

Matrix Multiplication:

The order in which matrices are multiplied affects the matrix product, e.g. $AB \neq BA$

For the product of two matrices to exist, the matrices must conform. For AB, the number of columns of A must equal the number of rows of B.

The matrix C = AB has the same number of rows as A and the same number of columns as B.

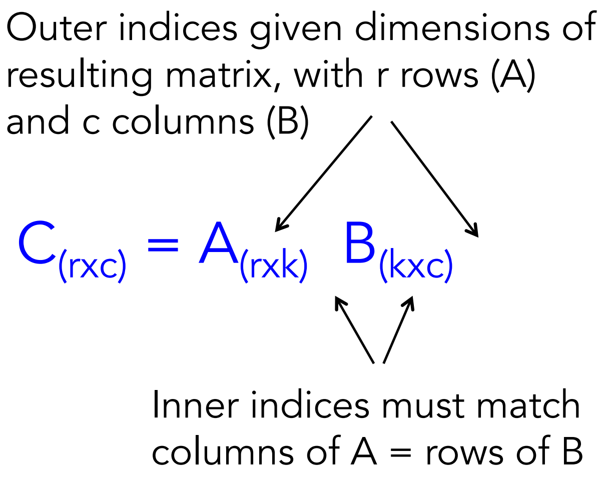
$$C_{(r \times c)} = A_{(r \times k)}\ B_{(k \times c)}$$

ij-th element of C is given by

$$C_{ij} = \sum_{l=1}^{k} A_{il} B_{lj}$$

Elements in the jth column of B

Elements in the ith row of matrix A



Outer indices given dimensions of resulting matrix, with r rows (A) and c columns (B)

$$C_{(rxc)} = A_{(rxk)} \ B_{(kxc)}$$

Inner indices must match columns of A = rows of B

Example: Is the product ABCD defined? If so, what is its dimensionality? Suppose

$$A_{3x5}\, B_{5x9}\, C_{9x6}\, D_{6x23}$$

Yes, defined, as inner indices match. Result is a 3 x 23 matrix (3 rows, 23 columns)

More formally, consider the product L = MN

Express the matrix M as a column vector of row vectors

$$\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_r} \end{pmatrix} \qquad \text{where} \qquad \mathbf{m_i} = (\, M_{i1} \quad M_{i2} \quad \cdots \quad M_{ic}\,)$$

Likewise express N as a row vector of column vectors

$$\mathbf{N} = (\, \mathbf{n_1} \quad \mathbf{n_2} \quad \cdots \quad \mathbf{n_b}\,) \qquad \text{where} \qquad \mathbf{n_j} = \begin{pmatrix} N_{1j} \\ N_{2j} \\ \vdots \\ N_{cj} \end{pmatrix}$$

The ij-th element of L is the inner product of M's row i with N's column j

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1 \cdot n_1} & \mathbf{m_1 \cdot n_2} & \cdots & \mathbf{m_1 \cdot n_b} \\ \mathbf{m_2 \cdot n_1} & \mathbf{m_2 \cdot n_2} & \cdots & \mathbf{m_2 \cdot n_b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_r \cdot n_1} & \mathbf{m_r \cdot n_2} & \cdots & \mathbf{m_r \cdot n_b} \end{pmatrix}$$

Example

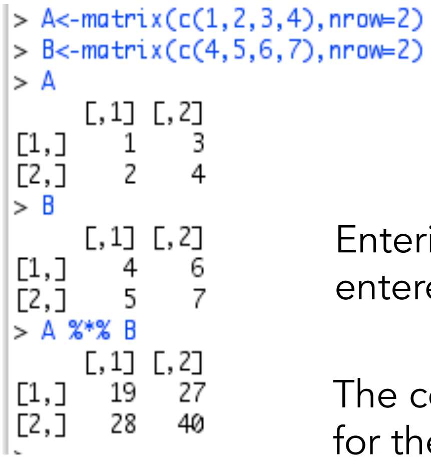
$$\mathbf{AB} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} e & f \\ g & h \end{pmatrix} = \begin{pmatrix} ae+bg & af+bh \\ ce+dg & cf+dh \end{pmatrix}$$

Likewise

$$\mathbf{BA} = \begin{pmatrix} ae+cf & eb+df \\ ga+ch & gd+dh \end{pmatrix}$$

ORDER of multiplication matters! Indeed, consider C_{3x5} D_{5x5} which gives a 3 x 5 matrix, versus D_{5x5} C_{3x5} , which is not defined.

Matrix multiplication in R



```
> A<-matrix(c(1,2,3,4),nrow=2)
> B<-matrix(c(4,5,6,7),nrow=2)
> A
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> B
     [,1] [,2]
[1,]    4    6
[2,]    5    7
> A %*% B
     [,1] [,2]
[1,]   19   27
[2,]   28   40
```

R fills in the matrix from the list c by filling in as columns, here with 2 rows (nrow=2)

Entering A or B displays what was entered (always a good thing to check)

The command %*% is the R code for the multiplication of two matrices

On your own: What is the matrix resulting from BA? What is A if nrow=1 or nrow=4 is used?

The Transpose of a Matrix

The transpose of a matrix exchanges the rows and columns, $A^T_{ij} = A_{ji}$

Useful identities

$(AB)^T = B^T A^T$

$(ABC)^T = C^T B^T A^T$

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Inner product = $a^T b = a^T_{(1\ X\ n)}\, b_{(n\ X\ 1)}$

Indices match, matrices conform

Dimension of resulting product is 1 X 1 (i.e. a scalar)

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

Note that $b^T a = (b^T a)^T = a^T b$

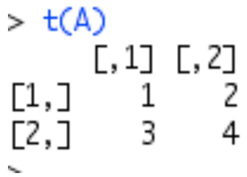
Outer product = $ab^T = a_{(n\ X\ 1)}\, b^T_{(1\ X\ n)}$

Resulting product is an n x n matrix

$$\begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & b_2 & \cdots & b_n \end{pmatrix}$$

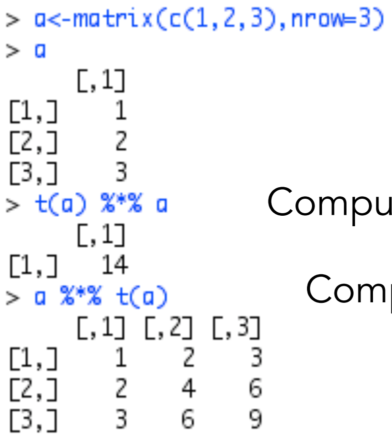
$$= \begin{pmatrix} a_1 b_1 & a_1 b_2 & \dots & a_1 b_n \\ a_2 b_1 & a_2 b_2 & \dots & a_2 b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_n b_1 & a_n b_2 & \dots & a_n b_{bn} \end{pmatrix}$$

R code for transposition



```
> t(A)
     [,1] [,2]
[1,]    1    2
[2,]    3    4
```

t(A) = transpose of A



```
> a<-matrix(c(1,2,3),nrow=3)
> a
     [,1]
[1,]    1
[2,]    2
[3,]    3
> t(a) %*% a
     [,1]
[1,]   14
> a %*% t(a)
     [,1] [,2] [,3]
[1,]    1    2    3
[2,]    2    4    6
[3,]    3    6    9
```

Enter the column vector a

Compute inner product $\mathbf{a}^T\mathbf{a}$

Compute outer product $\mathbf{a}\mathbf{a}^T$

Solving equations

- The identity matrix **I**
  - Serves the same role as 1 in scalar algebra, e.g., a*1=1*a =a, with **AI=IA= A**
- The inverse matrix $\mathbf{A}^{-1}$ (IF it exists)
  - Defined by $\mathbf{A}\,\mathbf{A}^{-1} = \mathbf{I}$, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$
  - Serves the same role as scalar division
    - To solve ax = c, multiply both sides by (1/a) to give:
    - (1/a)*ax = (1/a)c or (1/a)*a*x = 1*x = x,
    - Hence x = (1/a)c
    - To solve $\mathbf{Ax} = \mathbf{c}$, $A^{-1}\mathbf{Ax} = A^{-1}\,\mathbf{c}$
    - Or $\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{Ix} = \mathbf{x} = A^{-1}\,\mathbf{c}$

The Identity Matrix, I

The identity matrix serves the role of the number 1 in matrix multiplication: AI =A, IA = A

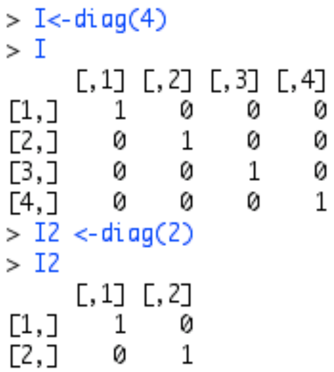
I is a square diagonal matrix, with all diagonal elements being one, all off-diagonal elements zero.

$$I_{ij} = \begin{cases} 1 \text{ for } i = j \\ 0 \text{ otherwise} \end{cases}$$

$$\mathbf{I}_{3x3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

The Identity Matrix in R

diag(k), where k is an integer, return the k x k I matix



```
> I<-diag(4)
> I
     [,1] [,2] [,3] [,4]
[1,]    1    0    0    0
[2,]    0    1    0    0
[3,]    0    0    1    0
[4,]    0    0    0    1
> I2 <-diag(2)
> I2
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

The Inverse Matrix, A^{-1}

For a square matrix A, define its Inverse A^{-1}, as the matrix satisfying

$$\mathbf{A^{-1}A = AA^{-1} = I}$$

$$\text{For } \mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \quad \mathbf{A}^{-1} = \frac{1}{\boxed{ad - bc}} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

If this quantity (the determinant) is zero, the inverse does not exist.

If det(A) is not zero, A^{-1} exists and A is said to be non-singular. If det(A) = 0, A is singular, and no *unique* inverse exists (generalized inverses do)

Generalized inverses, and their uses in solving systems of equations, are discussed in Appendix 3 of Lynch & Walsh

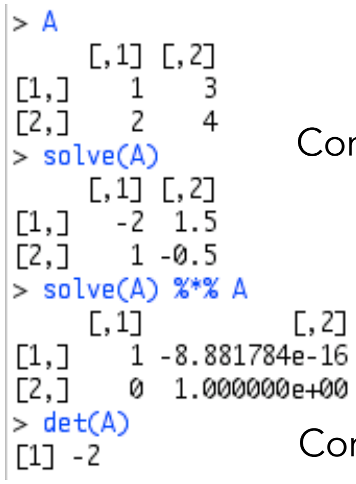
A^- is the typical notation to denote the G-inverse of a matrix

When a G-inverse is used, *provided* the system is consistent, then some of the variables have a family of solutions (e.g., $x_1 = 2$, but $x_2 + x_3 = 6$)

Inversion in R

solve(A) computes A^{-1}

det(A) computes determinant of A



```
> A
     [,1] [,2]
[1,]    1    3
[2,]    2    4
> solve(A)
     [,1] [,2]
[1,]   -2  1.5
[2,]    1 -0.5
> solve(A) %*% A
     [,1]          [,2]
[1,]    1 -8.881784e-16
[2,]    0  1.000000e+00
> det(A)
[1] -2
```

Using A entered earlier

Compute A^{-1}

Showing that $A^{-1} A = I$

Computing determinant of A

Homework

Put the following system of equations in matrix form, and solve using R

$$
\begin{aligned}
3x_1 + 4x_2 + 4\,x_3 + 6x_4 &= -10 \\
9x_1 + 2x_2 - x_3 - 6x_4 &= 20 \\
x_1 + x_2 + x_3 - 10x_4 &= 2 \\
2x_1 + 9x_2 + 2x_3 - x_4 &= -10
\end{aligned}
$$

Example: solve the OLS for $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$
in $y = \alpha + \beta_1 z_1 + \beta_2 z_2 + e$

$$\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c} \qquad \mathbf{c} = \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) \\ \sigma(z_1, z_2) & \sigma^2(z_2) \end{pmatrix}$$

It is more compact to use $\sigma(z_1, z_2) = \rho_{12}\,\sigma(z_1)\sigma(z_2)$

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix}$$

$$\begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1-\rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix} \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \end{pmatrix}$$

$$\beta_1 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y, z_1)}{\sigma^2(z_1)} - \rho_{12}\,\frac{\sigma(y, z_2)}{\sigma(z_1)\sigma(z_2)}\right]$$

$$\beta_2 = \frac{1}{1-\rho_{12}^2}\left[\frac{\sigma(y, z_2)}{\sigma^2(z_2)} - \rho_{12}\,\frac{\sigma(y, z_1)}{\sigma(z_1)\sigma(z_2)}\right]$$

If $\rho_{12} = 0$, these reduce to the two univariate slopes,

$$\beta_1 = \frac{\sigma(y, z_1)}{\sigma^2(z_1)} \quad \text{and} \quad \beta_2 = \frac{\sigma(y, z_2)}{\sigma^2(z_2)}$$

Likewise, if $\rho_{12} = 1$, this reduces to a univariate regression,

Useful identities

$$(A^T)^{-1} = (A^{-1})^T$$

$$(AB)^{-1} = B^{-1} A^{-1}$$

For a diagonal matrix **D**, then det (**D**), which is also denoted by |**D**|, = product of the diagonal elements

Also, the determinant of any square matrix **A**, det(**A**), is simply the product of the eigenvalues λ of **A,** which statisfy

$$\mathbf{Ae} = \lambda\mathbf{e}$$

If A is n x n, solutions to λ are an n-degree polynomial. **e** is the eigenvector associated with λ. If any of the roots to the equation are zero, A^{-1} is not defined. In this case, for some linear combination **b**, we have **Ab** = 0.

Variance-Covariance matrix

- A very important square matrix is the variance-covariance matrix **V** associated with a vector **x** of random variables.
- V_{ij} = Cov(x_i,x_j), so that the i-th diagonal element of **V** is the variance of x_i, and off -diagonal elements are covariances
- **V** is a symmetric, square matrix

The trace

The trace, tr(**A**) or trace(**A**), of a square matrix A is simply the sum of its diagonal elements

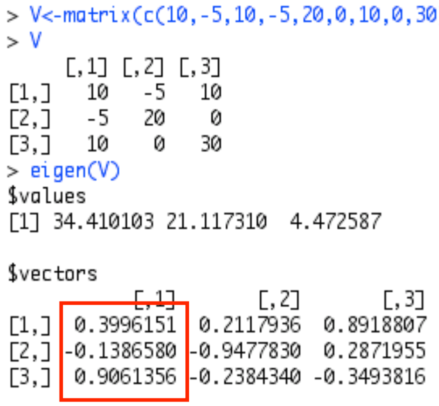
The importance of the trace is that it equals the sum of the eigenvalues of **A**, tr(**A**) = $\Sigma\ \lambda_i$

For a covariance matrix **V**, tr(V) measures the total amount of variation in the variables

λ_i / tr(**V**) is the fraction of the total variation in x contained in the linear combination $\mathbf{e}_i^T\mathbf{x}$, where $\mathbf{e}_i$, the i-th principal component of **V** is also the i-th eigenvector of **V** ($\mathbf{V}\mathbf{e}_i = \lambda_i\ \mathbf{e}_i$)

Eigenstructure in R

`eigen(A)` returns the eigenvalues and vectors of A



```
> V<-matrix(c(10,-5,10,-5,20,0,10,0,30),nrow=3)
> V
     [,1] [,2] [,3]
[1,]   10   -5   10
[2,]   -5   20    0
[3,]   10    0   30
> eigen(V)
$values
[1] 34.410103 21.117310  4.472587

$vectors
           [,1]       [,2]       [,3]
[1,]  0.3996151  0.2117936  0.8918807
[2,] -0.1386580 -0.9477830  0.2871955
[3,]  0.9061356 -0.2384340 -0.3493816
```

PC 1

Trace = 60

PC 1 accounts for 34.4/60 = 57% of all the variation

Quadratic and Bilinear Forms

Quadratic product: for $A_{n \times n}$ and $x_{n \times 1}$

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} x_i x_j$$ Scalar (1 x 1)

Bilinear Form (generalization of quadratic product)
for $A_{m \times n}$, $a_{n \times 1}$, $b_{m \times 1}$ their bilinear form is $b^T_{1 \times m} A_{m \times n} a_{n \times 1}$

$$\mathbf{b}^T \mathbf{A} \mathbf{a} = \sum_{i=1}^{m} \sum_{j=1}^{n} A_{ij} b_i a_j$$

Note that $b^T A\, a = a^T A^T b$

Covariance Matrices for Transformed Variables

What is the variance of the linear combination,
$c_1 x_1 + c_2 x_2 + \ldots + c_n x_n$? (note this is a scalar)

$$\sigma^2 \left(\mathbf{c}^T \mathbf{x}\right) = \sigma^2 \left(\sum_{i=1}^{n} c_i x_i \right) = \sigma \left(\sum_{i=1}^{n} c_i\, x_i \,, \sum_{j=1}^{n} c_j\, x_j \right)$$
$$= \sum_{i=1}^{n} \sum_{j=1}^{n} \sigma\left(c_i\, x_i, c_j\, x_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i\, c_j\, \sigma\left(x_i, x_j\right)$$
$$= \mathbf{c}^T \mathbf{V}\, \mathbf{c}$$

Likewise, the covariance between two linear combinations can be expressed as a bilinear form,

$$\sigma(\mathbf{a}^T \mathbf{x}, \mathbf{b}^T \mathbf{x}) = \mathbf{a}^T \mathbf{V}\, \mathbf{b}$$

Example: Suppose the variances of x_1, x_2, and x_3 are 10, 20, and 30. x_1 and x_2 have a covariance of -5, x_1 and x_3 of 10, while x_2 and x_3 are uncorrelated.

What are the variances of the new variables $y_1 = x_1-2x_2+5x_3$ and $y_2 = 6x_2-4x_3$?

$$\mathbf{V} = \begin{pmatrix} 10 & -5 & 10 \\ -5 & 20 & 0 \\ 10 & 0 & 30 \end{pmatrix}, \quad \mathbf{c}_1 = \begin{pmatrix} 1 \\ -2 \\ 5 \end{pmatrix}, \quad \mathbf{c}_2 = \begin{pmatrix} 0 \\ 6 \\ -4 \end{pmatrix}$$

$\text{Var}(y_1) = \text{Var}(c_1^T x) = c_1^T \text{Var}(x)\, c_1 = 960$

$\text{Var}(y_2) = \text{Var}(c_2^T x) = c_2^T \text{Var}(x)\, c_2 = 1200$

$\text{Cov}(y_1,y_2) = \text{Cov}(c_1^T x, c_2^T x) = c_1^T \text{Var}(x)\, c_2 = -910$

Homework: use R to compute the above values

Now suppose we transform one vector of random variables into another vector of random variables

Transform x into

(i) $y_{k \times 1} = A_{k \times n} x_{n \times 1}$

(ii) $z_{m \times 1} = B_{m \times n} x_{n \times 1}$

The covariance between the elements of these two transformed vectors is an k x m covariance matrix = AVB^T

For example, the covariance between y_i and y_j is given by the ij-th element of AVA^T

Likewise, the covariance between y_i and z_j is given by the ij-th element of AVB^T

Positive-definite matrix

- A matrix **V** is positive-definite if for all vectors **c** containing at least one non-zero member, $\mathbf{c}^T\mathbf{V}\mathbf{c} > 0$.
- A non-negative definite matrix satisfies $\mathbf{c}^T\mathbf{V}\mathbf{c} \geq 0$.
- Any covariance-matrix is (at least) non-negative definite, as $\mathrm{Var}(\mathbf{c}^T\mathbf{x}) = \mathbf{c}^T\mathbf{V}\mathbf{c} \geq 0$.
- Any nonsingular covariance matrix is positive-definite
  - Nonsingular means $\det(V) > 0$
  - Equivalently, all eigenvalues of V are positive, $\lambda_i > 0$.

The Multivariate Normal Distribution (MVN)

Consider the pdf for n independent normal random variables, the ith of which has mean μ_i and variance σ^2_i

$$p(\mathbf{x}) = \prod_{i=1}^{n} (2\pi)^{-1/2} \sigma_i^{-1} \exp\left(- \frac{(x_i - \mu_i)^2}{2\,\sigma_i^2} \right)$$

$$= (2\pi)^{-n/2} \left(\prod_{i=1}^{n} \sigma_i \right)^{-1} \exp\left(- \sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\,\sigma_i^2} \right)$$

This can be expressed more compactly in matrix form

Define the covariance matrix V for the vector x of the n normal random variable by

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n^2 \end{pmatrix} \qquad |\mathbf{V}| = \prod_{i=1}^{n} \sigma_i^2$$

Define the mean vector μ as $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$ gives

$$\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Hence in matrix from the MVN pdf becomes

$$p(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right]$$

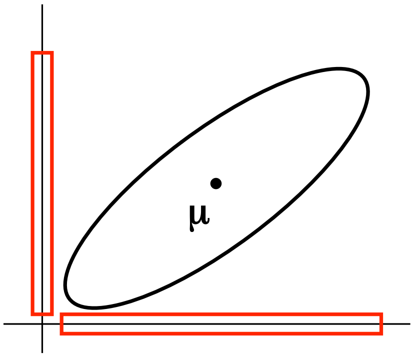
Notice this holds for any vector μ and symmetric positive -definite matrix V, as | V | > 0.

The multivariate normal

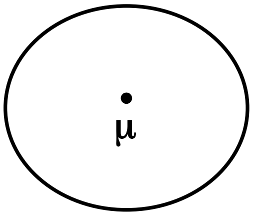
- Just as a univariate normal is defined by its mean and spread (variance), a multivariate normal is defined by its mean vector **μ** (also called the centroid) and variance-covariance matrix **V** (the distribution, or spread, of values around the centroid).

Vector of means $\boldsymbol{\mu}$ determines location

Spread (geometry) about $\boldsymbol{\mu}$ determined by **V**



x_1, x_2 equal variances, positively correlated

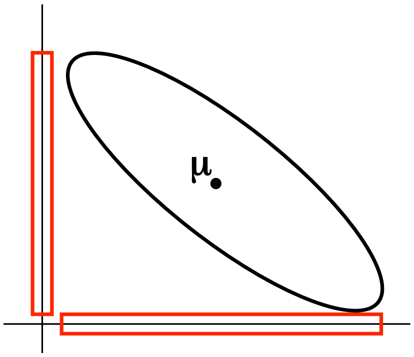


x_1, x_2 equal variances, uncorrelated

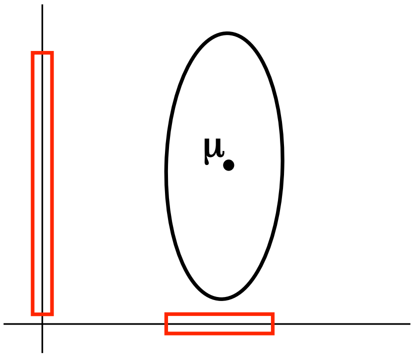
Eigenstructure (the eigenvectors and their corresponding eigenvalues) determines the geometry of **V**.

Vector of means $\boldsymbol{\mu}$ determines location

Spread (geometry) about $\boldsymbol{\mu}$ determined by **V**



x_1, x_2 equal variances, negatively correlated

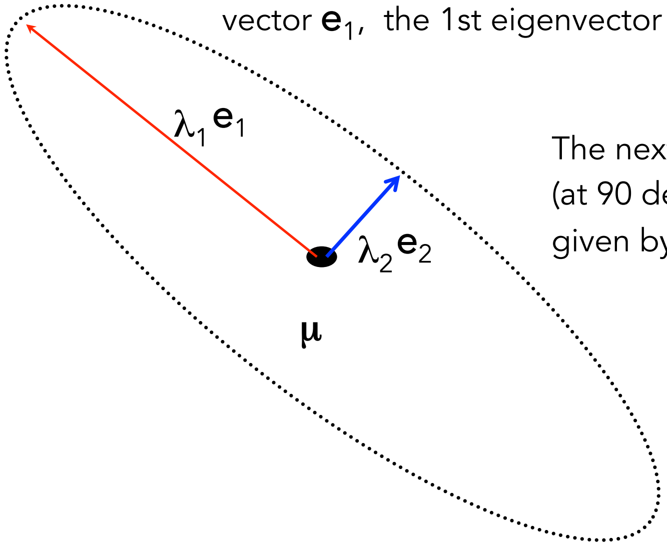


Var(x_1) < Var(x_2), uncorrelated

Positive tilt = positive correlations
Negative tilt = negative correlation
No tilt = uncorrelated

Eigenstructure of V

The direction of the largest axis of variation is given by the unit-length vector $\mathbf{e}_1$, the 1st eigenvector of V.



The next largest axis orthogonal (at 90 degrees from) to $\mathbf{e}_1$, is given by $\mathbf{e}_2$, the 2nd eigenvector

Principal components

- The principal components (or PCs) of a covariance matrix define the axes of variation.
  - PC1 is the direction (linear combination $\mathbf{c}^T\mathbf{x}$) that explains the most variation.
  - PC2 is the next largest direction (at 90degree from PC1), and so on
- PCi = ith eigenvector of V
- Fraction of variation accounted for by PCi = λ_i / trace(V)
- If V has a few large eigenvalues, most of the variation is distributed along a few linear combinations (axis of variation)

Properties of the MVN - I

1) If **x** is MVN, any subset of the variables in **x** is also MVN

2) If **x** is MVN, any linear combination of the elements of **x** is also MVN. If **x** ~ MVN(μ,V)

$$\text{for} \quad \mathbf{y} = \mathbf{x} + \mathbf{a}, \qquad \mathbf{y} \text{ is } \mathrm{MVN}_n(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V})$$

$$\text{for} \quad y = \mathbf{a}^T \mathbf{x} = \sum_{k=1}^{n} a_i x_i, \qquad y \text{ is } \mathrm{N}(\mathbf{a}^T \boldsymbol{\mu}, \mathbf{a}^T \mathbf{V}\, \mathbf{a})$$

$$\text{for} \quad \mathbf{y} = \mathbf{A}\mathbf{x}, \qquad \mathbf{y} \text{ is } \mathrm{MVN}_m \left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}^T \mathbf{V} \mathbf{A} \right)$$

Properties of the MVN - II

3) Conditional distributions are also MVN. Partition x into two components, x_1 (m dimensional column vector) and x_2 (n-m dimensional column vector)

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \qquad \text{and} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V_{x_1x_1}} & \mathbf{V_{x_1x_2}} \\ \mathbf{V^T_{x_1x_2}} & \mathbf{V_{x_2x_2}} \end{pmatrix}$$

x_1 I x_2 is MVN with m-dimensional mean vector

$$\boldsymbol{\mu}_{\mathbf{x_1|x_2}} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}} \mathbf{V^{-1}_{x_2x_2}} (\mathbf{x_2} - \boldsymbol{\mu_2})$$

and m x m covariance matrix

$$\mathbf{V_{x_1|x_2}} = \mathbf{V_{x_1x_1}} \text{ - } \mathbf{V_{x_1x_2}} \mathbf{V^{-1}_{x_2x_2}} \mathbf{V^T_{x_1x_2}}$$

Properties of the MVN - III

4) If x is MVN, the regression of any subset of x on another subset is linear and homoscedastic

$$\mathbf{x_1} = \boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} + \mathbf{e}$$

$$= \boldsymbol{\mu}_\mathbf{1} + \mathbf{V}_{\mathbf{x_1x_2}}\mathbf{V}_{\mathbf{x_2x_2}}^{-1}(\mathbf{x_2} - \boldsymbol{\mu}_\mathbf{2}) + \mathbf{e}$$

Where e is MVN with mean vector **0** and variance-covariance matrix $\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}}$

$$\boldsymbol{\mu}_\mathbf{1} + \mathbf{V}_{\mathbf{x_1x_2}}\mathbf{V}_{\mathbf{x_2x_2}}^{-1}(\mathbf{x_2} - \boldsymbol{\mu}_\mathbf{2}) + \mathbf{e}$$

The regression is linear because it is a linear function of x_2

The regression is homoscedastic because the variance-covariance matrix for e does not depend on the value of the x's

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V}_{\mathbf{x_1x_1}} - \mathbf{V}_{\mathbf{x_1x_2}}\mathbf{V}_{\mathbf{x_2x_2}}^{-1}\mathbf{V}_{\mathbf{x_1x_2}}^{T}$$

All these matrices are constant, and hence the same for any value of x

Example: Regression of Offspring value on Parental values

Assume the vector of offspring value and the values of both its parents is MVN. Then from the correlations among (outbred) relatives,

$$\begin{pmatrix} z_o \\ z_s \\ z_d \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_z^2 \begin{pmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{pmatrix} \right]$$

Let $\mathbf{x_1} = (\, z_o \,)$, $\mathbf{x_2} = \begin{pmatrix} z_s \\ z_d \end{pmatrix}$

$$\mathbf{V}_{\mathbf{x_1},\mathbf{x_1}} = \sigma_z^2, \quad \mathbf{V}_{\mathbf{x_1},\mathbf{x_2}} = \frac{h^2\sigma_z^2}{2}(\,1 \quad 1\,), \quad \mathbf{V}_{\mathbf{x_2},\mathbf{x_2}} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$= \boldsymbol{\mu}_{\mathbf{1}} + \mathbf{V}_{\mathbf{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu}_{\mathbf{2}}) + \mathbf{e}$$

Regression of Offspring value on Parental values (cont.)

$$= \boldsymbol{\mu}_{\mathbf{1}} + \mathbf{V}_{\mathbf{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} (\mathbf{x_2} - \boldsymbol{\mu}_{\mathbf{2}}) + \mathbf{e}$$

$$\mathbf{V}_{\mathbf{x_1},\mathbf{x_1}} = \sigma_z^2, \quad \mathbf{V}_{\mathbf{x_1},\mathbf{x_2}} = \frac{h^2\sigma_z^2}{2}(\,1 \quad 1\,), \quad \mathbf{V}_{\mathbf{x_2},\mathbf{x_2}} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

Hence,

$$\begin{aligned} z_o &= \mu_o + \frac{h^2\sigma_z^2}{2}\,(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_s - \mu_s \\ z_d - \mu_d \end{pmatrix} + e \\ &= \mu_o + \frac{h^2}{2}(z_s - \mu_s) + \frac{h^2}{2}(z_d - \mu_d) + e \end{aligned}$$

where e is normal with mean zero and variance

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V}_{\mathbf{x_1 x_1}} - \mathbf{V}_{\mathbf{x_1 x_2}} \mathbf{V}^{-1}_{\mathbf{x_2 x_2}} \mathbf{V}^{T}_{\mathbf{x_1 x_2}}$$

$$\begin{aligned} \sigma_e^2 &= \sigma_z^2 - \frac{h^2\sigma_z^2}{2}\,(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{h^2\sigma_z^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_z^2 \left(1 - \frac{h^4}{2}\right) \end{aligned}$$

Hence, the regression of offspring trait value given the trait values of its parents is

$$z_o = \mu_o + h^2/2(z_s - \mu_s) + h^2/2(z_d - \mu_d) + e$$

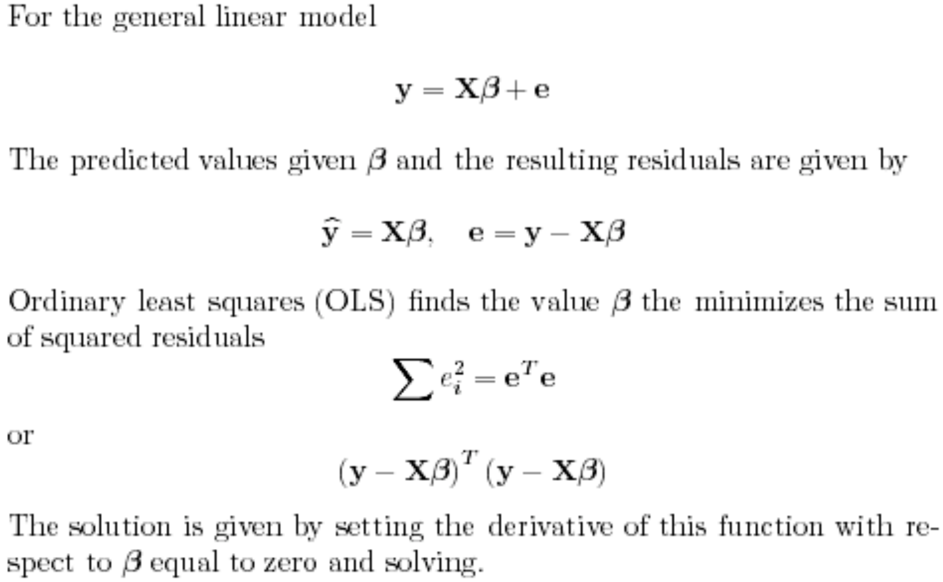
where the residual e is normal with mean zero and Var(e) = $\sigma_z^2(1-h^4/2)$

Similar logic gives the regression of offspring breeding value on parental breeding value as

$$A_o = \mu_o + (A_s - \mu_s)/2 + (A_d - \mu_d)/2 + e$$
$$= A_s/2 + A_d/2 + e$$

where the residual e is normal with mean zero and Var(e) = $\sigma_A^2/2$

Ordinary least squares



For the general linear model

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

The predicted values given $\boldsymbol{\beta}$ and the resulting residuals are given by

$$\widehat{\mathbf{y}} = \mathbf{X}\boldsymbol{\beta}, \quad \mathbf{e} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$$

Ordinary least squares (OLS) finds the value $\boldsymbol{\beta}$ the minimizes the sum of squared residuals

$$\sum e_i^2 = \mathbf{e}^T\mathbf{e}$$

or

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

The solution is given by setting the derivative of this function with respect to $\boldsymbol{\beta}$ equal to zero and solving.

Hence, we need to discuss vector/matrix derivatives

The gradient, the derivative of a vector-valued function

$$\nabla_{\mathbf{x}}[\,f\,] = \frac{\partial\, f}{\partial\, \mathbf{x}} = \begin{pmatrix} \frac{\partial\, f}{\partial\, x_1} \\ \frac{\partial\, f}{\partial\, x_2} \\ \vdots \\ \frac{\partial\, f}{\partial\, x_n} \end{pmatrix}$$

Compute the gradient for

$$f(\mathbf{x}) = \sum_{i=1}^{n} x_i^2 = \mathbf{x}^T\mathbf{x}$$

Since $\partial\, f / \partial\, x_i = 2x_i$, the gradient vector is just $\nabla_{\mathbf{x}}[\, f(\mathbf{x})\,] = 2\mathbf{x}$.

Some common derivatives

$$\nabla_{\mathbf{x}}\left[\,\mathbf{a}^T\mathbf{x}\right] = \nabla_{\mathbf{x}}\left[\,\mathbf{x}^T\mathbf{a}\,\right] = \mathbf{a}$$
$$\nabla_{\mathbf{x}}[\,\mathbf{A}\mathbf{x}] = \mathbf{A}^T$$

Turning to quadratic forms, if $\mathbf{A}$ is symmetric, then

$$\nabla_{\mathbf{x}}\left[\,\mathbf{x}^T\mathbf{A}\mathbf{x}\right] = 2\cdot\mathbf{A}\mathbf{x}$$
$$\nabla_{\mathbf{x}}\left[\,(\mathbf{x}-\mathbf{a})^T\mathbf{A}(\mathbf{x}-\mathbf{a})\,\right] = 2\cdot\mathbf{A}(\mathbf{x}-\ \mathrm{a})$$
$$\nabla_{\mathbf{x}}\left[\,(\mathbf{a}-\mathbf{x})^T\mathbf{A}(\mathbf{a}-\mathbf{x})\,\right] = -2\cdot\mathbf{A}(\mathbf{a}-\mathbf{x})$$

Taking $\mathbf{A} = \mathbf{I}$,

$$\nabla_{\mathbf{x}}\left[\,\mathbf{x}^T\mathbf{x}\right] = \nabla_{\mathbf{x}}\left[\,\mathbf{x}^T\,\mathbf{I}\,\mathbf{x}\right] = 2\cdot\mathbf{I}\,\mathbf{x} = 2\cdot\mathbf{x}$$

$$\begin{aligned}\sum_{i=1}^{n} e_i^2 &= \mathbf{e}^T\mathbf{e} = (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^T(\mathbf{y}-\mathbf{x}\boldsymbol{\beta}) \\ &= \mathbf{y}^T\mathbf{y} - \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^T\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}\end{aligned}$$

where the last step follows since the matrix product $\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{y}$ yields a scaler, and hence it equals its transpose,

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

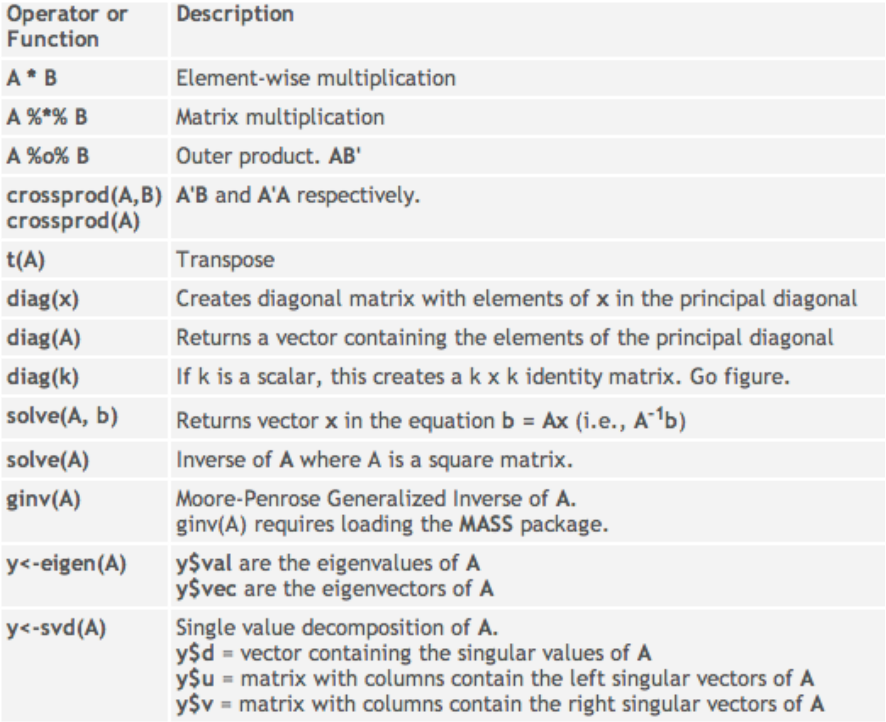
To find the vector $\boldsymbol{\beta}$ that minimizes $\mathbf{e}^T\mathbf{e}$, taking the derivative with respect to $\boldsymbol{\beta}$ and using Equations A5.1a/c gives

$$\frac{\partial\, \mathbf{e}^T\mathbf{e}}{\partial\, \boldsymbol{\beta}} = -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$$

Setting this equal to zero gives $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$ giving

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

Additional R matrix commands

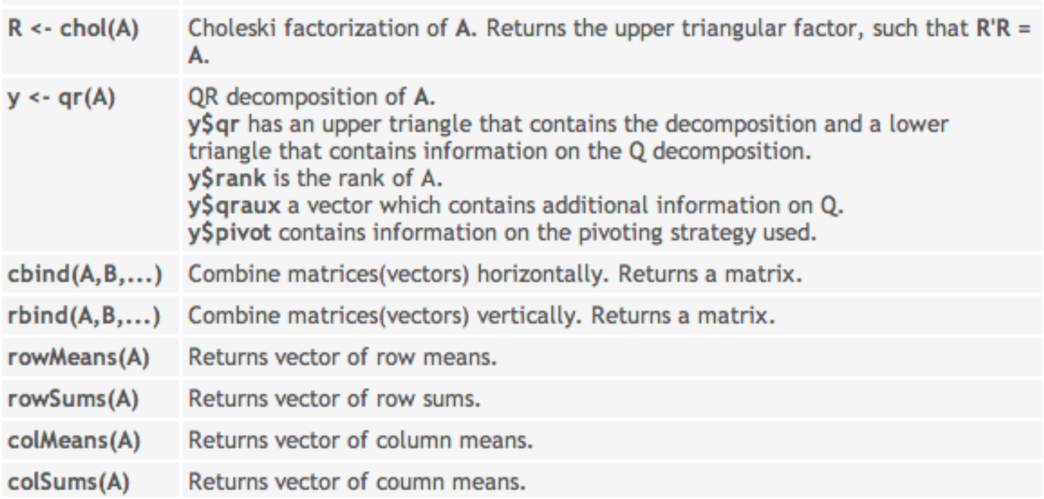


Operator or Function	Description
A * B	Element-wise multiplication
A %*% B	Matrix multiplication
A %o% B	Outer product. **AB'**

crossprod(A,B)
crossprod(A)
	A'B and **A'A** respectively.

t(A)	Transpose
diag(x)	Creates diagonal matrix with elements of **x** in the principal diagonal
diag(A)	Returns a vector containing the elements of the principal diagonal
diag(k)	If k is a scalar, this creates a k x k identity matrix. Go figure.
solve(A, b)	Returns vector **x** in the equation **b** = **Ax** (i.e., **A**$^{-1}$**b**)
solve(A)	Inverse of **A** where A is a square matrix.
ginv(A)	Moore-Penrose Generalized Inverse of **A**. ginv(A) requires loading the **MASS** package.
y<-eigen(A)	**y$val** are the eigenvalues of **A** **y$vec** are the eigenvectors of **A**
y<-svd(A)	Single value decomposition of **A**. **y$d** = vector containing the singular values of **A** **y$u** = matrix with columns contain the left singular vectors of **A** **y$v** = matrix with columns contain the right singular vectors of **A**

Additional R matrix commands (cont)



R <- chol(A)	Choleski factorization of **A.** Returns the upper triangular factor, such that **R'R** = **A.**
y <- qr(A)	QR decomposition of **A.** **y$qr** has an upper triangle that contains the decomposition and a lower triangle that contains information on the Q decomposition. **y$rank** is the rank of A. **y$qraux** a vector which contains additional information on Q. **y$pivot** contains information on the pivoting strategy used.
cbind(A,B,...)	Combine matrices(vectors) horizontally. Returns a matrix.
rbind(A,B,...)	Combine matrices(vectors) vertically. Returns a matrix.
rowMeans(A)	Returns vector of row means.
rowSums(A)	Returns vector of row sums.
colMeans(A)	Returns vector of column means.
colSums(A)	Returns vector of coumn means.

Additional references

- Lynch & Walsh Chapter 8 (intro to matrices)
- Online notes (Walsh & Lynch):
  - Appendix 4 (Matrix geometry)
  - Appendix 5 (Matrix derivatives)

Lecture 2:
Linear and Mixed Models

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

Quick Review of the Major Points

The general linear model can be written as

$$y = X\beta + e$$

- **y** = vector of observed response values
- X = Design matrix: observations of the explanatory variables in the assumed linear model
- β = vector of unknown parameters to estimate
- **e** = vector of residuals (deviation from model fit), **e** = y-X β

$y = X\beta + e$

Solution to $\boldsymbol{\beta}$ depends on the ***covariance structure***
(= covariance matrix) of the vector **e** of residuals

Ordinary least squares (OLS)

- OLS: $\mathbf{e} \sim \mathrm{MVN}(\mathbf{0}, \sigma^2\,\mathbf{I})$
- Residuals are homoscedastic and uncorrelated, so that we can write the cov matrix of **e** as $\mathrm{Cov}(e) = \sigma^2 I$
- the OLS estimate, $\mathrm{OLS}(\beta) = \mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{y}$

Generalized least squares (GLS)

- GLS: $e \sim \mathrm{MVN}(\mathbf{0},\ \mathbf{V})$
- Residuals are heteroscedastic and/or dependent,
- $\mathrm{GLS}(\beta) = (\mathbf{X}^T\,\mathbf{V}^{-1}\,\mathbf{X})^{-1}\,\mathbf{X}^T\,\mathbf{V}^{-1}\,\mathbf{y}$

BLUE

- Both the OLS and GLS solutions are also called the **Best Linear Unbiased Estimator** (or **BLUE** for short)
- Whether the OLS or GLS form is used depends on the assumed covariance structure for the residuals
  - Special case of $\mathrm{Var}(\mathbf{e}) = \sigma_e^2\,\mathbf{I}$ -- OLS
  - All others, i.e., $\mathrm{Var}(\mathbf{e}) = \mathbf{R}$ -- GLS

Linear Models

One tries to explain a response (or dependent) variable y as a linear function of a number of explanatory (or predictor) variables.

A multiple regression is a typical linear model,

$$y = \mu + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x + e$$

Here e is the residual, or deviation between the true value observed and the value predicted by the linear model.

The (partial) regression coefficients are interpreted as follows: a unit change in x_i while holding all other variables constant is associated with in a change of β_i in y

Linear Models

As with a univariate regression (y = a + bx + e), the model parameters are typically chosen by least squares, wherein they are chosen to minimize the sum of squared residuals, $\Sigma\ e_i^2$

This unweighted sum of squared residuals assumes an OLS error structure, so all residuals are equally weighted (homoscedastic) and uncorrelated

If the residuals differ in variances and/or some are correlated (GLS conditions), then we need to minimize the weighted sum $\mathbf{e}^T\mathbf{V}^{-1}\mathbf{e}$, which removes correlations and gives all residuals equal variance.

Predictor and Indicator Variables

Suppose we measure the offspring of p sires. One linear model would be

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

y_{ij} = trait value of offspring j from sire i

μ = overall mean. This term is included to give the s_i terms a mean value of zero, i.e., they are expressed as deviations from the mean

s_i = The effect for sire i (the mean of its offspring). Recall that variance in the s_i estimates Cov(half sibs) = $V_A/4$

e_{ij} = The deviation of the jth offspring from the family mean of sire i. The variance of the e's estimates the within-family variance.

Predictor and Indicator Variables

In a regression, the predictor variables are typically continuous, although they need not be.

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

Note that the predictor variables here are the s_i, (the value associated with sire i) something that we are trying to estimate

We can write this in linear model form, $y_{ij} = \mu + \Sigma_k \, x_{ik}s_i + e_{ij}$, by using indicator variables

$$x_{ik} = \begin{cases} 1 & \text{if sire } k = i \\ 0 & \text{otherwise} \end{cases}$$

Models consisting entirely of indicator variables are typically called ANOVA, or analysis of variance models

Models that contain no indicator variables (other than for the mean), but rather consist of observed values of continuous or discrete values are typically called regression models

Both are special cases of the General Linear Model (or GLM)

$$y_{ijk} = \mu + s_i + d_{ij} + \beta x_{ijk} + e_{ijk}$$

Example: Nested half sib/full sib design with an age correction β on the trait

Example: Nested half sib/full sib design with an age correction β on the trait

$$y_{ijk} = \mu + \overbrace{s_i + d_{ij}}^{\text{ANOVA model}} + \underbrace{\beta x_{ijk}}_{\text{Regression model}} + e_{ijk}$$

s_i = effect of sire i

d_{ij} = effect of dam j crossed to sire i

x_{ijk} = age of the kth offspring from i x j cross

Linear Models in Matrix Form

Suppose we have 3 variables in a multiple regression, with four (y,x) vectors of observations.

$$y_i = \mu + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + e_i$$

In matrix form, $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \mu \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \quad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} \\ 1 & x_{21} & x_{22} & x_{23} \\ 1 & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} \end{pmatrix} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \end{pmatrix}$$

The design (or incidence) matrix X. Details of both the experimental design and the observed values of the predictor variables all reside solely in X

In-class Exercise

Suppose you measure height and sprint speed for five individuals, with heights (x) of 9, 10, 11, 12, 13 and associated sprint speeds (y) of 60, 138, 131, 170, 221

1) Write in matrix form (i.e, the design matrix X and vector $\boldsymbol{\beta}$ of unknowns) the following models

- y = bx
- y = a + bx
- y = bx^2
- y = a + bx + cx^2

2) Using the X and y associated with these models, compute the OLS BLUE, $\mathbf{b} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ for each

Rank of the design matrix

- With n observations and p unknowns, X is an n x p matrix, so that $X^T X$ is p x p
- Thus, at most X can provide unique estimates for up to p < n parameters
- The rank of X is the number of independent rows of X. If X is of full rank, then rank = p
- A parameter is said to be **estimable** if we can provide a unique estimate of it. If the rank of X is k < p, then exactly k parameters are estimable (some as linear combinations, e.g. $\beta_1 - 3\beta_3 = 4$)
- if $\det(X^T X) = 0$, then X is not of full rank
- Number of nonzero eigenvalues of $X^T X$ gives the rank of X.

Experimental design and X

- The structure of **X** determines not only which parameters are estimable, but also the expected sample variances, as Var(**b**) = var(e)* $(\mathbf{X}^T\mathbf{X})^{-1}$
- Experimental design determines the structure of X before an experiment (of course, missing data almost always means the final **X** is different form the proposed **X**)
- Different criteria used for an optimal design. Let **V** = $(\mathbf{X}^T\mathbf{X})^{-1}$. The idea is to chose a design for **X** given the constraints of the experiment that:
  - A-optimality: minimizes tr(**V**)
  - D-optimality: minimizes det(**V**)
  - E-optimality: minimizes leading eigenvalue of **V**

Ordinary Least Squares (OLS)

When the covariance structure of the residuals has a certain form, we solve for the vector β using OLS

If residuals follow a MVN distribution, OLS = ML solution

If the residuals are homoscedastic and uncorrelated, $\sigma^2(e_i) = \sigma_e^2$, $\sigma(e_i,e_j) = 0$. Hence, each residual is equally weighted,

Sum of squared residuals can be written as

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Predicted value of the y's

Ordinary Least Squares (OLS)

$$\sum_{i=1}^{n} \widehat{e}_i^2 = \widehat{\mathbf{e}}^T \widehat{\mathbf{e}} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Taking (matrix) derivatives shows this is minimized by

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

This is the OLS estimate of the vector **β**

The variance-covariance estimate for the sample estimates is

$$\mathbf{V}_{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \sigma_e^2$$

The ij-th element gives the covariance between the estimates of β_i and β_j.

Sample Variances/Covariances

The residual variance can be estimated as

$$\widehat{\sigma_e^2} = \frac{1}{n - \text{rank}(X)} \sum_{i=1}^{n} \widehat{e_i^2}$$

The estimated residual variance can be substituted into

$$\mathbf{V}_{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$$

To give an approximation for the sampling variance and covariances of our estimates.

Confidence intervals follow since the vector of estimates ~ MVN($\boldsymbol{\beta}$, $\mathbf{V}_{\beta}$)

Example: Regression Through the Origin

$$y_i = \beta x_i + e_i$$

Here $\mathbf{X} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}$ $\quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$ $\quad \boldsymbol{\beta} = (\,\beta\,)$

$$\mathbf{X}^T\mathbf{X} = \sum_{i=1}^{n} x_i^2 \qquad \mathbf{X}^T\mathbf{y} = \sum_{i=1}^{n} x_i\, y_i$$

$$\boxed{\begin{array}{l} \beta = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} = \dfrac{\sum x_i\, y_i}{\sum x_i^2} \\[2ex] \sigma^2(\beta) = \dfrac{1}{n-1}\,\dfrac{\sum (y_i - \beta x_i)^2}{\sum x_i^2} \end{array}}$$

$$\sigma^2(b) = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\sigma_e^2 = \frac{\sigma_e^2}{\sum x_i^2}$$

$$\sigma_e^2 = \frac{1}{n\ \text{-}\ 1}\sum (y_i\ \text{-}\ \beta x_i)^2$$

Polynomial Regressions

GLM can easily handle any function of the observed predictor variables, provided the parameters to estimate are still linear, e.g. $y = \alpha + \beta_1 f(x) + \beta_2 g(x) + \cdots + e$

Quadratic regression:

$$y_i = \alpha + \beta_1 \, x_i + \beta_2 \, x_i^2 + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix}$$

Interaction Effects

Interaction terms (e.g. sex x age) are handled similarly

$$y_i = \alpha + \beta_1 \, x_{i1} + \beta_2 \, x_{i2} + \beta_3 \, x_{i1} x_{i2} + e_i$$

$$\boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{11}x_{12} \\ 1 & x_{21} & x_{22} & x_{21}x_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & x_{n1}x_{n2} \end{pmatrix}$$

With x_1 held constant, a unit change in x_2 changes y by $\beta_2 + \beta_3 x_1$ (i.e., the slope in x_2 depends on the current value of x_1)

Likewise, a unit change in x_1 changes y by $\beta_1 + \beta_3 x_2$

The GLM lets you build your own model!

- Suppose you want a quadratic regression forced through the origin where the slope of the quadratic term can vary over the sexes (pollen vs. seed parents)
- $Y_i = \beta_1 x_i + \beta_2 x_i^2 + \beta_3 s_i x_i^2$
- s_i is an indicator (0/1) variable for the sex (0 = male, 1 = female).
  - Male slope = β_2,
  - Female slope = $\beta_2 + \beta_3$

Generalized Least Squares (GLS)

Suppose the residuals no longer have the same variance (i.e., display heteroscedasticity). Clearly we do not wish to minimize the *unweighted* sum of squared residuals, because those residuals with smaller variance should receive more weight.

Likewise in the event the residuals are correlated, we also wish to take this into account (i.e., perform a suitable transformation to remove the correlations) before minimizing the sum of squares.

Either of the above settings leads to a GLS solution in place of an OLS solution.

In the GLS setting, the covariance matrix for the vector e of residuals is written as R where

$R_{ij} = \sigma(e_i, e_j)$

The linear model becomes $y = X\beta + e$, $\mathrm{cov}(e) = R$

The GLS solution for β is

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}$$

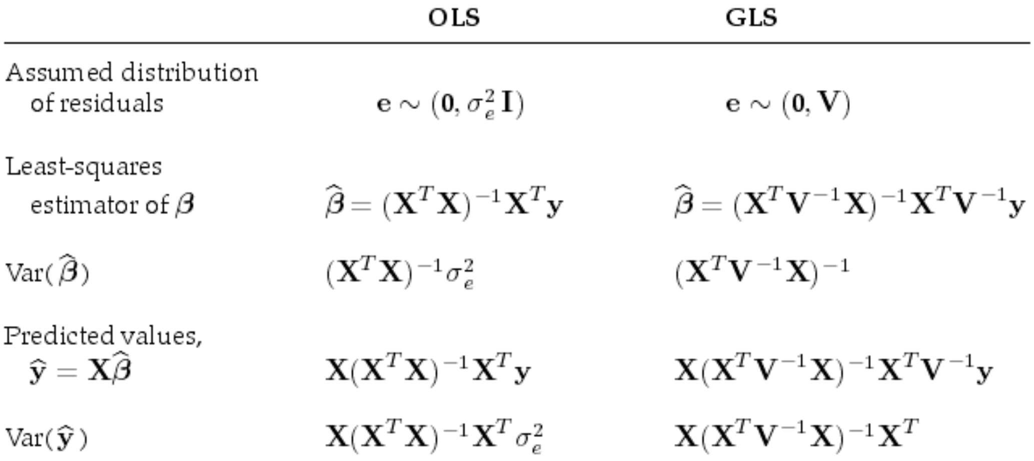
The variance-covariance of the estimated model parameters is given by

$$\mathbf{V_b} = \left(\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X}\right)^{-1}\sigma_e^2$$

Model diagnostics

- It's all about the residuals
- Plot the residuals
  - Quick and easy screen for outliers
- Test for normality among estimated residuals
  - Q-Q plot
  - Shapiro-Wilk test
  - If non-normal, try transformations, such as log

OLS, GLS summary



	OLS	GLS
Assumed distribution of residuals	$\mathbf{e} \sim (\mathbf{0}, \sigma_e^2 \mathbf{I})$	$\mathbf{e} \sim (\mathbf{0}, \mathbf{V})$
Least-squares estimator of $\boldsymbol{\beta}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Var($\widehat{\boldsymbol{\beta}}$)	$(\mathbf{X}^T\mathbf{X})^{-1}\sigma_e^2$	$(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$
Predicted values, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y}$
Var($\widehat{\mathbf{y}}$)	$\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\sigma_e^2$	$\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T$

Fixed vs. Random Effects

In linear models we are trying to accomplish two goals: estimation the values of model parameters and estimate any appropriate variances.

For example, in the simplest regression model, $y = \alpha + \beta x + e$, we estimate the values for α and β and also the variance of e. We, of course, can also estimate the $e_i = y_i - (\alpha + \beta x_i)$

Note that α/β are *fixed constants* we trying to estimate (fixed factors or fixed effects), while the e_i values are drawn from some probability distribution (typically Normal with mean 0, variance σ^2_e). The e_i are random effects.

This distinction between fixed and random effects is extremely important in terms of how we analyze a model. If a parameter is a fixed constant we wish to estimate, it is a fixed effect. If a parameter is drawn from some probability distribution and we are trying to make inferences on either the distribution and/or specific realizations from this distribution, it is a random effect.

We generally speak of estimating fixed factors (BLUE) and predicting random effects (BLUP -- best linear unbiased Predictor)

"Mixed" models (MM) contain both fixed and random factors

$$y = Xb + Zu + e, \quad u \sim MVN(0,R),\ e \sim MVN(0,\sigma^2_e I)$$

Key: need to specify covariance structures for MM

Example: Sire model

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

Here μ is a fixed effect, and e is a random effect

Is the sire effect s fixed or random ?

It depends. If we have (say) 10 sires, if we are ONLY interested in the values of these particular 10 sires and don't care to make any other inferences about the population from which the sires are drawn, then we can treat them as fixed effects. In the case, the model is fully specified by the covariance structure for the residuals. Thus, we need to estimate μ, s_1 to s_{10} and σ^2_e, and we write the model as $y_{ij} = \mu + s_i + e_{ij}$, $\sigma^2(e) = \sigma^2_e I$

Random effects models

- It is often useful to treat certain effects as random, as opposed to fixed
  - Suppose we have k effects. If we treat these as fixed, we spend k degrees of freedom
  - If we assume each of the k realizations are drawn from a normal with mean zero and unknown variance, only one degree of freedom lost --- that for estimating the variance
    - We can then predict the values of the k realizations

Environmental effects

- Consider yield data measured over several years in a series of plots.
- Standard to treat year-to-year variation at a specific site as being random effects
- Often the plot effects (mean value over years) are also treated as random.
- For example, consider plants group in growing region i, location k within that region, and year (season) k for that location-region effect
  - $E = R_i + L_{ik} + e_{ijk}$
  - Typically R can be a fixed effect, while L and e are random effects, $L_{ik} \sim N(0,\sigma^2_L)$ and $e_{ikj} \sim N(0,\sigma^2_e)$

Random models

- With a random model, one is assuming that all "levels" of a factor are not observed. Rather, some subset of values are drawn from some underlying distribution
  - For example, year to year variation in rainfall at a location. Each year is a random sample from the long-term distribution of rainfall values
  - Typically, assume a functional form for this underlying distribution (e.g., normal with mean 0) and then use observations to estimate the distribution parameters (here, the variance)

Random models (cont)

- Key feature:
  - Only one degree of freedom used (estimate of the variance)
  - Using the fixed effects and the estimated underlying distribution parameters, one then predicts the actual realizations of the individual values (i.e., the year effects)
  - Assumption: the covariance structure among the individual realizations of the realized effects. If only a variance is assumed, this implies each realization is independent. If realizations are assumed to be correlated, this structure must be estimated.

Random models

- Let’s go back to treating yearly effects as random
- If assume these are uncorrelated, only use one degree of freedom, but makes assumptions about covariance structure
  - Standard: Uncorrelated
  - Option: some sort of autocorrelation process, say with a yearly decay of r (must also be estimated)
- Conversely, could all be treated as fixed, but would use k degrees of freedom for k years, but no assumptions on their relationships (covariance structure)

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

Conversely, if we are not only interested in these 10 particular sires but also wish to make some inference about the population from which they were drawn (such as the additive variance, since $\sigma^2_A = 4\sigma^2_{s,}$), then the s_i are random effects. In this case we wish to estimate μ and the variances σ^2_s and σ^2_e. Since $2s_i$ also estimates (or predicts) the breeding value for sire i, we also wish to estimate (predict) these as well. Under a random-effects interpretation, we write the model as

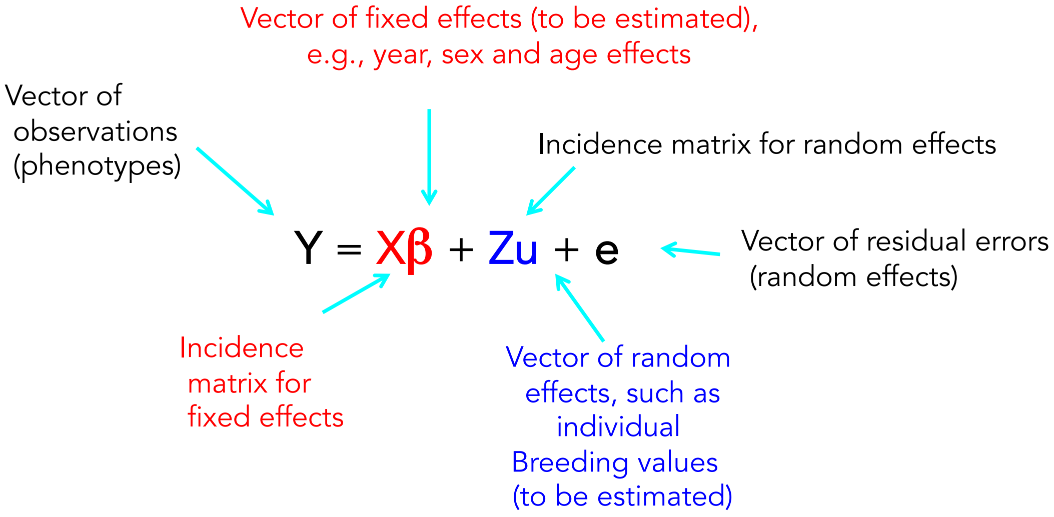
$y_{ij} = \mu + s_i + e_{ij}$, $\sigma^2(e) = \sigma^2_e I$, $\sigma^2(s) = \sigma^2_A A$

The relationship matrix A of know constants is given by the pedigree and is discussed later

Identifiability

- Recall that a fixed effect is said to be estimable if we can obtain a unique estimate for it (either because X is of full rank or when using a generalized inverse it returns a unique estimate)
  - Lack of estimable arises because the experiment design confounds effects
- The analogous term for random models is identifiability
  - The variance components have unique estimates

The general linear mixed model



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

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

Vector of observations (phenotypes)

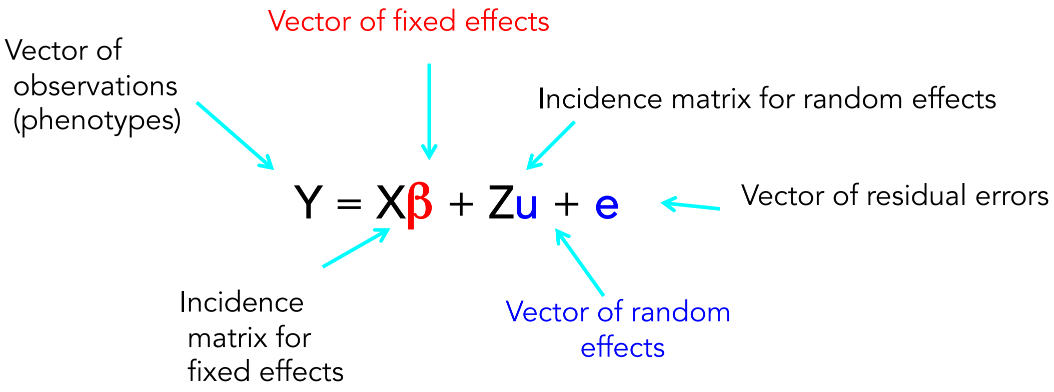
Incidence matrix for random effects

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

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

Vector of residual errors

Incidence matrix for fixed effects

Vector of random effects

Observe **y**, X, Z.

Estimate fixed effects $\boldsymbol{\beta}$

Estimate random effects **u, e**

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 vector **u** of random effects

The covariance matrix for y becomes

$$V = ZGZ^T + R$$

Hence, $y \sim MVN(X\beta, V)$

Mean $X\beta$ due to fixed effects
Variance V due to random effects

Chi-square and F distributions

Let $U_i \sim N(0,1)$, i.e., a unit normal

The sum $U_1^2 + U_2^2 + \cdots + U_k^2$ is a chi-square random variable with k degrees of freedom

Under appropriate normality assumptions, the sums of squares that appear in linear models are also chi-square distributed. In particular,

$$\sum_{i=1}^{n}(x_i - \overline{x})^2 \sim \chi^2_{n-1}$$

The ratio of two chi-squares is an **F distribution**

In particular, an F distribution with k numerator degrees of freedom, and n denominator degrees of freedom is given by

$$\frac{\chi^2_k/k}{\chi^2_n/n} \sim F_{k,n}$$

The expected value of a chi-square with k degrees of freedom is k, hence numerator and denominator both have expected value one

F distributions frequently arise in tests of linear models, as these usually involve ratios of sums of squares.

Sums of Squares in linear models

The total sums of squares (SST) of a linear model can be written as the sum of the error (or residual) sum of squares and the model (or regression) sum of squares

$$\begin{tikzcd} SS_T \arrow[d] & = SS_M \arrow[d] & + SS_E \arrow[d] \\ \sum(y_i-\overline{y})^2 & \sum(\widehat{y}_i-\overline{y})^2 & \sum(y_i-\widehat{y}_i)^2 \end{tikzcd}$$

r^2, the coefficient of determination, is the fraction of variation accounted for by the model

$$r^2 = \frac{SS_M}{SS_T} = 1 - \frac{SS_E}{SS_T}$$

Sums of Squares are quadratic products

$$SS_T = \sum_{i=1}^{n}(y_i-\overline{y})^2 = \sum_{i=1}^{n} y_i^2-\overline{y}^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n^2}\left(\sum_{i=1}^{n} y_i\right)^2$$

We can write this as a quadratic product as

$$SS_T = \mathbf{y}^T\mathbf{y} - \frac{1}{n}\mathbf{y}^T\mathbf{J}\mathbf{y} = \mathbf{y}^T\left(\mathbf{I}-\frac{1}{n}\mathbf{J}\right)\mathbf{y}$$

Where **J** is a matrix all of whose elements are 1's

$$SS_E = \sum_{i=1}^{n}(y_i-\widehat{y}_i)^2 = \sum_{i=1}^{n}\widehat{e}_i^2$$

$$SS_E = \mathbf{y}^T\left(\mathbf{I}-\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\right)\mathbf{y}$$

$$SS_M = SS_T - SS_E = \mathbf{y}^T\left(\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T - \frac{1}{n}\mathbf{J}\right)\mathbf{y}$$

Expected value of sums of squares

- In ANOVA tables, the E(MS), or expected value of the Mean Squares (scaled SS or Sum of Squares), often appears
- This directly follows from the quadratic product. If E(**x**) = **μ**, Var(**x**) = **V**, then
  - $E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \mathrm{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu}$

Hypothesis testing

Provided the residual errors in the model are MVN, then for a model with n observations and p estimated parameters,

$$\frac{\mathrm{SS}_E}{\sigma_e^2} \sim \chi^2_{n-p}$$

Consider the comparison of a full (p parameters) and reduced (q < p) models, where SSE_r = error SS for reduced model, SSE_f = error SS for full model

$$\left(\frac{\mathrm{SS}_{E_r} - \mathrm{SS}_{E_f}}{p-q}\right) \Big/ \left(\frac{\mathrm{SS}_{E_f}}{n-p}\right) = \left(\frac{n-p}{p-q}\right)\left(\frac{\mathrm{SS}_{E_r}}{\mathrm{SS}_{E_f}} - 1\right)$$

The difference in the error sum of squares for the full and reduced model provided a test for whether the model fit is the same

This ratio follows an $F_{p-q,n-p}$ distribution

Does our model account for a significant fraction of the variation?

Here the reduced model is just $y_i = u + e_i$

In this case, the error sum of squares for the reduced model is just the total sum of squares, and the F test ratio becomes

$$\left(\frac{n-p}{p-1}\right)\left(\frac{\mathrm{SS}_T}{\mathrm{SS}_{E_f}}-1\right)=\left(\frac{n-p}{p-1}\right)\left(\frac{r^2}{1-r^2}\right)$$

This ratio follows an $F_{p-1,n-p}$ distribution

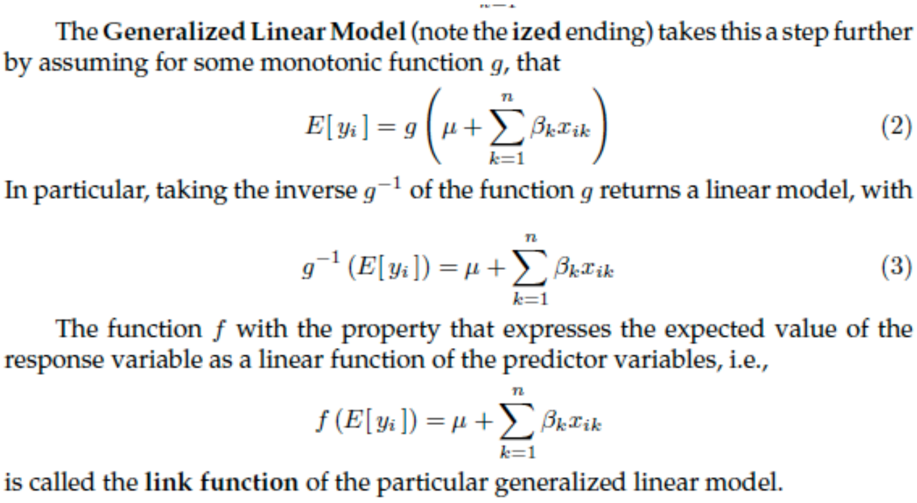
Different statistical models

- GLM = general linear model
  - OLS ordinary least squares: e ~ MVN(0,cI)
  - GLS generalized least squares: e ~ MVN(0,R)
- Non-linear models
  - Parametric growth curves
- Mixed models
  - Both fixed and random effects (beyond the residual)
- Mixture models
  - A weighted mixture of distributions
- Generalized linear models
  - Nonlinear functions, non-normality

Mixture models

- Under a mixture model, an observation potentially comes from one of several different distributions, so that the density function is $\pi_1\phi_1 + \pi_2\phi_2 + \pi_3\phi_3$
  - The mixture proportions π_i sum to one
  - The ϕ_i represent different distribution, e.g., normal with mean μ_i and variance σ^2
- Mixture models come up in QTL mapping -- an individual could have QTL genotype QQ, Qq, or qq
  - See Lynch & Walsh Chapter 13
- They also come up in codon models of evolution, were a site may be neutral, deleterious, or advantageous, each with a different distribution of selection coefficients
  - See Walsh & Lynch (volume 2A website), Chapters 10,11

Generalized linear models



The **Generalized Linear Model** (note the **ized** ending) takes this a step further by assuming for some monotonic function g, that

$$E[\,y_i\,] = g\left(\mu + \sum_{k=1}^{n} \beta_k x_{ik}\right) \qquad (2)$$

In particular, taking the inverse g^{-1} of the function g returns a linear model, with

$$g^{-1}\left(E[\,y_i\,]\right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik} \qquad (3)$$

The function f with the property that expresses the expected value of the response variable as a linear function of the predictor variables, i.e.,

$$f\left(E[\,y_i\,]\right) = \mu + \sum_{k=1}^{n} \beta_k x_{ik}$$

is called the **link function** of the particular generalized linear model.

Typically assume non-normal distribution for residuals, e.g., Poisson, binomial, gamma, etc

Likelihoods for GLMs

Under assumption of MVN, x ~ MVN(β ,**V**), the likelihood function becomes

$$L(\beta,\mathbf{V} \mid \mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}(\mathbf{x}-\beta)^T \mathbf{V}^{-1} (\mathbf{x}-\beta)\right]$$

Variance components (e.g., σ^2_A, σ^2_e, etc.) are included in **V**

REML = restricted maximum likelihood. Method of choice for variance components, as it maximizes that part of the likelihood function that is independent of the fixed effects, β.

Overview And Introduction to Mixed Models

- References
  - Searle, S.R. 1971 Linear Models, Wiley
  - Schaefer, L.R., Linear Models and Computer Strategies in Animal Breeding
  - Lynch and Walsh Chapter 8

Linear vs non-linear

Linear
2nd order Polynomial

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{1i}^2 + b_3 X_{3i} + b_4 X_{4i}^2 + b_5 \left(X_{1i} X_{2i} \right) .. + \varepsilon_\iota$$

Non-linear

$$Y_i = b_0 e^{-b_1 X_i} \varepsilon_\iota$$

log-linear

$$\ln\left(Y_i\right) = \ln(b_0) - b_1 X_i + \ln(\varepsilon_\iota)$$

Why Linear: Life is Non-Linear

Taylor Expansion

$$Y = f(X)$$

$$Y \approx f(a) + \frac{f^{'}(a)(x-a)}{1!} + \frac{f^{''}(a)(x-a)^2}{2!} + \ldots + \frac{f^{n}(a)(x-a)^n}{n!}$$

$$Y = e^{-X} \qquad Y^{'} = -e^{-X} \qquad Y^{''} = e^{-X}$$

At a=0 $$Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots.$$

Lower Order Terms Are more Important than higher

Works for other values of ‘a’ but not as exact, example a=.1

$$Y = e^{-X} \qquad Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} + \ldots.$$

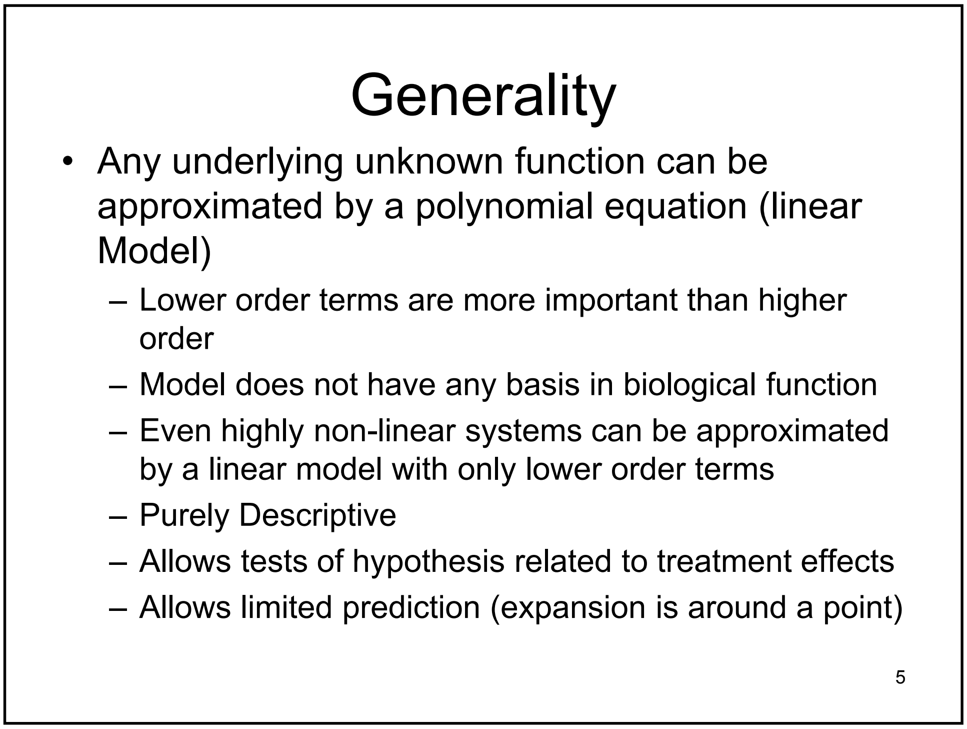
X=.1

$$Y \approx 1$$

$$Y = e^{-.1} = .904837 \qquad Y \approx 1 - x = 1 - .1 = .9$$

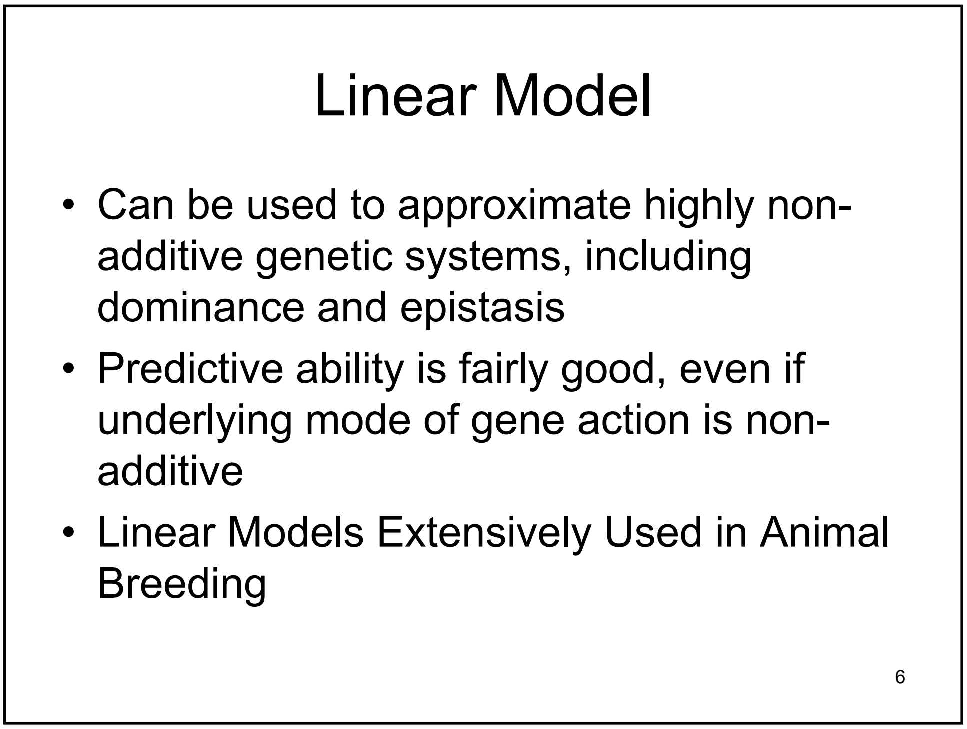
$$Y \approx 1 - x + \frac{x^2}{2!} = 1 - .1 + \frac{.1^2}{2!} = .905$$

$$Y \approx 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} = 1 - .1 + \frac{.1^2}{2!} - \frac{.1^3}{3!} = .904833$$



Generality

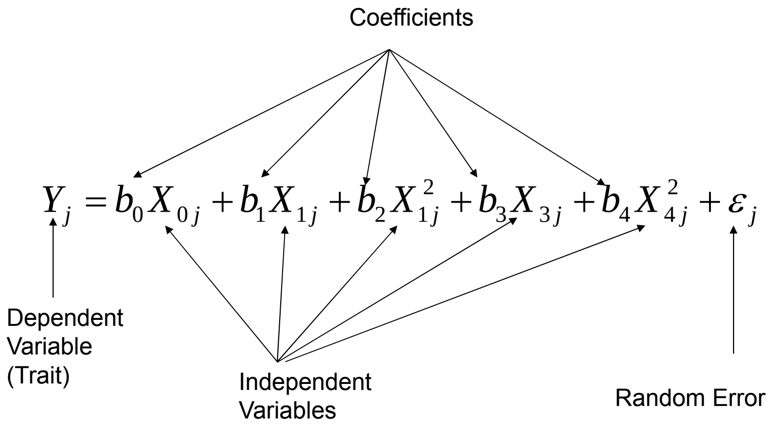
- Any underlying unknown function can be approximated by a polynomial equation (linear Model)
  - Lower order terms are more important than higher order
  - Model does not have any basis in biological function
  - Even highly non-linear systems can be approximated by a linear model with only lower order terms
  - Purely Descriptive
  - Allows tests of hypothesis related to treatment effects
  - Allows limited prediction (expansion is around a point)



Linear Model

- Can be used to approximate highly non-additive genetic systems, including dominance and epistasis
- Predictive ability is fairly good, even if underlying mode of gene action is non-additive
- Linear Models Extensively Used in Animal Breeding

One Random effect Linear Model



Coefficients

$$Y_j = b_0 X_{0j} + b_1 X_{1j} + b_2 X_{1j}^2 + b_3 X_{3j} + b_4 X_{4j}^2 + \varepsilon_j$$

Dependent Variable (Trait)

Independent Variables

Random Error

Matrix Notation

$$Y_1 = b_0 X_{01} + b_1 X_{11} + b_2 X_{11}^2 + b_3 X_{31} + b_4 X_{41}^2 + \varepsilon_1$$

$$Y_2 = b_0 X_{02} + b_1 X_{12} + b_2 X_{12}^2 + b_3 X_{32} + b_4 X_{42}^2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = b_0 X_{0n} + b_1 X_{1n} + b_2 X_{1n}^2 + b_3 X_{3n} + b_4 X_{4n}^2 + \varepsilon_n$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \underbrace{\begin{bmatrix} X_{01} & X_{11} & X^2{}_{11} & X_{21} & X^2{}_{21} \\ X_{02} & X_{12} & X^2{}_{12} & X_{22} & X^2{}_{22} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ X_{0n} & X_{1n} & X^2{}_{1n} & X_{2n} & X^2{}_{2n} \end{bmatrix}} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$$

Estimation

$$\mathbf{Y} = \mathbf{XB} + \boldsymbol{\varepsilon}$$

Ordinary Least Squares

- Independent variables (X)
  - fixed
  - measured without error
- Residuals
  - Random
  - Independently and Identically Distributed (IID) with Mean 0 and variance σ^2

Independently and Identically Distributed with Mean 0 and variance σ^2

$$V(\varepsilon) = E\left[\varepsilon - E(\varepsilon)\right]^2$$

The error distribution from which each observation is sampled is the same

$$V(\varepsilon) = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

$$V(\varepsilon) = \mathbf{I}\sigma_e^2$$

No Environmental Correlations

When would these assumptions be violated?

Ordinary Least Squares Estimator

Find Solutions such that the sum of the residuals squared is minimum

$$\varepsilon_j = Y_j - E(Y_j)$$

$$\sum_{j=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 & \varepsilon_2 & \cdots & \varepsilon_n \end{bmatrix} \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$E(Y_j) = \sum_{i=0}^{k} b_i X_{ij}$$

$$\varepsilon_j = Y_j - \sum_{i=0}^{k} b_i X_{ij}$$

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} \left(Y_j - \sum_{i=0}^{k} b_j X_{ij} \right)^2$$

Least Square Estimators

$$\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \sum_{j=1}^{n} \varepsilon_j^2 = \sum_{j=1}^{n} \left(Y_j - \sum_{i=0}^{m} b_i X_{ij} \right)^2$$

Find all bi such that sum of residuals squared is minimum

$$\frac{\partial(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon})}{\partial b_i} = 2\sum_{j=1}^{n} \left(Y_j - \sum_{i=0}^{m} b_i X_{ij} \right) \left[- X_{ij} \right]$$

Set=0 for each i and solve system

Normal Equations

$$\begin{bmatrix} \sum x_{0j}^2 & \sum x_{0j}x_{1j} & \cdots & \sum x_{0j}x_{kj} \\ \sum x_{0j}x_{1j} & \sum x_{1j}^2 & \cdots & \sum x_{1j}x_{kj} \\ \vdots & \vdots & & \vdots \\ \sum x_{0j}x_{kj} & \sum x_{1j}x_{kj} & \cdots & \sum x_{kj}^2 \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} \sum x_{0j}y_j \\ \sum x_{1j}y_j \\ \vdots \\ \sum x_{kj}y_j \end{bmatrix}$$

$$\mathbf{X'XB} = \mathbf{X'Y}$$

$$\hat{\mathbf{B}} = \left(\mathbf{X'X}\right)^{-\mathbf{1}}\left(\mathbf{X'Y}\right)$$

$$V(\hat{\mathbf{B}}) = \sigma_e^2\left(\mathbf{X'X}\right)^{-\mathbf{1}}$$

Prediction

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\mathbf{B}}$$

$$V(\hat{\mathbf{Y}}) = V(\mathbf{X}\hat{\mathbf{B}})$$

$$V(\hat{\mathbf{Y}}) = \mathbf{X}V(\hat{\mathbf{B}})\mathbf{X'}$$

$$V(\hat{\mathbf{Y}}) = \mathbf{X}\left(\mathbf{X'X}\right)^{-\mathbf{1}}\mathbf{X'}\sigma_e^2$$

$$\hat{\mathbf{B}} = \left(\mathbf{X'X}\right)^{-\mathbf{1}}\left(\mathbf{X'Y}\right)$$

$$V(\hat{\mathbf{B}}) = \sigma_e^2\left(\mathbf{X'X}\right)^{-\mathbf{1}}$$

Example Factor Affecting Fatty Acid
From Gill, J. Design and Analysis of experiments

Fatty Acid	Amount over Weight (Kg)	Age
10	6	28
20	12	40
17	10	32
12	8	36
11	9	34

R Code Example 1

```
Y = matrix( c(10,
            20,
            17,
            12,
            11  ), 5,1)

X = matrix(c( 1, 6,  28,
                 1, 12, 40,
                 1, 10, 32,
                 1,  8, 36,
                 1,  9, 34 ),5,3, byrow = TRUE)
LHS =(t(X) %*% X  )
RHS =(t(X) %*% Y)
C = solve(LHS)
B = C %*% RHS
B
```

```
> B
         [,1]
[1,]  2.3333333
[2,]  2.0833333
[3,] -0.2083333
>
```

BY GLM

```
• data one;
• input fatty_acid over_wt age;
• cards;
• 10  6  28
• 20 12 40
• 17 10 32
• 12  8 36
• 11  9 34
• ;
• proc glm;
• model fatty_acid=over_wt age
  / solution;
• run;
• quit;
```

- Compare results from IML to GLM

Generalized Least Squares (GLS)

- Ordinary Least Squares
  - Independent variables
    - fixed
    - measured without error
  - Residuals
    - Random
    - Independently and Identically Distributed (IID) with Mean 0 and variance σ^2

- Generalized Least Squares
  - Independent variables
    - fixed
    - measured without error
  - Residuals
    - Random

$$V(\varepsilon) = \mathbf{V}$$

GLS

Minimize weighted SS

$$(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})$$

Weighting by the inverse of the variance

$$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$

If

$$\mathbf{V}=\mathbf{I}\sigma_e^2$$

$$\hat{\mathbf{b}}=(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y})$$

Maximum Likelihood (ML) Solution to Same Problem

- Generalized Least Squares
  - Independent variables
    - fixed
    - measured without error
  - Residuals
    - Random

$$V(\varepsilon)=\mathbf{V}$$

- Maximum Likelihood
  - Independent variables
    - fixed
    - measured without error
  - Residuals
    - Random

$$V(\varepsilon)=\mathbf{V}$$

$$\varepsilon\approx N(\mathbf{0},\mathbf{V})$$

ML

$$L = \frac{1}{(2\pi)^{\frac{N}{2}} \left|\mathbf{V}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{y}-\mathbf{Xb})' \mathbf{V}^{-1} (\mathbf{y}-\mathbf{Xb})}$$

Maximize w.r.t **b** $\frac{\partial(\ln L)}{\partial \mathbf{b}} = 0$

$$\ln L = \ln(C) - \tfrac{1}{2}(\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1} (\mathbf{y} - \mathbf{Xb})$$

$$\frac{\partial(\ln L)}{\partial \mathbf{b}} = -\tfrac{1}{2}(\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1}(-\mathbf{X}) - \tfrac{1}{2}(-\mathbf{X})' \mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

$$\frac{\partial(\ln L)}{\partial \mathbf{b}} = (\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1}\mathbf{X}$$

$$(\mathbf{y} - \mathbf{Xb})' \mathbf{V}^{-1}\mathbf{X} = 0$$

$$\left(\mathbf{y}' - (\mathbf{Xb})'\right)\mathbf{V}^{-1}\mathbf{X} = 0$$

$$(\mathbf{y}' - \mathbf{b}'\mathbf{X}')\mathbf{V}^{-1}\mathbf{X} = 0$$

$$\mathbf{b}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}$$

$$\hat{\mathbf{b}}' = (\mathbf{y}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}$$

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})$$

Same as GLS
Just because one approach has an assumption does not mean this assumption is necessary in general

Variance of b

$$V(\mathbf{b}) = \left(\mathbf{X'V^{-1}X}\right)^{-1}$$

Note if $\mathbf{V} = \mathbf{I}\sigma_e^2$

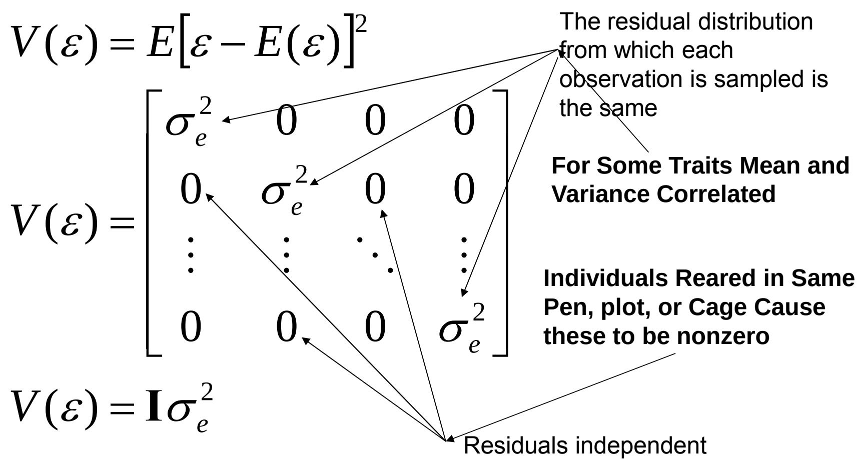
$$V(\mathbf{b}) = \sigma_e^2\left(\mathbf{X'X}\right)^{-1}$$

•This is not the distribution of **b**, but rather is the variance of the estimate
•**b** is considered a fixed effect and as such does not have a distribution

BLUP Best Linear Unbiased Prediction-Estimation

- References
- Searle, S.R. 1971 Linear Models, Wiley
- Schaefer, L.R., Linear Models and Computer Strategies in Animal Breeding
- Lynch and Walsh Chapter 26

OLS Independently and Identically Distributed Errors with Mean 0 and variance σ²



$$V(\varepsilon) = E\left[\varepsilon - E(\varepsilon)\right]^2$$

$$V(\varepsilon) = \begin{bmatrix} \sigma_e^2 & 0 & 0 & 0 \\ 0 & \sigma_e^2 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \sigma_e^2 \end{bmatrix}$$

$$V(\varepsilon) = \mathbf{I}\sigma_e^2$$

The residual distribution from which each observation is sampled is the same

For Some Traits Mean and Variance Correlated

Individuals Reared in Same Pen, plot, or Cage Cause these to be nonzero

Residuals independent

Solutions

- GLS
  - Fixes problem with changing variances and correlations in the data
- What about fixed effects?
  - How does one correct for
    - Environmental trend without a control
    - Herd effects
    - Year effects
    - Hatch effects
    - Confounding

Confounding of data

- Herd effects
  - Balanced design no problem
  - Require sample of every family in every herd
  - Old solution was within herd deviations
  - What if better herds have better genetics
- Fixed effects must be adjusted for genetic differences
- Random effects must be adjusted for fixed effects
- Requires simultaneous solutions

Mixed Model

Simultaneous Adjustment of Fixed and Random effects

- Separates Independent variable into those that are
  - Fixed $\mathbf{Xb}$
  - Random $\mathbf{Zu}$

X=value of each fixed effect
b=linear regression coefficients
Z=incidence matrix of random effect, usually a 1 corresponding to each animal
u=estimate of random effects (breeding value)

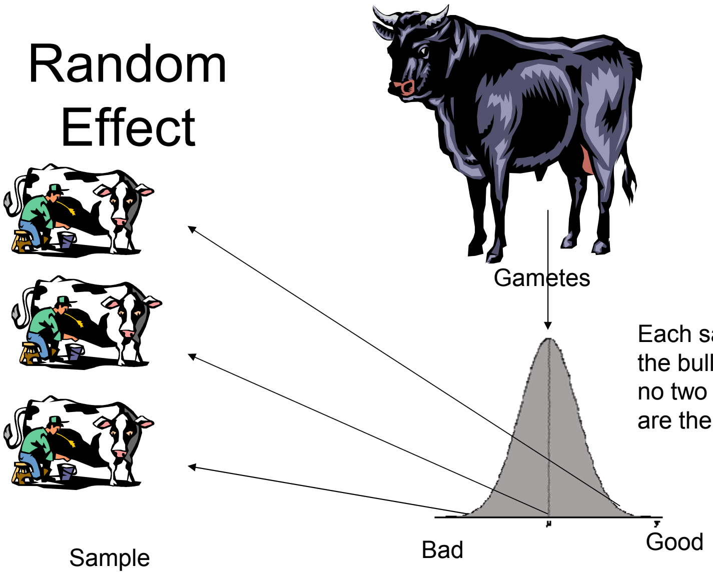
$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

More importantly model’s the variance structure

Fixed and Random Effects

- Fixed Effect
  - Inference Space only to those levels
  - Age, Hatch, Location, Parity, and Sex effects
- Random Effect
  - Effect Sampled From a Distribution of Effects
  - Inference Space To The Population From Which The Random Effect Was Sampled
  - If a new sample of observations were made (a new experiment), and the levels were completely different between the two samples, then the factors is usually random

Random Effect



Gametes

Each sample from the bull is different, no two gametes are the same

Sample

Bad

Good

Inference is to the genetic worth of the bull (breeding value)

Variances In Mixed Models

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

$$V(\mathbf{b}) = \mathbf{0}$$

$$V(\mathbf{u}) = E(\mathbf{uu}') = \mathbf{G}$$

$$V(\mathbf{e}) = E(\mathbf{ee}') = \mathbf{R}$$

$$V(\mathbf{Y}) = V(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) = \mathbf{ZGZ'} + \mathbf{R}$$

Estimate the breeding values “**u”** and fixed effects simultaneously

Old concept was to first adjust for the fixed effects, output the residuals and estimate the random effects

Resulted in Biased Estimates of Both Fixed and Random Effects

ML Derivation of Solutions

Joint density of **y** and **u** $\quad f(\mathbf{y},\mathbf{u}) = g(\mathbf{y}/\mathbf{u})h(\mathbf{u})$

$$g(\mathbf{y}/\mathbf{u}) = g(\mathbf{e})$$

$$g(\mathbf{e}) = \frac{1}{(2\pi)^{\frac{1}{2}N}\sigma_e} e^{-\frac{1}{2}\mathbf{e}'V(\mathbf{e})^{-1}\mathbf{e}} \qquad h(\mathbf{u}) = \frac{1}{(2\pi)^{\frac{1}{2}N}\sigma_u} e^{-\frac{1}{2}\mathbf{u}'V(\mathbf{u})^{-1}\mathbf{u}}$$

$$f(\mathbf{y},\mathbf{u}) = \frac{1}{(2\pi)^{\frac{1}{2}N}\sigma_e} e^{-\frac{1}{2}\mathbf{e}'\mathbf{R}^{-1}\mathbf{e}} \frac{1}{(2\pi)^{\frac{1}{2}N}\sigma_u} e^{-\frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}}$$

$$f(\mathbf{y},\mathbf{u}) = c_1 e^{-\frac{1}{2}\mathbf{e}'\mathbf{R}^{-1}\mathbf{e}} c_2 e^{-\frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}}$$

$$f(\mathbf{y},\mathbf{u}) = c e^{-\frac{1}{2}\mathbf{e}'\mathbf{R}^{-1}\mathbf{e}} e^{-\frac{1}{2}\mathbf{u}'\mathbf{G}^{-1}\mathbf{u}}$$

$$f(\mathbf{y},\mathbf{u}) = L = ce^{-\frac{1}{2}\mathbf{e'R}^{-1}\mathbf{e}}e^{-\frac{1}{2}\mathbf{u'G}^{-1}\mathbf{u}}$$

Maximize w.r.t **b and u**

$$\ln(L) = \ln(c) - \tfrac{1}{2}\mathbf{e'R}^{-1}\mathbf{e} - \tfrac{1}{2}\mathbf{u'G}^{-1}\mathbf{u}$$

$$\mathbf{e} = \mathbf{Y} - \mathbf{Xb} - \mathbf{Zu}$$

$$\ln(L) = \ln(c) - \tfrac{1}{2}(\mathbf{Y} - \mathbf{Xb} - \mathbf{Zu})'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{Xb} - \mathbf{Zu}) - \tfrac{1}{2}\mathbf{u'G}^{-1}\mathbf{u}$$

SIMPLIFY FIRST THEN TAKE DERIVATIVES

$$(\mathbf{Y} - \mathbf{Xb} - \mathbf{Zu})'\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{Xb} - \mathbf{Zu}) + \mathbf{u'G}^{-1}\mathbf{u}$$
$$= \left[\mathbf{Y}' - (\mathbf{Xb})' - (\mathbf{Zu})'\right]\mathbf{R}^{-1}(\mathbf{Y} - \mathbf{Xb} - \mathbf{Zu}) + \mathbf{u'G}^{-1}\mathbf{u}$$

$$= \mathbf{Y}'\mathbf{R}^{-1}\mathbf{Y} - \mathbf{Y}'\mathbf{R}^{-1}\mathbf{Xb} - \mathbf{Y}'\mathbf{R}^{-1}\mathbf{Zu}$$
$$-(\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Y} + (\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Xb} + (\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Zu}$$
$$-(\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Y} + (\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Xb} + (\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Zu} + \mathbf{u'G}^{-1}\mathbf{u}$$

$$\frac{\partial(\ln L)}{\partial\mathbf{b}} = 0 \qquad -\mathbf{Y}'\mathbf{R}^{-1}\mathbf{X} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} + (\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{X} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Xb} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Zu} + (\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{X} = 0$$

$$-2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Xb} + 2\mathbf{X}'\mathbf{R}^{-1}\mathbf{Zu} = 0$$

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Xb} + \mathbf{X}'\mathbf{R}^{-1}\mathbf{Zu} = \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}$$

Take Derivative w.r.t **u**

$$(\mathbf{Y}-\mathbf{Xb}-\mathbf{Zu})'\mathbf{R}^{-1}(\mathbf{Y}-\mathbf{Xb}-\mathbf{Zu})+\mathbf{u'}\mathbf{G}^{-1}\mathbf{u}$$

$$\begin{aligned}&=\mathbf{Y}'\mathbf{R}^{-1}\mathbf{Y}-\mathbf{Y}'\mathbf{R}^{-1}\mathbf{Xb}-\mathbf{Y}'\mathbf{R}^{-1}\mathbf{Zu}\\&-(\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Y}+(\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Xb}+(\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Zu}\\&-(\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Y}+(\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Xb}+(\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Zu}+\mathbf{u'}\mathbf{G}^{-1}\mathbf{u}\end{aligned}$$

$$\frac{\partial(\ln L)}{\partial\mathbf{u}}=0 \qquad \begin{aligned}&-\mathbf{Y}'\mathbf{R}^{-1}\mathbf{Z}+(\mathbf{Xb})'\mathbf{R}^{-1}\mathbf{Z}-(\mathbf{Z})'\mathbf{R}^{-1}\mathbf{Y}+(\mathbf{Z})'\mathbf{R}^{-1}\mathbf{Xb}\\&+(\mathbf{Z})'\mathbf{R}^{-1}\mathbf{Zu}+(\mathbf{Zu})'\mathbf{R}^{-1}\mathbf{Z}+2\mathbf{G}^{-1}\mathbf{u}=0\end{aligned}$$

$$-2\mathbf{Y}'\mathbf{R}^{-1}\mathbf{Z}+2\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Xb}+2\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Zu}+2\mathbf{G}^{-1}\mathbf{u}=0$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Xb}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Zu}+\mathbf{G}^{-1}\mathbf{u}=\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y}$$

Mixed Model Equations

$$\mathbf{X}'\mathbf{R}^{-1}\mathbf{Xb}+\mathbf{X}'\mathbf{R}^{-1}\mathbf{Zu}=\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}$$

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Xb}+\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Zu}+\mathbf{G}^{-1}\mathbf{u}=\mathbf{Z'}\mathbf{R}^{-1}\mathbf{Y}$$

$$\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{u}\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{Y}\\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y}\end{bmatrix}$$

Simplifications If $\mathbf{R}=\mathbf{I}\sigma_e^2$

$$\begin{bmatrix}\mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\sigma_e^2\mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\mathbf{b}\\ \mathbf{u}\end{bmatrix}=\begin{bmatrix}\mathbf{X}'\mathbf{Y}\\ \mathbf{Z'}\mathbf{Y}\end{bmatrix}$$

Alternative derivations are possible that do not require Normal Dist’n Assumptions, resulting in these same solutions and are therefore also Best Linear Unbiased Predictors (BLUP)

BLUP Breeding Values

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\sigma_e^2\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

With Diploid Organisms and Assuming Additivity

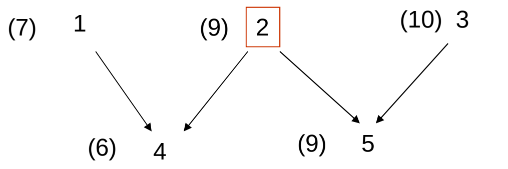
$$\mathbf{G} = \mathbf{A}\sigma_a^2 \qquad \mathbf{G}^{-1} = \frac{1}{\sigma_a^2}\mathbf{A}^{-1}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

Only Estimate of Ratio is Needed

Only inverse is needed

Example 2


(7) 1

(9) 2

(10) 3

(6) 4

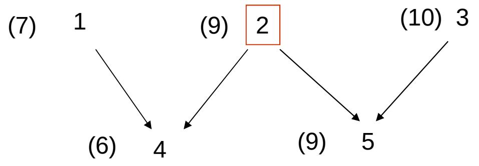
(9) 5

$$\begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}[\mu] + \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

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

Example 2

$$\mathbf{Y} = \begin{bmatrix} 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$


(7) 1

(9) 2

(10) 3

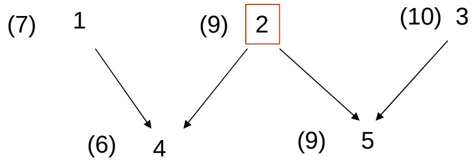
(6) 4

(9) 5

$$b = [\mu]$$

$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \mathbf{u} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \end{bmatrix}$$

Find **u**


(7) 1

(9) 2

(10) 3

(6) 4

(9) 5

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$$

Assume heritability=.5

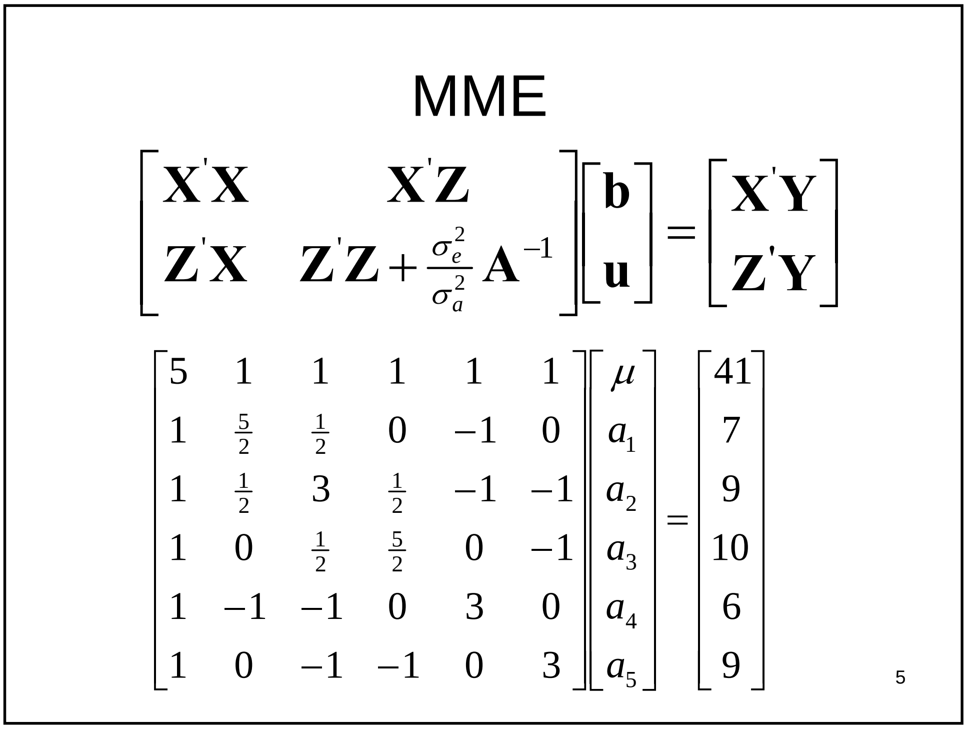
$$\sigma_a^2 = 2$$

$$\sigma_e^2 = 2$$

$$h^2 = \frac{2}{2+2} = .5$$

$$\frac{\sigma_e^2}{\sigma_a^2} = 1$$

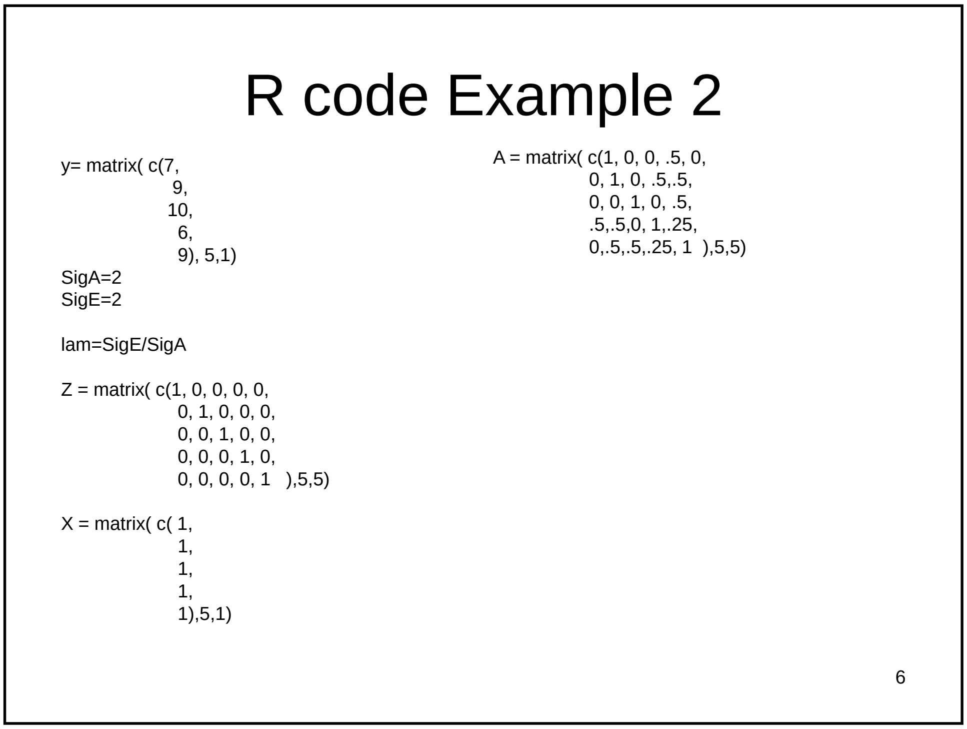
$$\mathbf{Z}^{'}\mathbf{Z} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} = \begin{bmatrix} \frac{5}{2} & \frac{1}{2} & 0 & -1 & 0 \\ \frac{1}{2} & 3 & \frac{1}{2} & -1 & -1 \\ 0 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ -1 & -1 & 0 & 3 & 0 \\ 0 & -1 & -1 & 0 & 3 \end{bmatrix}$$



MME

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

$$\begin{bmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & \frac{5}{2} & \frac{1}{2} & 0 & -1 & 0 \\ 1 & \frac{1}{2} & 3 & \frac{1}{2} & -1 & -1 \\ 1 & 0 & \frac{1}{2} & \frac{5}{2} & 0 & -1 \\ 1 & -1 & -1 & 0 & 3 & 0 \\ 1 & 0 & -1 & -1 & 0 & 3 \end{bmatrix} \begin{bmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix} = \begin{bmatrix} 41 \\ 7 \\ 9 \\ 10 \\ 6 \\ 9 \end{bmatrix}$$



R code Example 2

```
y= matrix( c(7,
                    9,
                   10,
                     6,
                     9), 5,1)
SigA=2
SigE=2

lam=SigE/SigA

Z = matrix( c(1, 0, 0, 0, 0,
                     0, 1, 0, 0, 0,
                     0, 0, 1, 0, 0,
                     0, 0, 0, 1, 0,
                     0, 0, 0, 0, 1   ),5,5)

X = matrix( c( 1,
                     1,
                     1,
                     1,
                     1),5,1)
```

```
A = matrix( c(1, 0, 0, .5, 0,
                  0, 1, 0, .5,.5,
                  0, 0, 1, 0, .5,
                  .5,.5,0, 1,.25,
                  0,.5,.5,.25, 1  ),5,5)
```

R code

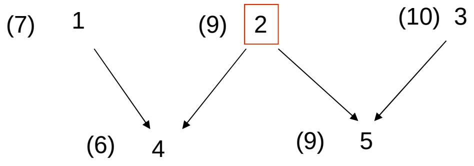
```
LHS = rbind( cbind(t(X) %*% X ,  t(X) %*% Z ),
              cbind( t(Z) %*% X  , ( t(Z) %*% Z ) + (lam * solve(A)) ))

RHS = rbind(t(X) %*% y,
              t(Z) %*% y)

C = solve(LHS)

BU = C %*% RHS

BU

yhat=X*BU[1]+BU[2:6]
yhat
```

```
[1,]  8.30
[2,] -0.96
[3,]  0.07
[4,]  0.88
[5,] -1.06
[6,]  0.55
```

$$b = [\hat{\mu}]$$

$$\mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$

Compare predicted value with phenotype

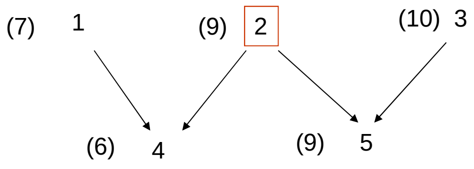


$$\hat{\mathbf{Y}} =$$

```
[1,] 7.34
 [2,] 8.37
 [3,] 9.18
 [4,] 7.23
 [5,] 8.85
```

- •Values were regressed partially to the mean **u**=8.30
- •Note that simple average of phenotypic values gives **u**=`8.20`
- •The fixed effects were adjusted for the random effects and random effect were adjusted for fixed effects simultaneously

Assume heritability=.01



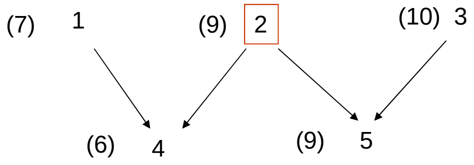
$$\sigma_a^2 = 2$$

$$\sigma_e^2 = 200$$

$$h^2 = \frac{2}{202} \cong .01$$

- •What do you expect the breeding values to be?
- •In terms of deviation from overall mean?
- •In terms of deviation from observed phenotype?

Rerun R code



$$b = [\hat{\mu}]$$

```
[1,] 8.20
[2,] -0.02
[3,] 0.00
[4,] 0.02
[5,] -0.02
[6,] 0.02
```

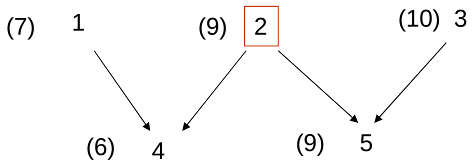
$$\mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$

$$\hat{\mathbf{Y}} =$$

```
[1,] 8.17
[2,] 8.20
[3,] 8.22
[4,] 8.17
[5,] 8.21
```

- •All values were regressed to the mean **u**=`8.20`
- •In this case **u** is the average of the phenotypic values because there were no genetic effects to adjust for

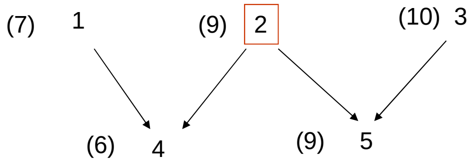
Assume heritability=.99



$$\sigma_a^2 = 200$$
$$\sigma_e^2 = .002$$
$$h^2 = \frac{200}{202} \cong .99$$

- What do you expect the breeding values to be?
- In terms of deviation from overall mean?
- In terms of deviation from observed phenotype?

Rerun R code



$$b = \begin{bmatrix} \hat{\mu} \end{bmatrix} \qquad \mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$

```
[1,] 8.65
[2,] -1.65
[3,] 0.32
[4,] 1.33
[5,] -2.61
[6,] 0.35
```

$\hat{\mathbf{Y}} =$

```
[1,] 6.99
[2,] 8.97
[3,] 9.98
[4,] 6.03
[5,] 9.00
```

The phenotypic and genotypic means are the same

Variance of the Estimates

$$\begin{bmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{'}\mathbf{X} & \mathbf{X}^{'}\mathbf{Z} \\ \mathbf{Z}^{'}\mathbf{X} & \mathbf{Z}^{'}\mathbf{Z} + \frac{\sigma_e^2}{\sigma_a^2}\mathbf{A}^{-1} \end{bmatrix}^{-1}$$

$$V(\hat{\mathbf{b}}) = \mathbf{C}_{11}\sigma_e^2$$

$$V(\hat{\mathbf{u}} - \mathbf{u}) = C_{22}\sigma_e^2$$

Prediction Error Variance

$$V(\hat{\mathbf{u}}) = \mathbf{A}\sigma_a^2 + \mathbf{C}_{22}\sigma_e^2$$

Prediction Error Variance
Including Drift Variance

Kennedy and Sorensen *Quantitative Genetics*

PEV

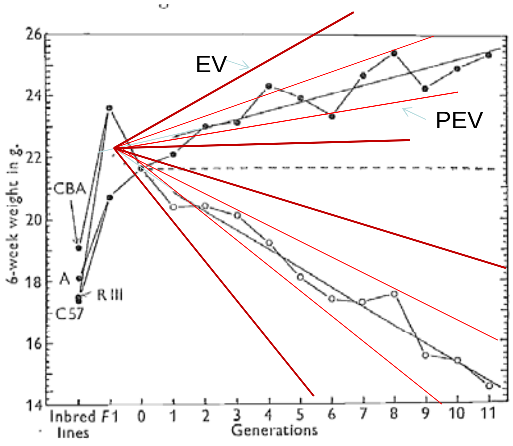
1.12236	0.29509	0.32030	0.65827	0.39093
0.29509	1.14758	0.29509	0.68854	0.68854
0.32030	0.29509	1.12236	0.39093	0.65827
0.65827	0.68854	0.39093	1.2686	0.60026
0.39093	0.68854	0.65827	0.60026	1.2686

EV

2.86014	0.29509	0.32030	1.52716	0.39093
0.29509	2.88536	0.29509	1.5574	1.5574
0.32030	0.29509	2.86014	0.39093	1.52716
1.52716	1.5574	0.39093	3.00643	1.03471
0.39093	1.5574	1.52716	1.03471	3.00643

Selection Experiments and Replication

Falconer,D.S. 1953. Selection for Large and Small Size in Mice. Journal Of Genetics 51:470-501



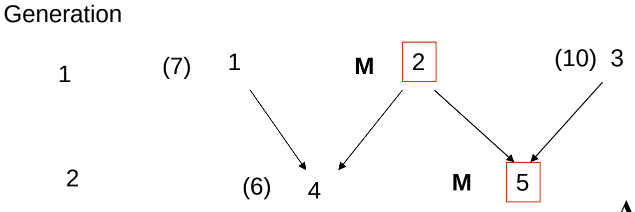
Is there significant asymmetrical response to selection?

Replicated experiment needed to find variation in selection response including random genetic drift

Alternative: no replication find EV:
includes variation in response due to drift.
Assumes additive infinitesimal model

Up is not significantly different from 0 while down is significantly less, thus asymmetry remains even after correcting for genetic drift

Missing Values (Sex Limited Traits)



Generation

1 (7) 1 M $\boxed{2}$ (10) 3

2 (6) 4 M $\boxed{5}$

$$\mathbf{Y} = \begin{bmatrix} 7 \\ 10 \\ 6 \end{bmatrix} \quad \mathbf{e} = \begin{bmatrix} e_1 \\ e_3 \\ e_4 \end{bmatrix} \quad b = [\mu] \quad \mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{2} & 0 \\ 0 & 1 & 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 1 & \frac{1}{4} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{4} & 1 \end{bmatrix}$$

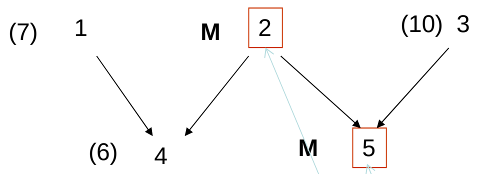
$$\mathbf{X} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \quad Z = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix} \quad \mathbf{U} = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{bmatrix}$$

Assume h^2=.5

R code Example 3

```
Y = matrix( c( 7,
               10,
              6), 3,1)
SigA=2
SigE=2
lam=SigE/SigA

Z = matrix( c(1, 0, 0, 0, 0,
               0, 0, 1, 0, 0,
               0, 0, 0, 1, 0), 3,5, byrow = TRUE)

X = matrix( c( 1,
               1,
               1), 3,1)

A = matrix( c(1, 0, 0, .5, 0,
               0, 1, 0, .5,.5,
               0, 0, 1, 0, .5,
               .5,.5,0, 1,.25,
               0,.5,.5,.25, 1  ), 5,5)
```

```
LHS = rbind( cbind(t(X) %*% X ,  t(X) %*% Z ),
              cbind( t(Z) %*% X  , ( t(Z) %*% Z ) + (lam * solve(A)) ))

RHS = rbind(t(X) %*%Y,
              t(Z) %*% Y)

C = solve(LHS)
BU = C %*% RHS
BU
X1 = matrix( c( 1,
          1,
          1,
          1,
          1),5,1)

yhat=X1*BU[1]+BU[2:6]
yhat
```



$$b = [\hat{\mu}]$$

```
[ 1,] 7.846
 [2,] -0.641
 [3,] -0.435
 [4,] 1.076
 [5,] -0.974
 [6,] 0.320
```

$$\mathbf{U} = \begin{bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \end{bmatrix}$$

$$\hat{\mathbf{Y}} =$$

```
[1,] 7.205
[2,] 7.410
[3,] 8.923
[4,] 6.871
[5,] 8.166
```

Extensions of Model

- Inclusion of Dominance and Epistasis
  - Dominance
    - Dominance effects are the result of interaction of alleles within a locus
    - Dominance relationship matrix needed
    - Reflects the probability that individuals have the same pair of alleles in common at a locus
  - Epistasis
    - Epistatic genetic effects are the result of interactions between alleles at different loci
    - Epistatic relationship matrix needed
    - Reflects the probability that individuals have the same pair of alleles in common at different loci (4 possible pairings of 2 alleles at 2 loci)
  - Useful in crossbreeding programs but generally not useful in pure breeding programs
    - An individual does not pass on dominance or epistatic effects (without inbreeding or cloning), which are a function of both parents
    - Exception is Additive x Additive epistasis is a function of 2 alleles at different loci in the same gamete, but dissipates with recombination and/or segregation

Estimation of Variances Using all Data in a Pedigree

- REML
  - – EM-REML iterative process whereby
    - A value is assumed for additive variance
    - Estimates of breeding values found
    - Additive variance V(A) is estimated as variance of breeding values V(A)=(u'A^{-1}u +stuff)/n
    - The new value of V(A) is substituted into the MME
    - Estimates of breeding values (u) are found
    - The process repeated until convergence
  - – DF-REML work by trial and error finding a value of V(A) that maximize the likelihood

WMM1

Estimation of Effects and Parameters via Iteration and MCMC

Distributions

$$b_i \mid \mathbf{b_{i'}}, \mathbf{a}, \sigma_a^2, \sigma_\varepsilon^2, y \sim N\left(\hat{b}_i, \frac{\sigma_\varepsilon^2}{LHS_{ii}}\right)$$

$$a_i \mid \mathbf{b}, \mathbf{a_{i'}}, \sigma_a^2, \sigma_\varepsilon^2, y \sim N\left(\hat{a}_i, \frac{\sigma_\varepsilon^2}{LHS_{ii}}\right)$$

$$\sigma_a^2 \mid \mathbf{b}, \mathbf{a}, \sigma_\varepsilon^2, y \sim \hat{v}_a \hat{S}_a^2 \chi_{\hat{v}_a}^{-2}$$

$$\sigma_e^2 \mid \mathbf{b}, \mathbf{a}, \sigma_a^2, y \sim \hat{v}_\varepsilon \hat{S}_\varepsilon^2 \chi_{\hat{v}_\varepsilon}^{-2}$$

Estimates

$$\hat{S}_a^2 = \left(\mathbf{a'}\,\mathbf{A}^{-\mathbf{1}}\mathbf{a} + v_a S_a^2\right) / \hat{v}_a$$

$$\hat{v}_a = q + v_a$$

$$\hat{S}_\varepsilon^2 = \left(\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} + v_\varepsilon S_\varepsilon^2\right) / \hat{v}_\varepsilon$$

$$\hat{v}_\varepsilon = N + v_\varepsilon$$

$$\boldsymbol{\varepsilon} = \mathbf{Y} - \mathbf{XB} - \mathbf{Za}$$

1. Solutions to MME are found using iterative approach (Gauss-Seidel)
2. With each Iteration a random amount is added to each solution based on the expected distribution
3. After processing all equations in the MME, new variances are computed and a random amount is added to each solution based on the expected distribution
4. After a burn in period, and many 1000 iterations, the average value of each parameter, with the empirical standard error is the best estimate of the effects and variances

Slide 22

WMM1 Sa is a prior guess about sig(a)
Va is the degree of belief in that prior

Se is prior guess about sig(e)
Ve is degrees of belief in that prior

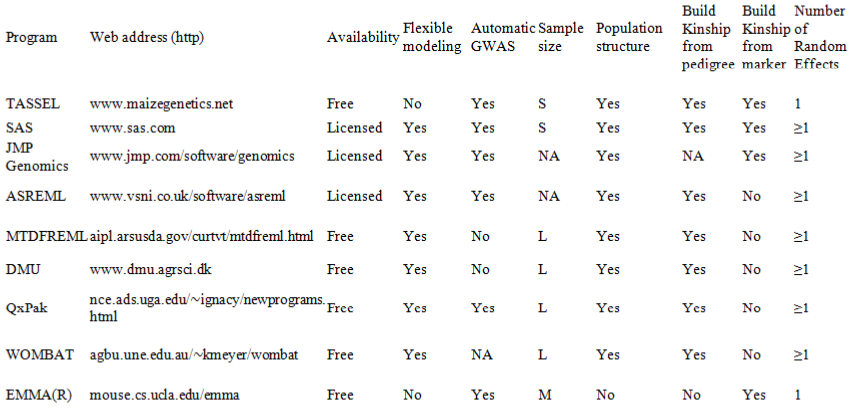
q is number of random effects
N is the number of phenotypes
William Muir, 5/20/2009

Appendix 1

Software packages for estimating EBVs, Variance Components, GWAS and genomic selection

Software engineering the mixed model for genome-wide association studies on large samples

http://bib.oxfordjournals.org/content/10/6/664/T1.expansion.html



Program	Web address (http)	Availability	Flexible modeling	Automatic GWAS	Sample size	Population structure	Build Kinship from pedigree	Build Kinship from marker	Number of Random Effects
TASSEL	www.maizegenetics.net	Free	No	Yes	S	Yes	Yes	Yes	1
SAS	www.sas.com	Licensed	Yes	Yes	S	Yes	Yes	Yes	≥1
JMP Genomics	www.jmp.com/software/genomics	Licensed	Yes	Yes	NA	Yes	NA	Yes	≥1
ASREML	www.vsni.co.uk/software/asreml	Licensed	Yes	Yes	NA	Yes	Yes	No	≥1
MTDFREML	aipl.arsusda.gov/curtvt/mtdfreml.html	Free	Yes	No	L	Yes	Yes	No	≥1
DMU	www.dmu.agrsci.dk	Free	Yes	No	L	Yes	Yes	No	≥1
QxPak	nce.ads.uga.edu/~ignacy/newprograms.html	Free	Yes	Yes	L	Yes	Yes	No	≥1
WOMBAT	agbu.une.edu.au/~kmeyer/wombat	Free	Yes	NA	L	Yes	Yes	No	≥1
EMMA(R)	mouse.cs.ucla.edu/emma	Free	No	Yes	M	No	No	Yes	1

Software
Ignacy Misztal UGA

- Overview
  - http://nce.ads.uga.edu/~ignacy/newprograms.html
- General Documents
  - http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=blupf90.pdf
  - http://nce.ads.uga.edu/wiki/lib/exe/fetch.php?media=remlf90.pdf
- Binaries (UNIX, Windows, Max)
  - http://nce.ads.uga.edu/html/projects/programs/

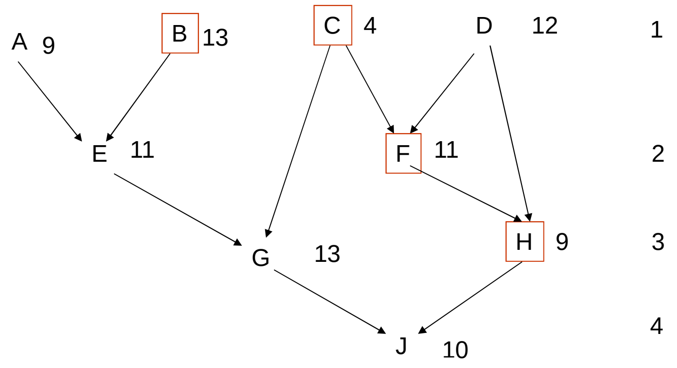
R packages

- QTL mapping
  - ***onemap*** – It is used to generate or rearrange genetic maps
  - ***rqtl*** – performs QTL mapping for bi-parental populations
  - **GAPIT** – most common package for Genome-Wide Association Mapping
- BLUP (Animal Model)
  - ***pedigree*** – Generates A matrix from sparse pedigree
  - ***MCMCglmm*** – Generalized Mixed Models incorporating pedigrees
  - ***pedigreemm*** - Fit mixed-effects models incorporating pedigrees
- Genomic Selection
  - ***rrBLUP*** – classic package to perform ridge regression BLUP and GBLUP
  - ***BGLR*** – whole genome regressions methods of genomic selection
  - ***randomForest*** – Random Forest Regression (non-parametric GS)
  - ***brnn*** – Bayesian Regularized Neural Network (non-parametric GS)
  - ***parallel*** – Allows the use of multiple cores for faster computation

Appendix 2

Problems and Solutions

Problem 1



Find the best estimate of the genetic worth of each animal. Assume a heritability of .5.

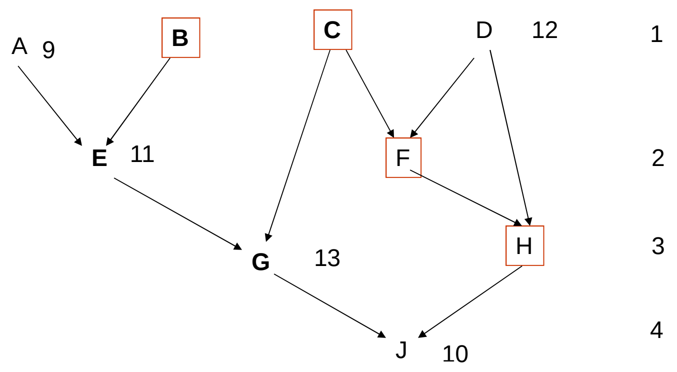
Answer Problem 1

```
proc iml;
start main;
                 A={1     0      0     0     0.5  0    0.25   0      0.125,
                    0     1      0     0     0.5  0    0.25   0      0.125,
                    0     0      1     0     0    0.5  0.5    0.25   0.375,
                    0     0      0     1     0    0.5  0      0.75   0.375,
                    0.5   0.5    0     0     1    0    0.5    0      0.25,
                    0     0      0.5   0.5   0    1    0.25   0.75   0.5,
                    0.25  0.25   0.5   0     0.5  0.25 1      0.125  0.5625,
                    0     0      0.25  0.75  0    0.75 0.125  1.25   0.6875,
                    0.125 0.125  0.375 0.375 0.25 0.5  0.5625 0.6875 1.0625};

y={9,
   13,
   4,
   12,
   11,
   11,
   13,
    9,
   10};

X={1,
   1,
   1,
   1,
   1,
   1,
   1,
   1,
   1};

            AINV=INV(A);
            lam=1;

            Z={1 0 0 0 0 0 0 0 0,
               0 1 0 0 0 0 0 0 0,
               0 0 1 0 0 0 0 0 0,
               0 0 0 1 0 0 0 0 0,
               0 0 0 0 1 0 0 0 0,
               0 0 0 0 0 1 0 0 0,
               0 0 0 0 0 0 1 0 0,
               0 0 0 0 0 0 0 1 0,
               0 0 0 0 0 0 0 0 1};

            LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*Z+AINV#LAM));
            RHS=(X`*Y)//(Z`*Y);
            C=INV(LHS);
            BU=C*RHS;
```

Answer

```
       10.07
       -0.31
       1.689
       -2.28
       0.905
BU=    1.145
       -0.31
       0.564
       -0.19
       0.105
```

Problem 2: Sex Limited Trait



Estimate breeding values for the males.
Assume a heritability of .5.

Answer Problem 2

```
proc iml;       A={1    0      0     0     0.5  0    0.25   0      0.125,
start main;       0     1      0     0     0.5  0    0.25   0      0.125,
                  0     0      1     0     0    0.5  0.5    0.25   0.375,
y={9,             0     0      0     1     0    0.5  0      0.75   0.375,
   12,            0.5   0.5    0     0     1    0    0.5    0      0.25,
   11,            0     0      0.5   0.5   0    1    0.25   0.75   0.5,
   13,            0.25  0.25   0.5   0     0.5  0.25 1      0.125  0.5625,
   10};           0     0      0.25  0.75  0    0.75 0.125  1.25   0.6875,
                  0.125 0.125  0.375 0.375 0.25 0.5  0.5625 0.6875 1.0625};

X={1,
   1,
   1,        AINV=INV(A);
   1,        lam=1;
   1};
             Z={1 0 0 0 0 0 0 0 0,
                0 0 0 1 0 0 0 0 0,
                0 0 0 0 1 0 0 0 0,
                0 0 0 0 0 0 1 0 0,
                0 0 0 0 0 0 0 0 1};

             LHS=((X`*X)||(X`*Z))//((Z`*X)||(Z`*Z+AINV#LAM));
             RHS=(X`*Y)//(Z`*Y);
             C=INV(LHS);
             BU=C*RHS;
```

Answer

```
       11.03
        -0.89
         0.247
         0.338
         0.307
BU=     -0.075
         0.206
         0.587
         0.023
        -0.102
```

Genomic Selection

References

Goddard (2008) Genomic selection: prediction of accuracy and maximization of long term response Genetica DOI 10.1007/s10709-008-9308-0

NejatiJavaremi, A., C. Smith, and J. P. Gibson, 1997 Effect of total allelic relationship on accuracy of evaluation and response to selection. Journal Of Animal Science **75**: 1738-1745.

VanRaden, P. M., 2008 Efficient Methods to Compute Genomic Predictions. Journal Of Dairy Science **91**: 4414-4423

Meuwissen, T. H. E., B. J. Hayes, and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**: 1819-1829

Genomic Selection

- Assumes
  - Dense markers evenly spaced across the genome
  - Assumes markers are in LD with QTL affecting trait(s) of interest
  - Each marker accounts for an equal proportion of genetic variance (infinitesimal model)
  - Genetic Effects are Normally Distributed

Model

$$\mathbf{Y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

$$V(\mathbf{Y}) = V(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) = \mathbf{ZGZ'} + \mathbf{R}$$

$$\mathbf{u}_{n,1} = \mathbf{M}_{n,p}\mathbf{a}_{p,1}$$

M is the marker matrix
a is a vector of SNP effects
Note **Ma** is a vector of summed marker effects

$$V(\mathbf{u}) = E(\mathbf{u}_{n,1}\mathbf{u}_{1,n}') = \mathbf{G}_{n,n}$$

$$\mathbf{G}_{n,n} = \sigma^2_{A*}\mathbf{M}_{n,p}\mathbf{M}_{p,n}' / L$$

Genomic Relationship Matrix (GRM)

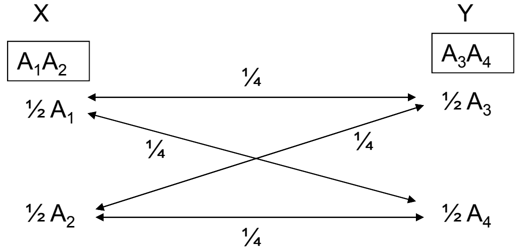
$$\mathbf{R} = \mathbf{I}\sigma^2_e$$

Genomic Relationship Matrix

- Assumes
  - Alike in State (AIS) alleles were at one time a result of a single mutation, thus IBD when traced back in evolutionary time

AIS relationships

TA_k=total allelic relationship at kth locus
TA_k=2x coefficient of relationship(Malecot. 1948)



$$TA_k = 2\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{4}$$

Compute (AIS) relationship matrix (G)

$$TA_k = 2\frac{\sum_{i=1}^{2}\sum_{j=1}^{2} I_{ij}}{4}$$

TA_k=total allelic relationship at kth locus
TA_k=2x coefficient of relationship
(Malecot. 1948)

$$G^*_{xy} = \frac{\sum_{k=1}^{L} TA_k}{L}$$

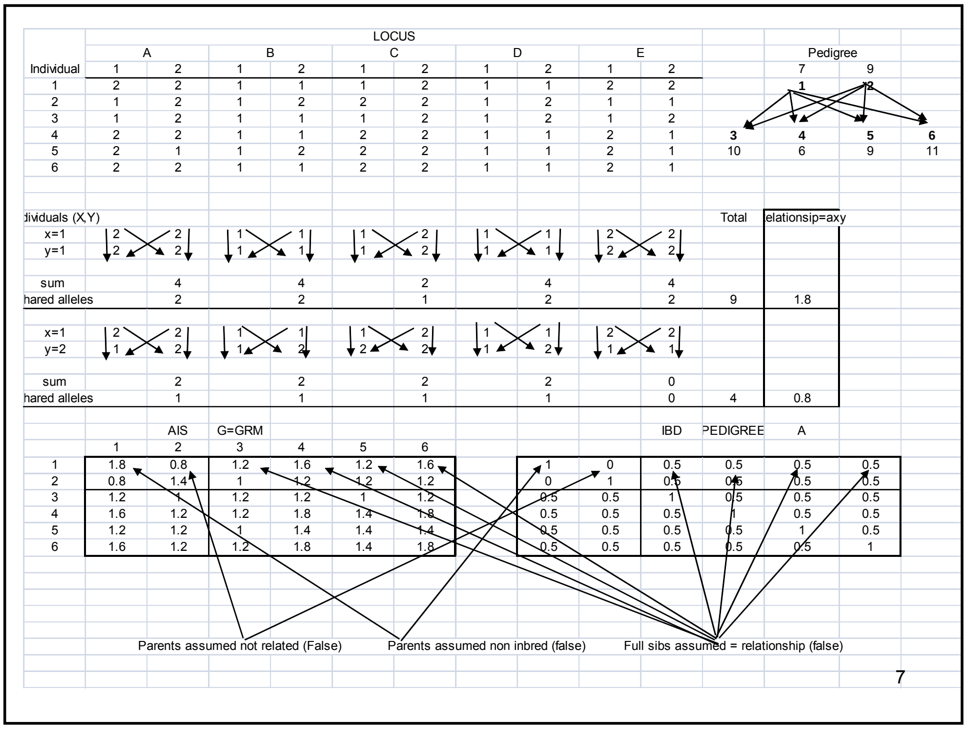
$$\mathbf{G} = \sigma^2_{A*}\mathbf{G}^*$$

$$\sigma^2_{A*}$$

Is the additive genetic variance associated with the markers for the trait

$$\sigma^2_{A*} < \sigma^2_A$$

Note: with low marker density the markers may not capture any genetic variance

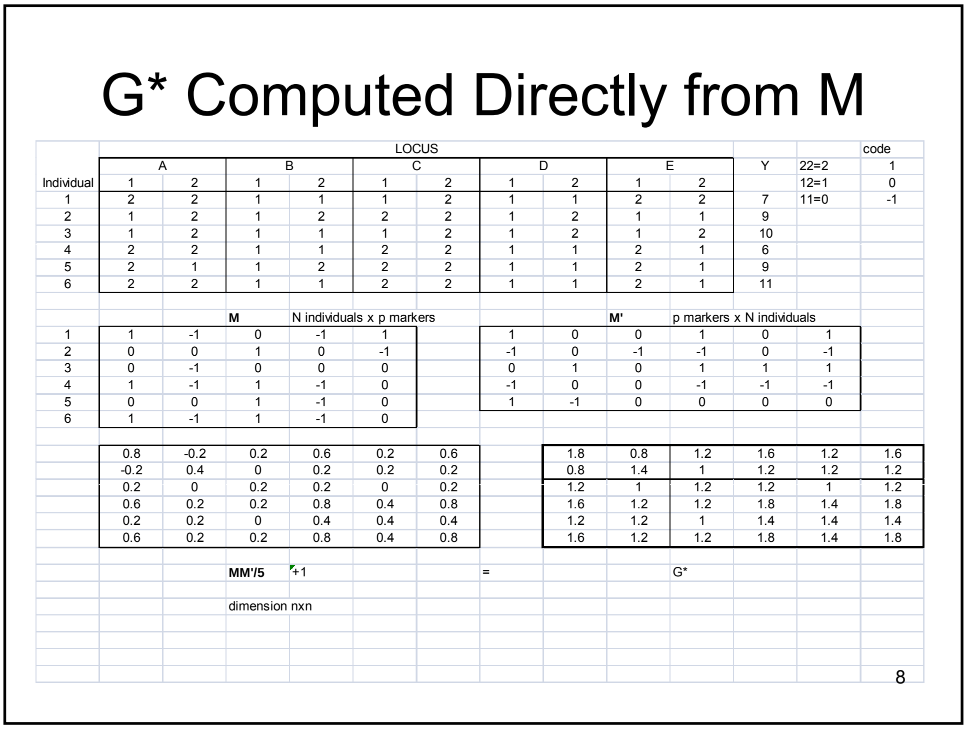


	LOCUS									
	A		B		C		D		E	
Individual	1	2	1	2	1	2	1	2	1	2
1	2	2	1	1	1	2	1	1	2	2
2	1	2	1	2	2	2	1	2	1	1
3	1	2	1	1	1	2	1	2	1	2
4	2	2	1	1	2	2	1	1	2	1
5	2	1	1	2	2	2	1	1	2	1
6	2	2	1	1	2	2	1	1	2	1

dividuals (X,Y)											Total	elationsip=axy
x=1	2	2	1	1	1	2	1	1	2	2		
y=1	2	2	1	1	1	2	1	1	2	2		
sum		4		4		2		4		4		
hared alleles		2		2		1		2		2	9	1.8
x=1	2	2	1	1	1	2	1	1	2	2		
y=2	1	2	1	2	2	2	1	2	1	1		
sum		2		2		2		2		0		
hared alleles		1		1		1		1		0	4	0.8

		AIS	G=GRM							IBD	PEDIGREE	A	
	1	2	3	4	5	6							
1	1.8	0.8	1.2	1.6	1.2	1.6		1	0	0.5	0.5	0.5	0.5
2	0.8	1.4	1	1.2	1.2	1.2		0	1	0.5	0.5	0.5	0.5
3	1.2	1	1.2	1.2	1	1.2		0.5	0.5	1	0.5	0.5	0.5
4	1.6	1.2	1.2	1.8	1.4	1.8		0.5	0.5	0.5	1	0.5	0.5
5	1.2	1.2	1	1.4	1.4	1.4		0.5	0.5	0.5	0.5	1	0.5
6	1.6	1.2	1.2	1.8	1.4	1.8		0.5	0.5	0.5	0.5	0.5	1

Parents assumed not related (False)   Parents assumed non inbred (false)   Full sibs assumed = relationship (false)



G* Computed Directly from M

	LOCUS												code
	A		B		C		D		E		Y	22=2	1
Individual	1	2	1	2	1	2	1	2	1	2		12=1	0
1	2	2	1	1	1	2	1	1	2	2	7	11=0	-1
2	1	2	1	2	2	2	1	2	1	1	9		
3	1	2	1	1	1	2	1	2	1	2	10		
4	2	2	1	1	2	2	1	1	2	1	6		
5	2	1	1	2	2	2	1	1	2	1	9		
6	2	2	1	1	2	2	1	1	2	1	11		

			M	N individuals x p markers				**M'**	p markers x N individuals			
1	1	-1	0	-1	1		1	0	0	1	0	1
2	0	0	1	0	-1		-1	0	-1	-1	0	-1
3	0	-1	0	0	0		0	1	0	1	1	1
4	1	-1	1	-1	0		-1	0	0	-1	-1	-1
5	0	0	1	-1	0		1	-1	0	0	0	0
6	1	-1	1	-1	0							

0.8	-0.2	0.2	0.6	0.2	0.6		1.8	0.8	1.2	1.6	1.2	1.6
-0.2	0.4	0	0.2	0.2	0.2		0.8	1.4	1	1.2	1.2	1.2
0.2	0	0.2	0.2	0	0.2		1.2	1	1.2	1.2	1	1.2
0.6	0.2	0.2	0.8	0.4	0.8		1.6	1.2	1.2	1.8	1.4	1.8
0.2	0.2	0	0.4	0.4	0.4		1.2	1.2	1	1.4	1.4	1.4
0.6	0.2	0.2	0.8	0.4	0.8		1.6	1.2	1.2	1.8	1.4	1.8

MM'/5 +1 = G*

dimension nxn

Mixed Model Equations

$$\mathbf{G} = \mathbf{MM}'/L$$

M (n individuals x p markers)
M(n,p)M'(p,n)
MM'(n,n)

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \frac{1}{\sigma^2_{A*}}\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

n effects to estimate

Simplifications If $\mathbf{R} = \mathbf{I}\sigma^2_e$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma^2_e}{\sigma^2_{A*}}\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

G may not have an inverse

- G may not be positive definite
  - The G matrix is an estimate of the true genetic variance co-variance matrix
  - Genotyping errors and possible inclusion of individuals without all parents creates inconsistency
  - Solution: Add small constant to diagonal elements (ridge)

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \frac{\sigma^2_e}{\sigma^2_{A*}}(\mathbf{G} + \boldsymbol{\lambda})^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

$$\boldsymbol{\lambda} = c\mathbf{I}$$

Example

$$\mathbf{Y}=\begin{bmatrix}7\\9\\10\\6\\9\\11\end{bmatrix}\quad \mathbf{X}=\begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}\quad b=\left[\mu_0\right]\quad Z=\begin{bmatrix}1&0&0&0&0&0\\0&1&0&0&0&0\\0&0&1&0&0&0\\0&0&0&1&0&0\\0&0&0&0&1&0\\0&0&0&0&0&1\end{bmatrix}\quad \mathbf{u}=\begin{bmatrix}u_1\\u_2\\u_3\\u_4\\u_5\\u_6\end{bmatrix}$$

$$\mathbf{M}=\begin{bmatrix}1&-1&0&-1&1\\0&0&1&0&-1\\0&-1&0&0&0\\1&-1&1&-1&0\\0&0&1&-1&0\\1&-1&1&-1&0\end{bmatrix}\quad \mathbf{MM'}/L=\begin{bmatrix}.8&-.2&.2&.6&.2&.6\\-.2&.4&0&.2&.2&.2\\.2&0&.2&.2&0&.2\\.6&.2&.2&.8&.4&.8\\.2&.2&0&.4&.4&.4\\.6&.2&.2&.8&.4&.8\end{bmatrix}$$

$\sigma_A^2=10$

$\sigma_{A*}^2=5$

$\sigma_\varepsilon^2=20$

Note, only ½ the additive genetic variance was captured by the markers (missing heritability issue)

R code Example 4

```
NL=5
SigA=5
SigE=20
Lam=SigE/SigA

Y = matrix( c(          7,
                        9,
                        10,
                        6,
                        9,
                        11), 6,1)

Z = matrix( c(          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),6,6)
```

```
X = matrix( c(     1,
                   1,
                   1,
                   1,
                   1,
                   1  ), 6,1)

M = matrix( c( 1,-1,0,-1,1,
               0,0,1,0,-1,
               0,-1,0,0,0,
               1,-1,1,-1,0,
               0,0,1,-1,0,
               1,-1,1,-1,0),6,5,  byrow = TRUE)

G=(1/NL)*M%*%t(M)
```

Check G for inverse
GI=solve(G)

R code

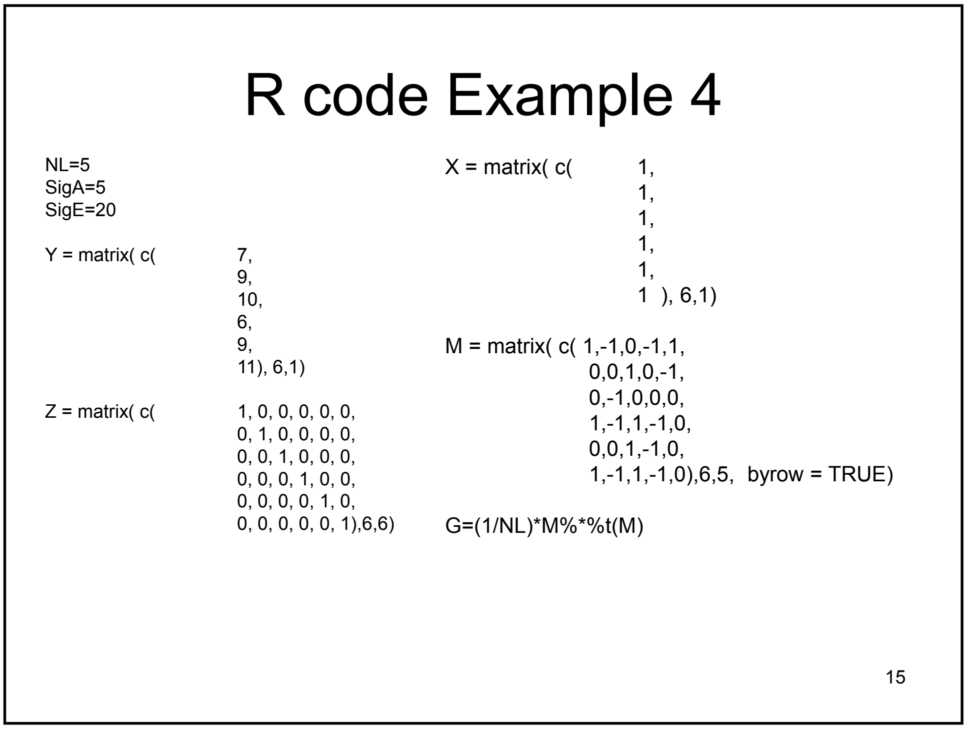
```
r=.00001
I = matrix( c(     1, 0, 0, 0, 0, 0,                ridge=r*I
                   0, 1, 0, 0, 0, 0,                G1=G+ridge
                   0, 0, 1, 0, 0, 0,                INVG=solve(G1)
                   0, 0, 0, 1, 0, 0,
                   0, 0, 0, 0, 1, 0,
                   0, 0, 0, 0, 0, 1),6,6)


LHS = rbind( cbind(t(X) %*% X                  ,  t(X) %*%Z ),
             cbind(t(Z) %*% X                  ,  t(Z)%*%Z +Lam*INVG))


RHS = rbind(t(X)%*%Y,
              t(Z)%*%Y)               [1,] 8.76
                                      [2,] -0.25
C = solve(LHS)                        [3,] 0.09
                                      [4,] -0.02        gEBV
BU = C %*% RHS                        [5,] -0.16
                                      [6,] -0.05
BU                                    [7,] -0.16
```

G may not be positive definite (2nd solution)

- multiply both sides of the second equation by $\mathbf{G}\sigma^2_{A*}$

$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}+\frac{\sigma_e^2}{\sigma_{A*}^2}\mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \mathbf{Z}'\mathbf{Y} \end{bmatrix}$$

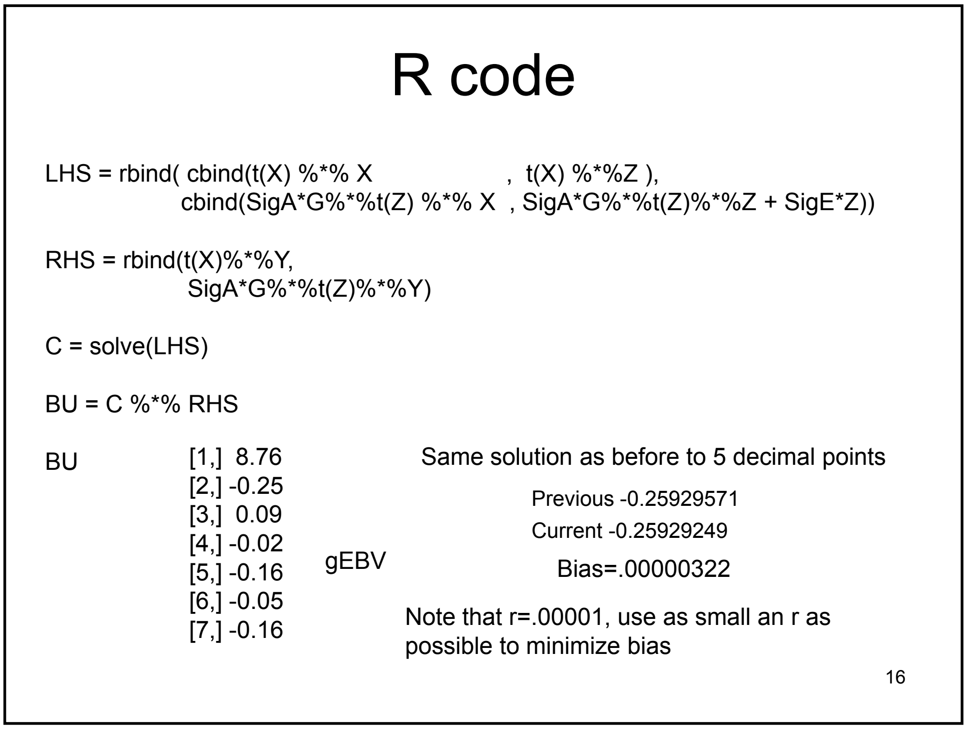
$$\begin{bmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \sigma^2_{A*}\mathbf{G}\mathbf{Z}'\mathbf{X} & \sigma^2_{A*}\mathbf{G}\mathbf{Z}'\mathbf{Z}+\sigma_e^2\mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{Y} \\ \sigma^2_{A*}\mathbf{G}\mathbf{Z}'\mathbf{Y} \end{bmatrix}$$



R code Example 4

```
NL=5
SigA=5
SigE=20

Y = matrix( c(       7,
                     9,
                     10,
                     6,
                     9,
                     11), 6,1)

Z = matrix( c(       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),6,6)
```

```
X = matrix( c(          1,
                        1,
                        1,
                        1,
                        1,
                        1  ), 6,1)

M = matrix( c( 1,-1,0,-1,1,
               0,0,1,0,-1,
               0,-1,0,0,0,
               1,-1,1,-1,0,
               0,0,1,-1,0,
               1,-1,1,-1,0),6,5,  byrow = TRUE)

G=(1/NL)*M%*%t(M)
```



R code

```
LHS = rbind( cbind(t(X) %*% X                  ,  t(X) %*%Z ),
               cbind(SigA*G%*%t(Z) %*% X  , SigA*G%*%t(Z)%*%Z + SigE*Z))

RHS = rbind(t(X)%*%Y,
               SigA*G%*%t(Z)%*%Y)

C = solve(LHS)

BU = C %*% RHS
```

```
BU          [1,]  8.76
            [2,] -0.25
            [3,]  0.09
            [4,] -0.02
            [5,] -0.16     gEBV
            [6,] -0.05
            [7,] -0.16
```

Same solution as before to 5 decimal points

Previous -0.25929571

Current -0.25929249

Bias=.00000322

Note that r=.00001, use as small an r as possible to minimize bias

Equivalent Model

Estimation of Marker effects

$$\mathbf{Y}_{n,1} = \mathbf{Xb} + \mathbf{M}_{n,p}\mathbf{'a}_{p,1} + \mathbf{e}$$

$$V(\mathbf{Y}) = V(\mathbf{Xb} + \mathbf{M'a} + \mathbf{e})$$

$$V(\mathbf{Y}) = \mathbf{M'}V(\mathbf{a})\mathbf{M} + \mathbf{R}$$

$$V(\mathbf{a}) = E(\mathbf{a}_{p,1}\mathbf{a}_{1,p}') = \sigma_g^2\mathbf{I}_{p,p}$$

$$\mathbf{R} = \mathbf{I}\sigma_e^2$$

Equivalent Model

Estimation of Marker effects

$$\begin{bmatrix} \mathbf{X'}_{1,N}\,\mathbf{X}_{N,1} & \mathbf{X}_{1,N}\mathbf{'M}_{N,p} \\ \mathbf{M'}_{p,N}\,\mathbf{X}_{N,1} & \mathbf{M}'_{p,N}\mathbf{M}_{N,p} + \frac{\sigma_e^2}{\sigma_g^2}\mathbf{I} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_{1,1} \\ \mathbf{g}_{p,1} \end{bmatrix} = \begin{bmatrix} \mathbf{X'}_{1,N}\,\mathbf{Y}_{N,1} \\ \mathbf{M'}_{p,N}\,\mathbf{Y}_{N,1} \end{bmatrix}$$

Assumption depends on method
1) (GBLUP, ssGBLUP) Genetic variance associated with each marker is equal $\sigma_g^2 = \left(\frac{\sigma_{A^*}^2}{L}\right)$
2) (Bayes A) sampled from a t distribution
3) (Bayes B and Bayes C π) from a mixture of distributions (null and t)

$$\hat{u}_i = GEBV_i = \mathbf{Mg} = \sum_j M_{ij}\hat{g}_j$$

Example 5 SNP BLUP

```
NL=5
SigA=5
Sigg=SigA/NL
SigE=20

y = matrix( c(          7,
                        9,
                        10,
                        6,
                        9,
                        11), 6,1)

I = matrix( c(          1, 0, 0, 0, 0,
                        0, 1, 0, 0, 0,
                        0, 0, 1, 0, 0,
                        0, 0, 0, 1, 0,
                        0, 0, 0, 0, 1),5,5)
```

```
X = matrix( c(      1,
                    1,
                    1,
                    1,
                    1,
                    1  ), 6,1)

M = matrix( c( 1,-1,0,-1,1,
               0,0,1,0,-1,
               0,-1,0,0,0,
               1,-1,1,-1,0,
               0,0,1,-1,0,
               1,-1,1,-1,0),6,5,  byrow = TRUE)
```

GWAS

```
LHS = rbind( cbind(t(X) %*% X ,  t(X) %*%M ),
             cbind( t(M) %*% X  , t(M)%*%M + (SigE/Sigg)*I))

RHS = rbind(t(X)%*% y,
             t(M)%*%y)

C = solve(LHS)

Bg = C %*% RHS
                          [1,]  8.76
Bg                        [2,] -0.08
                          [3,]  0.02
                          [4,]  0.01       Marker effects
                          [5,]  0.07
                          [6,] -0.08
```

gEBV

```
g=Bg[2:6]
U=M%*%g
U
```

```
 [1,] -0.25
[2,]  0.09
[3,] -0.02
[4,] -0.16
[5,] -0.05
[6,] -0.16
```

Compare to GBLUP

Single Step ssGBLUP

- Merge G matrix into regular A matrix
  - Aguilar, I., I. Misztal, D. L. Johnson, A. Legarra, S. Tsuruta *et al.* 2010 Hot topic: A unified approach to utilize phenotypic, full pedigree, and genomic information for genetic evaluation of Holstein final score. Journal Of Dairy Science **93**: 743-752.
  - Corrects for multi-trait selection bias
    - Vitezica, Z. G., I. Aguilar, I. Misztal, and A. Legarra, 2011 Bias in genomic predictions for populations under selection. Genetics Research **93**: 357-366.
  - Uses all information
    - Phenotypes of animals without genotypes
      - J. Anim Sci. 2011. 89:23-28. doi:10.2527/jas.2010-3071
- Software
  - http://nce.ads.uga.edu/~ignacy/genomic-blupf90/
  - http://snp.toulouse.inra.fr/~alegarra/

ssGBLUP

GBLUP only animals genotyped included in analysis

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

ssGBLUP all animals with phenotypes included

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{H}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

$$\mathrm{H}^{-1}=\mathrm{A}^{-1}+\begin{bmatrix} 0 & 0 \\ 0 & \mathrm{G}^{-1}\text{-}\mathrm{A}_{22}^{-1} \end{bmatrix}$$

g=Z(ZZ')$^{-1}$u

GWAS

GBLUP

SNP effects using only genotyped individuals

$$\begin{bmatrix} \mathbf{X'}_{1,N}\ \mathbf{X}_{N,1} & \mathbf{X}_{1,N}\mathbf{'M}_{N,p} \\ \mathbf{M'}_{p,N}\ \mathbf{X'}_{N,1} & \mathbf{M}'_{p,N}\mathbf{M}_{N,p}+\frac{\sigma_e^2}{\sigma_g^2}\mathbf{I} \end{bmatrix}\begin{bmatrix} \boldsymbol{\beta}_{1,1} \\ \mathbf{g}_{p,1} \end{bmatrix}=\begin{bmatrix} \mathbf{X'}_{1,N}\ \mathbf{Y}_{N,1} \\ \mathbf{M'}_{p,N}\ \mathbf{Y}_{N,1} \end{bmatrix}$$

ssGBLUP

SNP effects using all information

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{H}^{-1} \end{bmatrix}\begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix}=\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

$$g=\frac{\sigma_u^2}{\sigma_a^2}\mathbf{DZ'H^{-1}u}$$

Wang,H, I. Misztal, I. Aguilar, A. Legarra, and W. M. Muir Genome-wide association mapping including phenotypes from relatives without genotypes. 2012. Genetics Research 94:73-83

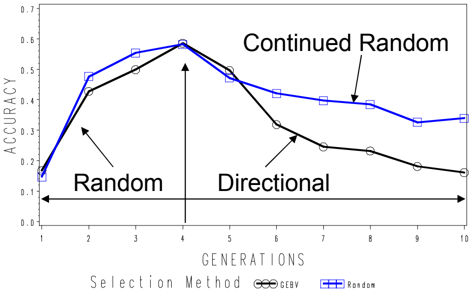
Issues

Genomic Selection and GWAS

- Admixture (Walsh Lectures)
  - Major problem
  - False Positives
  - Spurious Correlations
  - Correlation does not mean Causation
  - Partial Solution
    - Use of Igenstrat to correct for structure
    - Use of Structure to correct for structure
    - Scaling of G to combine all populations and cross in common relationship matrix
- Pedigree errors
  - More costly in terms of accuracy with G than A matrix
- Cost
  - Use dense SNP genotyping all breeders (parents)
  - Low Density on all candidates
    - Reducing the number of markers down to those that are most predictive
    - Going from 60,000 SNP to 384 SNP for genotyping
    - GWAS SNP selection
- Selection (See next slide)
  - Allele frequencies change
  - Older data becomes a liability

Effect of Random vs. Directional Selection on Accuracy

Starting in HWE

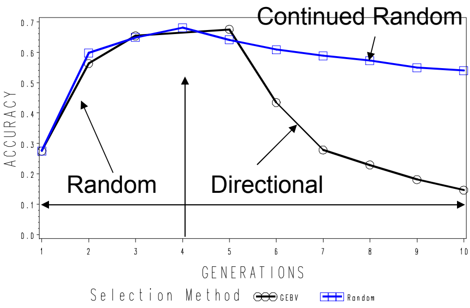


Continued Random

Random

Directional

Starting in MDE



Continued Random

Random

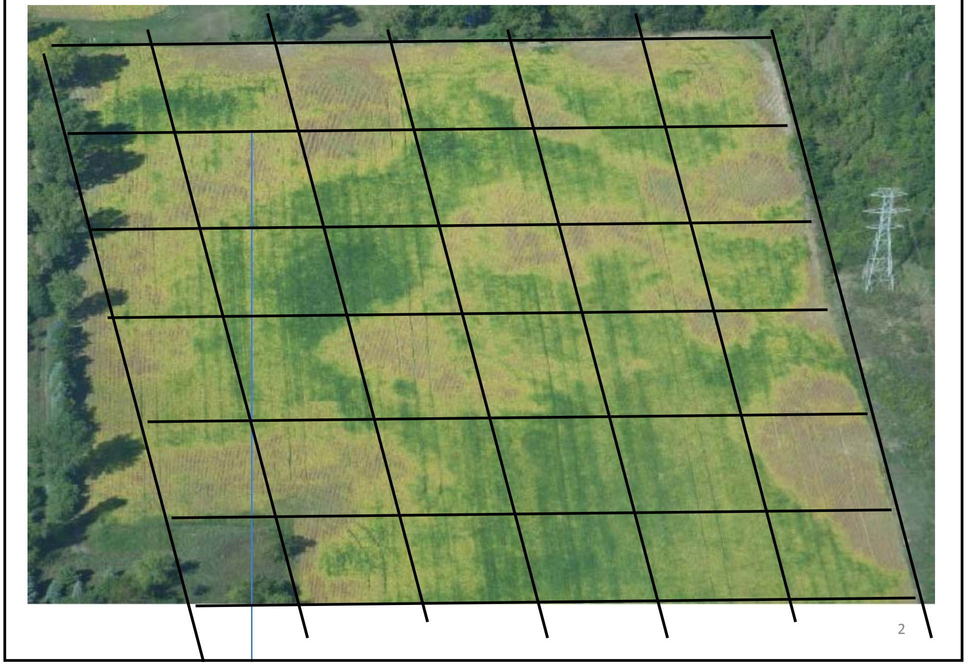
Directional

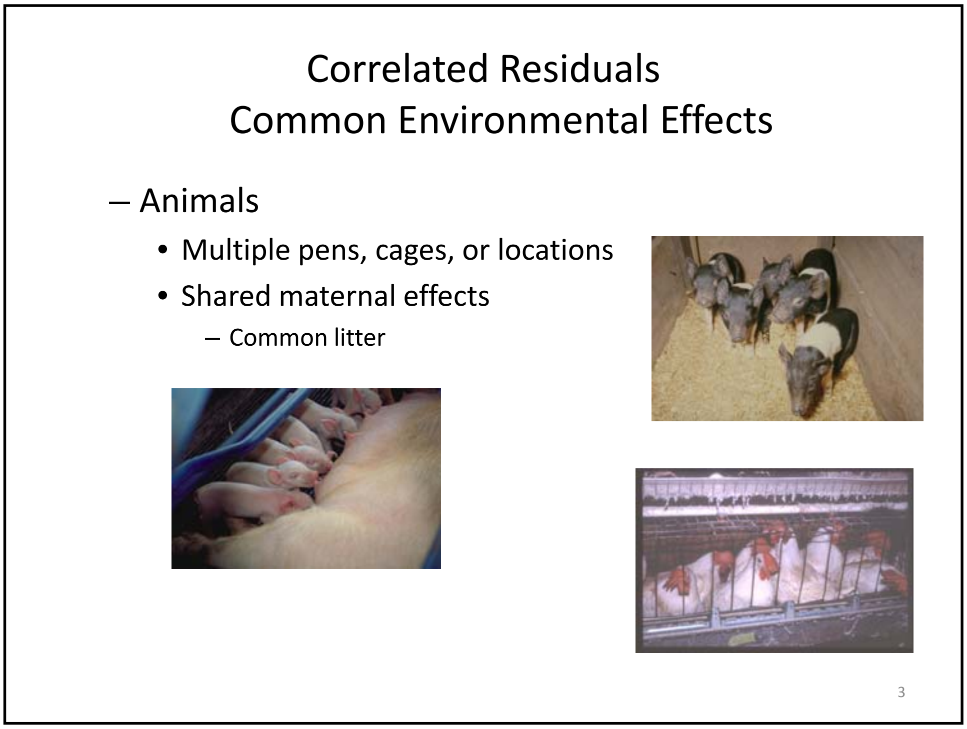
h^2=.1 N=256, Ne=32, 100/100 Marker/QTL loci distributed on 100cM.

(average over 60 replicates, SEM=.02).

Correlated Residuals
Common Environmental Effects

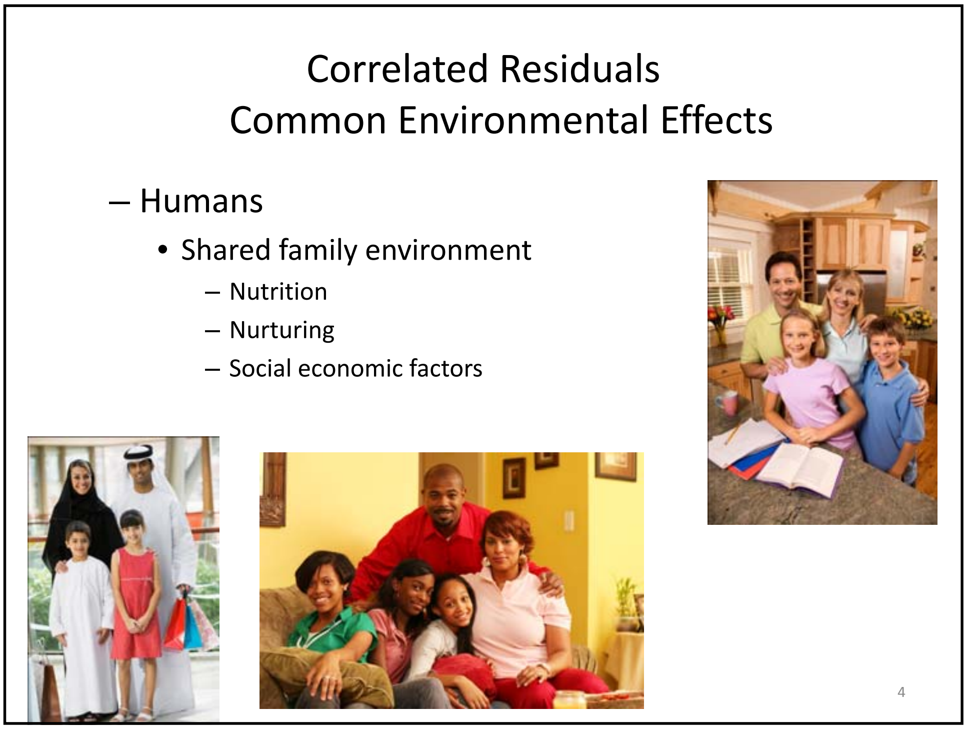
- Environmental effects common within a group partial between groups
  - Agronomy
    - Plots in fields





Correlated Residuals
Common Environmental Effects

- Animals
  - Multiple pens, cages, or locations
  - Shared maternal effects
    - Common litter



Correlated Residuals
Common Environmental Effects

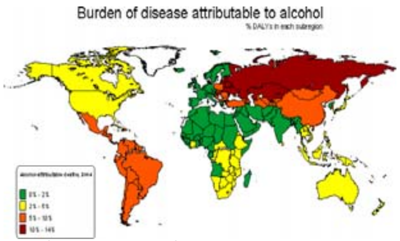
- Humans
  - Shared family environment
    - Nutrition
    - Nurturing
    - Social economic factors

Common Environmental Effects

- In humans, confounds genetic effects with social economic factors
- In plant or animal breeding, reduces response to selection
  - Common environmental effects are included with the phenotype
  - Errors in selection decisions

Alcoholism

- Is this disease the result of nature or nurture?
  - The most accurate predictor of alcoholism is
    - Parents drinking habits
    - Ethnicity


Burden of disease attributable to alcohol
% DALY's in each subregion
0% - 2%
2% - 5%
5% - 10%
10% - 14%

  - Is drinking behavior learned (Nurture)?
  - Is it inherited (Nature)?

How to separate Nature from Nurture

- Nurture imposes a correlated environment
- Nature imposes shared IBD alleles

Solution

- Experimental design
  - Randomized Complete Block (RCB)
    - Block =common environment effect
    - all treatments in all blocks
    - Best design
  - Not possible with human (no randomization) and most plant and animal breeding programs (not practical)
- Breeders in the past
  - Performed within and between family selection
  - Tried to adjust for herd/Y/S as fixed effects then solved for breeding values
  - Problem: best genetics confounded with herd (adjusting for fixed removed some genetic effects)
- Mixed models
  - Empirical Bayesian approach to estimate and adjust for the effect
    - First use of mixed models
    - recovery of inter-block information
    - Yates, 1939; Cox, 1958

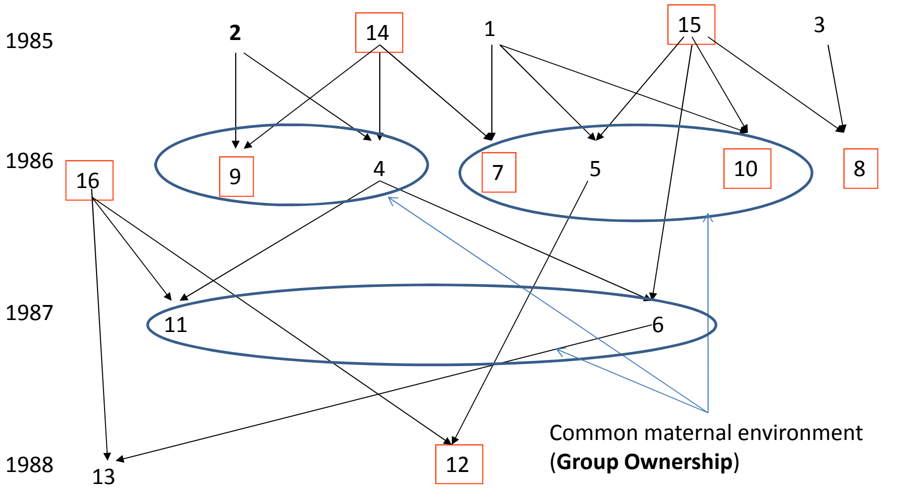
Mixed Model Solution

- Needs
  - Phenotypes
    - The more confounded the data the more data that is needed to get clear results
  - Group Ownership
    - Households
  - Relationships
    - Pedigree
      - Or
    - Genotypes to create Genomic Relationship Matrix (GRM)
  - Variances of Random Effects (Given or estimated from the data)

Example 1

- 10 calves, 5 male and 5 female, from 3 sires and 6 dams, were sampled over 3 years and weaning weight recorded. Some dams were used more than once.
- Remove the fixed effects of year and sex and the random effect of common maternal environment. Then estimate the breeding values of all animals for weaning weight.

Pedigree (**Relationships**)



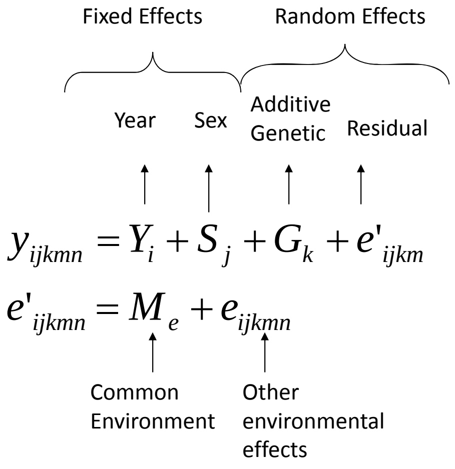
Common maternal environment
(**Group Ownership**)

Phenotypes

Schaeffer Table 8.7

Animal	Sire	Dam	Year	Sex	Wean Wt
7	14	1	86	M	400
4	14	2	86	F	380
8	15	3	86	M	410
5	15	1	87	F	350
9	14	2	87	M	420
6	15	4	87	F	360
10	15	1	88	M	390
11	16	4	88	F	390
12	16	5	88	M	430
13	16	6	88	F	370

Model solution 1



$$y_{ijkmn} = Y_i + S_j + G_k + e'_{ijkm}$$

$$e'_{ijkmn} = M_e + e_{ijkmn}$$

Solution 1
Model Variance-covariance Structure

$$\mathrm{V}\begin{bmatrix} \mathbf{G} \\ \mathbf{e'} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_G^2 & 0 \\ 0 & \mathbf{R}\sigma_{e'}^2 \end{bmatrix}$$

$$\sigma_{e'}^2 = \sigma_m^2 + \sigma_e^2$$

Correlated Residuals

Animals in a common group (pen, herd, or **mother**) share a common environmental effect. Let ρ be the correlation between residuals due to shared environment.

A covariance within groups is reflected in a between group variance
Principle for estimating heritability via ANOVA (between and within sire variances)

ρ is the intra-class environmental correlation

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}$$

$$\mathbf{R} = \left(\sigma_b^2 + \sigma_e^2\right)\begin{bmatrix} 1 & \rho & \rho & \cdots & 0 & 0 & 0 \\ \rho & 1 & \rho & \cdots & 0 & 0 & 0 \\ \rho & \rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \rho & \rho \\ 0 & 0 & 0 & \cdots & \rho & 1 & \rho \\ 0 & 0 & 0 & \cdots & \rho & \rho & 1 \end{bmatrix} = \begin{bmatrix} \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 & \cdots & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \cdots & 0 & 0 & 0 \\ \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_b^2+\sigma_e^2 & \sigma_b^2 & \sigma_b^2 \\ 0 & 0 & 0 & \cdots & \sigma_b^2 & \sigma_b^2+\sigma_e^2 & \sigma_b^2 \\ 0 & 0 & 0 & \cdots & \sigma_b^2 & \sigma_b^2 & \sigma_b^2+\sigma_e^2 \end{bmatrix}$$

Variances of Random Effects

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

$$\sigma_G^2 = 2000$$

$$\sigma_{m_e}^2 = 500$$

$$\sigma_e^2 = 6500$$

$$\rho = \frac{500}{500 + 6500} = .0714$$

The shared environmental effect is small, but real

Environmental Correlation Matrix (R)

Shared Maternal
Environment

An	Sire	Dam
7	14	1
4	14	2
8	15	3
5	15	1
9	14	2
6	15	4
10	15	1
11	16	4
12	16	5
13	16	6

animals 7, 5, 10 shared mother 1

$\rho = .0714$ $\sigma_e^2 = 6500 + 500$

$$R = \begin{array}{c|cccccccccc} & 7 & 4 & 8 & 5 & 9 & 6 & 10 & 11 & 12 & 13 \\ \hline 7 & 1 & 0 & 0 & \rho & 0 & 0 & \rho & 0 & 0 & 0 \\ 4 & 0 & 1 & 0 & 0 & \rho & 0 & 0 & 0 & 0 & 0 \\ 8 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 5 & \rho & 0 & 0 & 1 & 0 & 0 & \rho & 0 & 0 & 0 \\ 9 & 0 & \rho & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 6 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \rho & 0 & 0 \\ 10 & \rho & 0 & 0 & \rho & 0 & 0 & 1 & 0 & 0 & 0 \\ 11 & 0 & 0 & 0 & 0 & 0 & \rho & 0 & 1 & 0 & 0 \\ 12 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 13 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \sigma_e^2$$

MME Residual Correlation Structure

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

$$\begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{b} \\ \mathbf{u} \end{bmatrix} = \begin{bmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{Y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Y} \end{bmatrix}$$

All animals have a genetic effect but only animals with a record contribute to the phenotype

Note that these animals did not have records and are missing

An	Sire	Dam
7	14	1
4	14	2
8	15	3
5	15	1
9	14	2
6	15	4
10	15	1
11	16	4
12	16	5
13	16	6

$$\mathbf{Z} = \begin{array}{cccccccccccccccc} 14 & 1 & 2 & 15 & 3 & 16 & 7 & 4 & 8 & 5 & 9 & 6 & 10 & 11 & 12 & 13 \\ \end{array}$$

$$\mathbf{Z} = \begin{bmatrix} 0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0 \\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1 \end{bmatrix}$$

Year sex
86 87 88 m

$$\mathbf{X} = \begin{bmatrix} 1&0&0&1 \\ 1&0&0&0 \\ 1&0&0&1 \\ 0&1&0&0 \\ 0&1&0&1 \\ 0&1&0&0 \\ 0&0&1&1 \\ 0&0&1&0 \\ 0&0&1&1 \\ 0&0&1&0 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \quad \begin{matrix} \left.\begin{matrix} b_1 \\ b_2 \\ b_3 \end{matrix}\right\} \text{Herd Year} \\ b_4 \leftarrow \text{Sex} \end{matrix}$$

$$\mathbf{Y} = \begin{bmatrix} 400 \\ 380 \\ 410 \\ 350 \\ 420 \\ 360 \\ 390 \\ 390 \\ 430 \\ 370 \end{bmatrix}$$

R code Example 6

```
Sig_m=500
Sig_e=6500
Sig_g=2000
m=Sig_m/(Sig_m+Sig_e)

Y=matrix(c(400,
           380,
           410,
           350,
           420,
           360,
           390,
           390,
           430,
           370),10,1);

 X=matrix(c(1, 0, 0, 1,
            1, 0, 0, 0,
            1, 0, 0, 1,
            0, 1, 0, 0,
            0, 1, 0, 1,
            0, 1, 0, 0,
            0, 0, 1, 1,
            0, 0, 1, 0,
            0, 0, 1, 1,
           0, 0, 1, 0),10,4, byrow = TRUE )
```

```
A=matrix(c(
1, 0, 0, 0, 0, 0, .5, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, 0, 0, .5, 0, .25 0,
 0, 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 0, 0, 1, 0, 0, 0, 0, .5, .5, 0, .5, .5, 0, .25 .25,
 0, 0, 0, 0, 1, 0, 0, 0, .5, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, .5, .5, .5,
 .5, .5, 0, 0, 0, 0, 1, .25 0, .25 .25 .125, .25 .125, .125, 0.0625,
 .5, 0, .5, 0, 0, 0, .25 1, 0, 0, .5, .5, 0, .5, 0, .25,
  0, 0, 0, .5, .5, 0, 0, 0, 1, .25 0, .25 .25 0, .125, .125,
  0, .5, 0, .5, 0, 0, .25 0, .25 1, 0, .25 .5, 0, .5, .125,
 .5, 0, .5, 0, 0, 0, .25 .5, 0, 0, 1, .25 0, .25 0, .125,
 .25 0, .25 .5, 0, 0, .125, .5, .25 .25 .25 1, .25 .25 .125, .5,
   0, .5, 0, .5, 0, 0, .25 0, .25 .5, 0, .25 1, 0, .25 .125,
 .25 0, .25 0, 0, .5, .125, .5, 0, 0, .25 .25 0, 1, .25 0.3750,
   0, .25 0, .25 0, .5, .125, 0, .125, .5, 0, .125, .25 .25 1, 0.3125,
  .125, 0, .125, .25 0, .5, 0.0625, .25 .125, .125, .125, .5, .125, 0.375, 0.3125, 1.0
),16,16)

Ainv=solve(A)
```

```
Z=  matrix(c
(0,  0, 0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1),
10,16,byrow=TRUE);

R=matrix(c(1,  0,   0,   m,     0,     0,     m,     0,      0,     0,
           0,  1,   0,   0,     m,     0,     0,     0,      0,     0,
           0,  0,   1,   0,     0,     0,     0,     0,      0,     0,
           m, 0,   0,    1,     0,     0,     m,     0,      0,     0,
           0, m,   0,    0,     1,     0,     0,     0,      0,     0,
           0,  0,   0,   0,     0,     1,     0,     m,      0,     0,
           m, 0,   0,   m,     0,     0,     1,     0,      0,     0,
           0,  0,   0,   0,     0,     m,     0,     1,      0,     0,
           0,  0,   0,   0,     0,     0,     0,     0,      1,     0,
           0,  0,   0,   0,     0,     0,     0,     0,      0,    1),10,10)
```

```
Sig_EM=Sig_e+Sig_m

R=R*Sig_EM

RINV=solve(R)

LHS=rbind(
cbind( t(X)%*%RINV%*%X, t(X) %*%RINV%*%Z) ,
cbind( t(Z)%*%RINV%*%X, t(Z) %*%RINV%*%Z+Ainv*(1/Sig_g)) )

RHS=rbind(
t(X)%*%RINV%*%Y,
t(Z)%*%RINV%*%Y)

C=solve(LHS)
BU=C %*% RHS
BU
```

369.87422		Herd Year
363.57807		
375.03977		
40.764716		Sex
1.8135021	14	BV
-3.805516	1	
2.7837732	2	
-3.560112	15	
0.1342493	3	
2.6341034	16	
-1.966278	7	
3.4648406	4	
-1.578682	8	
-3.883955	5	
3.9162079	9	
-0.906753	6	
-6.316917	10	
4.5689512	11	
1.227629	12	
0.1257446	13	

Solution 2: model the data structure

Add another random effect due to shared environment (mother, cage, herd)

$$y_{ijkmn} = \underbrace{\underset{\uparrow\atop\text{Year}}{Y_i} + \underset{\uparrow\atop\text{Sex}}{S_j}}_{\text{Fixed Effects}} + \underbrace{\underset{\nwarrow\atop\text{Additive Genetic}}{G_k} + \underset{\nwarrow\atop\text{Common Group Environment}}{M_e} + \underset{\nwarrow\atop\text{Random error}}{e_{ijkmn}}}_{\text{Random Effects}}$$

MME Correlated Residuals

$$\mathbf{y} = \mathbf{Xb} + \underset{\nearrow\atop\text{Genetic effect}}{\mathbf{Z_1}}\mathbf{G} + \underset{\nwarrow\atop\text{Shared Group Environmental effect}}{\mathbf{Z_2}}\mathbf{m} + \mathbf{e}$$

$$V\begin{bmatrix}\mathbf{G}\\ \mathbf{m}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_G^2 & 0 & 0\\ 0 & \mathbf{I}\sigma_{m_e}^2 & 0\\ 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix}$$

Year sex
86 87 88 m

$$\mathbf{X}=\begin{bmatrix}1&0&0&1\\1&0&0&0\\1&0&0&1\\0&1&0&0\\0&1&0&1\\0&1&0&0\\0&0&1&1\\0&0&1&0\\0&0&1&1\\0&0&1&0\end{bmatrix}$$

$$\mathbf{B}=\begin{bmatrix}b_1\\b_2\\b_3\\b_4\end{bmatrix}$$

Herd
Year

Sex

$$\mathbf{Y}=\begin{bmatrix}400\\380\\410\\350\\420\\360\\390\\390\\430\\370\end{bmatrix}$$

Animal Direct Genetic Effect

Note that sires and dams did not have records and are missing

An	Sire	Dam
7	14	1
4	14	2
8	15	3
5	15	1
9	14	2
6	15	4
10	15	1
11	16	4
12	16	5
13	16	6

$$\mathbf{Z}_1=\begin{array}{cccccccccccccccc}14&1&2&15&3&16&7&4&8&5&9&6&10&11&12&13\\ \hline 0&0&0&0&0&0&1&0&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&1&0&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&1&0&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&1&0&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&1&0&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&1&0&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&1&0&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&1&0&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&1&0\\0&0&0&0&0&0&0&0&0&0&0&0&0&0&0&1\end{array}$$

Common Maternal Environmental effect

Animal 1 was the mother of animals 7, 5, 10

Shared groups (Mothers)

An	Sire	Dam
7	14	1
4	14	2
8	15	3
5	15	1
9	14	2
6	15	4
10	15	1
11	16	4
12	16	5
13	16	6

$$\mathbf{Z}_2 = \begin{array}{c} \begin{matrix} 1 & 2 & 3 & 4 & 5 & 6 \end{matrix} \\ \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

MME

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1} + \mathbf{A}^{-1}k_{11} & \mathbf{Z_1'Z_2} \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{I}k_{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{G}} \\ \hat{\mathbf{M}}_e \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} \sigma_G^2 & 0 \\ 0 & \sigma_{m_e}^2 \end{bmatrix}^{-1} \sigma_e^2$$

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} 2000 & 0 \\ 0 & 500 \end{bmatrix}^{-1} 6500 = \begin{bmatrix} 3.25 & 0 \\ 0 & 13 \end{bmatrix}$$

R code Example 7

```
Sig_m=500
Sig_e=6500
Sig_g=2000

Y=matrix(c(400,
             380,
             410,
             350,
             420,
             360,
             390,
             390,
             430,
             370),10,1);

 X=matrix(c(1, 0, 0, 1,
             1, 0, 0, 0,
             1, 0, 0, 1,
             0, 1, 0, 0,
             0, 1, 0, 1,
             0, 1, 0, 0,
             0, 0, 1, 1,
             0, 0, 1, 0,
             0, 0, 1, 1,
             0, 0, 1, 0),10,4, byrow = TRUE )
```

```
A=matrix(c(
1, 0, 0, 0, 0, 0, .5, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, 0, 0, .5, 0, .25 0,
 0, 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 0, 0, 1, 0, 0, 0, 0, .5, .5, 0, .5, .5, 0, .25 .25,
 0, 0, 0, 0, 1, 0, 0, 0, .5, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, .5, .5, .5,
 .5, .5, 0, 0, 0, 0, 1, .25 0, .25 .25 .125, .25 .125, .125, 0.0625,
 .5, 0, .5, 0, 0, 0, .25 1, 0, 0, .5, .5, 0, .5, 0, .25,
  0, 0, 0, .5, .5, 0, 0, 0, 1, .25 0, .25 .25 0, .125, .125,
  0, .5, 0, .5, 0, 0, .25 0, .25 1, 0, .25 .5, 0, .5, .125,
 .5, 0, .5, 0, 0, 0, .25 .5, 0, 0, 1, .25 0, .25 0, .125,
 .25 0, .25 .5, 0, 0, .125, .5, .25 .25 .25 1, .25 .25 .125, .5,
   0, .5, 0, .5, 0, 0, .25 0, .25 .5, 0, .25 1, 0, .25 .125,
 .25 0, .25 0, 0, .5, .125, .5, 0, 0, .25 .25 0, 1, .25 0.3750,
   0, .25 0, .25 0, .5, .125, 0, .125, .5, 0, .125, .25 .25 1, 0.3125,
  .125, 0, .125, .25 0, .5, 0.0625, .25 .125, .125, .125, .5, .125, 0.375, 0.3125, 1.0
),16,16)
Ainv=solve(A)
```

```
Z1=  matrix(c
(0,  0, 0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1),
10,16,byrow=TRUE)
```

```
 Z2=matrix(c
 (1 , 0,  0,  0,  0,  0,
   0,  1,  0,  0,  0,  0,
   0,  0,  1,  0,  0,  0,
   1,  0,  0,  0,  0,  0,
   0,  1,  0,  0,  0,  0,
   0,  0,  0,  1,  0,  0,
   1,  0,  0,  0,  0,  0,
   0,  0,  0,  1,  0,  0,
   0,  0,  0,  0,  1,  0,
   0,  0,  0,  0,  0,  1),
 10,6,byrow=TRUE)
```

```
I=matrix(c
(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),6,6)
```

```
P=matrix(c(Sig_g,   0,
           0, Sig_m),2,2,)

K=solve(P)*Sig_e

LHS=rbind(
cbind(t(X)%*%X,      t(X) %*%Z1,                          t(X) %*%Z2)   ,
cbind(t(Z1) %*%X ,   t(Z1) %*%Z1+Ainv*K[1,1] ,  t(Z1) %*%Z2) ,
cbind(t(Z2) %*%X,   t(Z2) %*%Z1,                        t(Z2) %*%Z2+I*K[2,2]))

RHS=rbind(t(X) %*%Y,
          t(Z1) %*%Y,
          t(Z2) %*%Y)
C=solve(LHS)
BU=C %*% RHS
BU
```

Solutions

B

$\mathbf{B}$	
369.87	Year
363.57	Year
375.03	Year
40.76	Sex

$\hat{\mathbf{G}}$	Animal
1.81	14
-3.80	1
2.783	2
-3.56	15
0.13	3
2.63	16
-1.96	7
3.46	4
-1.57	8
-3.88	5
3.91	9
-0.90	6
-6.31	10
4.56	11
1.22	12
0.12	13

$\hat{\mathbf{M}}_e$	Dam
-2.365897	1
1.226796	2
0.0671247	3
0.5146638	4
0.9262775	5
-0.368965	6

Note:results are same as using correlated residual matrix

How are the results used?

- The model separates Nature from Nurture
  - Human experimenters maybe interested in both effects
    - Maternal Care: How much of the variation in infant weight at 4 weeks post delivery, is due to the genes of the child vs. the nurture of the mother and perhaps the cause of the nurture effects
    - Disease Risk: Alcoholism: risk due to drinking environment (nurture) separated from risk due to nature (genes)
  - Breeders are only interested in making maximal genetic improvement
    - Use the additive genetic effects to select best animals

Impact of Fixed vs. Random Effect

- Example: Correlated residuals due to years
- Data was collected over 3 **years**, the researcher was concerned about a common environmental effect due to years but could not decide if years should be a fixed or random effect. What difference does it really make?
  1. Model 1: Include fixed effect for sex and year; Animal as random
     - Additive genetic (2000)
     - Residual (6500).
  2. Model 2: Same as above but now assume the effect of years is random
     - Between year variance (500)
     - Residual (6500)
     - What are the best estimates of the breeding value of each animal? What are the year effects?
  3. Model 3: Same as Model 2 but increase between year variance to (100,000)
- What impact does fixed vs. random year effect have on the results?

Years Fixed

$$\mathbf{y} = \underset{\substack{\uparrow \\ \text{Fixed effects}}}{\mathbf{Xb}} + \underset{\substack{\uparrow \\ \text{Additive Genetic effects}}}{\mathbf{Z_1G}} + \mathbf{e}$$

$$V\begin{bmatrix} \mathbf{G} \\ \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} \mathbf{A}\sigma_G^2 & 0 \\ \\ 0 & \mathbf{I}\sigma_e^2 \end{bmatrix}$$

$$\mathbf{X} = \begin{array}{cccc} \text{Year} & & & \text{sex} \\ 86 & 87 & 88 & \text{m} \\ \hline 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{array}$$

$$\mathbf{B} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix} \begin{array}{l} \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \begin{array}{l}\text{Herd} \\ \text{Year}\end{array} \\ \leftarrow \text{Sex} \end{array}$$

$$\mathbf{Y} = \begin{bmatrix} 400 \\ 380 \\ 410 \\ 350 \\ 420 \\ 360 \\ 390 \\ 390 \\ 430 \\ 370 \end{bmatrix}$$

R code Example 8

```
Sig_e=6500
Sig_g=2000

Y=matrix(c(400,
              380,
              410,
              350,
              420,
              360,
              390,
              390,
              430,
              370),10,1);

 X=matrix(c(1, 0, 0, 1,
               1, 0, 0, 0,
               1, 0, 0, 1,
               0, 1, 0, 0,
               0, 1, 0, 1,
               0, 1, 0, 0,
               0, 0, 1, 1,
               0, 0, 1, 0,
               0, 0, 1, 1,
              0, 0, 1, 0),10,4, byrow = TRUE )
```

```
A=matrix(c(
1, 0, 0, 0, 0, 0, .5, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, 0, 0, .5, 0, .25 0,
 0, 0, 1, 0, 0, 0, 0, .5, 0, 0, .5, .25 0, .25 0, .125,
 0, 0, 0, 1, 0, 0, 0, 0, .5, .5, 0, .5, .5, 0, .25 .25,
 0, 0, 0, 0, 1, 0, 0, 0, .5, 0, 0, 0, 0, 0, 0, 0,
 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, .5, .5, .5,
 .5, .5, 0, 0, 0, 0, 1, .25 0, .25 .25 .125, .25 .125, .125, 0.0625,
 .5, 0, .5, 0, 0, 0, .25 1, 0, 0, .5, .5, 0, .5, 0, .25,
  0, 0, 0, .5, .5, 0, 0, 0, 1, .25 0, .25 .25 0, .125, .125,
  0, .5, 0, .5, 0, 0, .25 0, .25 1, 0, .25 .5, 0, .5, .125,
 .5, 0, .5, 0, 0, 0, .25 .5, 0, 0, 1, .25 0, .25 0, .125,
 .25 0, .25 .5, 0, 0, .125, .5, .25 .25 .25 1, .25 .25 .125, .5,
   0, .5, 0, .5, 0, 0, .25 0, .25 .5, 0, .25 1, 0, .25 .125,
 .25 0, .25 0, 0, .5, .125, .5, 0, 0, .25 .25 0, 1, .25 0.3750,
   0, .25 0, .25 0, .5, .125, 0, .125, .5, 0, .125, .25 .25 1, 0.3125,
  .125, 0, .125, .25 0, .5, 0.0625, .25 .125, .125, .125, .5, .125, 0.375, 0.3125, 1.0
),16,16)

Ainv=solve(A)
```

```
Z1=  matrix(c
(0,  0, 0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1),
10,16,byrow=TRUE)

LHS=rbind(cbind(t(X)%*%X,    t( X) %*%Z1) ,
               cbind(t(Z1) %*%X, t(Z1) %*%Z1+Ainv*(Sig_e/Sig_g)) )

RHS=rbind(t( X) %*%Y,  t( Z1) %*%Y)
C=solve(LHS)
BU=C%*%RHS
BU
```

Solutions: Years Fixed

Fixed effects

year

369.59

363.39

374.79

sex

40.63

(y1+y2+y3)/3=u=369.26

Y1-u=.33

Y2-u=-5.86

Y3-u=5.53

Come back to this

Additive Genetic Effect

1.96

-4.39

3.16

-3.78

0.20

2.84

-2.41

3.94

-1.58

-4.44

4.35

-0.72

-6.93

4.97

1.24

0.28

Years Random

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z_1}\underset{\substack{\uparrow\\ \text{Genetic effect}}}{\mathbf{G}} + \mathbf{Z_2}\underset{\substack{\uparrow\\ \text{Year effect}}}{\mathbf{T}} + \mathbf{e}$$

$$V\begin{bmatrix}\mathbf{G}\\ \mathbf{T}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{A}\sigma_G^2 & 0 & 0\\ 0 & \mathbf{I}\sigma_T^2 & 0\\ 0 & 0 & \mathbf{I}\sigma_e^2\end{bmatrix}$$

$$\mathbf{X} = \begin{array}{cc} \text{overall mean} & \text{sex} \\ \downarrow & \downarrow \\ \end{array} \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 0 \end{bmatrix}$$

$$\mathbf{B} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} \begin{array}{l} \text{Mean} \\ \text{Sex effect} \end{array}$$

$$\mathbf{Y} = \begin{bmatrix} 400 \\ 380 \\ 410 \\ 350 \\ 420 \\ 360 \\ 390 \\ 390 \\ 430 \\ 370 \end{bmatrix}$$

Animal Additive Genetic Effect

Same as before

An	Sire	Dam
7	14	1
4	14	2
8	15	3
5	15	1
9	14	2
6	15	4
10	15	1
11	16	4
12	16	5
13	16	6

$$\mathbf{Z}_1 = \begin{array}{cccccccccccccccc} 14 & 1 & 2 & 15 & 3 & 16 & 7 & 4 & 8 & 5 & 9 & 6 & 10 & 11 & 12 & 13 \end{array}$$
$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Common Year Effect

An	Year	
7	86	
4	86	
8	86	
5	87	
9	87	
6	87	
10	88	
11	88	
12	88	
13	88	

Year(T)

$$\mathbf{Z}_2 = \begin{array}{ccc} 86 & 87 & 88 \\ \end{array} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

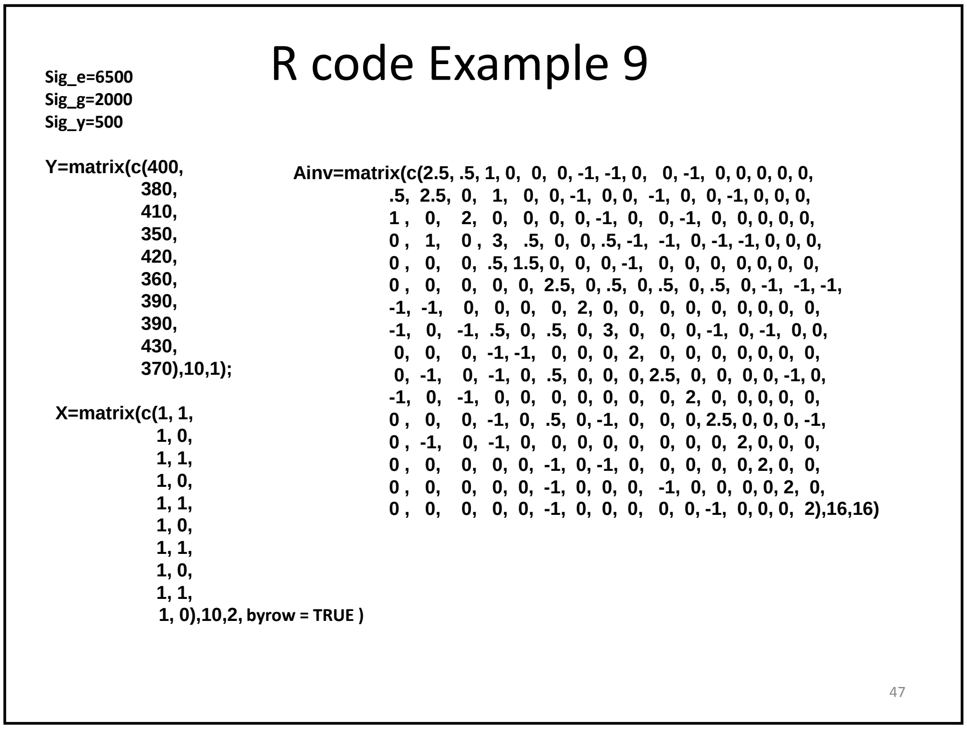
$$\mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

MME

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z_1} & \mathbf{X'Z_2} \\ \mathbf{Z_1'X} & \mathbf{Z_1'Z_1} + \mathbf{A}^{-1}k_{11} & \mathbf{Z_1'Z_2} \\ \mathbf{Z_2'X} & \mathbf{Z_2'Z_1} & \mathbf{Z_2'Z_2} + \mathbf{I}k_{22} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{G}} \\ \hat{\mathbf{T}} \end{bmatrix} = \begin{bmatrix} \mathbf{X'y} \\ \mathbf{Z_1'y} \\ \mathbf{Z_2'y} \end{bmatrix}$$

$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} \sigma_G^2 & 0 \\ 0 & \sigma_T^2 \end{bmatrix}^{-1} \sigma_e^2$$

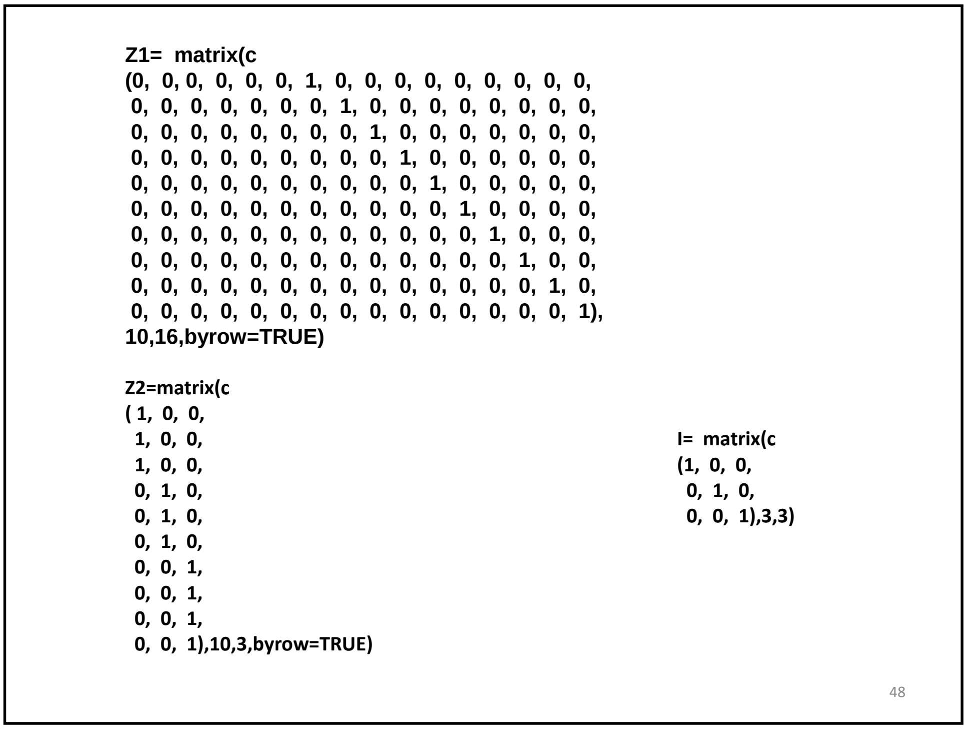
$$\begin{bmatrix} k_{11} & k_{12} \\ k_{21} & k_{22} \end{bmatrix} = \begin{bmatrix} 2000 & 0 \\ 0 & 500 \end{bmatrix}^{-1} 6500 = \begin{bmatrix} 3.25 & 0 \\ 0 & 13 \end{bmatrix}$$



R code Example 9

```
Sig_e=6500
Sig_g=2000
Sig_y=500

Y=matrix(c(400,
           380,
           410,
           350,
           420,
           360,
           390,
           390,
           430,
           370),10,1);

 X=matrix(c(1, 1,
            1, 0,
            1, 1,
            1, 0,
            1, 1,
            1, 0,
            1, 1,
            1, 0,
            1, 1,
            1, 0),10,2, byrow = TRUE )
```

```
Ainv=matrix(c(2.5, .5, 1, 0,  0,  0, -1, -1, 0,   0, -1,  0, 0, 0, 0, 0,
        .5,  2.5,  0,   1,   0,  0, -1,  0, 0,  -1,  0,  0, -1, 0, 0, 0,
        1 ,   0,    2,   0,   0,  0,  0, -1,  0,   0, -1,  0,  0, 0, 0, 0,
        0 ,   1,    0 ,  3,   .5,  0,  0, .5, -1,  -1,  0, -1, -1, 0, 0, 0,
        0 ,   0,    0,  .5, 1.5, 0,  0,  0, -1,   0,  0,  0,  0, 0, 0,  0,
        0 ,   0,    0,   0,  0,  2.5,  0, .5,  0, .5,  0, .5,  0, -1,  -1, -1,
        -1,  -1,    0,   0,  0,   0,  2,  0,  0,   0,  0,  0,  0, 0, 0,  0,
        -1,   0,   -1,  .5,  0,  .5,  0,  3,  0,   0,  0, -1,  0, -1,  0, 0,
         0,   0,    0,  -1, -1,   0,  0,  0,  2,   0,  0,  0,  0, 0, 0,  0,
         0,  -1,    0,  -1,  0,  .5,  0,  0,  0, 2.5,  0,  0,  0, 0, -1, 0,
        -1,   0,   -1,   0,  0,   0,  0,  0,  0,   0,  2,  0,  0, 0, 0,  0,
        0 ,   0,    0,  -1,  0,  .5,  0, -1,  0,   0,  0, 2.5, 0, 0, 0, -1,
        0 ,  -1,    0,  -1,  0,   0,  0,  0,  0,   0,  0,  0,  2, 0, 0,  0,
        0 ,   0,    0,   0,  0,  -1,  0, -1,  0,   0,  0,  0,  0, 2, 0,  0,
        0 ,   0,    0,   0,  0,  -1,  0,  0,  0,  -1,  0,  0,  0, 0, 2,  0,
        0 ,   0,    0,   0,  0,  -1,  0,  0,  0,   0,  0, -1,  0, 0, 0,  2),16,16)
```



```
Z1=  matrix(c
(0,  0, 0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1,  0,
 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  1),
10,16,byrow=TRUE)
```

```
Z2=matrix(c
( 1,  0,  0,
  1,  0,  0,
  1,  0,  0,
  0,  1,  0,
  0,  1,  0,
  0,  1,  0,
  0,  0,  1,
  0,  0,  1,
  0,  0,  1,
  0,  0,  1),10,3,byrow=TRUE)
```

```
I=  matrix(c
(1,  0,  0,
  0,  1,  0,
  0,  0,  1),3,3)
```

```
P=matrix(c(Sig_g ,  0,
                 0,    Sig_y=500),2,2)
K=solve(P)*Sig_e

LHS=rbind(
cbind(t(X)%*%X,      t(X) %*%Z1,                          t(X) %*%Z2)   ,
cbind(t(Z1) %*%X ,  t(Z1) %*%Z1+Ainv*K[1,1] ,  t(Z1) %*%Z2) ,
cbind(t(Z2) %*%X,   t(Z2) %*%Z1,                        t(Z2) %*%Z2+I*K[2,2]))

RHS=rbind(t(X) %*%Y,
                t(Z1) %*%Y,
                t(Z2) %*%Y)
C=solve(LHS)
BU=C %*% RHS
BU
```

Solutions
Years random (var=500)

Fixed effects

369.07 mean
41.94 Sex effect

Year Effects

-0.008
-0.988
0.996

Additive Genetic Effect

1.40
-4.41
2.67
-4.14
0.13
4.35
-2.77
3.65
-1.87
-5.06
3.09
-1.38
-6.64
6.12
2.09
1.27

Years random (var=100000)

```
P=matrix(c(Sig_g ,  0,
               0,    Sig_y=100000),2,2)
K=solve(P)*Sig_e

LHS=rbind(
cbind(t(X)%*%X,      t(X) %*%Z1,                          t(X) %*%Z2)   ,
cbind(t(Z1) %*%X ,  t(Z1) %*%Z1+Ainv*K[1,1] ,  t(Z1) %*%Z2) ,
cbind(t(Z2) %*%X,   t(Z2) %*%Z1,                          t(Z2) %*%Z2+I*K[2,2]))

RHS=rbind(t(X) %*%Y,
              t(Z1) %*%Y,
              t(Z2) %*%Y)
C=solve(LHS)
BU=C %*% RHS
BU
```

Solutions

Fixed effects

369.25 mean
40.67 Sex effect

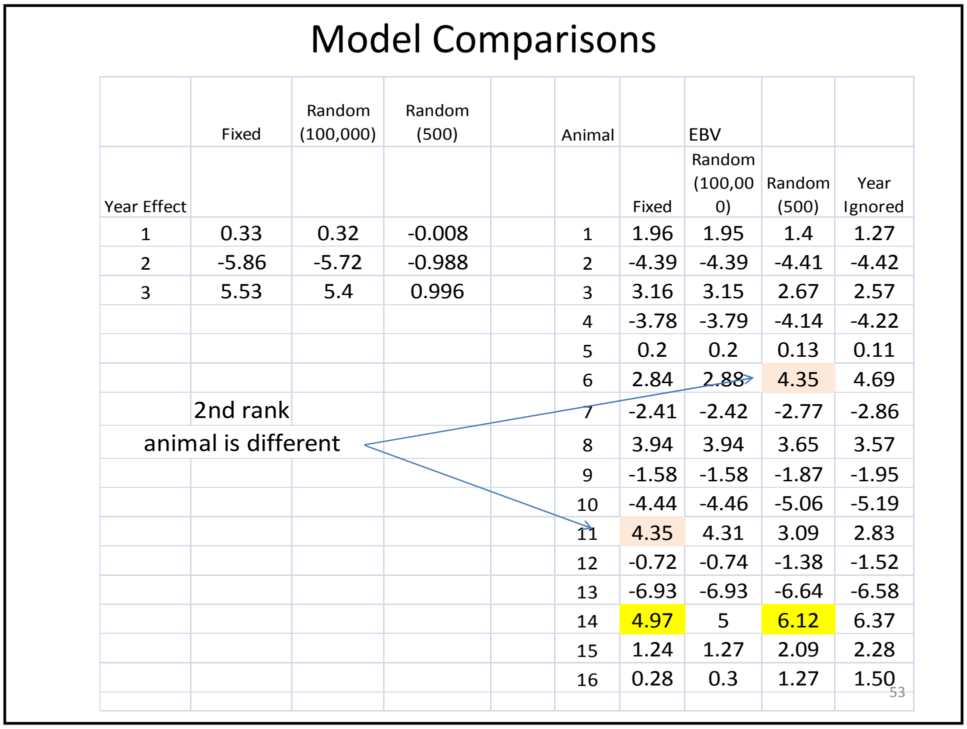
Year Effects
0.32
-5.72
5.40

Note the effect of years increased and near identical to slide with years fixed

Additive Genetic Effect

1.95
-4.39
3.15
-3.79
0.20
2.88
-2.42
3.94
-1.58
-4.46
4.31
-0.74
-6.93
5.00
1.27
0.30

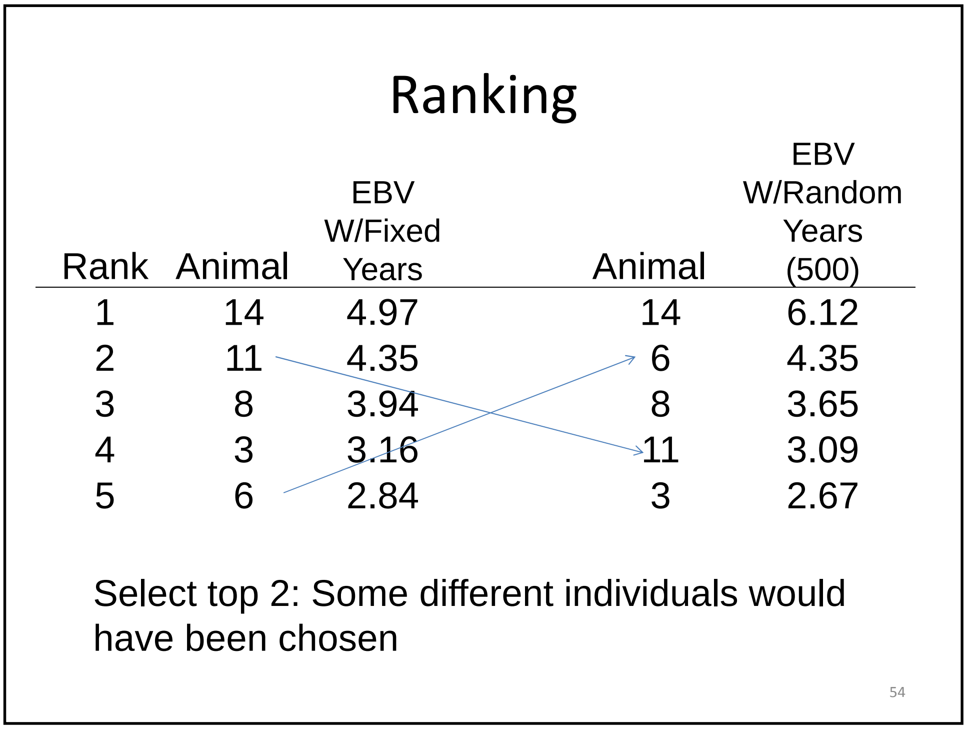
Additive effects near identical to previous example where years were fixed



Model Comparisons

	Fixed	Random (100,000)	Random (500)		Animal		EBV		
Year Effect						Fixed	Random (100,000)	Random (500)	Year Ignored
1	0.33	0.32	-0.008		1	1.96	1.95	1.4	1.27
2	-5.86	-5.72	-0.988		2	-4.39	-4.39	-4.41	-4.42
3	5.53	5.4	0.996		3	3.16	3.15	2.67	2.57
					4	-3.78	-3.79	-4.14	-4.22
					5	0.2	0.2	0.13	0.11
					6	2.84	2.88	4.35	4.69
					7	-2.41	-2.42	-2.77	-2.86
					8	3.94	3.94	3.65	3.57
					9	-1.58	-1.58	-1.87	-1.95
					10	-4.44	-4.46	-5.06	-5.19
					11	4.35	4.31	3.09	2.83
					12	-0.72	-0.74	-1.38	-1.52
					13	-6.93	-6.93	-6.64	-6.58
					14	4.97	5	6.12	6.37
					15	1.24	1.27	2.09	2.28
					16	0.28	0.3	1.27	1.50

2nd rank
animal is different

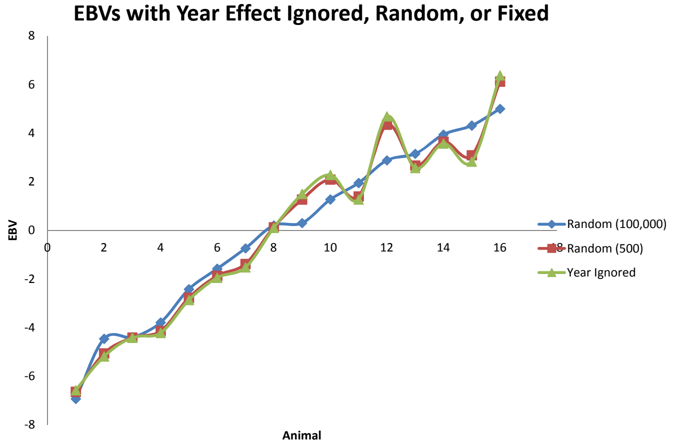


Ranking

Rank	Animal	EBV W/Fixed Years	Animal	EBV W/Random Years (500)
1	14	4.97	14	6.12
2	11	4.35	6	4.35
3	8	3.94	8	3.65
4	3	3.16	11	3.09
5	6	2.84	3	2.67

Select top 2: Some different individuals would have been chosen

How much should the EBV be adjusted for year?



When Year is fit as a random effect, the data tells us how much to adjust

Setting a factor as fixed is equivalent to assuming the intra-class correlation is 1

$$\rho = \frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2}$$

Or variance of the effect is infinite

$$\rho = \frac{500}{500 + 6500} = .07$$

First Example

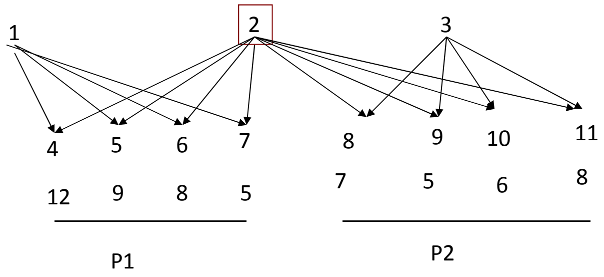
$$\rho = \frac{100000}{100000 + 6500} \cong 1$$

Second Example

Setting a factor at random allows the residual correlation, and amount of adjustment for that factor, to be estimated from the data

For a random factor the intra-class correlation is used as a shrinkage factor, how much to adjust for the factor; a correlation of 1, totally adjusts for the factor, a correlation of 0, the factor is ignored Y’=Y-ρ(T-u)

Example 10 Impact of Common Maternal Environment of Ranking of Selection Candidates



Choose genetically the best individuals for breeding from among the offspring

Genetic Parameters

Case A
minor common
environmental effect

$$\sigma_G^2 = 10$$
$$\sigma_{m_e}^2 = 1$$
$$\sigma_e^2 = 100$$

Case B
moderate common
environmental effect

$$\sigma_G^2 = 10$$
$$\sigma_{m_e}^2 = 100$$
$$\sigma_e^2 = 100$$

Case C
Major common
environmental effect

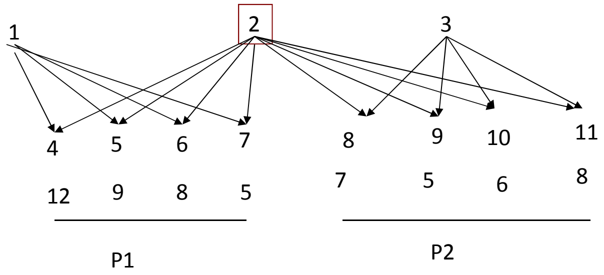
$$\sigma_G^2 = 10$$
$$\sigma_{m_e}^2 = 1000$$
$$\sigma_e^2 = 100$$

Example 10 common maternal environment.R

Solutions Case A

Animal	Family	EBV	rank
4	1	0.29	1
5	1	0.15	2
6	1	0.10	3
7	1	-0.04	4
11	2	-0.05	5
8	2	-0.10	6
10	2	-0.15	7
9	2	-0.20	8

Example Impact of Common Maternal Environment of Ranking of Selection Candidates



$$\sigma_G^2 = 10$$

$$\sigma_{m_e}^2 = 1$$

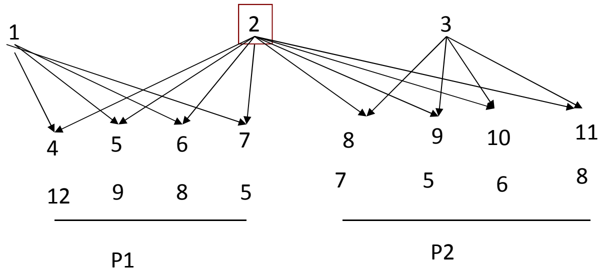
$$\sigma_e^2 = 100$$

Individuals 4, 5, 6, 7 highest ranking all from the same mother
Example between family selection
Heritability low and common family effects small
No Competitive effects

Solutions Case **B**

Animal	Family	EBV	rank
4	1	0.26	1
5	1	0.12	2
6	1	0.07	3
11	2	-0.03	4
7	1	-0.07	5
8	2	-0.07	6
10	2	-0.12	7
9	2	-0.17	8

Moderate Common (Family) Environmental Effects



$$\sigma_G^2 = 10$$

$$\sigma_{m_e}^2 = 100$$

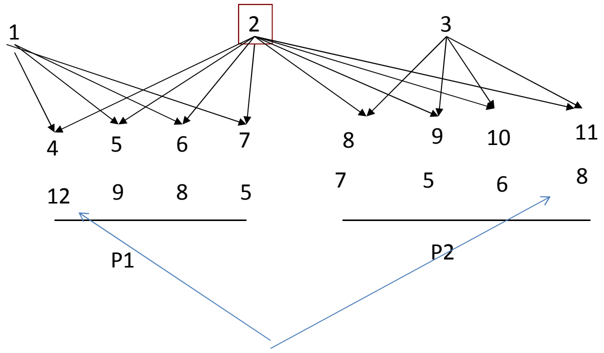
$$\sigma_e^2 = 100$$

Individuals 4 , 5, 6 and 11 would be chosen, some from both families

Solutions Case C

Animal	Family	EBV	rank
4	1	0.17	1
11	2	0.07	2
5	1	0.03	3
8	2	0.02	4
6	1	-0.02	5
10	2	-0.03	6
9	2	-0.08	7
7	1	-0.16	8

Major Common (Family) Environmental Effects



$$\sigma_G^2 = 10$$

$$\sigma_{m_e}^2 = 1000$$

$$\sigma_e^2 = 100$$

- •Top individuals are highest ranking within each family
- •Major Common environmental effects

Within Family Selection

Between and within family selection

$$I = b_1\left(Y_{i.} - \overline{Y}_{..}\right) + b_2\left(Y_{ij} - \overline{Y}_{i.}\right)$$

Between family deviation Within family deviation

•b1=0 is within family selection
•b2=0 is between family selection
•If both are >0 then finding optimal weight was difficult
•The mixed model approach solves this problem

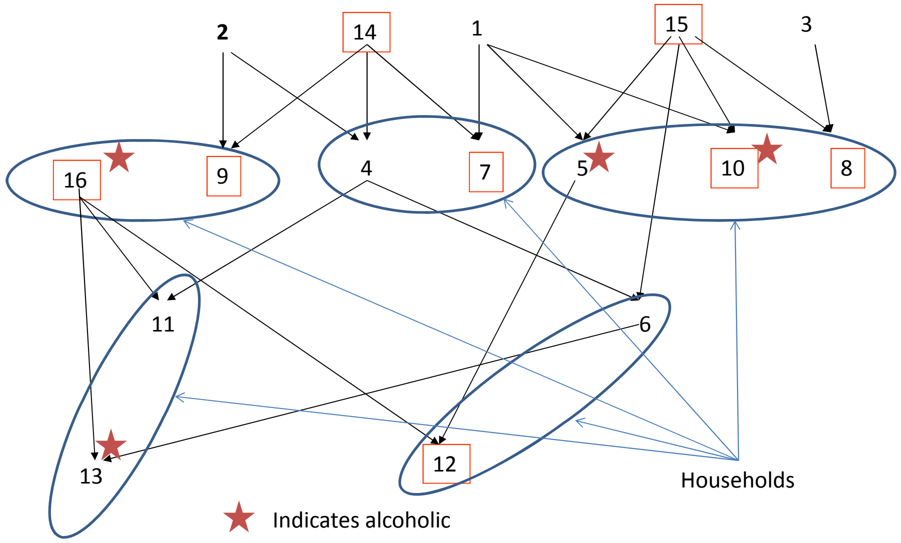
Negative Environmental Correlations

- If the intraclass correlation is truly negative, then the only way to model the data is with a correlation matrix rather than a 2nd random effect
- A negative intraclass correlation implies there is greater variation with a group than between groups
- If modeling between and within population variation it is possible to get a true negative Fis or Fit if one of the sub-populations is the result of out-crossing. There will be more heterozygotes within a population than expected.
- Unstable competition can also result in greater variability within groups (likes compete more than dislikes)

Genetics of Disease Resistance or Susceptibility

- Influenced by Nature and Nurture
- Alcoholism, or other learned behaviors, such as smoking
- Assume Nurture is determined by adolescent household

Example Households and relationships



Problem

- Define a mixed model that would separate the effects of Nature from that of Nurture
- What variance components need to be estimated
- How could these be estimated
- Can you use these results to predict risk of alcoholism if individuals from this population produced offspring? How?

Lecture 7: Models with multiple random effects: Repeated Measures and Maternal effects

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

Often there are several vectors of random effects

- Repeatability models
  - Multiple measures
- Common family effects
  - Cleaning up residual covariance structure
- Maternal effects models
  - Maternal effect has a genetic (i.e., breeding value) component

Multiple random effects

$$y = X\beta + Za + Wu + e$$

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

$$y = X\beta + Za + Wu + e$$

Defining the covariance structure key in any mixed-model

Suppose **e** ~ $(0, \sigma_e^2 \text{ I})$, **u** ~ $(0, \sigma_u^2 \text{ I})$, **a** ~ $(0, \sigma_A^2 \text{ A})$,
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:

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

y = Xβ + Za + Wu + 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**

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

Note that if we ignored the second vector u of random effects, and assumed **y = Xβ + Za + e***, then **e* = Wu + e,** with Var(**e***) = σ_e^2 I + σ_u^2 **WW**^{T}

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

Mixed-model Equations

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

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 $\sigma_e^2 = (1\text{-}r)\ \sigma_z^2$

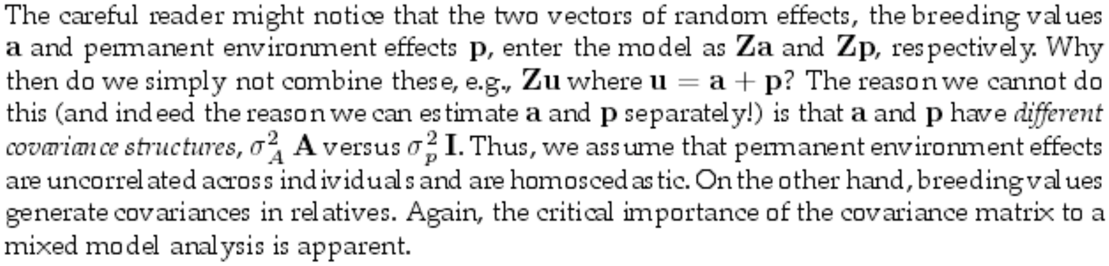
Resulting mixed model

$$y = X\beta + Za + Zp + e$$

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

Notice that we could also write this model as
$y = X\beta + Z(a + p) + e = y = X\beta + Zv + e, v = a+p$

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*, $\sigma^2_A\,\mathbf{A}$ versus $\sigma^2_p\,\mathbf{I}$. 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:

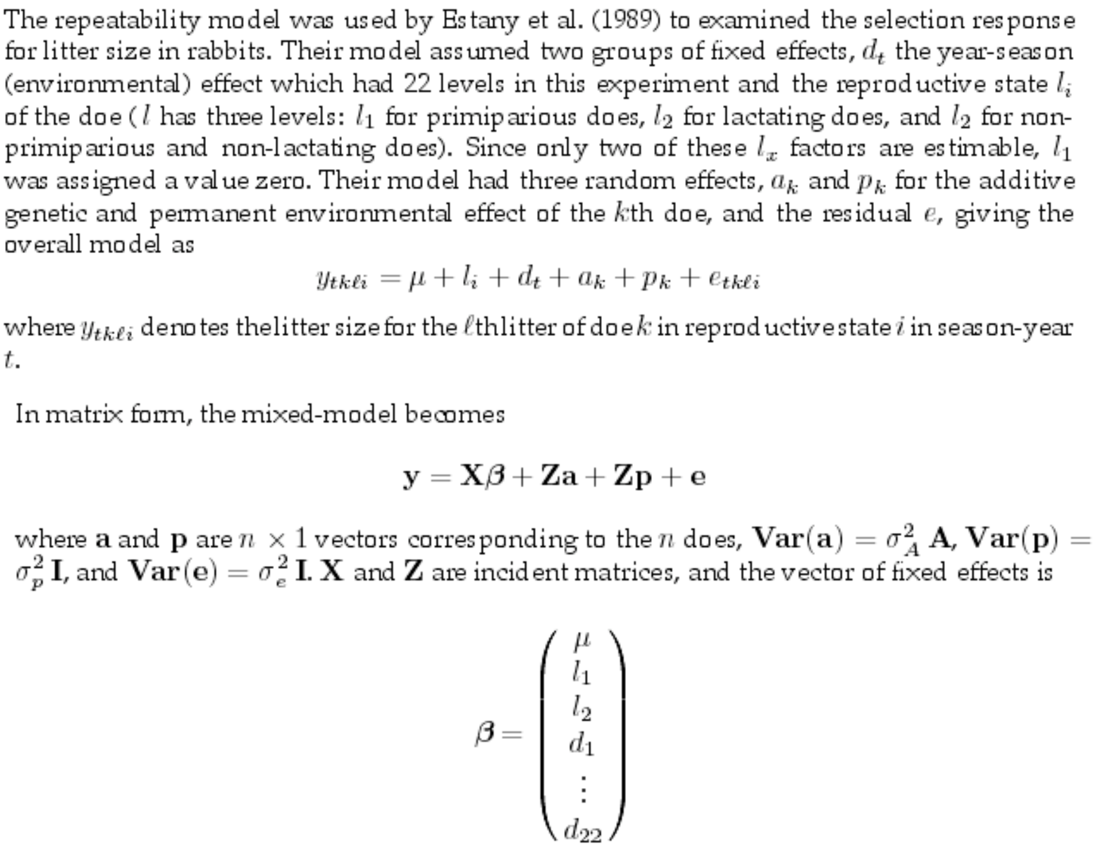
$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{31} \\ y_{32} \end{pmatrix}, \qquad \mathbf{Z} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix}$$

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 **y = Xβ + Za + e***
  – Then **e* = Zp + e,**
  – Var(**e***) = σ_e^2 I + σ_p^2 ZZ^T
- Assuming that Var(**e***) = σ_e^2 I gives an incorrect model
- We could either
  – use **y = Xβ + Za + e*** with the correct error structure (covariance) for e* = σ_e^2 I + σ_p^2 ZZ^T
  – Or use **y = Xβ + Za +Zp + e,** where e = σ_e^2 I



The repeatability model was used by Estany et al. (1989) to examined the selection response for litter size in rabbits. Their model assumed two groups of fixed effects, d_t the year-season (environmental) effect which had 22 levels in this experiment and the reproductive state l_i of the doe (l has three levels: l_1 for primiparious does, l_2 for lactating does, and l_2 for non-primiparious and non-lactating does). Since only two of these l_x factors are estimable, l_1 was assigned a value zero. Their model had three random effects, a_k and p_k for the additive genetic and permanent environmental effect of the kth doe, and the residual e, giving the overall model as

$$y_{tk\ell i} = \mu + l_i + d_t + a_k + p_k + e_{tk\ell i}$$

where $y_{tk\ell i}$ denotes the litter size for the ℓth litter of doe k in reproductive state i in season-year t.

In matrix form, the mixed-model becomes

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

where $\mathbf{a}$ and $\mathbf{p}$ are $n \times 1$ vectors corresponding to the n does, $\mathbf{Var}(\mathbf{a}) = \sigma^2_A\,\mathbf{A}$, $\mathbf{Var}(\mathbf{p}) = \sigma^2_p\,\mathbf{I}$, and $\mathbf{Var}(\mathbf{e}) = \sigma^2_e\,\mathbf{I}$. $\mathbf{X}$ and $\mathbf{Z}$ are incident matrices, and the vector of fixed effects is

$$\boldsymbol{\beta} = \begin{pmatrix} \mu \\ l_1 \\ l_2 \\ d_1 \\ \vdots \\ d_{22} \end{pmatrix}$$

Resulting mixed-model equations

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

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} = \frac{1-r}{h^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_p^2} = \frac{1-r}{r-h^2}$$

Common family effects

- Sibs in the same family also share a common environment
  - Cov(full sibs) = $\sigma_A^2/2 + \sigma_D^2/4 + \sigma_{ce}^2$
- Hence, if the model assumes $y_i = \mu + a_i + c_i + e_i$, with **a** ~ 0, σ_A^2**A**, c ~ 0, σ_{cf}^2I. If there are records for different sibs from the same family, Var(**e**) is no longer σ_e^2 **I**
- **y = Xβ + Za + Wc + e**
- Again, if common family effect ignored (we assume **y = Xβ + Za + e***) the error structure is e* = σ_e^2 I + σ_{cf}^2 **WW**^{T}
  - Where $\sigma_{cf}^2 = \sigma_D^2/4 + \sigma_{ce}^2$
  - The common family effect may contain both environment and non-additive genetic components

Example: Measure 7 individuals, first five are from family one, last two from family 2

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

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \quad \mathbf{Z} = \mathbf{I}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Z = I as every individual has a single record.
If there are missing and/or repeated records,
Z does not have this simple structure

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

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \quad \mathbf{Z} = \mathbf{I}, \quad \mathbf{a} = \begin{pmatrix} A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \\ A_6 \\ A_7 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{c} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix}$$

Again, matrix multiplication gives us the form of the **Z** and **W** matrices. Consider y_6:

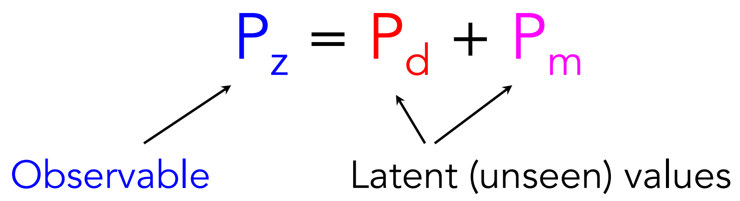
$$y_6 = \mu + A_6 + c_2 + e_6$$

Maternal effects with genetic components

- The phenotype of an offspring can be influenced by its mother beyond her genetic contribution
- For example, two offspring with identical genotypes will still show potentially significant differences in size if they receive different amounts of milk from their mothers
- Such maternal effects can be quite important
- While we have just discussed models with common family effects, these are potentially rather different that maternal effects models
  - Common family environmental effects are assumed not to be inherited across generations.

- Consider milk yield. The heritability for this trait is around 30% and the milk yield of the mother has a significant impact on the weight of her offspring
- Offspring with high breeding values for milk will tend to have daughters with above -average milk yield, and hence above -average maternal effects
- The value of an offspring can be considered to consist of two components
  - A direct effect (intrinsic breeding value)
  - A maternal contribution

Phenotypic value = direct value + maternal value



$$P_z = P_d + P_m$$

Observable

Latent (unseen) values

Both of the latent values can be further decomposed into breeding plus residual (environmental + non- additive genetic) values

$$P_d = \mu + A_d + E_d, \qquad P_m = \mu + A_m + E_m,$$

The direct breeding value A_d appears in the phenotype of its carrier

The maternal breeding value A_m DOES NOT appear in the phenotype of its carrier, but **rather in the phenotype of her offspring**

Direct vs. maternal breeding values

- The direct and maternal contributions are best thought of as **two separate, but potentially correlated**, traits.
  - Hence, we need to consider $\sigma(A_d,A_m)$ in addition to $\sigma^2(A_d)$ and $\sigma^2(A_m)$. This changes the form of the mixed-model equations
- The direct BV (A_d) is expressed in the individual carrying it
- The maternal BV (A_m) is only expressed in the offspring trait value (and only mom's A_m appears)

Covariance structure

$$\begin{pmatrix} \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2(A_d)\,\mathbf{A} & \sigma(A_d, A_m)\,\mathbf{A} \\ \sigma(A_d, A_m)\,\mathbf{A} & \sigma^2(A_m)\,\mathbf{A} \end{pmatrix}$$

This is often written using the Kronecker (or direct) product:

$$\mathbf{A} \otimes \mathbf{B} = \begin{pmatrix} a_{11}\mathbf{B} & \cdots & a_{1n}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{m1}\mathbf{B} & \cdots & {}_{mn}\mathbf{B} \end{pmatrix}$$

Giving

$$\begin{pmatrix} \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \mathbf{G} \otimes \mathbf{A} \qquad \mathbf{G} = \begin{pmatrix} \sigma^2(A_d) & \sigma(A_d, A_m) \\ \sigma(A_d, A_m) & \sigma^2(A_m) \end{pmatrix}$$

The mixed-model becomes

Direct effects
breeding values

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_d\mathbf{a}_d + \mathbf{Z}_m\mathbf{a}_m + \mathbf{e}$$

Maternal effects
breeding values

The error structure needs a little care, as the direct E_d and maternal E_m residual values can be correlated*. Initially, we will assume Var(**e**) ~ $\sigma_e^2\mathbf{I}$

*See Bijma 2006 J. Anim. Sci. 84:800-806 for treatment of correlated environmental residuals under this model

The resulting mixed-model equations become

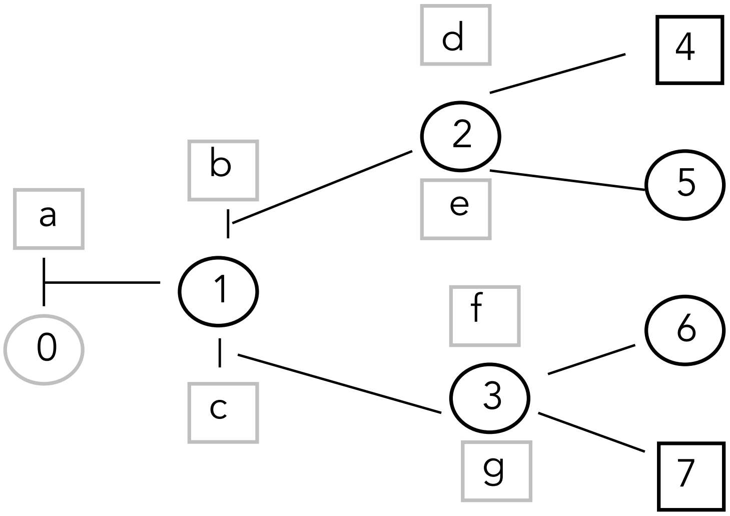
$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d+\lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m+\lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d+\lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m+\lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

where the weights λ_i are related to elements in the inverse of $\mathbf{G}$, viz.,

$$\begin{pmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_3 \end{pmatrix} = \sigma_e^2\,\mathbf{G}^{-1} = \sigma_e^2 \begin{pmatrix} \sigma^2(A_d) & \sigma(A_d,A_m) \\ \sigma(A_d,A_m) & \sigma^2(A_m) \end{pmatrix}^{-1}$$

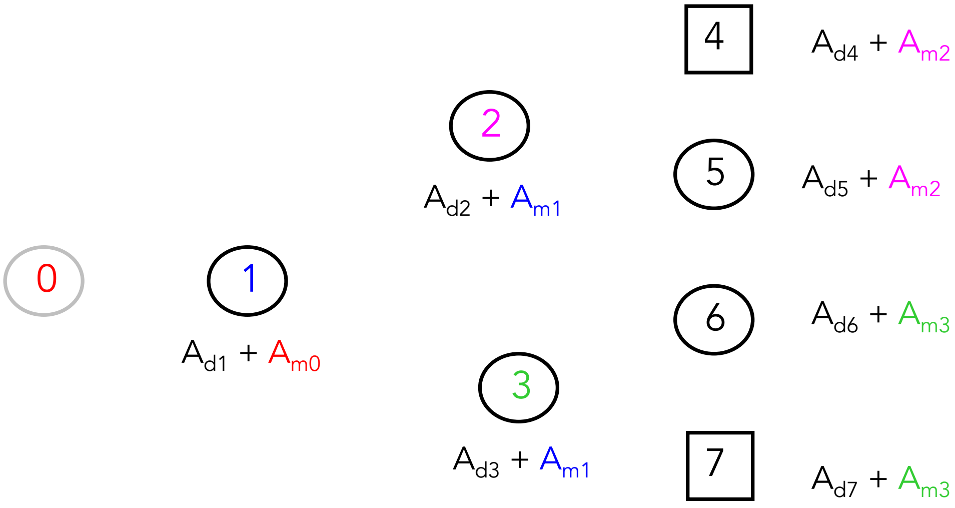
Filling out the maternal effects incident matrix Z_m

A little bookkeeping care is needed when filling out Z_m, because the A_m associated with a record (measured individual) is that of their mother.



1-7 have records

All sires unrelated



The observed values are y_1 through y_7.
What we can estimate are A_{d1} through A_{d7},
A_{m0} through A_{m3}

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \end{pmatrix}, \qquad \mathbf{a}_d = \begin{pmatrix} A_{d,1} \\ A_{d,2} \\ A_{d,3} \\ A_{d,4} \\ A_{d,5} \\ A_{d,6} \\ A_{d,7} \end{pmatrix}, \qquad \mathbf{Z}_d = \mathbf{I}, \qquad \mathbf{a}_m = \begin{pmatrix} A_{m.o} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \end{pmatrix}$$

Note that we estimate A_{m0} even though we don't have a record (observation) on her.

Since $\mathbf{Z}_m\mathbf{a}_m$ must be a 7 x 1 matrix, $\mathbf{Z}_m$ is 7 x 4 (as $\mathbf{a}_m$ is 4 x 1)

Record 1 is associated with A_{m0}

Records 2 and 3 are associated with A_{m1}

Records 4 and 5 are associated with A_{m2}

Records 6 and 7 are associated with A_{m3}

Record 1 is associated with A_{m0}

Records 2 and 3 are associated with A_{m1}

Records 4 and 5 are associated with A_{m2}

Records 6 and 7 are associated with A_{m3}

$$\mathbf{Z}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \quad \textbf{as} \quad \mathbf{Z}_m \mathbf{a}_m = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \end{pmatrix} = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,1} \\ A_{m,2} \\ A_{m,2} \\ A_{m,3} \\ A_{m,3} \end{pmatrix}$$

What about A_{m4} through A_{m7}?

Although we have records that only directly relate A_{m0} to A_{m3}, through the use of A we can (in theory) also estimate the maternal breeding values for individuals 4 through 7. Note this includes the maternal BVs for the two males (5 & 7), as they can pass this onto their daughters.

$$\mathbf{Z}_m^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \mathbf{a}_m^* = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \\ A_{m,4} \\ A_{m,5} \\ A_{m,6} \\ A_{m,7} \end{pmatrix}$$

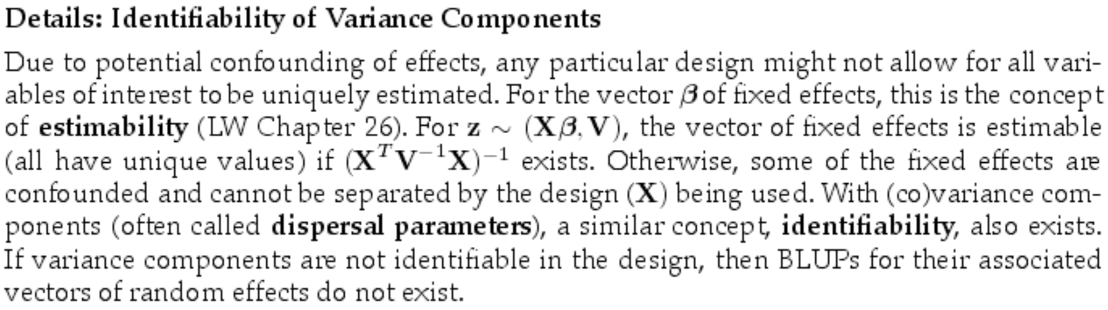
Note that

$$\mathbf{Z}_m^* \mathbf{a}_m^* = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,2} \\ A_{m,3} \\ A_{m,4} \\ A_{m,5} \\ A_{m,6} \\ A_{m,7} \end{pmatrix} = \begin{pmatrix} A_{m,0} \\ A_{m,1} \\ A_{m,1} \\ A_{m,2} \\ A_{m,2} \\ A_{m,3} \\ A_{m,3} \end{pmatrix}$$

All this raises the question about what can, and cannot, be estimated from the data (**y**) and the design (Z_m, Z_d)?

First issue: Is the structure of the design such that we can estimate all of the variance components. This is the issue of identifiability

Estimability vs. Identifiability



Details: Identifiability of Variance Components

Due to potential confounding of effects, any particular design might not allow for all variables of interest to be uniquely estimated. For the vector β of fixed effects, this is the concept of **estimability** (LW Chapter 26). For $\mathbf{z} \sim (\mathbf{X}\beta, \mathbf{V})$, the vector of fixed effects is estimable (all have unique values) if $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ exists. Otherwise, some of the fixed effects are confounded and cannot be separated by the design $(\mathbf{X})$ being used. With (co)variance components (often called **dispersal parameters**), a similar concept, **identifiability**, also exists. If variance components are not identifiable in the design, then BLUPs for their associated vectors of random effects do not exist.

Conditions for identifiability of REML estimates of (co)variance components are given by Rothenberg (1971), Jiang (1996), and Cantet and Cappa (2008). Before presenting these, we first review a few details about REML. Recall (LW Chapter 27) that REML estimates are those that maximize that part of the likelihood function that is independent of the fixed effects (this is often stated as being the **translation invariant** part). Let $\mathbf{V}$ be the covariance matrix of $\mathbf{z}$, which is a function of its variance components. As detailed in LW Chapter 27, Harville (1977) shows that (if it exists) the transformation provided by the matrix

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} \tag{1a}$$

plays a critical role in REML estimates. That this matrix can remove fixed effects can be seen by noting that

$$\mathbf{Pz} = \mathbf{V}^{-1}\left(\mathbf{z} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) \tag{1b}$$

yields a vector that is the data vector adjusted by the (estimated) fixed effects. Now consider covariance structures of the form

$$\mathbf{V} = \sum_{i=1}^{n} \mathbf{V}_i\theta_i \tag{2a}$$

where $\mathbf{V}_i$ is a matrix of known constants and the θ_i are unknown variances and covariances to be estimated.

The equations to maximize the likelihood over the restricted space (the REML estimates) are given by LW Equations 27.18 and 27.19, and are solved iteratively. These equations involve the **trace** (sum of the diagonal elements) of matrix products involving $\mathbf{P}$ and the $\mathbf{V}_i$. Recall (LW Appendix 4) that for a vector $\boldsymbol{\Theta}$ of n unknowns, the Fisher information matrix $\mathbf{F}$ (the matrix of second partial derivatives of the likelihood with respect to the parameters) can be used to provide large-sample standard errors. The resulting $n \times n$ information matrix for REML estimates of the unknown θ_i in Equation 2a is

$$F_{ij} = \text{trace}\,(\mathbf{PV}_i\mathbf{PV}_j) \tag{2b}$$

Much in the same fashion that the existence of $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ informs us that all fixed effects are estimable in a given design, all variance components θ_i are identifiable if all of the eigenvalues of $\mathbf{F}$ are positive, that is, that $\mathbf{F}$ is positive-definite (Rothenberg 1971, Jiang 1996). For the maternal effects mixed model, Equation 2a becomes

$$\mathbf{V} = \mathbf{V}_1\,\sigma^2(A_d) + \mathbf{V}_2\,\sigma(A_d, A_s) + \mathbf{V}_3\,\sigma^2(A_s) + \mathbf{V}_4\,\sigma_e^2 \tag{3a}$$

where

$$\mathbf{V}_1 = \mathbf{Z}_d\mathbf{A}\mathbf{Z}_d^T, \quad \mathbf{V}_2 = \left(\mathbf{Z}_d\mathbf{A}\mathbf{Z}_m^T + \mathbf{Z}_m\mathbf{A}\mathbf{Z}_d^T\right), \quad \mathbf{V}_3 = \mathbf{Z}_m\mathbf{A}\mathbf{Z}_s^T, \quad \mathbf{V}_4 = \mathbf{I} \tag{3b}$$

Substituting Equations 1a and 3b into Equation 2b fills out the $\mathbf{F}$ matrix (which is only 4×4 in this case given the four unknown variance components). For any particular design, the eigenvalues of this matrix can be computed to determine if the variance components are all identifiable.

Second issue, connectivity

Even if the design is such that we can estimate all the genetic variances, whether we can estimate all of the $\boldsymbol{\beta}$, $\mathbf{a}_d$, and $\mathbf{a}_m$ in the model depends on whether a unique inverse exists for the MME

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d+\lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m+\lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d+\lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m+\lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

Unique estimates of all the $\boldsymbol{\beta}$ require $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ exists

If $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ does not exist, a generalized inverse is used which can uniquely estimate k linear combinations of the $\boldsymbol{\beta}$ where k is the rank of $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}$

Likewise, if the MME equation does not have an inverse (and this is not due to constraints on $\boldsymbol{\beta}$), then a generalized inverse can be used to estimate unique estimates of certain linear combinations of the $\mathbf{a}_d$ and $\mathbf{a}_m$.

$$\begin{pmatrix} \mathbf{X}^T\mathbf{X} & \mathbf{X}^T\mathbf{Z}_d & \mathbf{X}^T\mathbf{Z}_s \\ \mathbf{Z}_d\mathbf{X}^T & \mathbf{Z}_d^T\mathbf{Z}_d+\lambda_1\mathbf{A}^{-1} & \mathbf{Z}_d^T\mathbf{Z}_m+\lambda_2\mathbf{A}^{-1} \\ \mathbf{Z}_m\mathbf{X}^T & \mathbf{Z}_m^T\mathbf{Z}_d+\lambda_2\mathbf{A}^{-1} & \mathbf{Z}_m^T\mathbf{Z}_m+\lambda_3\mathbf{A}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta} \\ \mathbf{a}_d \\ \mathbf{a}_m \end{pmatrix} = \begin{pmatrix} \mathbf{X}^T\mathbf{y} \\ \mathbf{Z}_d^T\mathbf{y} \\ \mathbf{Z}_m^T\mathbf{y} \end{pmatrix}$$

A key role in ensuring that unique estimates of $\mathbf{a}_d$ and $\mathbf{a}_m$ exist is played by the relationship matrix A. If individuals with records and individuals without records are sufficiently well connected (non-zero entries in A for their pair-wise relatedness), then we usually can estimate values of un-observed individuals (although their precision is another issue)

Indirect Genetic Effects

- Inherited Genetic effect of one animals measured another
  - Inherited Social interactions
    - Competition
    - Mutualism
    - Pack behavior
  - Theory for evolution of social effects and how to estimate effects

The Problem: Competitive Interactions

- Active (Social)
  - Dominance
  - Peck Order
- Passive (Shared Limited Resources)
  - Plants or Animals
    - Space
    - Food Supply
  - Movement

Results of Antagonist Social Interactions

- Reduced Gain
- Increased Mortality
  - Direct
    - Injuries
  - Indirect
    - Immune response
    - Diseases susceptibility
- Reduced Feed Efficiency
  - Energy lost in fighting
  - Increased Fat deposition
  - Disproportionate Feed Consumption

Animal Well Being Concerns

Addressing Social Interactions in Animal Breeding Programs

- Two Selection Methods
  - Direct selection against undesirable behaviors
  - Multilevel selection
    - Kin
    - Group
    - Optimal

Direct Selection

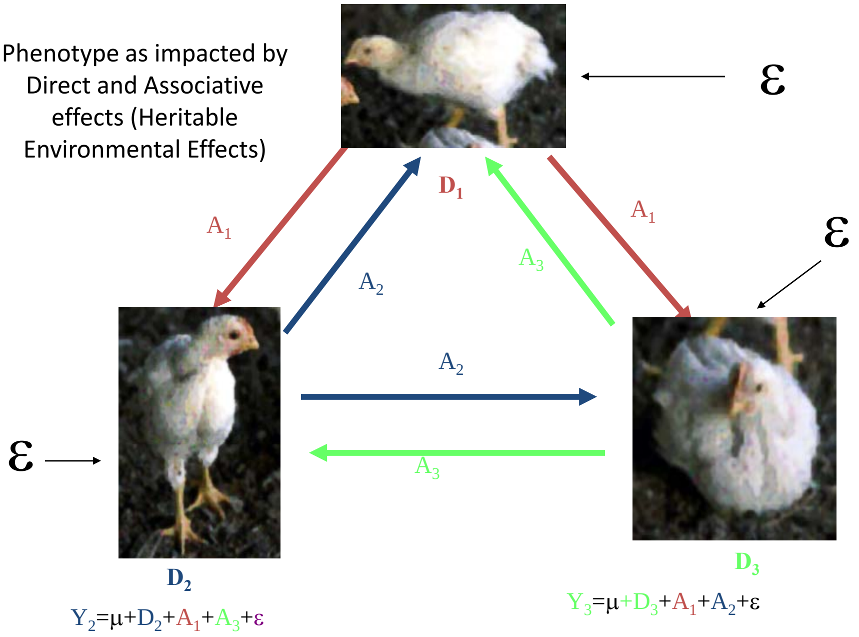
- Feather pecking, Tail biting, Skin lesions, Biomarkers, Tonic immobility (Kjaer and Hocking, 2004; Muir and Craig, 1998; Turner *et al.*, 2008)
  - Highly Successful (Craig and Muir, 1993; Kjaer *et al.*, 2001)
- Requires Quantification
  - Can Be Costly and labor intensive
  - Diverts selection intensity
  - Possible Undesirable Genetic Correlations

Framework Social Evolution Context

- Hamilton (1963, 1964a,b)
  - Altruism Can Evolve Under Individual Selection
  - Introduced “inclusive fitness”
  - Kin Selection (br-c > 0)

Framework Plant and Animal Breeding Context

- Bruce Griffing (1967, 1968a, 1968b, 1969, 1976a, 1976b, 1977)
  - Introduced Associative Effects
    - Heritable Environmental Effects
  - Generalized Multilevel Selection Theory
    - Focuses on merit relative to levels of organization
    - Extension of between and within family deviations for non-interacting genotypes developed by Lush (1947)
    - More Extensively Developed by Bijma et al. 2007a

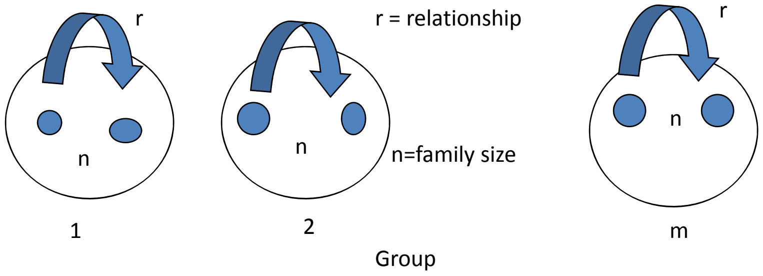


Phenotype as impacted by Direct and Associative effects (Heritable Environmental Effects)

$Y_2=\mu+D_2+A_1+A_3+\varepsilon$

$Y_3=\mu+D_3+A_1+A_2+\varepsilon$

Multilevel Selection



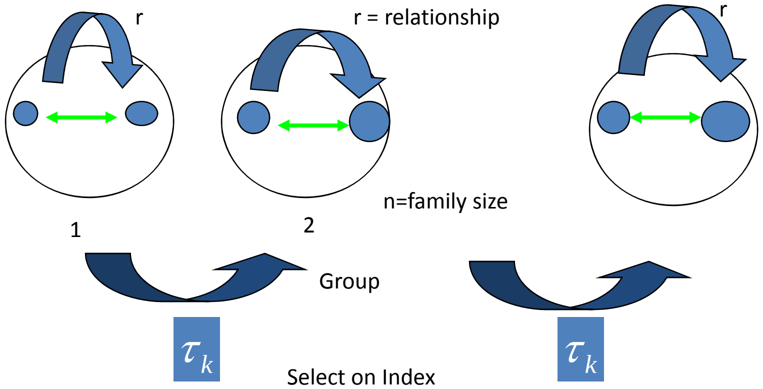
Phenotypic Deviations=within family deviation+ between family deviation

$$Y_{kl} - \overline{Y}_{..} = \left(\overline{Y}_{k.} - \overline{Y}_{..}\right) + \left(Y_{kl} - \overline{Y}_{k.}\right)$$

Same as Lush' derivation except animals are now grouped by family

Grouping introduces covariance (genetic and environmental)

Multilevel Selection



Select on Index

$$I_{kl} = B_1\tau_k + B_2\gamma_{(k)l}$$

Multi-level selection models specify fitness as a function of the mean trait value of the group and the individual deviation thereof:

Multilevel Selection

Type of Selection	B_1	B_2	Expected Response
Kin	1	1	$\Delta u = \left(\frac{i}{c}\right)\left[\mathrm{cov}(G\tau)+\mathrm{cov}(G\gamma)\right]$
Group	1	0	$\Delta u = \left(\frac{i}{c}\right)\left[\mathrm{cov}(G\tau)\right]$
Within	0	1	$\Delta u = \left(\frac{i}{c}\right)\left[\mathrm{cov}(G\gamma)\right]$
Optimal	$B_1 = \left(\frac{\mathrm{cov}(G\tau)}{\sigma_\tau^2}\right)$	$B_2 = \left(\frac{\mathrm{cov}(G\gamma)}{\sigma_\gamma^2}\right)$	$\Delta u = \left(\frac{i}{c}\right)\left[B_1\,\mathrm{cov}(G\tau)+B_2\,\mathrm{cov}(G\gamma)\right]$

$$\mathrm{cov}(G\tau) = \left[\frac{1+(n-1)r}{n}\right]\left[\sigma_D^2 + 2(n-1)\sigma_{DA} + (n-1)^2\sigma_A^2\right]$$

$$\mathrm{cov}(G\gamma) = \left[\frac{(1-r)}{n}\right]\left[(n-1)\sigma_D^2 + (n-1)(n-2)\sigma_{DA} - (n-1)^2\sigma_A^2\right]$$

Estimation of Parameters

Mixed Model Equations

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_d\mu_d + \mathbf{Z}_a\mu_a + \varepsilon$$

Muir and Schinckel (2002)

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_d\mu_d + \mathbf{Z}_a\mu_a + \varepsilon$$

Include correlated residual in R matrix

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_d\mu_d + \mathbf{Z}_a\mu_a + \underset{\uparrow}{\mathbf{Z}_c}\mu_c + e$$

Random effect for shared group

Variances

$$G = \begin{bmatrix} \sigma_D^2 & \sigma_{ad} \\ \sigma_{ad} & \sigma_a^2 \end{bmatrix} \otimes \mathbf{A}$$

σ_d^2 Additive Direct Effects

σ_a^2 Additive Associate Effects=Indirect Genetic Effect (IGE)

σ_{ad} Additive Covariance Between Direct and Indirect Effects

σ_E^2 Environmental

$\nearrow \sigma_c^2$ Between Group

$\searrow \sigma_e^2$ Within Group

MME with correlated residuals

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z_d & X'R^{-1}Z_a \\ Z_dR^{-1}X' & Z_d^{'}R^{-1}Z_d + k_1A^{-1} & Z_d^{'}R^{-1}Z_a + k_2A^{-1} \\ Z_aR^{-1}X' & Z_a^{'}R^{-1}Z_d + k_2A^{-1} & Z_a^{'}R^{-1}Z_a + k_3A^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mu_d \\ \mu_a \end{bmatrix} = \begin{bmatrix} X'R^{-1}X \\ X'R^{-1}Z_d \\ X'R^{-1}Z_a \end{bmatrix}$$

$$\mathbf{R} = \left(\sigma_c^2 + \sigma_e^2\right) \begin{bmatrix} 1 & \rho & \rho & \cdots & 0 & 0 & 0 \\ \rho & 1 & \rho & \cdots & 0 & 0 & 0 \\ \rho & \rho & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & \rho & \rho \\ 0 & 0 & 0 & \cdots & \rho & 1 & \rho \\ 0 & 0 & 0 & \cdots & \rho & \rho & 1 \end{bmatrix}$$

$$\begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix} = \begin{bmatrix} \sigma_d^2 & \sigma_{ad} \\ \sigma_{ad} & \sigma_a^2 \end{bmatrix}^{-1}$$

ρ= intra-class environmental correlation

$$\rho = \frac{\sigma_c^2}{\sigma_c^2 + \sigma_e^2}$$

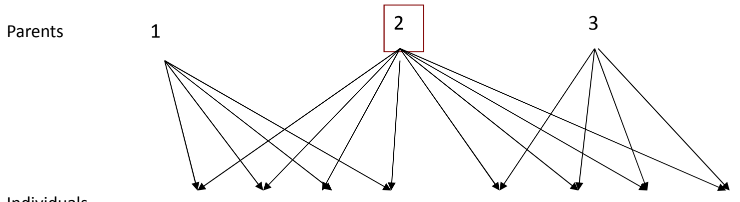
Note if ρ < 0, then then $\sigma_c^2 < 0$

MME with random effect for shared environmental effect

$$\begin{bmatrix} \mathbf{X'X} & \mathbf{X'Z}_D & \mathbf{X'Z}_A & \mathbf{X'Z}_c \\ \mathbf{Z}_D^{'}\mathbf{X} & \mathbf{Z}_D^{'}\mathbf{Z}_D + \mathbf{A}^{-1}\mathbf{k}_{11} & \mathbf{Z}_D^{'}\mathbf{Z}_A + \mathbf{A}^{-1}\mathbf{k}_{12} & \mathbf{Z}_D^{'}\mathbf{Z}_c \\ \mathbf{Z}_A^{'}\mathbf{X} & \mathbf{Z}_A^{'}\mathbf{Z}_D + \mathbf{A}^{-1}\mathbf{k}_{21} & \mathbf{Z}_A^{'}\mathbf{Z}_A + \mathbf{A}^{-1}\mathbf{k}_{22} & \mathbf{Z}_A^{'}\mathbf{Z}_c \\ \mathbf{Z}_c^{'}\mathbf{X} & \mathbf{Z}_c^{'}\mathbf{Z}_D & \mathbf{Z}_c^{'}\mathbf{Z}_A & \mathbf{Z}_c^{'}\mathbf{Z}_c + \mathbf{I}k_{33} \end{bmatrix} \begin{bmatrix} \hat{\mathbf{b}} \\ \boldsymbol{\mu}_{\mathbf{d}} \\ \boldsymbol{\mu}_{\mathbf{a}}^{g} \\ \boldsymbol{\mu}_{a}^{e} \end{bmatrix} = \begin{bmatrix} \mathbf{X}^{'}\mathbf{y} \\ \mathbf{Z}_D^{'}\mathbf{y} \\ \mathbf{Z}_A^{'}\mathbf{y} \\ \mathbf{Z}_A^{'}\mathbf{y} \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix} = \sigma_\varepsilon^2 \begin{bmatrix} \sigma_D^2 & \sigma_{ad} \\ \sigma_{ad} & \sigma_a^2 \end{bmatrix}^{-1} \qquad k_{33} = \sigma_\varepsilon^2 / \sigma_c^2$$

Example 13



Individuals	4	5	6	7	8	9	10	11
Phenotype	12	9	8	5	7	5	6	8
Pen	1	1	2	2	2	1	1	2

Associative Genetic Matrix

Sir	Da	Anim	Pen	Y
2	1	4	1	12
2	1	5	1	9
2	1	6	2	8
2	1	7	2	5
2	3	8	2	7
2	3	9	1	5
2	3	10	1	6
2	3	11	2	8

Animal

$$Z_a = \begin{array}{ccccccccccc} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\ \hline 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{array}$$

Who was in the same pen as animal 4

Genetic Parameters

$$\sigma_\varepsilon^2 \begin{bmatrix} \sigma_d^2 & \sigma_{ad}^2 \\ \sigma_d^2 & \sigma_a^2 \end{bmatrix} = 60 \begin{bmatrix} 30 & -4 \\ -4 & 10 \end{bmatrix}$$

$$\sigma_c^2 = 6$$

R code Example 13

```
Y=matrix(c(
12,
9,
8,
5,
7,
5,
6,
8),8,1)
```

```
X=matrix(c(
1,
1,
1,
1,
1,
1,
1,
1),8,1)
```

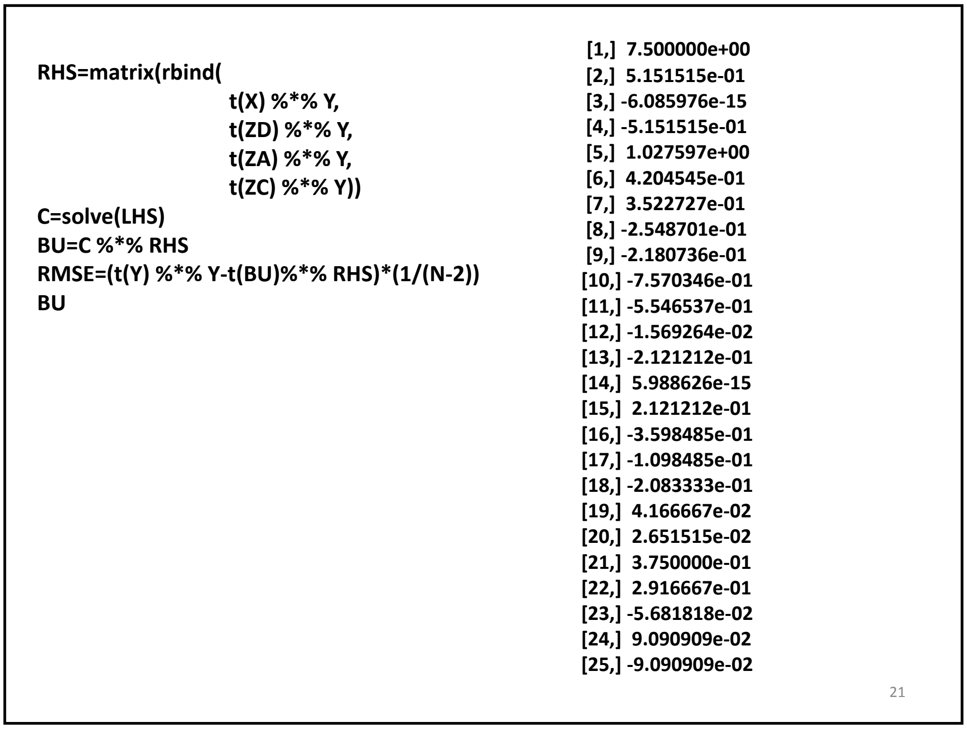
```
A=matrix(c
(1,  0,  0,   .5,   .5,  .5,  .5,   0,   0,    0,   0,
 0,  1,  0,   .5,  .5,  .5,  .5,   .5,  .5,   .5,  .5,
 0,  0,  1,    0,   0,   0,    0,  .5,  .5,   .5,  .5,
.5, .5,  0,    1,  .5,  .5,  .5, .25, .25,  .25, .25,
.5, .5,  0,   .5,   1,  .5,  .5, .25, .25,  .25, .25,
.5, .5,  0,   .5,  .5,   1,  .5, .25, .25,  .25, .25,
.5, .5,  0,   .5,  .5,  .5,   1, .25, .25,  .25, .25,
 0, .5, .5, .25, .25, .25, .25,   1,  .5,   .5,  .5,
 0, .5, .5, .25, .25, .25, .25,  .5,   1,   .5,  .5,
 0, .5, .5, .25, .25, .25, .25,  .5,  .5,    1,  .5,
 0, .5, .5, .25, .25, .25, .25,  .5,  .5,   .5,   1),11,11)
```

```
ZD=matrix(c(
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1)
,8,11,byrow=TRUE)
```

```
ZA=matrix(c(
0, 0, 0, 0, 1, 0, 0, 0, 1, 1, 0,
0, 0, 0, 1, 0, 0, 0, 0, 1, 1, 0,
0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 1,
0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,
0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1,
0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 0,
0, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0,
 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0)
,8,11,byrow=TRUE)
```

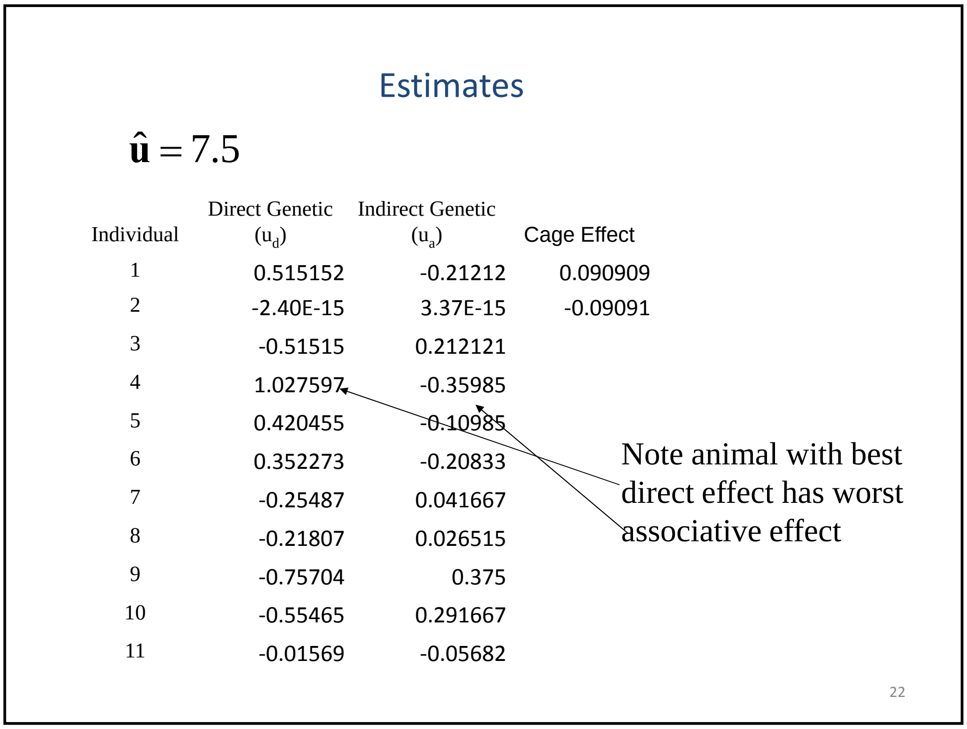
```
ZC=matrix(c(
1, 0,
1, 0,
0, 1,
0, 1,
0, 1,
1, 0,
1, 0,
0, 1)
,8,2,byrow=TRUE)
```

```
Sig_d=30
Sig_a=10
Sig_ad=-4
Sig_e=60
Sig_c=6
N=8
P=matrix(c(Sig_d, Sig_ad,
           Sig_ad, Sig_a),2,2)
K=solve(P)*Sig_e
AINV=solve(A)
I2=matrix(c(1,0,
            0,1),2,2)
K33=Sig_e/Sig_c
LHS= rbind(
cbind(t(X)%*%X,  t(X)%*%ZD,                t(X)%*%ZA,                 t(X)%*%ZC),
cbind(t(ZD)%*%X, t(ZD)%*%ZD+AINV*K[1,1],t(ZD)%*%ZA+AINV*K[1,2], t(ZD)%*%ZC),
cbind(t(ZA)%*%X, t(ZA)%*%ZD+AINV*K[2,1], t(ZA)%*%ZA+AINV*K[2,2], t(ZA)%*%ZC),
cbind(t(ZC)%*%X, t(ZC)%*%ZD,               t(ZC)%*%ZA,      t(ZC)%*%ZC+I2*K33))
```



```
RHS=matrix(rbind(
                    t(X) %*% Y,
                    t(ZD) %*% Y,
                    t(ZA) %*% Y,
                    t(ZC) %*% Y))
C=solve(LHS)
BU=C %*% RHS
RMSE=(t(Y) %*% Y-t(BU)%*% RHS)*(1/(N-2))
BU
```

```
 [1,]  7.500000e+00
 [2,]  5.151515e-01
 [3,] -6.085976e-15
 [4,] -5.151515e-01
 [5,]  1.027597e+00
 [6,]  4.204545e-01
 [7,]  3.522727e-01
 [8,] -2.548701e-01
 [9,] -2.180736e-01
[10,] -7.570346e-01
[11,] -5.546537e-01
[12,] -1.569264e-02
[13,] -2.121212e-01
[14,]  5.988626e-15
[15,]  2.121212e-01
[16,] -3.598485e-01
[17,] -1.098485e-01
[18,] -2.083333e-01
[19,]  4.166667e-02
[20,]  2.651515e-02
[21,]  3.750000e-01
[22,]  2.916667e-01
[23,] -5.681818e-02
[24,]  9.090909e-02
[25,] -9.090909e-02
```



Estimates

$$\hat{\mathbf{u}} = 7.5$$

Individual	Direct Genetic (u_d)	Indirect Genetic (u_a)	Cage Effect
1	0.515152	-0.21212	0.090909
2	-2.40E-15	3.37E-15	-0.09091
3	-0.51515	0.212121	
4	1.027597	-0.35985	
5	0.420455	-0.10985	
6	0.352273	-0.20833	
7	-0.25487	0.041667	
8	-0.21807	0.026515	
9	-0.75704	0.375	
10	-0.55465	0.291667	
11	-0.01569	-0.05682	

Note animal with best direct effect has worst associative effect

Index Selection

$$I = b_1 \hat{\mu}_d + b_2 \hat{\mu}_a$$

The total breeding value (TBV) is the sum of the direct and all IGE effects

$$TBV = \hat{\mu}_d + (n-1)\hat{\mu}_a$$

How Important Are Associative Effects In Breeding Programs?

- Total Breeding Value (TBV) (Bijma et al. 2007a)
- V(TBV) associative effects are scaled by (n-1)^2
- Phenotypic Variance associative effects are scaled by (n-1)
- “Heritability” can be >1

$$TBV_i = A_{D_i} + (n-1)A_{S_i}$$

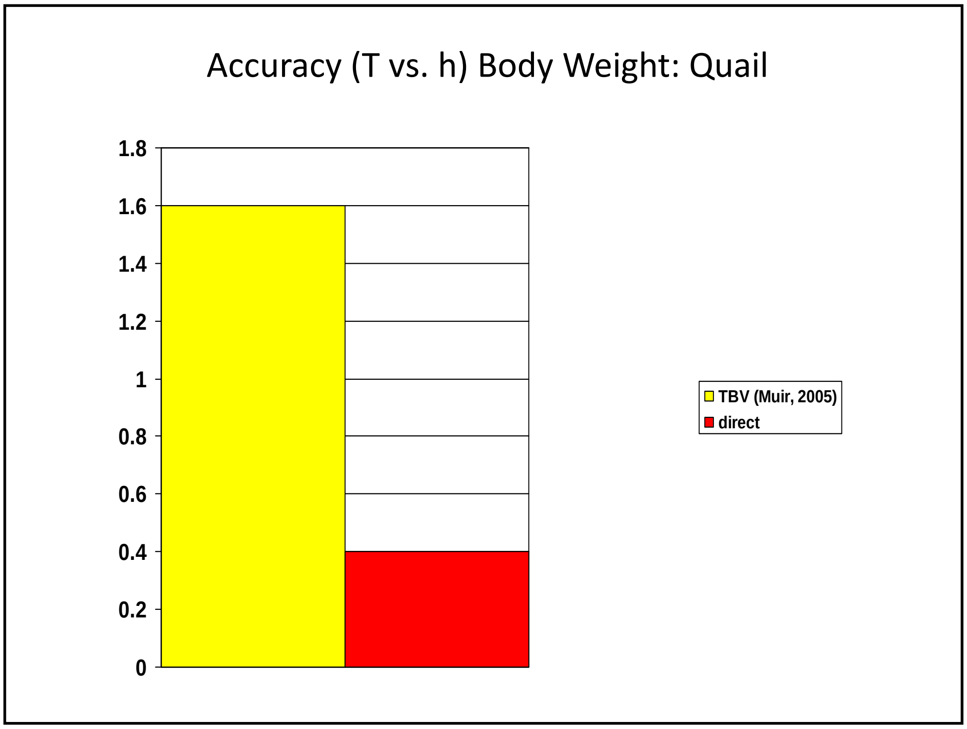
$$V(TBV) = \sigma^2_{A_d} + 2(n-1)\sigma_{A_d A_s} + (n-1)^2 \sigma^2_{A_s}$$

$$V(Y) = \sigma^2_{A_d} + (n-1)\sigma^2_{A_s} + \sigma^2_{\varepsilon}$$

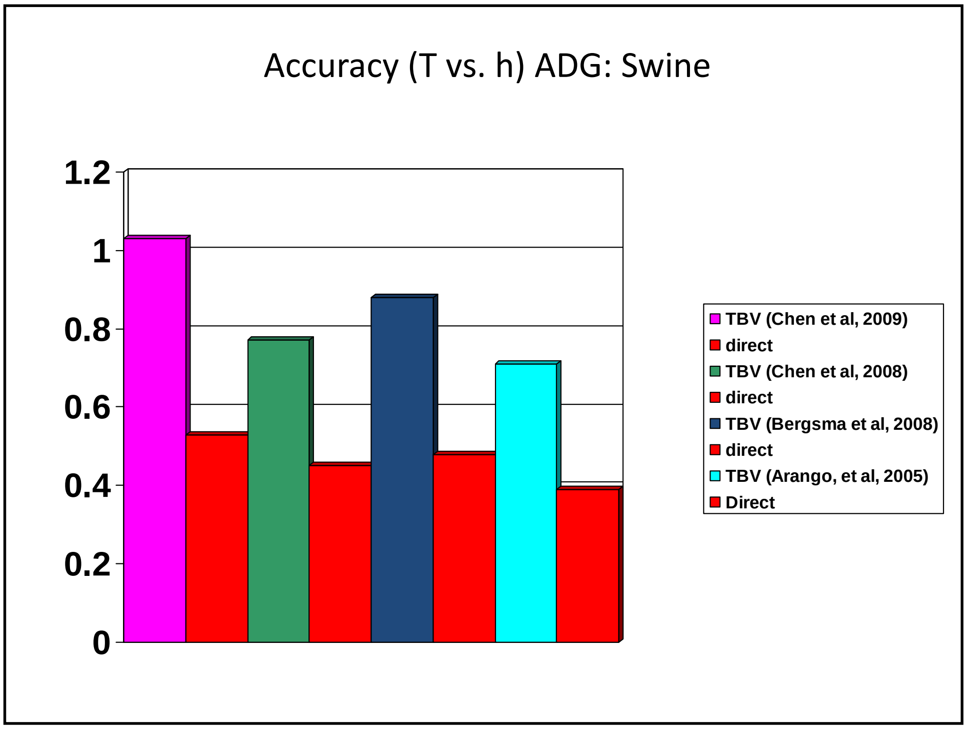
$$T^2 = \frac{V(TBV)}{V(Y)} = \frac{\sigma^2_{A_d} + 2(n-1)\sigma_{A_d A_s} + (n-1)^2 \sigma^2_{A_s}}{\sigma^2_{A_d} + (n-1)\sigma^2_{A_s} + \sigma^2_{\varepsilon}}$$

$$R = iT\sqrt{V(TVB)}$$

accuracy

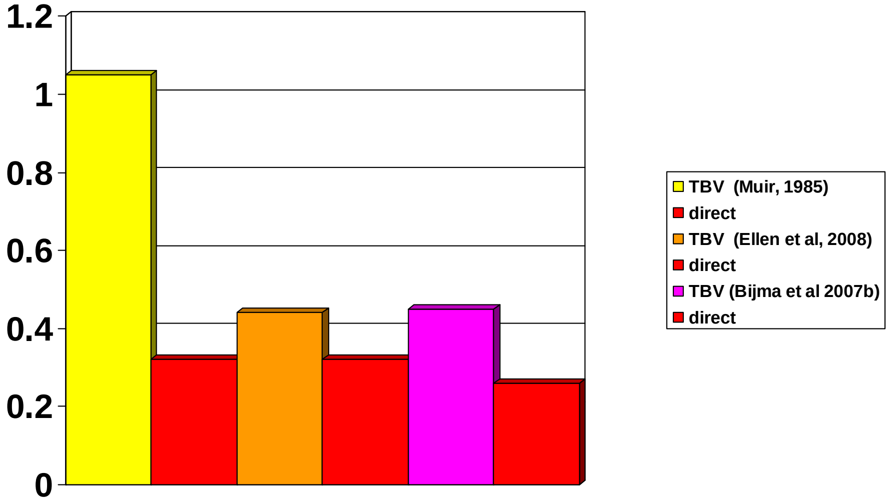


Accuracy (T vs. h) Body Weight: Quail

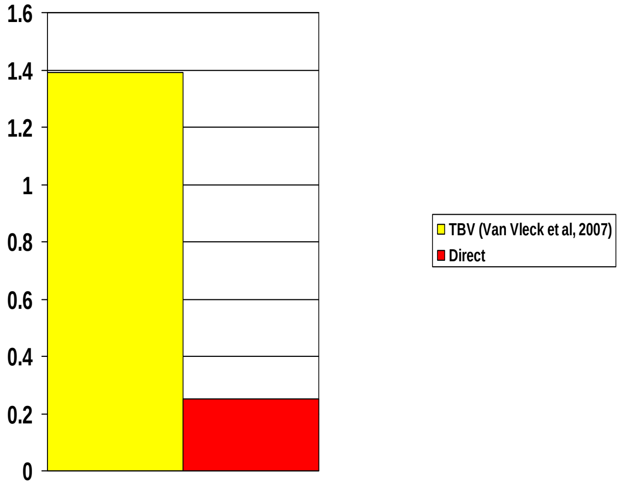


Accuracy (T vs. h) ADG: Swine

Accuracy (T vs. h) Survival Days: Layers



Accuracy (T vs. h) ADG: Beef Cattle



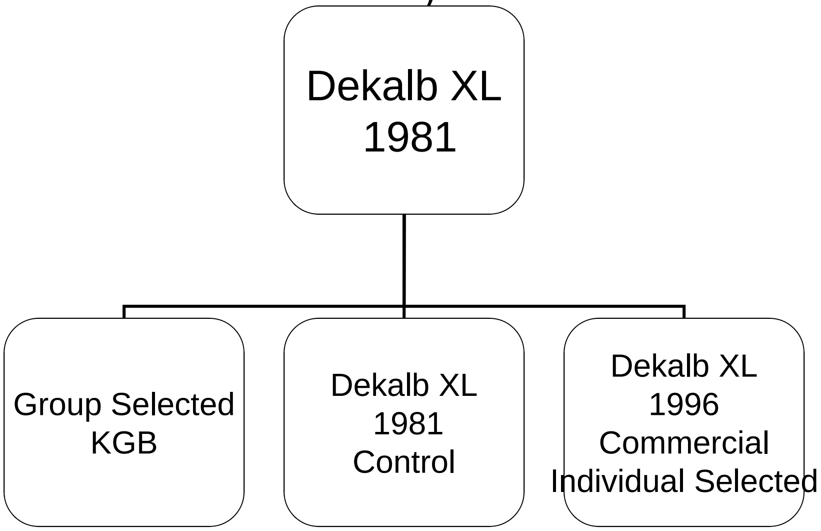
Selection Experiments

Model Organisms

Poultry

Swine

Layers: Group vs. Individual Selection (Muir, 1996)



Dekalb XL
1981

Group Selected
KGB

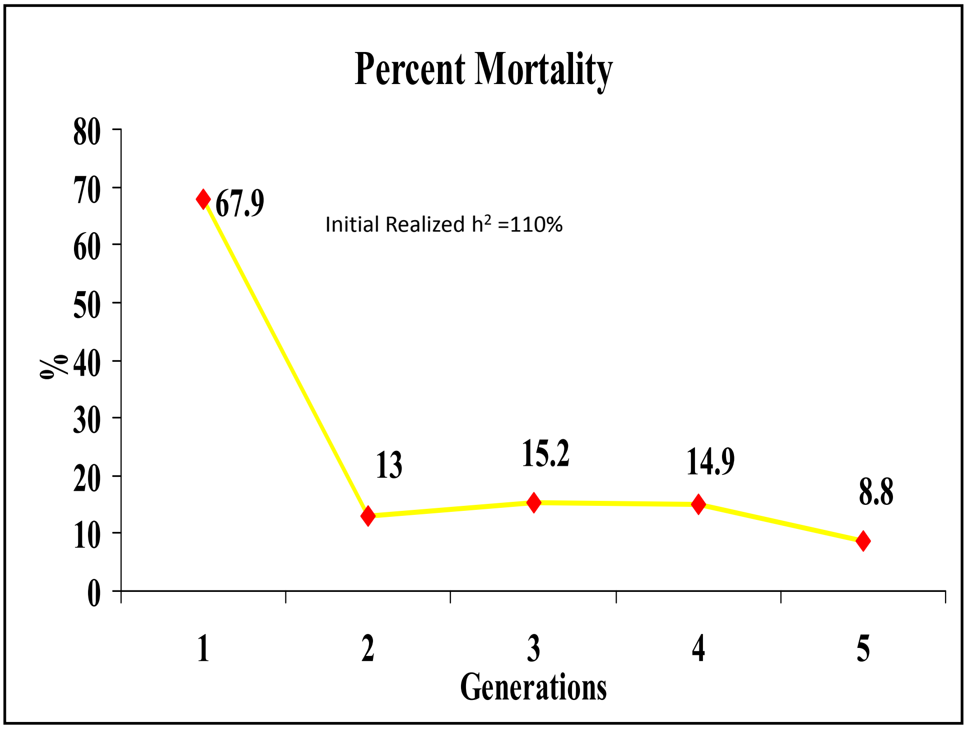
Dekalb XL
1981
Control

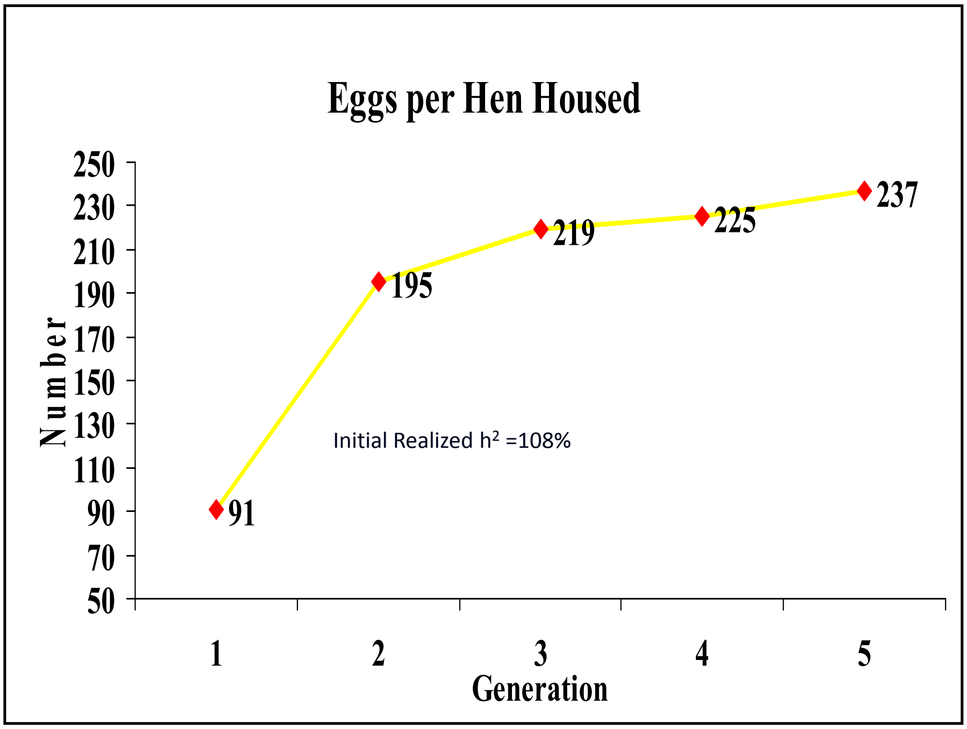
Dekalb XL
1996
Commercial
Individual Selected

Group Selection

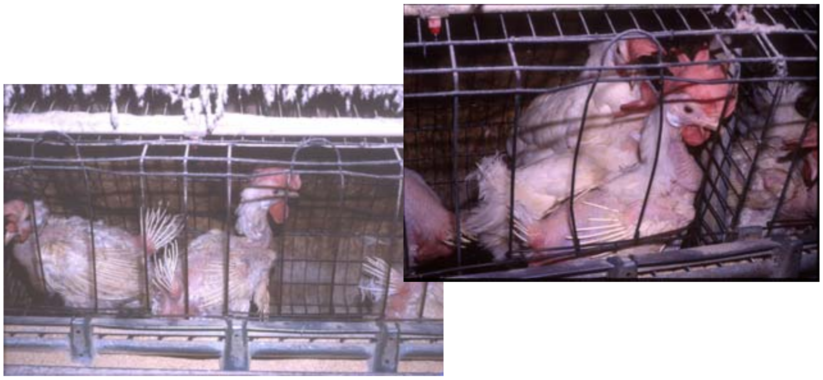
- Selected Index
  - Total Days Survival and Rate of Lay
  - Full Record (12 months of production)
  - Groups 12 Bird Half Sib Family One Colony Cage (56 sq in/bird)
  - Saved Birds from the Best 24/384 Colony Cages
  - Repeated for 6 generations

Control (Dekalb 1981)
Randomly Selected From Single Bird Cages


Percent Mortality
80
70
60
50
40
30
20
10
0
%
67.9
Initial Realized h² =110%
13
15.2
14.9
8.8
1
2
3
4
5
Generations

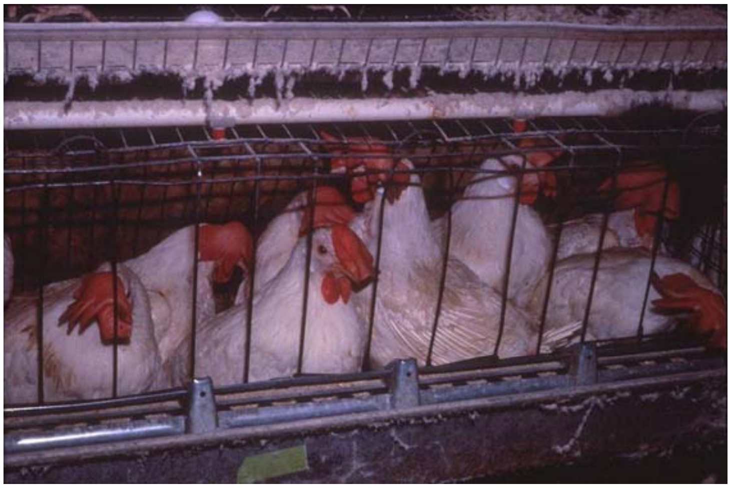

Eggs per Hen Housed
250
230
210
190
170
150
130
110
90
70
50
Number
237
225
219
195
Initial Realized h² =108%
91
1
2
3
4
5
Generation

Control Bird (DXL) After 12 Months of Production



6 Alive

KGB Bird After 12 Months

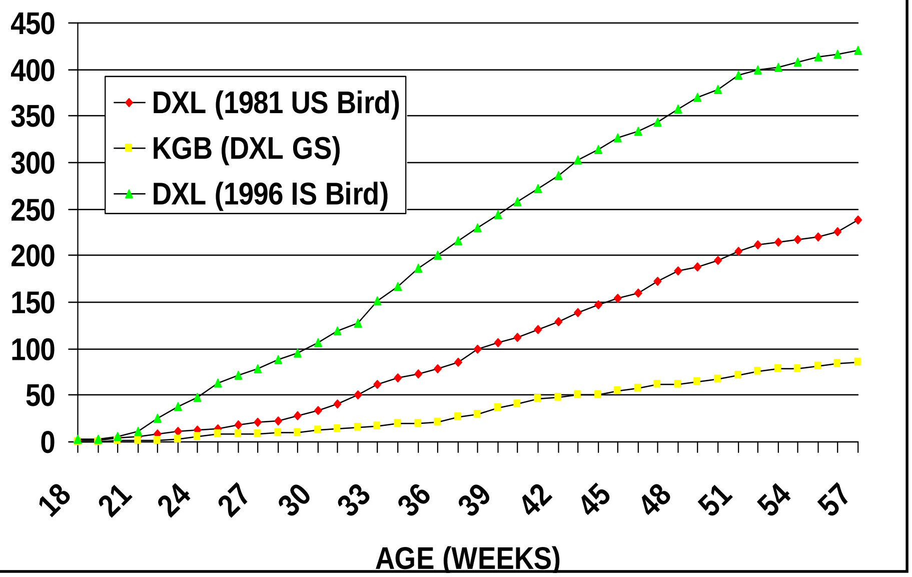


12 Alive

7th Generation (Craig and Muir, 1996)

- 3 Lines Were Compared
  - – Group Selected (KGB)
  - – Control (Dekalb, 1981)
  - – Individual Selection (Dekalb, 1996)
- Housed
  - – Single
  - – 12-bird Cages

Cumulative Mortality



Comparison of Kin, Individual, and TBV Selection

MUIR, W. M., 2005 Incorporation of competitive effects in forest tree or animal breeding programs. Genetics **170:** 1247-1259

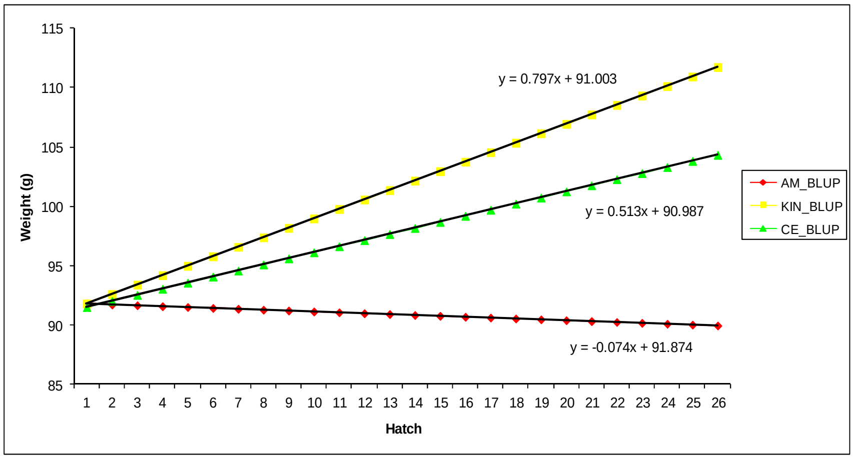
- Experimental Model
  - Quail
  - Trait: 6 Week Weight (wt)
- Methods tested:
  - Individual selection unrelated groups: (AM-BLUP)
  - Multi-level in related groups: (Kin-BLUP)
  - TBV Index direct and indirect (non-kin groups, CE-BLUP)
- Selected for 25 Hatches

Estimates of Genetic Parameters Based on Random Matings First 2 Generations

$$\mathbf{G} = \begin{bmatrix} \sigma_D^2 & \sigma_{AD} \\ \sigma_{AD} & \sigma_A^2 \end{bmatrix} = \begin{bmatrix} 33.7 & -5.5 \\ -5.5 & 2.8 \end{bmatrix}$$

$$\sigma_e^2 = 124.5$$

Genetic Trends



Conclusion: Selection on TBV

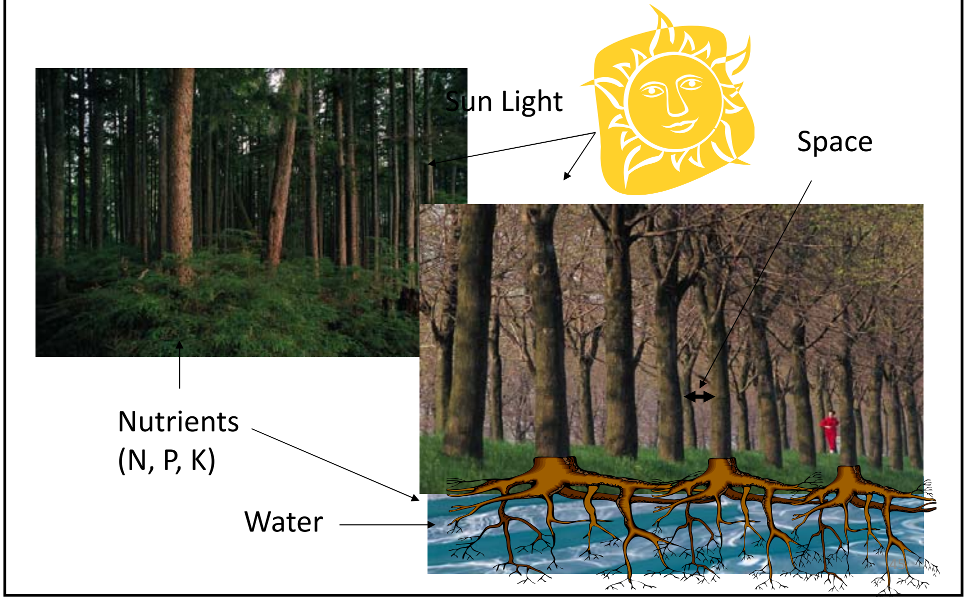
- – Effective but did not achieve theoretic gains
  - • Errors in parameter estimation
  - • Variances and covariance's change with selection
- – Implementation
  - • Management
    - – Easy to fill pens with same aged pigs (random)

Multi-level selection
individual selection in family groups

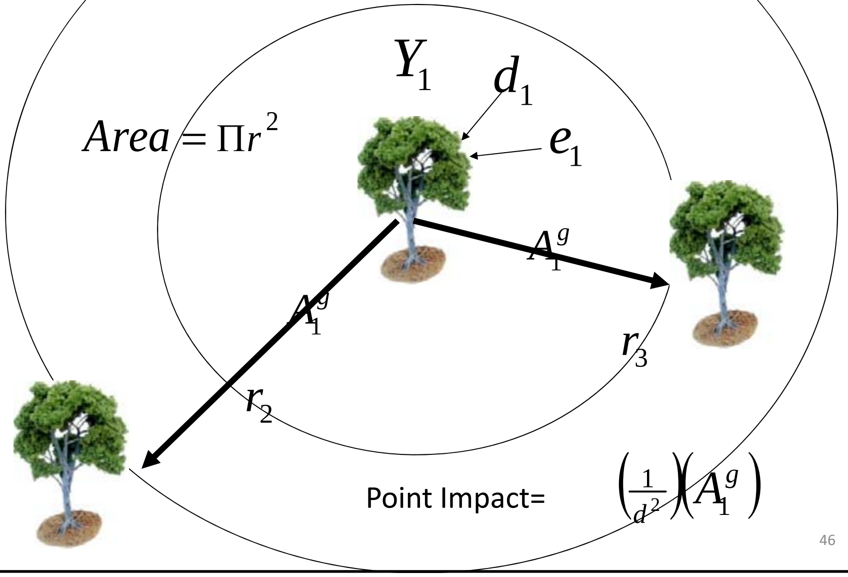
- Most Effective
- Achieved theoretic gains
  - Robust to errors in parameter estimation
  - No concerns for covariance's changing with selection
- Implementation
  - Programming : none (same model)
  - Management
    - Difficult
    - Filling of cages with same age and number of pigs

Competitive
Effects in Tree Breeding Programs
Competition by Distance

Trees Compete For Limited Resources

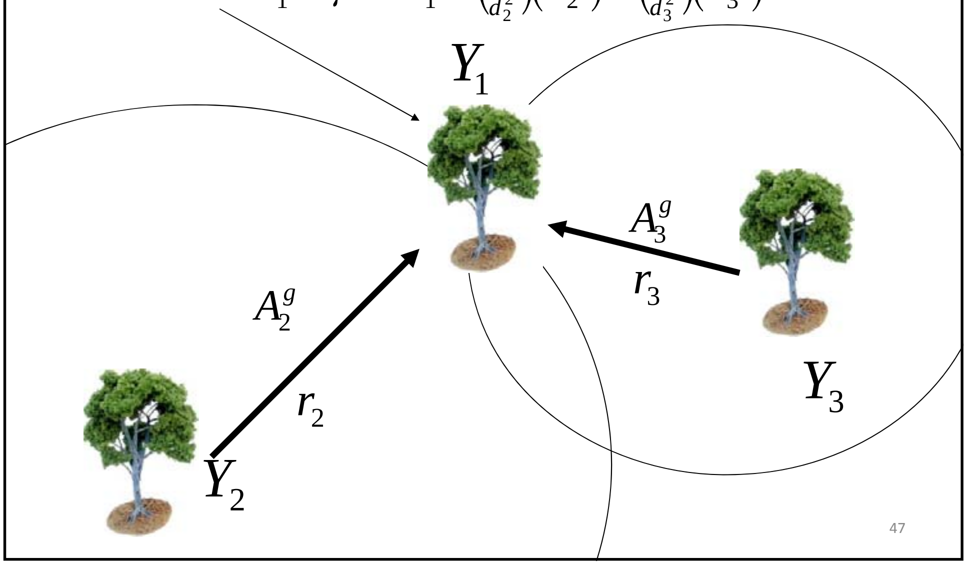


Circle of Influence


Y_1
d_1
$Area = \Pi r^2$
e_1
A_1^g
A_1^g
r_3
r_2
Point Impact= $\left(\frac{1}{d^2}\right)\left(A_1^g\right)$

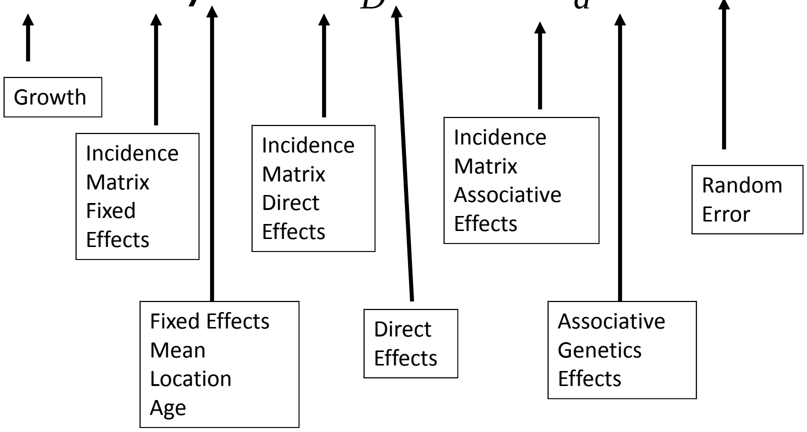
Interacting Genotype Model

$$Y_1 = \mu + D_1 + \left(\frac{1}{d_2^2}\right)\left(A_2^g\right) + \left(\frac{1}{d_3^2}\right)\left(A_3^g\right) + \varepsilon$$



Mixed Model Equations

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{Z}_D D + \mathbf{Z}_a a + \varepsilon$$



Growth

Incidence Matrix Fixed Effects

Fixed Effects Mean Location Age

Incidence Matrix Direct Effects

Direct Effects

Incidence Matrix Associative Effects

Associative Genetics Effects

Random Error

Associative Effects Incidence Matrix

$$\mathbf{Z} = \begin{bmatrix} 0 & \frac{1}{d_2^2} & \frac{1}{d_3^2} \\ \frac{1}{d_2^2} & 0 & \frac{1}{d_4^2} \\ \frac{1}{d_3^2} & \frac{1}{d_4^2} & 0 \end{bmatrix}$$

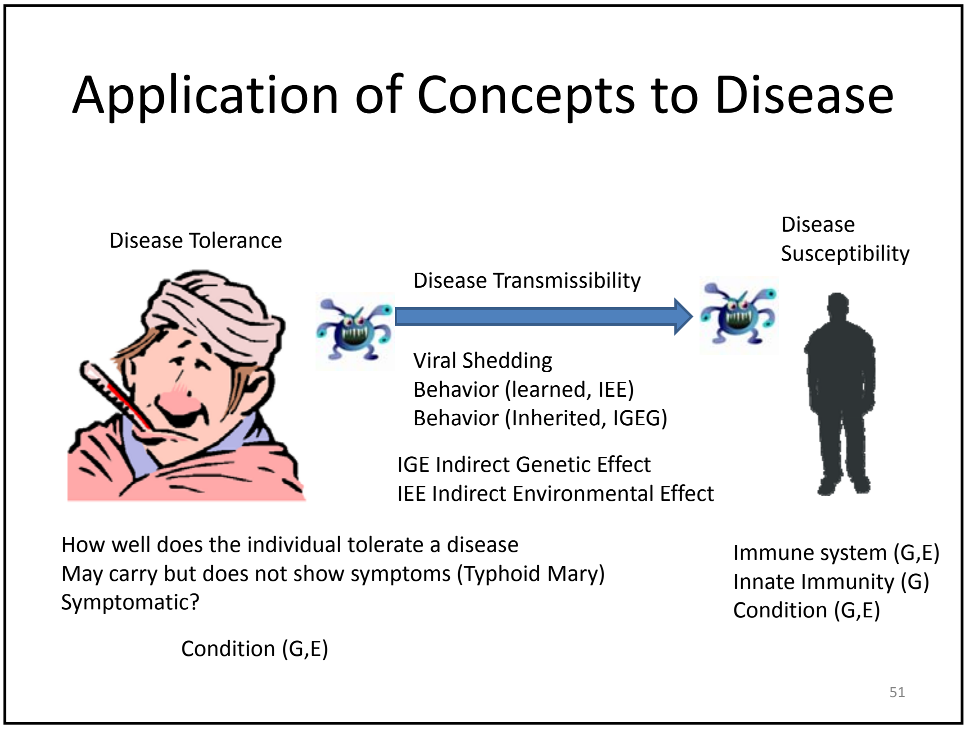
MME Competition

$$\begin{bmatrix} X'R^{-1}X & X'R^{-1}Z_d & X'R^{-1}Z_a \\ Z_dR^{-1}X' & Z_d'R^{-1}Z_d + k_1A^{-1} & Z_d'R^{-1}Z_a + k_2A^{-1} \\ Z_aR^{-1}X' & Z_a'R^{-1}Z_d + k_2A^{-1} & Z_a'R^{-1}Z_a + k_3A^{-1} \end{bmatrix} \begin{bmatrix} \beta \\ \mu_d \\ \mu_a \end{bmatrix} = \begin{bmatrix} X'R^{-1}X \\ X'R^{-1}Z_d \\ X'R^{-1}Z_a \end{bmatrix}$$

$$\begin{bmatrix} k_1 & k_2 \\ k_2 & k_3 \end{bmatrix} = \begin{bmatrix} \sigma_d^2 & \sigma_{ad} \\ \sigma_{ad} & \sigma_a^2 \end{bmatrix}^{-1}$$

Note that **R,** the residual covariance matrix, is a spatial correlation matrix and maybe defined similar to Za or by plot

$$R = \sigma_e^2 \begin{bmatrix} 1 & \frac{1}{d_2^2} & \frac{1}{d_3^2} \\ \frac{1}{d_2^2} & 1 & \frac{1}{d_4^2} \\ \frac{1}{d_3^2} & \frac{1}{d_4^2} & 1 \end{bmatrix}$$



Application of Concepts to Disease

Disease Tolerance

Disease Transmissibility

Disease Susceptibility

Viral Shedding
Behavior (learned, IEE)
Behavior (Inherited, IGEG)

IGE Indirect Genetic Effect
IEE Indirect Environmental Effect

How well does the individual tolerate a disease
May carry but does not show symptoms (Typhoid Mary)
Symptomatic?

Condition (G,E)

Immune system (G,E)
Innate Immunity (G)
Condition (G,E)

Lecture 8
QTL and Association Mapping with Mixed Models

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

QTL & Association mapping

- We would like to know both the genomic locations (map positions) and effects (either genotypic means or variances) for genes underlying quantitative trait variation
- QTL mapping
  - Using linkage information on a set of known relatives
- Association mapping
  - Using very fine scale LD to map genes in a set of random individuals from a population

Outline

- Basics of QTL mapping
  - Line crosses
    - typically fixed effects models
  - Outbred populations
    - Random effects family models
    - General pedigree methods
- High parameter models
  - Shrinkage approaches for detecting epistasis
- Association mapping

Inbred Line Cross QTL mapping

- Most powerful design
  - Cross two fully inbred lines, look at marker-trait segregation in the F_2 (or other, such as F_n) generations
  - P1: MMQQ, P2:mmqq
  - All F_1 same genotype/phase: MQ/mq
  - Hence, in the F1, all parents have the same genotype
  - At most only two alleles, each with freq 1/2
  - Idea: Does the mean trait value of (say) MM individuals differ from (say) mm
    - Different marker genotypes have different mean trait values

Expected Marker Means

The expected trait mean for marker genotype M_j is just

$$\mu_{M_j} = \sum_{k=1}^{N} \mu_{Q_k} \Pr(\, Q_k \,|\, M_j \,)$$

For example, if QQ = 2a, Qq = a(1+k), qq = 0, then in the F2 of an MMQQ/mmqq cross,

$$(\mu_{MM} - \mu_{mm})/2 = a(1-2c)$$

- If the trait mean is significantly different for the genotypes at a marker locus, it is linked to a QTL

- A small MM-mm difference could be (i) a tightly-linked QTL of small effect or (ii) loose linkage to a large QTL

Linear Models for QTL Detection

The use of differences in the mean trait value for different marker genotypes to detect a QTL and estimate its effects is a use of linear models.

One-way ANOVA.

Value of trait in kth individual of marker genotype i

$$z_{ik} = \mu + b_i + e_{ik}$$

Effect of marker genotype i on trait value

$$z_{ik} = \mu + b_i + e_{ik}$$

Detection: a QTL is linked to the marker if at least one of the b_i is significantly different from zero

Estimation: (QTL effect and position): This requires relating the b_i to the QTL effects and map position

Detecting epistasis

One major advantage of linear models is their flexibility. To test for epistasis between two QTLs, use ANOVA with an interaction term

$$z = \mu + a_i + b_k + d_{ik} + e$$

Effect from marker genotype at first marker set (can be > 1 loci)

Effect from marker genotype at second marker set

Interaction between marker genotypes i in 1st marker set and k in 2nd marker set

Detecting epistasis

$$z = \mu + a_i + b_k + d_{ik} + e$$

- At least one of the a_i significantly different from 0
---- QTL linked to first marker set

- At least one of the b_k significantly different from 0
---- QTL linked to second marker set

- At least one of the d_{ik} significantly different from 0
---- interactions between QTL in sets 1 and two

Problem: Huge number of potential interaction terms (order m^2, where m = number of markers)

Model selection

- With (say) 300 markers, we have (potentially) 300 single-marker terms and 300*299/2 = 44,850 epistatic terms
  - Hence, a model with up to p= 45,150 possible parameters
  - 2^p possible submodels = $10^{13,600}$ ouch!
- The issue of Model selection becomes very important.
- How do we find the best model?
  - Stepwise regression approaches
    - Forward selection (add terms one at a time)
    - Backwards selection (delete terms one at a time)
  - Try all models, assess best fit
  - Mixed-model approaches (Stochastic Search Variable Selection, or SSVS)

Model Selection

Model Selection: Use some criteria to chose among a number of candidate models. Weight goodness-of-fit (L, value of the likelihood at the MLEs) vs. number of estimated parameters (k)

AIC = Akaike's information criterion
AIC = 2k - 2 Ln(L)

BIC = Bayesian information criterion (Schwarz criterion)
BIC = k*ln(n)/n - 2 Ln(L)/n
BIC penalizes free parameters more strongly than AIC

Other measures. For these (and AIVC, BIC) smaller score indicates better model fit

Model averaging

Model averaging: Generate a composite model by weighting (averaging) the various models, using AIC, BIC, or other

Idea: Perhaps no "best" model, but several models all extremely close. Better to report this "distribution" rather than the best one

One approach is to average the coefficients on the "best-fitting" models using some scheme to return a composite model

Supersaturated Models

A problem with many QTL approaches is that there are far more parameters (p) to estimate than there are independent samples (n). Case in point: epistasis

Such supersaturated models arise commonly in Genomics. How do we deal with them?

One approach is to have all parameters included, but some are shrunk back (regressed) towards zero by assigning them a very small posterior variance

Shrinkage estimators

Shrinkage estimates: Rather than adding interaction terms one at a time, a shrinkage method starts with all interactions included, and then shrinks most back to zero.

Under a Bayesian analysis, any effect is *random*. One can assume the effect for (say) interaction *ij* is drawn from a normal with mean zero and variance σ^2_{ij}

Further, the interaction-specific variances are themselves random variables drawn from a hyperparameter distribution, such as an inverse chi-square.

One then estimates the hyperparameters and uses these to predict the variances, with effects with small variances shrinking back to zero, and effects with large variances remaining in the model.

What is a "QTL"

- A detected "QTL" in a mapping experiment is a region of a chromosome detected by linkage.
- Usually large (typically 10-40 cM)
- When further examined, most "large" QTLs turn out to be a linked collection of locations with increasingly smaller effects
- The more one localizes, the more subregions that are found, and the smaller the effect in each subregion
- This is called fractionation

Limitations of QTL mapping

- Poor resolution (~20 cM or greater in most designs with sample sizes in low to mid 100's)
  - Detected "QTLs" are thus large chromosomal regions
- Fine mapping requires either
  - Further crosses (recombinations) involving regions of interest (i.e., RILs, NILs)
  - Enormous sample sizes
    - If marker-QTL distance is 0.5cM, require sample sizes in excess of 3400 to have a 95% chance of 10 (or more) recombination events in sample
    - 10 recombination events allows one to separate effects that differ by ~ 0.6 SD

Limitations of QTL mapping (cont)

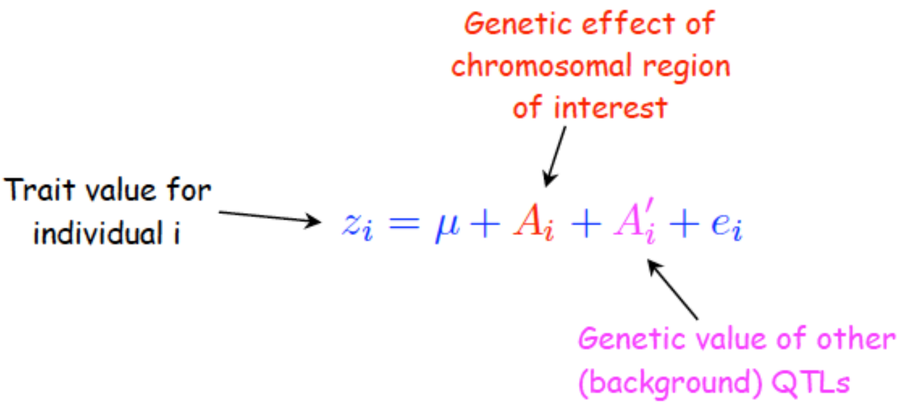
- "Major" QTLs typically fractionate
  - QTLs of large effect (accounting for > 10% of the variance) are routinely discovered.
  - However, a large QTL peak in an initial experiment generally becomes a series of smaller and smaller peaks upon subsequent fine-mapping.
- The Beavis effect:
  - When power for detection is low, marker-trait associations declared to be statistically significant significantly overestimate their true effects.
  - This effect can be very large (order of magnitude) when power is low.

Outbred populations

- When we move from the simple framework of an inbred line cross QTL design to a set of parents from an outbred population, complications arise as the parents don't all have the same genotypes
  - Differences in linkage phase
  - Many uninformative as to linkage (varies over makers)
  - Possibility of multiple alleles
- Result: express marker effects in terms of the variance in trait value it explains, rather than in terms of mean marker effects

General Pedigree Methods

Random effects (hence, variance component) method for detecting QTLs in general pedigrees



Genetic effect of chromosomal region of interest

Trait value for individual i

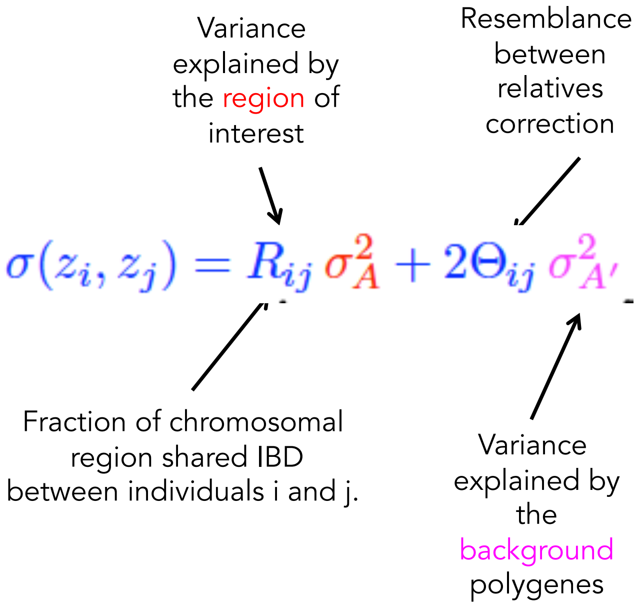
$$z_i = \mu + A_i + A'_i + e_i$$

Genetic value of other (background) QTLs

The model is rerun for each marker

$$z_i = \mu + A_i + A'_i + e_i$$

The covariance between individuals i and j is thus



Variance explained by the region of interest

Resemblance between relatives correction

$$\sigma(z_i, z_j) = R_{ij}\,\sigma^2_A + 2\Theta_{ij}\,\sigma^2_{A'}$$

Fraction of chromosomal region shared IBD between individuals i and j.

Variance explained by the background polygenes

Assume z is MVN, giving the covariance matrix as

$$\mathbf{V} = \mathbf{R}\,\sigma_A^2 + \mathbf{A}\,\sigma_{A'}^2 + \mathbf{I}\,\sigma_e^2$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker data

Estimated from the pedigree

The resulting likelihood function is

$$\ell(\mathbf{z} \,|\, \mu, \sigma_A^2, \sigma_{A'}^2, \sigma_e^2) = \frac{1}{\sqrt{(2\pi)^n \,|\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{z} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{z} - \boldsymbol{\mu})\right]$$

A significant σ_A^2 indicates a linked QTL.

Association & LD mapping

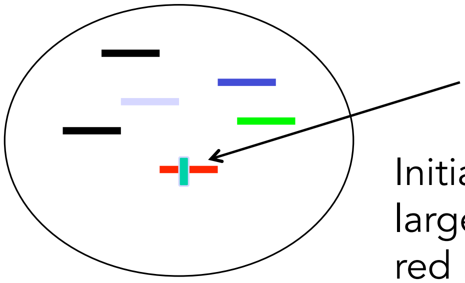
Mapping major genes (LD mapping) vs. trying to Map QTLs (Association mapping)

Idea: Collect random sample of individuals, contrast trait means over marker genotypes

If a dense enough marker map, likely population level linkage disequilibrium (LD) between closely-linked genes

Fine-mapping genes

Suppose an allele causing an effect on the trait arose as a single mutation in a closed population



New mutation arises on red chromosome

Initially, the new mutation is largely associated with the red haplotype

Hence, markers that define the red haplotype are likely to be associated (i.e. in LD) with the mutant allele

Background: Association mapping

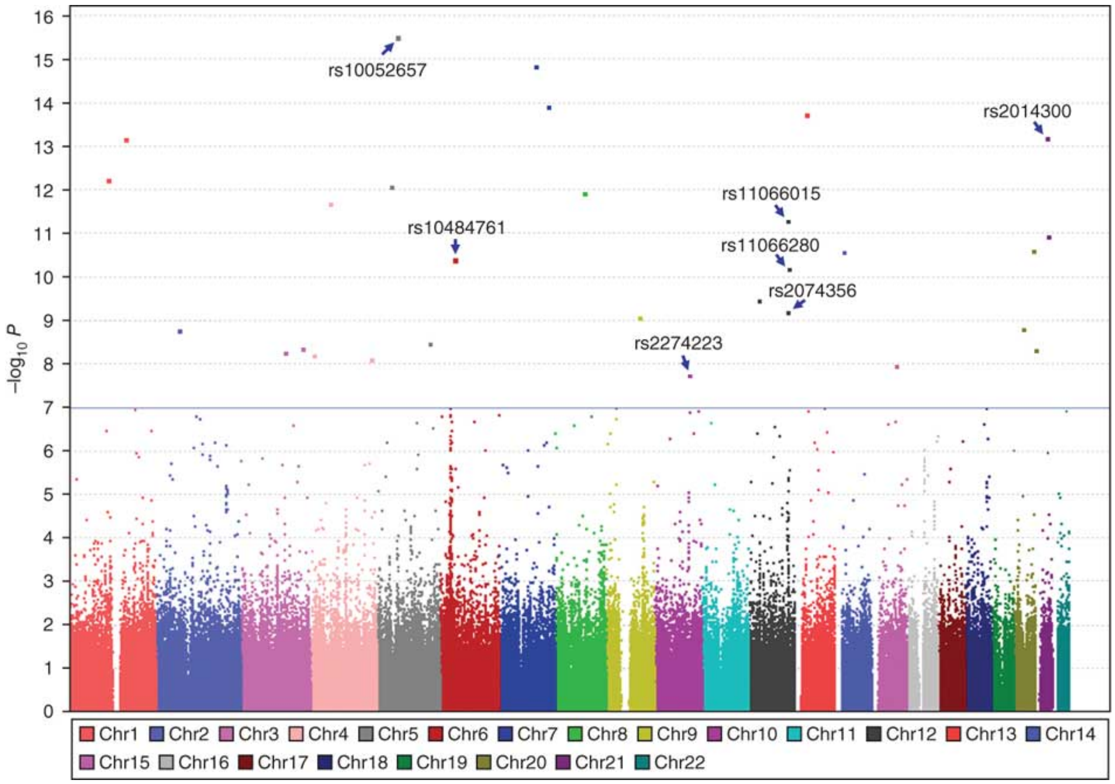
- If one has a very large number of SNPs, then new mutations (such as those that influence a trait) will be in LD with very close SNPs for hundreds to thousands of generations, generating a marker-trait association.
  - Association mapping looks over all sets of SNPs for trait -SNP associations. GWAS = genome-wide association studies.
  - This is also the basis for genomic selection
- Main point from extensive human association studies
  - Almost all QTLs have very small effects
  - Marker-trait associations do not fully recapture all of the additive variance in the trait (due to incomplete LD)
  - This has been called the "missing heritability problem" by human geneticists, but not really a problem at all (more shortly).

Association mapping

- Marker-trait associations within a population of unrelated individuals
- Very high marker density (~ 100s of markers/cM) required
  - Marker density no less than the average track length of linkage disequilibrium (LD)
- Relies on very slow breakdown of initial LD generated by a new mutation near a marker to generate marker-trait associations
  - LD decays very quickly unless very tight linkage
  - Hence, resolution on the scale of LD in the population(s) being studied (1 ~ 40 kB)
- Widely used since mid 1990's. Mainstay of human genetics, strong inroads in breeding, evolutionary genetics
- Power a function of the genetic variance of a QTL, not its mean effects

Manhattan plots

- The results for a Genome-wide Association study (or GWAS) are typically displayed using a Manhattan plot.
  - At each SNP, -ln(p), the negative log of the p value for a significant marker-trait association is plotted. Values above a threshold indicate significant effects
  - Threshold set by Bonferroni-style multiple comparisons correction
  - With n markers, an overall false-positive rate of p requires each marker be tested using p/n.
  - With n = 10^6 SNPs, p must exceed $0.01/10^6$ or 10^{-8} to have a control of 1% of a false-positive



Population Stratification

When population being sampled actually consists of several distinct subpopulations we have lumped together, marker alleles may provide information as to which group an individual belongs. If there are other risk factors in a group, this can create a false association btw marker and trait

Example. The Gm marker was thought (for biological reasons) to be an excellent candidate gene for diabetes in the high-risk population of Pima Indians in the American Southwest. Initially a very strong association was observed:

Gm^+	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Gm$^+$	Total	% with diabetes
Present	293	8%
Absent	4,627	29%

Problem: freq(Gm$^+$) in Caucasians (lower-risk diabetes Population) is 67%, Gm$^+$ rare in full-blooded Pima

The association was re-examined in a population of Pima that were 7/8th (or more) full heritage:

Gm$^+$	Total	% with diabetes
Present	17	59%
Absent	1,764	60%

Linkage vs. Association

The distinction between linkage and association is subtle, yet critical

Marker allele M is associated with the trait if

$\text{Cov}(M,y) \neq 0$

While such associations can arise via linkage, they can also arise via population structure.

Thus, association DOES NOT imply linkage, and linkage is not sufficient for association

Accounting for population structure

- Three classes of approaches proposed
  - 1) Attempts to correct for common pop structure signal (regression/PC methods)
  - 2) Attempts to first assign individuals into subpopulations and then perform association mapping in each set (Structure)
  - 3) Mixed models that use all of the marker information (Tassle, EMMA, many others)
    - These can also account for cryptic relatedness in the data set, which also causes false-positives.

Regression Approaches

One approach to control for structure is simply to include a number of markers, outside of the SNP of interest, chosen because they are expected to vary over any subpopulations

How might you choose these in a sample? Try those markers (read STRs) that show the largest departure from Hardy-Weinberg, as this is expected in markers that vary the most over subpopulations.

Indicator (0 / 1) Variable
for SNP genotype k. Typically
k = 3, i.e. AA, Aa aa

$$y = \mu + \sum_{k=1}^{n} \beta_k \, M_k + \sum_{j=1}^{m} \gamma_j \, b_j + e$$

Significant β indicates
marker-trait association

m unlinked markers that
vary across subpopulations.
b_j = marker genotype indicator
variable

SNP marker
under consideration

Variations on this theme (eigenstrat) --- use all of the marker information to extract a set of significant PCs, which are then included in the model as cofactors

Structured Association Mapping

Pritchard and Rosenberg (1999) proposed Structured Association Mapping, wherein one assumes k subpopulations (each in Hardy-Weinberg).

Given a large number of markers, one then attempts to assign individuals to groups using an MCMC Bayesian classifier

Once individuals assigned to groups, association mapping without any correction can occur in each group.

Mixed-model approaches

- Mixed models use marker data to
  - Account for population structure
  - Account for cryptic relatedness
- Three general approaches:
  - Treat a single SNP as fixed
    - TASSLE, EMMA
  - Treat a single SNP as random
    - General pedigree method
  - Fit all of the SNPs at once as random
    - GBLUP

Structure plus Kinship Methods

Association mapping in plants offer occurs by first taking a large collection of lines, some closely related, others more distantly related. Thus, in addition to this collection being a series of subpopulations (derivatives from a number of founding lines), there can also be additional structure within each subpopulation (groups of more closely related lines within any particular lineage).

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

Fixed effects in blue, random effects in red

This is a mixed-model approach. The program TASSEL runs this model.

Q-K method

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

β = vector of fixed effects

a = SNP effects (fits SNPs one at a time)

v = vector of subpopulation effects (STRUCTURE)
Q_{ij} = Prob(individual i in group j). Determined
from STRUCTURE output

u = shared polygenic effects due to kinship.
Cov(u) = var(A)*A, where the relationship matrix
A estimated from marker data matrix K, also called a
GRM – a genomic relationship matrix

Which markers to include in K?

- Best approach is to leave out the marker being tested (and any in LD with it) when construction the genomic relationship matrix
  - LOCO approach – leave out one chromosome (which the tested marker is linked to)
- Best approach seems to be to use most of the markers
- Other mixed-model approaches along these lines

Treat Single SNP as random: General Pedigree method

$$\mathbf{V} = \mathbf{R}\,\sigma_A^2 + \mathbf{A}\,\sigma_{A'}^2 + \mathbf{I}\,\sigma_e^2$$

Here

$$\mathbf{R}_{ij} = \begin{cases} 1 & \text{for } i = j \\ \widehat{R}_{ij} & \text{for } i \neq j \end{cases}, \qquad \mathbf{A}_{ij} = \begin{cases} 1 & \text{for } i = j \\ 2\Theta_{ij} & \text{for } i \neq j \end{cases}$$

Estimated from marker data

Estimated from the pedigree

The resulting likelihood function is

$$\ell(\mathbf{z}\,|\,\mu, \sigma_A^2, \sigma_{A'}^2, \sigma_e^2) = \frac{1}{\sqrt{(2\pi)^n\,|\mathbf{V}|}} \exp\left[-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})^T\,\mathbf{V}^{-1}\,(\mathbf{z}-\boldsymbol{\mu})\right]$$

A significant σ_A^2 indicates a linked QTL.

GBLUP

- The Q-K method tests SNPs one at a time, treating them as fixed effects
- The general pedigree method (slides 24-26) also tests one marker at a time, treating them as random effects
- Genomic selection can be though of as estimating all of the SNP effects at once and hence can also be used for GWAS

BLUP, GBLUP, and GWAS

- Pedigree information gives EXPECTED value of shared sites (i.e., ½ for full-sibs)
  - A matrix in BLUP
  - The actual realization of the fraction of shared genes for a particular pair of relatives can be rather different, due to sampling variance in segregation of alleles
  - GRM (or K or marker matrix M)
  - Hence "identical" relatives can differ significantly in faction of shared regions
  - Dense marker information can account for this

The general setting

- Suppose we have n measured individuals (the n x 1 vector **y** of trait values)
- The n x n relationship matrix **A** gives the relatedness among the sampled individuals, where the elements of **A** are obtained from the pedigree of measured individuals
- We may also have p (>> n) SNPs per individual, where the n x p marker information matrix **M** contains the marker data, where M_{ij} = score for SNP j (i.e., 0 for 00, 1 for 10, 2 for 11) in individual i.

Covariance structure of random effects

- A critical element specifying the mixed model is the covariance structure (matrix) of the vector **u** of random effects
- Standard form is that Cov(**u**) = variance component * matrix of known constants
  - This is the case for pedigree data, where **u** is typically the vector of breeding values, and the pedigree defines a relationship matrix **A**, with Cov(**u**) = Var(A) * **A,** the additive variance times the relationship matrix
  - With marker data, the covariance of random effects are functions of the marker information matrix **M**.
    - If **u** is the vector of p marker effects, then Cov(**u**) = Var(m) * $\mathbf{M}^T\mathbf{M}$, the marker variance times the covariance structure of the markers.

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

Pedigree-based BV estimation: (BLUP)
u_{nx1} = vector of BVs, Cov(u) = Var(A) $\mathbf{A}_{nxn}$

Marker-based BV estimation: (GBLUP)
u_{nx1} = vector of BVs, Cov(u) = Var(m) $\mathbf{M}^T\mathbf{M}$ (n x n)

GWAS: u_{px1} = vector of marker effects,
Cov(u) = Var(m) $\mathbf{M}\mathbf{M}^T$ (p x p)

Genomic selection: predicted vector of breeding values from marker effects, $GBV_{nx1} = \mathbf{M}_{nxp}u_{px1}$.
Note that Cov(GBV) = Var(m) $\mathbf{M}^T\mathbf{M}$ (n x n)

Lots of variations of these general ideas by adding additional assumptions on covariance structure.

GWAS Model diagnostics

The “Genomic Control” parameter λ

Devlin and Roeder (1999). Basic idea is that association tests (marker presence/absence vs. trait presence/absence) is typically done with a standard 2 x 2 χ^2 test.

When population structure is present, the test statistic now follows a scaled χ^2, so that if S is the test statistic, then $S/\lambda \sim \chi^2_1$ (so $S \sim \lambda\chi^2_1$) . Hence, population structure should inflate all of the tests (on average) by a common amount λ.

Hence, if we have suitably corrected for population structure, the estimated inflation factor λ among tests should be ~ 1.

A robust estimator for λ is offered from the medium (50% value) of the test statistics, so that for m tests

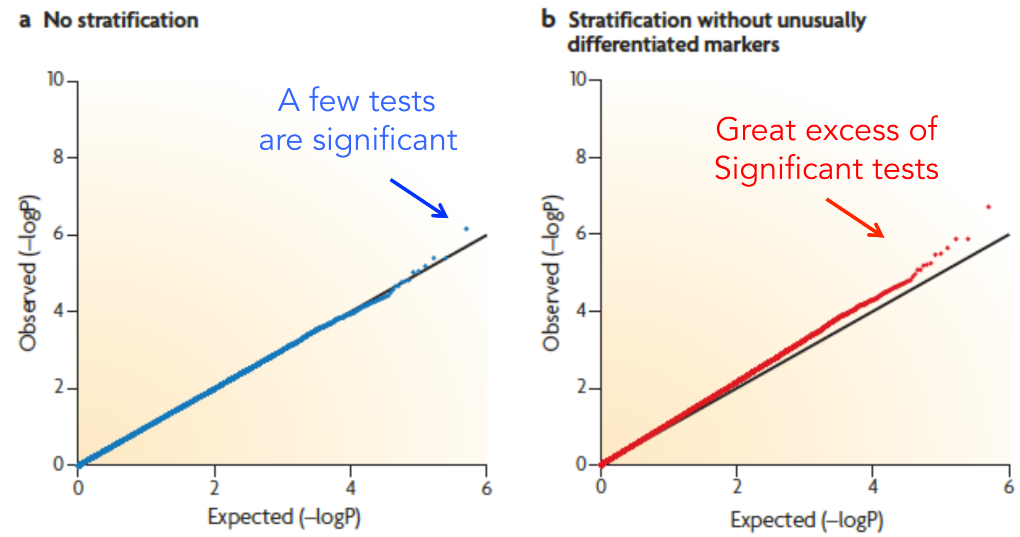
$$\widehat{\lambda} = \frac{\text{medium}(S_1, \cdots; S_m)}{0.456}$$

Genomic control λ as a diagnostic tool

- Presence of population structure will inflate the λ parameter
- A value above 1 is considered evidence of additional structure in the data
  - Could be population structure, cryptic relatedness, or both
  - A lambda value less that 1.05 is generally considered benign
- One issue is that if the true polygenic model holds (lots of sites of small effect), then a significant fraction will have inflated p values, and hence an inflated λ value.
- Hence, often one computes the λ following attempts to remove population structure. If the resulting value is below 1.05, suggestion that structure has been largely removed.

P – P plots

- Another powerful diagnostic tool is the p-p plot.
- If all tests are drawn from the null, then the distribution of p values should be uniform.
  - There should be a slight excess of tests with very low p indicating true positives
- This gives a straight line of a log-log plot of observed (seen) and expected (uniform) p values with a slight rise near small values
  - If the fraction of true positives is high (i.e., many sites influence the trait), this also bends the p-p plot



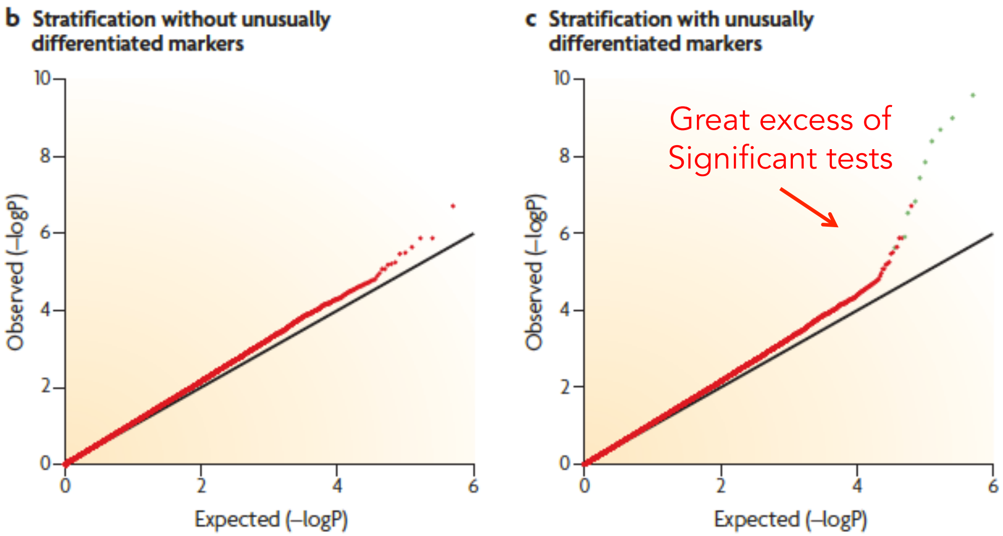
a No stratification

b Stratification without unusually differentiated markers

A few tests are significant

Great excess of Significant tests

Price et al. 2010 Nat Rev Gene 11: 459



b Stratification without unusually differentiated markers

c Stratification with unusually differentiated markers

Great excess of Significant tests

As with using λ, one should construct p-p following some approach to correct for structure & relatedness to see if they look unusual.

Association mapping (power)

Q/q is the polymorphic site contributing to trait variation, M/m alleles (at a SNP) used as a marker

Let p be the frequency of M, and assume that Q only resides on the M background (complete disequilibrium)

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Haloptype	Frequency	effect
QM	rp	a
qM	(1-r)p	0
qm	1-p	0

Effect of m = 0

Effect of M = ar

Genetic variation associated with Q = $2(rp)(1-rp)a^2$ ~ $2rpa^2$ when Q rare. Hence, little power if Q rare

Genetic variation associated with marker M is $2p(1-p)(ar)^2$ ~ $2pa^2r^2$

Ratio of marker/true effect variance is ~ r

Hence, if Q rare within the A class, even less power, as M only captures a fraction of the associated QTL.

Common variants

- Association mapping is only powerful for common variants
  - freq(Q) moderate
  - freq (r) of Q within M haplotypes modest to large
- Large effect alleles (a large) can leave small signals.
- The fraction of the actual variance accounted for by the markers is no greater than ~ ave(r), the average frequency of Q within a haplotype class
- Hence, don't expect to capture all of Var(A) with markers, esp. when QTL alleles are rare but markers are common (e.g. common SNPs, p > 0.05)
- Low power to detect G x G, G x E interactions

"How wonderful that we have met with a paradox. Now we have some hope of making progress" -- Neils Bohr

The case of the missing heritability

Infamous figure from *Nature* on the angst of human geneticists over the finding that all of their discovered SNPs still accounted for only a fraction of relative-based heritability estimates of human disease.

The "missing heritability" pseudo paradox

- A number of GWAS workers noted that the sum of their significant marker variances was much less (typically 10%) than the additive variance estimated from biometrical methods
- The "missing heritability" problem was birthed from this observation.
- Not a paradox at all
  - Low power means small effect (i.e. variance) sites are unlikely to be called as significant, esp. given the high stringency associated with control of false positives over tens of thousands of tests
  - Further, even if all markers are detected, only a fraction ~ r (the frequency of the causative site within a marker haplotype class) of the underlying variance is accounted for.

Lecture 9: G x E: Genotype-environment interaction

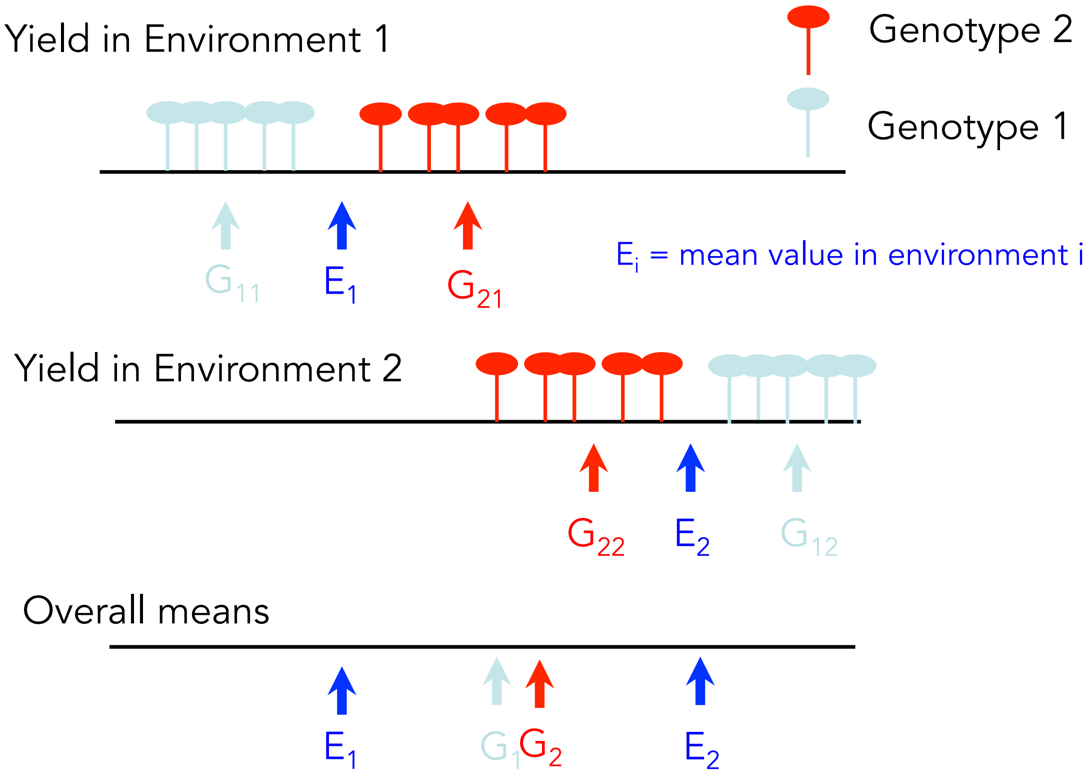
Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

G x E

- Introduction to G x E
  - Basics of G x E
  - Some suggested rules
  - Treating G x E as a correlated-trait problem
- Estimation of G x E terms
  - Finlay-Wilkinson regressions
- SVD-based methods
  - The singular value decomposition (SVD)
  - AMMI models
- Factorial regressions
- Mixed-Model approaches
  - BLUP
  - Structured covariance models

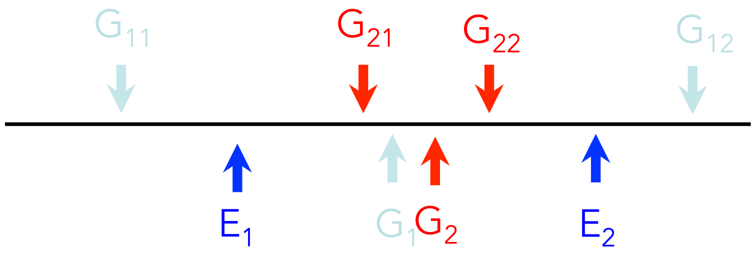
Genotypes vs. individuals

- Much of the G x E theory is developed for plant breeders who are using pure (= fully inbred) lines, so that every individual has the same genotype
- The same basic approaches can be used by taking family members as the replicates for outbred species. Here the “genotype” over the family members is some composite value (the mean breeding value of the family).



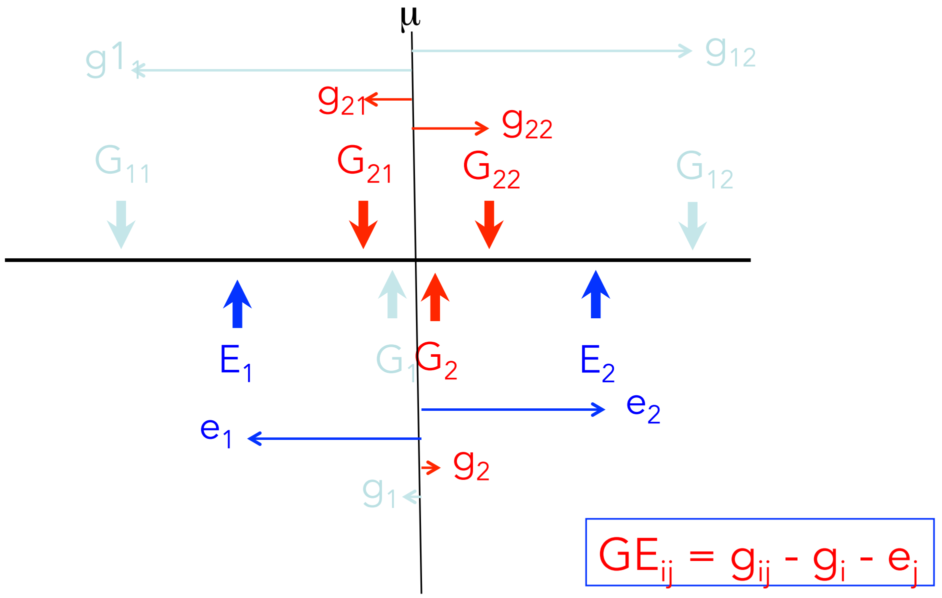
E_i = mean value in environment i

G_{ij} = mean of genotype i in environment j



Under base model of Quantitative Genetics,
$G_{ij} = \mu + G_i + E_j$

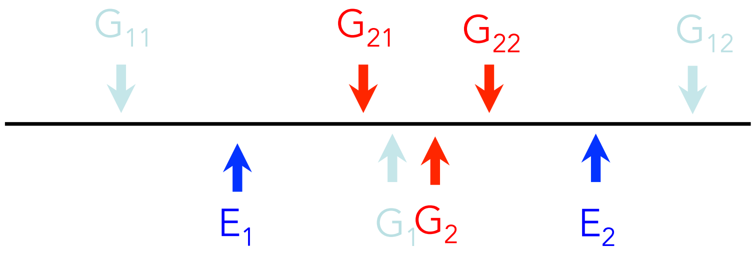
When G x E present, there is an interaction between a particular genotype and a particular environment so that G_{ij} is no longer additive, $G_{ij} = \mu + G_i + E_i + GE_{ij}$



$GE_{ij} = g_{ij} - g_i - e_j$

Components measured as deviations
from the mean μ

Which genotype is the best?



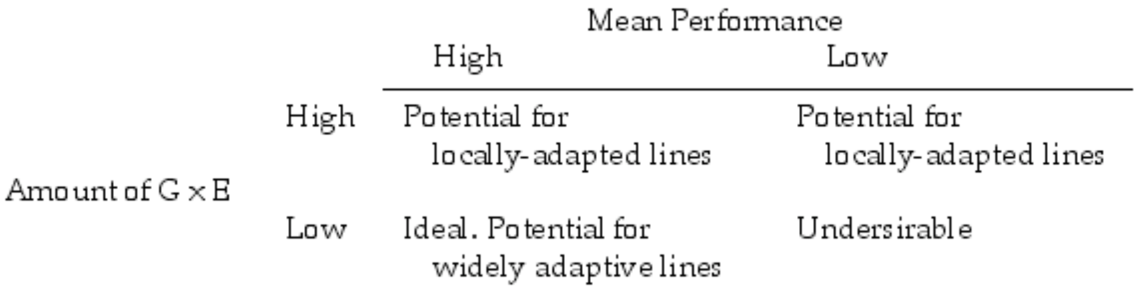
Depends: If the genotypes are grown in both environments, G_2 has a higher mean

If the genotypes are only grown in environment 1, G_2 has a higher mean

If the genotypes are only grown in environment 2, G_1 has a higher mean

G x E: Both a problem and an opportunity

- A line with little G x E has stability across environments.
- However, a line with high G x E may outperform all others in specific environments.
- G x E implies the opportunity to fine-tune specific lines to specific environments
- High $\sigma^2(GE)$ implies high G x E in at least some lines in the sample.



		Mean Performance	
		High	Low
Amount of G x E	High	Potential for locally-adapted lines	Potential for locally-adapted lines
	Low	Ideal. Potential for widely adaptive lines	Undersirable

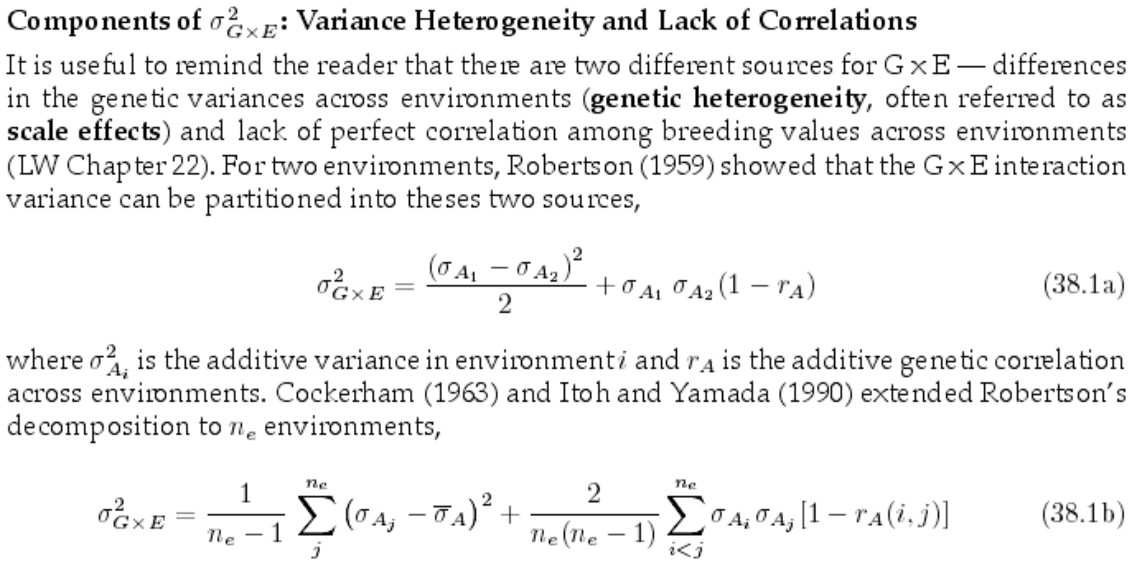
Ideal: high mean performance, low G x E

Low G x E = widely adaptive lines/genotypes

High G x E = locally adaptive lines/genotypes

Major vs. minor environments

- An identical genotype will display slightly different traits values even over apparently identical environments due to micro-environmental variation and developmental noise
- However, macro-environments (such as different locations or different years <such as a wet vs. a dry year>) can show substantial variation, and genotypes (pure lines) may differentially perform over such macro-environments (G x E).
- Problem: The mean environment of a location may be somewhat predictable (e.g., corn in the tropics vs. temperate North American), but year-to-year variation at the same location is essentially unpredictable.
- Decompose G x E into components
  - $G \times E_{\text{locations}} + G \times E_{\text{years}} + G \times E_{\text{years x locations}}$
  - Ideal: strong G x E over locations, high stability over years.



Components of $\sigma^2_{G\times E}$: Variance Heterogeneity and Lack of Correlations

It is useful to remind the reader that there are two different sources for G × E — differences in the genetic variances across environments (**genetic heterogeneity**, often referred to as **scale effects**) and lack of perfect correlation among breeding values across environments (LW Chapter 22). For two environments, Robertson (1959) showed that the G × E interaction variance can be partitioned into theses two sources,

$$\sigma^2_{G\times E} = \frac{(\sigma_{A_1} - \sigma_{A_2})^2}{2} + \sigma_{A_1}\,\sigma_{A_2}(1 - r_A) \tag{38.1a}$$

where $\sigma^2_{A_i}$ is the additive variance in environment i and r_A is the additive genetic correlation across environments. Cockerham (1963) and Itoh and Yamada (1990) extended Robertson's decomposition to n_e environments,

$$\sigma^2_{G\times E} = \frac{1}{n_e - 1}\sum_j^{n_e}\left(\sigma_{A_j} - \overline{\sigma}_A\right)^2 + \frac{2}{n_e(n_e-1)}\sum_{i<j}^{n_e}\sigma_{A_i}\sigma_{A_j}\left[1 - r_A(i,j)\right] \tag{38.1b}$$

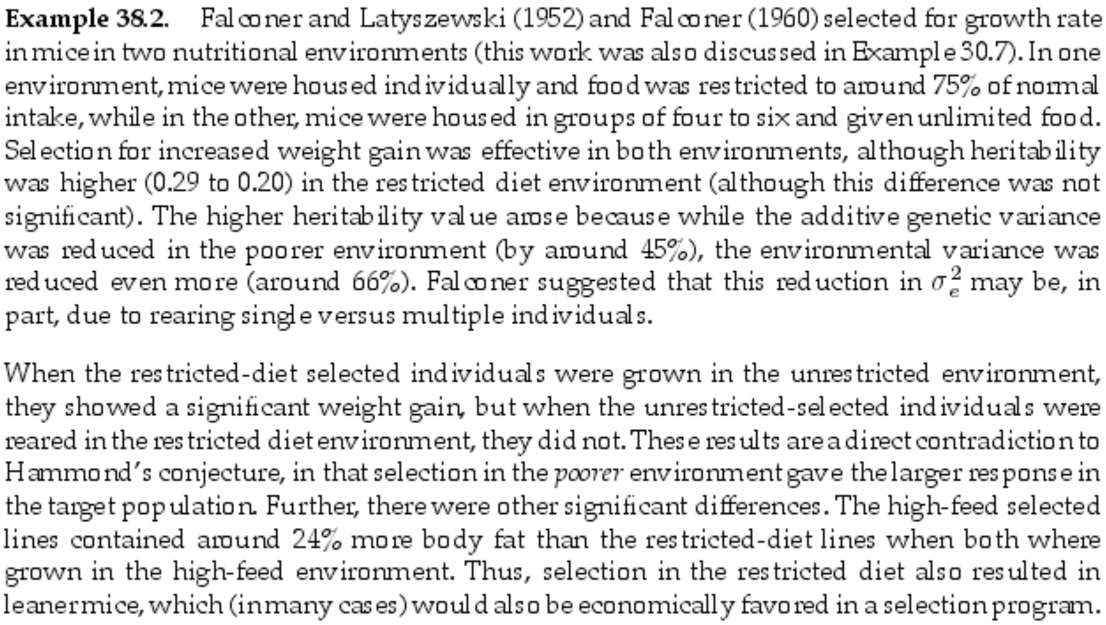
Key: differences in scale and lack of perfect correlation over environments both generate G x E

Falconer: G x E

- The modern treatment of G x E starts with Falconer (1952)
  - Measures of the same trait in different environments are correlated traits
  - Hence, if measured in k environments, it's a k-dimensional trait
  - Thus results from direct and correlated responses apply to selection on G x E
- If selection in environment i, expected change in environment j is
  - $CR_j = i_i\, h_i\, h_j\, r_A\, \sigma_P\,(j)$

Hammond’s Conjecture

- Hammond (1947) suggested that selection be undertaken in a more favorable environment to maximize progress in a less favorable one.
- Idea: perhaps more genetic variation, and hence greater discrimination, between genotypes.
- Downside: don’t know if Var(G) greater in ”better” environments. Even if it is, between-environment correlation can be small.

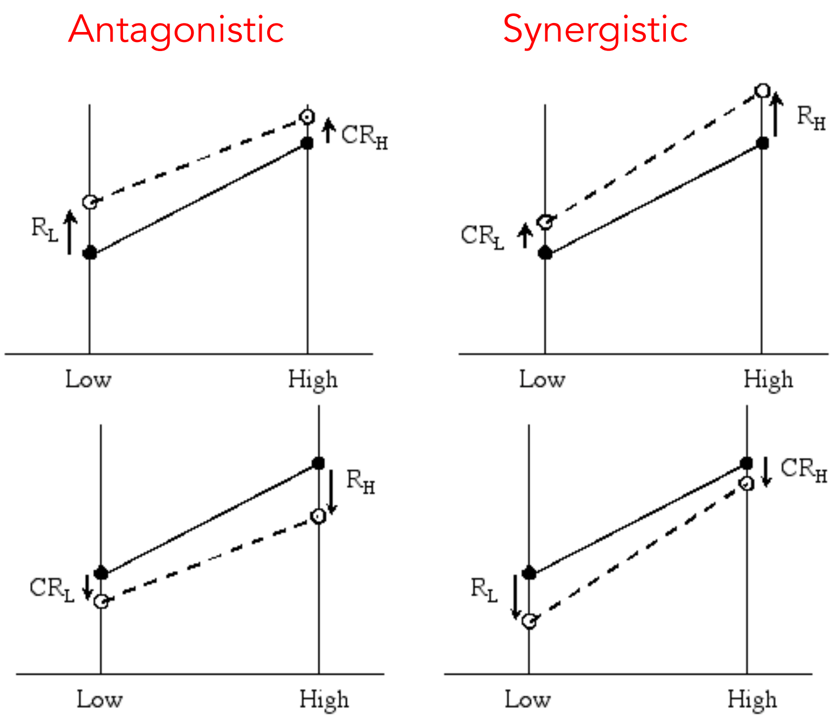


Example 38.2. Falconer and Latyszewski (1952) and Falconer (1960) selected for growth rate in mice in two nutritional environments (this work was also discussed in Example 30.7). In one environment, mice were housed individually and food was restricted to around 75% of normal intake, while in the other, mice were housed in groups of four to six and given unlimited food. Selection for increased weight gain was effective in both environments, although heritability was higher (0.29 to 0.20) in the restricted diet environment (although this difference was not significant). The higher heritability value arose because while the additive genetic variance was reduced in the poorer environment (by around 45%), the environmental variance was reduced even more (around 66%). Falconer suggested that this reduction in σ_e^2 may be, in part, due to rearing single versus multiple individuals.

When the restricted-diet selected individuals were grown in the unrestricted environment, they showed a significant weight gain, but when the unrestricted-selected individuals were reared in the restricted diet environment, they did not. These results are a direct contradiction to Hammond’s conjecture, in that selection in the *poorer* environment gave the larger response in the target population. Further, there were other significant differences. The high-feed selected lines contained around 24% more body fat than the restricted-diet lines when both where grown in the high-feed environment. Thus, selection in the restricted diet also resulted in leaner mice, which (in many cases) would also be economically favored in a selection program.

Jinks-Connolly rule

- Stability of the genotypic value over environments is a measure of G x E sensitivity.
  - High stability = low sensitivity
- Antagonistic G x E selection
  - Up-selecting in the bad environment
- Synergistic G x E selection
  - Up-selecting in the good environment
- Jinks-Connolly rule:
  - Antagonistic selection improves stability (decreases environmental sensitivity), while synergistic selection decreases stability



Slope = measure of sensitivity. Reducing the slope increases stability

While Jinks-Connolly suggests a general trend and is expected to hold more often than not, Falconer (1990) noted that a modification of this rule held in all 24 experimental cases he examined, namely that the *sensitivity is less after antagonistic selection than after synergistic selection*. Since the sensitivity is a slope, this means that the change in the numerator of Equation 38.5 is greater under antagonistic selection than under synergistic selection. When selecting to decrease a trait, this requires

$$(R_H - CR_L) - (CR_H - R_L) > 0 \tag{38.6a}$$

which rearranges to recover

$$R_H + R_L > CR_H + CR_L \tag{38.6b}$$

with this same condition holding for selection to increase a trait. Hence, for Falconer's modification to hold, the less restrictive assumption that the sum of the direct responses is greater than the sum of correlated responses must hold.

What about Falconer's (1989) suggestion that *mean performance* over the two environments is best improved by antagonistic selection? If the mean change is equally weighted in both environments, then when selecting to increase a trait, under antagonistic selection direct response occurs in the low environment, while under synergistic selection direct response occurs in the high environment. Thus, Falconer's (1989) suggestion holds when the average of the direct response in low and the correlated response in high exceeds the direct response in high and the correlated response in low,

$$R_L + CR_H > R_H + CR_L \tag{38.7a}$$

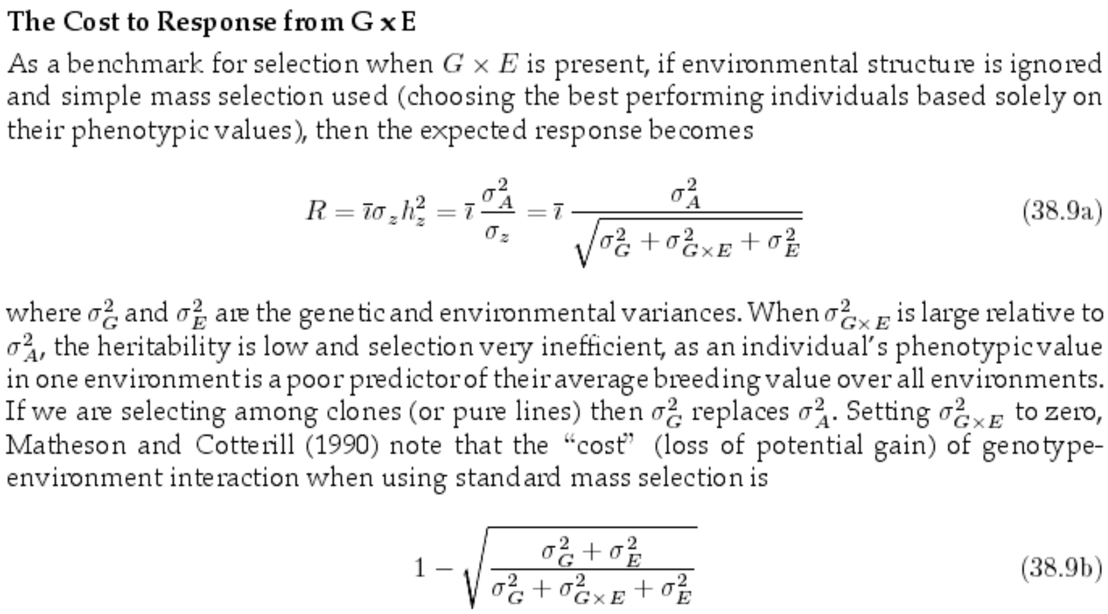
Assuming equal selection in both environments, then from Equation 38.3a, this reduces to

$$h_L\left(\sigma_{A_L} + r_A\sigma_{A_H}\right) > h_H\left(\sigma_{A_H} + r_A\sigma_{A_L}\right) \tag{38.7b}$$

Conversely, when selecting to decrease trait value, this condition becomes

$$R_H + CR_L > R_L + CR_H \tag{38.7c}$$

Note that Equations 38.7a and 38.7c are mutually exclusive, so that if antagonistic selection is better in one direction, it will be worse in the opposite direction. Thus, as Falconer (1990) pointed out, there is little theoretical justification for his earlier (1989) suggestion.



The Cost to Response from G x E

As a benchmark for selection when $G \times E$ is present, if environmental structure is ignored and simple mass selection used (choosing the best performing individuals based solely on their phenotypic values), then the expected response becomes

$$R = \bar{\imath}\sigma_z h_z^2 = \bar{\imath}\,\frac{\sigma_A^2}{\sigma_z} = \bar{\imath}\,\frac{\sigma_A^2}{\sqrt{\sigma_G^2 + \sigma_{G\times E}^2 + \sigma_E^2}} \tag{38.9a}$$

where σ_G^2 and σ_E^2 are the genetic and environmental variances. When $\sigma_{G\times E}^2$ is large relative to σ_A^2, the heritability is low and selection very inefficient, as an individual's phenotypic value in one environment is a poor predictor of their average breeding value over all environments. If we are selecting among clones (or pure lines) then σ_G^2 replaces σ_A^2. Setting $\sigma_{G\times E}^2$ to zero, Matheson and Cotterill (1990) note that the "cost" (loss of potential gain) of genotype-environment interaction when using standard mass selection is

$$1 - \sqrt{\frac{\sigma_G^2 + \sigma_E^2}{\sigma_G^2 + \sigma_{G\times E}^2 + \sigma_E^2}} \tag{38.9b}$$

Replication over environments can reduce effect of G x E in selection response

If members of the same genotype/line are replicated over n_e random environments, response to selection based on line (or family) means is

$$R = \bar{\imath}\,\frac{\sigma_G^2}{\sigma_{\bar{z}}} = \bar{\imath}\,\frac{\sigma_G^2}{\sqrt{\sigma_G^2 + (\sigma_E^2 + \sigma_{G\times E}^2)/n_e + \sigma_e^2/(n_r\, n_e)}}$$

Estimating the GE term

- While GE can be estimated directly from the mean in a cell (i.e., G_i in E_j) we can usually get more information (and a better estimate) by considering the entire design and exploiting structure in the GE terms
- This approach also allows us to potentially predict the GE terms in specific environments
- Basic idea: replace GE_{ij} by $\alpha_i\gamma_j$ or more generally by $\Sigma_k\ \alpha_{ki}\gamma_{kj}$ These are called biadditive or bilinear models. This (at first sight) seems more complicated. Why do this?
- With n_G genotypes and n_E environments, we have
  - $n_G\, n_E$ GE terms (assuming no missing values)
  - $n_G + n_E$ α_i and γ_j unique terms
  - $k(n_G + n_E)$ unique terms in $\Sigma_k\ \alpha_{ki}\gamma_{kj}$.
- Suppose 50 genotypes in 10 environments
  - 500 GE_{ij} terms, 60 unique α_i and γ_J terms, and (for k=3), 180 unique α_{ki} and γ_{ki} terms.

Finlay-Wilkinson Regression

Also called a joint regression or regression on an environmental index.

Let $\mu + G_i$ be the mean of the ith genotype over all environments, and $\mu + E_j$ be the average yield of all genotypes in environment j

$$\mu_{ij} = \mu + G_i + E_j(1 + \beta_i) + \delta_{ij}$$

The FW regression estimates GE_{ij} by the regression $GE_{ij} = \beta_i E_j + \delta_{ij}$. The regression coefficient is obtained for each genotype from the slope of the regression of the G_{ij} over the E_j. δ_{ij} is the residual (lack of fit). If $\sigma^2(GE) >> \sigma^2(\delta)$, then the regression accounted for most of the variation in GE.

Slope a stability measure

$$\mu_{ij} = \mu + G_i + E_j(1+\beta_i) + \delta_{ij}$$

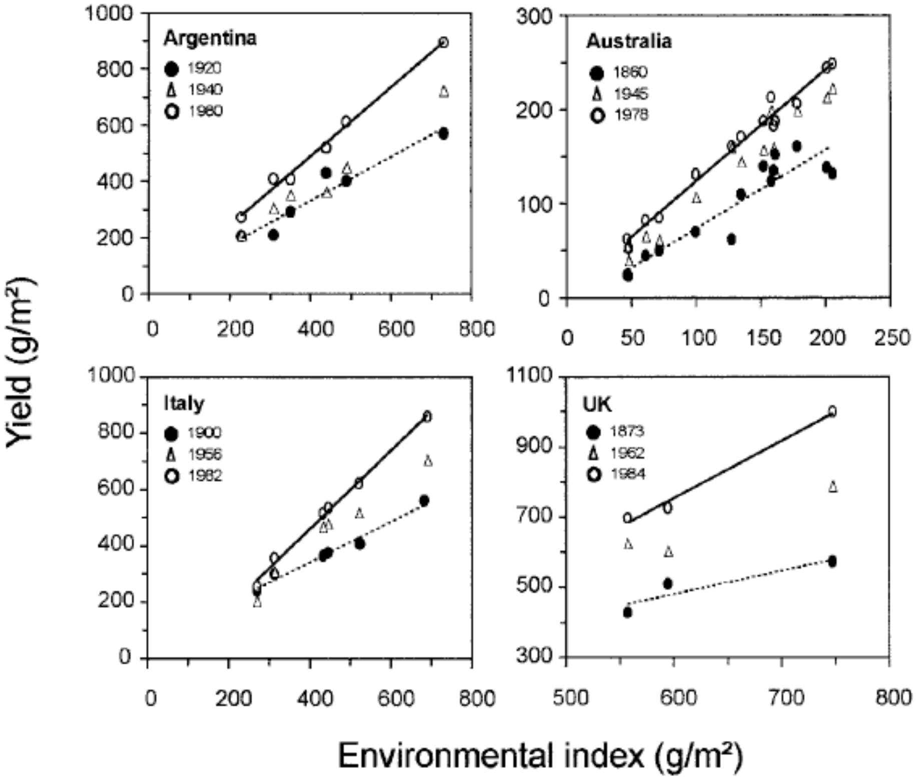
If β_i = -1, strong G x E, with genotype i having identical performance over all environments (good and bad).

If β_i = 0, no G x E.

If β_i > 0, G x E, magnifying the effect of the environment. Over-performs in good environments, under-performs in bad environments.

Application

- Yield in lines of wheat over different environments was examined by Calderini and Slafer (1999). The lines examined were from different eras of breeding (for four different countries)
- Newer lines had larger values, but also had higher slopes (large β_i values), indicating less stability over mean environmental conditions than see in older lines



Regression slope for each genotype is β_i

SVD approaches

- In Finlay-Wilkinson, the GE_{ij} term was estimated by $\beta_i E_j$, where E_j was observed. We could also have used $\gamma_j G_i$, where γ_j is the regression of genotype values over the j-th environment. Again G_i is observable.
- Singular-value decomposition (SVD) approaches consider a more general approach, approximating GE_{ij} by $\Sigma_k\, \alpha_{ki}\gamma_{kj}$ where the α_{ki} and γ_{kj} are determined by the first k terms in the SVD of the matrix of GE terms.
- The SVD is a way to obtain the best approximation of a full matrix by some matrix of lower dimension.

The Singular-Value Decomposition (SVD)

An $n \times p$ matrix $\mathbf{A}$ can always be decomposed as the product of three matrices: an $n \times p$ diagonal matrix $\boldsymbol{\Lambda}$ and two unitary matrices, $\mathbf{U}$ which is $n \times n$ and $\mathbf{V}$ which is $p \times p$. The resulting **singular value decomposition** (**SVD**) of $\mathbf{A}$ is given by

$$\mathbf{A}_{n\times p} = \mathbf{U}_{n\times n}\boldsymbol{\Lambda}_{n\times p}\mathbf{V}^T_{p\times p} \tag{39.16a}$$

We have indicated the dimensionality of each matrix to allow the reader to verify that each matrix multiplication conforms. The diagonal elements $\lambda_1, \cdots, \lambda_s$ of $\boldsymbol{\Lambda}$ correspond to the **singular values** of $\mathbf{A}$ and are ordered by decreasing magnitude. Returning to the unitary matrices $\mathbf{U}$ and $\mathbf{V}$, we can write each as a row vector of column vectors,

$$\mathbf{U} = (\mathbf{u}_1, \cdots, \mathbf{u}_i, \cdots \mathbf{u}_n), \qquad \mathbf{V} = (\mathbf{v}_1, \cdots, \mathbf{v}_i, \cdots \mathbf{v}_p) \tag{39.16b}$$

where $\mathbf{u}_i$ and $\mathbf{v}_i$ are n and p-dimensional column vectors (often called the **left** and **right singular vectors**, respectively). Since both $\mathbf{U}$ and $\mathbf{V}$ are unitary, by definition (Appendix 4) each column vector has length one and are mutually orthogonal (i.e., if $i \neq j$, $\mathbf{u}_i\mathbf{u}_j^T = \mathbf{v}_i\mathbf{v}_j^T = 0$). Since $\boldsymbol{\Lambda}$ is diagonal, it immediately follows from matrix multiplication that we can write any element in $\mathbf{A}$ as

$$A_{ij} = \sum_{k=1}^{s} \lambda_k \, u_{ik} \, v_{kj} \tag{39.16c}$$

where λ_k is the kth singular value and $s \leq \min(p, n)$ is the number of non-zero singular values.

The importance of the singular value decomposition in the analysis of G × E arises from the **Eckart-Young theorem** (1938), which relates the best approximation of a matrix by some lower-rank (say k) matrix with the SVD. Define as our measure of goodness of fit between a matrix $\mathbf{A}$ and a lower rank approximation $\widehat{\mathbf{A}}$ as the sum of squared differences over all elements,

$$\sum_{ij}(A_{ij} - \hat{A}_{ij})^2$$

Eckart and Young show that the best fitting approximation $\widehat{\mathbf{A}}$ of rank $m < s$ is given from the first m terms of the singular value decomposition (the **rank-m SVD**),

$$\hat{A}_{ij} = \sum_{k=1}^{m} \lambda_k \, u_{ik} \, v_{kj} \tag{39.17a}$$

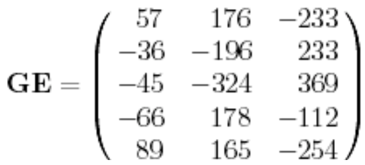
For example, the best rank-2 approximation for the G × E interaction is given by

$$GE_{ij} \simeq \lambda_1 \, u_{i1} \, v_{j1} + \lambda_2 \, u_{i2} \, v_{j2} \tag{39.17b}$$

where λ_i is the ith singular value of the **GE** matrix, $\mathbf{u}$ and $\mathbf{v}$ are the associated singular vectors (see Example 39.3). The fraction of total variation of a matrix accounted for by taking the first m terms in its SVD is

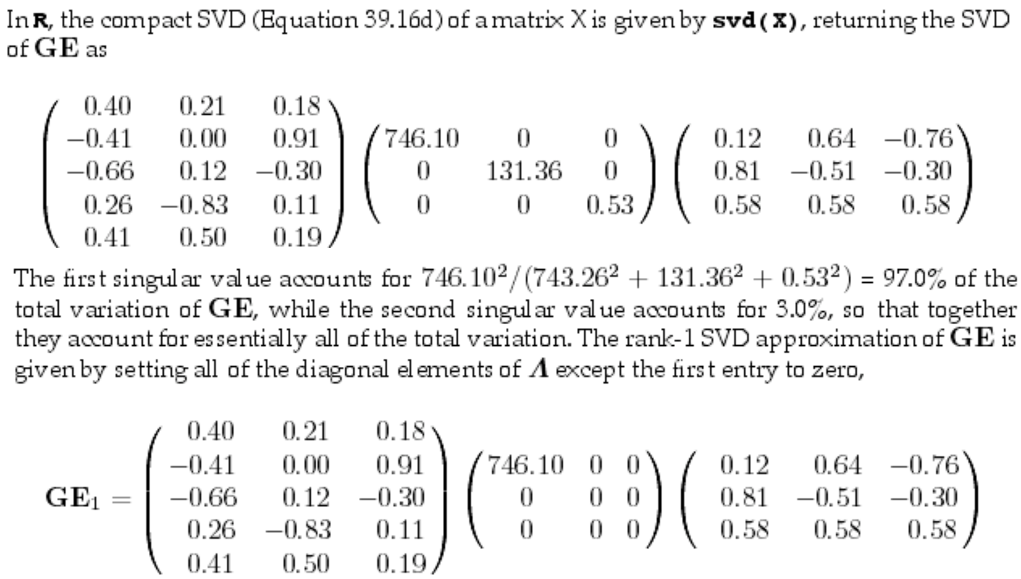
$$\sum_{k=1}^{m} \lambda_k^2 / \sum_{ij} A_{ij}^2 = \frac{\lambda_1^2 + \cdots + \lambda_m^2}{\lambda_1^2 + \cdots + \lambda_s^2}$$

A data set for soybeans grown in New York (Gauch 1992) gives the GE matrix as



$$\mathbf{GE} = \begin{pmatrix} 57 & 176 & -233 \\ -36 & -196 & 233 \\ -45 & -324 & 369 \\ -66 & 178 & -112 \\ 89 & 165 & -254 \end{pmatrix}$$

Where GE_{ij} = value for Genotype i in envir. j

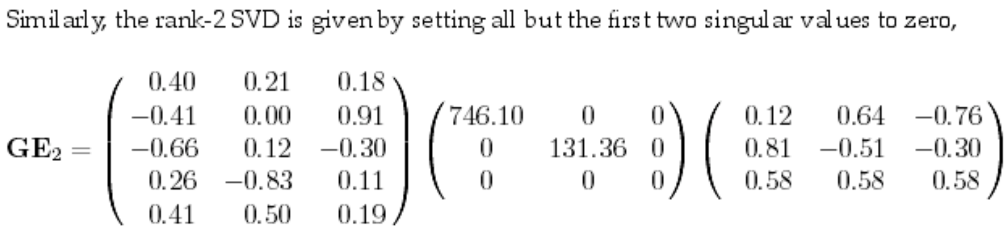


In **R**, the compact SVD (Equation 39.16d) of a matrix X is given by **svd(X)**, returning the SVD of **GE** as

$$\begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 131.36 & 0 \\ 0 & 0 & 0.53 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

The first singular value accounts for $746.10^2/(743.26^2 + 131.36^2 + 0.53^2)$ = 97.0% of the total variation of **GE**, while the second singular value accounts for 3.0%, so that together they account for essentially all of the total variation. The rank-1 SVD approximation of **GE** is given by setting all of the diagonal elements of $\boldsymbol{\Lambda}$ except the first entry to zero,

$$\mathbf{GE}_1 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$



Similarly, the rank-2 SVD is given by setting all but the first two singular values to zero,

$$\mathbf{GE}_2 = \begin{pmatrix} 0.40 & 0.21 & 0.18 \\ -0.41 & 0.00 & 0.91 \\ -0.66 & 0.12 & -0.30 \\ 0.26 & -0.83 & 0.11 \\ 0.41 & 0.50 & 0.19 \end{pmatrix} \begin{pmatrix} 746.10 & 0 & 0 \\ 0 & 131.36 & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.12 & 0.64 & -0.76 \\ 0.81 & -0.51 & -0.30 \\ 0.58 & 0.58 & 0.58 \end{pmatrix}$$

For example, the rank-1 SVD approximation for GE_{32} is $g_{31}\lambda_1 e_{12}$ = 746.10*(-0.66)*0.64 = -315

The rank-2 SVD approximation is $g_{31}\lambda_1 e_{12} + g_{32}\lambda_2 e_{22}$ = 746.10*(-0.66)*0.64 + 131.36* 0.12*(-0.51) = -323

Actual value is -324

Generally, the rank-2 SVD approximation for GE_{ij} is $g_{i1}\lambda_1 e_{1j} + g_{i2}\lambda_2 e_{2j}$

AMMI models

Additive main effects, multiplicative interaction (AMMI) models use the first m terms in the SVD of GE:

$$GE_{ij} = \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij}$$

Giving

$$\mu_{ij} = \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij}$$

AMMI is actually a *family* of models, with AMMI$_m$ denoting AMMI with the first m SVD terms.

AMMI models

$$\mu_{ij} = \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij}$$

Fit main effects

Fit principal components to the interaction term (SVD is a generalization of PC methods) →

$$GE_{ij} = \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij}$$

Factorial Regressions

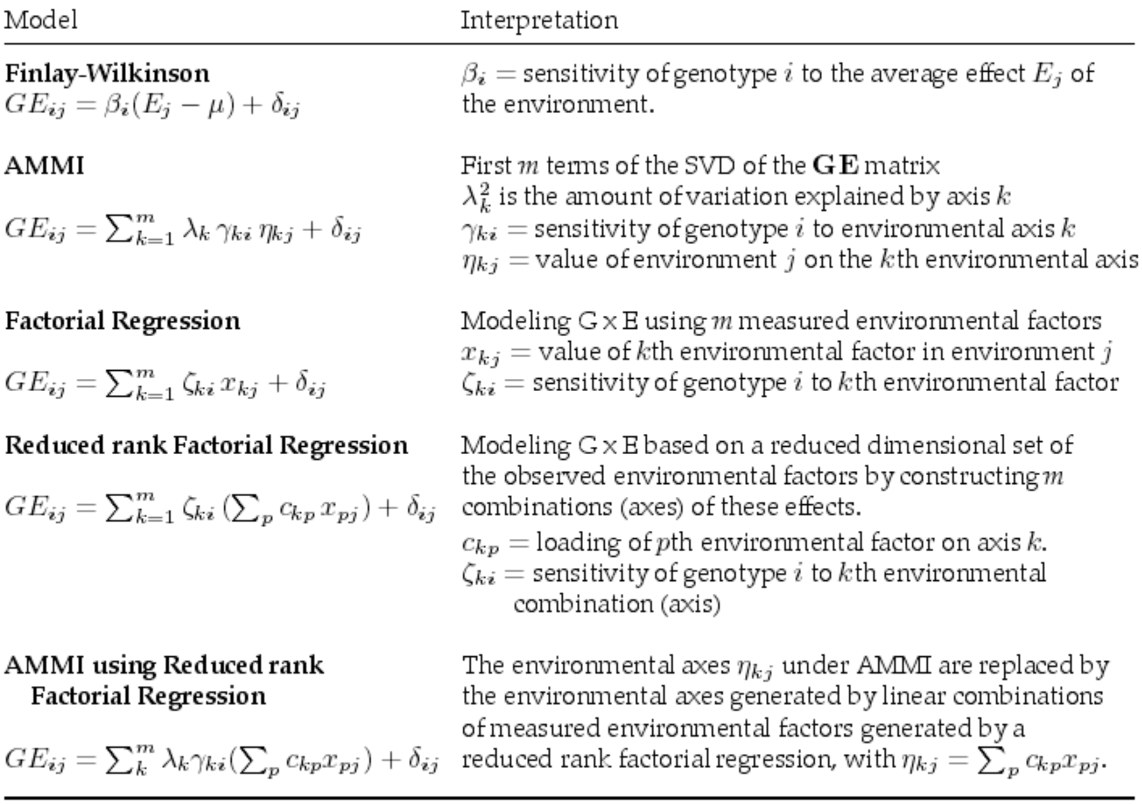
- While AMMI models attempt to extract information about how G x E interactions are related across sets of genotypes and environments, factorial regressions incorporate direct measures of environmental factors in an attempt to account for the observed pattern of G x E.
- The power of this approach is that if we can determine which genotypes are more (or less) sensitive to which environmental features, the breeder may be able to more finely tailor a line to a particular environment without necessarily requiring trials in the target environment.

Suppose we have a series of *m* measured values from the environments of interest (such as average rainfall, maximum temperature, etc.)   Let x_{kj} denote the value of the k-th environmental variable in environment j

Factorial regressions model the GE term as the sensitivity ζ_{ki} of environmental value k to genotype i, (this is a regression slope to be estimated from the data)

$$GE_{ij} = \sum_{k=1}^{m} \zeta_{ki}\, x_{kj} + \delta_{ij}$$

Note that the Finlay-Wilkinson regression is a special case where m = 1 and x_j is the mean trait value (over all genotypes) in that environment.



Model	Interpretation
Finlay-Wilkinson $GE_{ij} = \beta_i(E_j - \mu) + \delta_{ij}$	β_i = sensitivity of genotype i to the average effect E_j of the environment.
AMMI $GE_{ij} = \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij}$	First m terms of the SVD of the $\mathbf{GE}$ matrix λ_k^2 is the amount of variation explained by axis k γ_{ki} = sensitivity of genotype i to environmental axis k η_{kj} = value of environment j on the kth environmental axis
Factorial Regression $GE_{ij} = \sum_{k=1}^{m} \zeta_{ki} \, x_{kj} + \delta_{ij}$	Modeling G x E using m measured environmental factors x_{kj} = value of kth environmental factor in environment j ζ_{ki} = sensitivity of genotype i to kth environmental factor
Reduced rank Factorial Regression $GE_{ij} = \sum_{k=1}^{m} \zeta_{ki} \left(\sum_p c_{kp} \, x_{pj}\right) + \delta_{ij}$	Modeling G x E based on a reduced dimensional set of the observed environmental factors by constructing m combinations (axes) of these effects. c_{kp} = loading of pth environmental factor on axis k. ζ_{ki} = sensitivity of genotype i to kth environmental combination (axis)
AMMI using Reduced rank Factorial Regression $GE_{ij} = \sum_{k}^{m} \lambda_k \gamma_{ki} (\sum_p c_{kp} x_{pj}) + \delta_{ij}$	The environmental axes η_{kj} under AMMI are replaced by the environmental axes generated by linear combinations of measured environmental factors generated by a reduced rank factorial regression, with $\eta_{kj} = \sum_p c_{kp} x_{pj}$.

Mixed model analysis of G x E

- Thus far, our discussion of estimating GE has be set in terms of fixed effects.
- Mixed models are a powerful alternative, as they easily handle missing data (i.e., not all combinations of G and E explored).
- As with all mixed models, key is the assumed covariance structure
  - Structured covariance models
    - Compound symmetry
    - Finlay-Wilkinson
    - Factor-analytic models (closely related to AMMI)

Basic GxE Mixed model

- Typically, we assume either G or E is fixed, and the other random (making GE random)
- Taking E as fixed, basic model becomes
- $\mathbf{z} = \mathbf{X\beta} + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{ge} + \mathbf{e}$
  - The vector **β** of fixed effects includes estimates of the E_j. The vector **g** contains estimates of the G_i values, while the vector **ge** contains estimates of all the GE_{ij}.
  - Typically we assume **e** ~ **0**, σ_e^2 **I**, and independent of **g** and **ge**.
  - Models significantly differ on the <u>variance/covariance structure</u> of **g** and **ge**.

Example

We have two genotypes and three environments. Let z_{ijk} denote the k-th replicate of genotype i in environment j. Suppose we have single replicates of genotype 1 in all three environments, two replicates of genotype 2 in environment 1, and one in environment 3

$$\mathbf{z} = \begin{pmatrix} z_{111} \\ z_{121} \\ z_{131} \\ z_{211} \\ z_{212} \\ z_{231} \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} E_1^* \\ E_2^* \\ E_3^* \end{pmatrix}, \quad \mathbf{g} = \begin{pmatrix} G_1 \\ G_2 \end{pmatrix}, \quad \mathbf{ge} = \begin{pmatrix} GE_{11} \\ GE_{12} \\ GE_{13} \\ GE_{21} \\ GE_{22} \\ GE_{23} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \epsilon_{111} \\ \epsilon_{121} \\ \epsilon_{131} \\ \epsilon_{211} \\ \epsilon_{212} \\ \epsilon_{231} \end{pmatrix}$$

Here $E_i^* = \mu + E_i$, with the E_i constrained to sum to zero. The resulting design matrices are

$$\mathbf{X} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{Z}_1 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Z}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

Compound Symmetry assumption

- To proceed further on the analysis of the mixed model, we need covariance assumptions on **g** and **ge**.
- The compound symmetry assumption is
  - $\sigma^2(G_i) = \sigma_G^2$, $\sigma^2(GE_{ij}) = \sigma^2_{GE}$
  - Plus no covariances across effects
  - Under these assumptions, the covariance of any genotype across any two (different) environments is the same.
  - Likewise, the genetic variance within any environment is constant across environments
  - Net result, the genetic covariance is the same between any two environments

Under the compound symmetry assumption, the genetic variance and covariances become as follows:

Expected genetic variance within a given environment

$$\begin{aligned}\sigma(z_{ijk}, z_{ij\ell}) &= \sigma(G_i + GE_{ij}, G_i + GE_{ij}) \\ &= \sigma(G_i, G_i) + \sigma(GE_{ij}, GE_{ij}) \\ &= \sigma_G^2 + \sigma_{GE}^2\end{aligned}$$

Genetic covariance of the same genotype across environments

$$\sigma(z_{ij}, z_{ik}) = \sigma(G_i + GE_{ij}, G_i + GE_{ik}) = \sigma(G_i, G_i)$$

Genetic correlation across environments is constant

$$\rho_G = \frac{\sigma_G^2}{\sigma_G^2 + \sigma_{GE}^2}$$

For our example, the resulting covariance matrix becomes

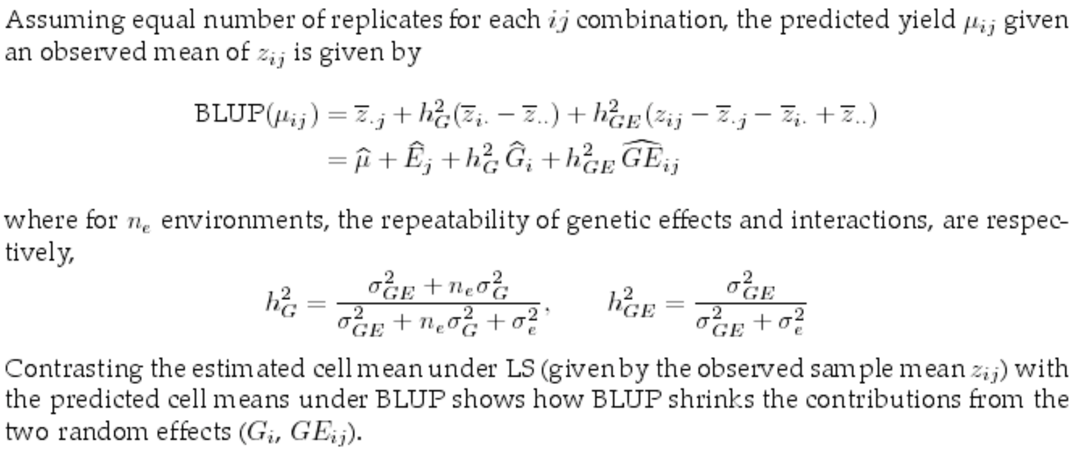
$$\mathbf{V_z} = \mathbf{Z}_1\,\mathbf{V_g}\,\mathbf{Z}_1^T + \mathbf{Z}_2\,\mathbf{V_{ge}}\,\mathbf{Z}_2^T + \mathbf{V_e}$$

$$\mathbf{V_g} = \sigma_G^2\,\mathbf{I}_{2\times2}, \quad \mathbf{V_{ge}} = \sigma_{G\times E}^2\,\mathbf{I}_{6\times6}, \quad \mathbf{V_e} = \sigma_e^2\,\mathbf{I}_{6\times6}$$

Mixed-model allows for missing values.

Under fixed-effect model, estimate of μ_{ij} = $\overline{z_{ij}}$.

BLUP estimates under mixed-model (E fixed, G random)



Assuming equal number of replicates for each ij combination, the predicted yield μ_{ij} given an observed mean of z_{ij} is given by

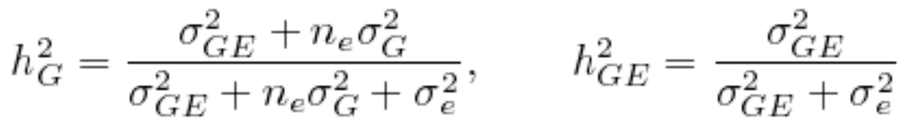
$$\begin{aligned}\text{BLUP}(\mu_{ij}) &= \overline{z}_{\cdot j} + h^2_G(\overline{z}_{i\cdot} - \overline{z}_{\cdot\cdot}) + h^2_{GE}(z_{ij} - \overline{z}_{\cdot j} - \overline{z}_{i\cdot} + \overline{z}_{\cdot\cdot}) \\ &= \widehat{\mu} + \widehat{E}_j + h^2_G\,\widehat{G}_i + h^2_{GE}\,\widehat{GE}_{ij}\end{aligned}$$

where for n_e environments, the repeatability of genetic effects and interactions, are respectively,

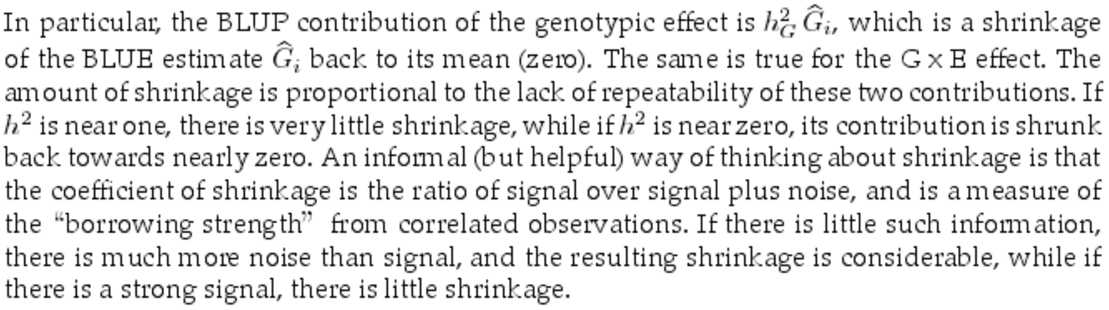
$$h^2_G = \frac{\sigma^2_{GE} + n_e\sigma^2_G}{\sigma^2_{GE} + n_e\sigma^2_G + \sigma^2_e}, \qquad h^2_{GE} = \frac{\sigma^2_{GE}}{\sigma^2_{GE} + \sigma^2_e}$$

Contrasting the estimated cell mean under LS (given by the observed sample mean z_{ij}) with the predicted cell means under BLUP shows how BLUP shrinks the contributions from the two random effects (G_i, GE_{ij}).

BLUP shrinks (regresses) the BLUE estimate back towards zero

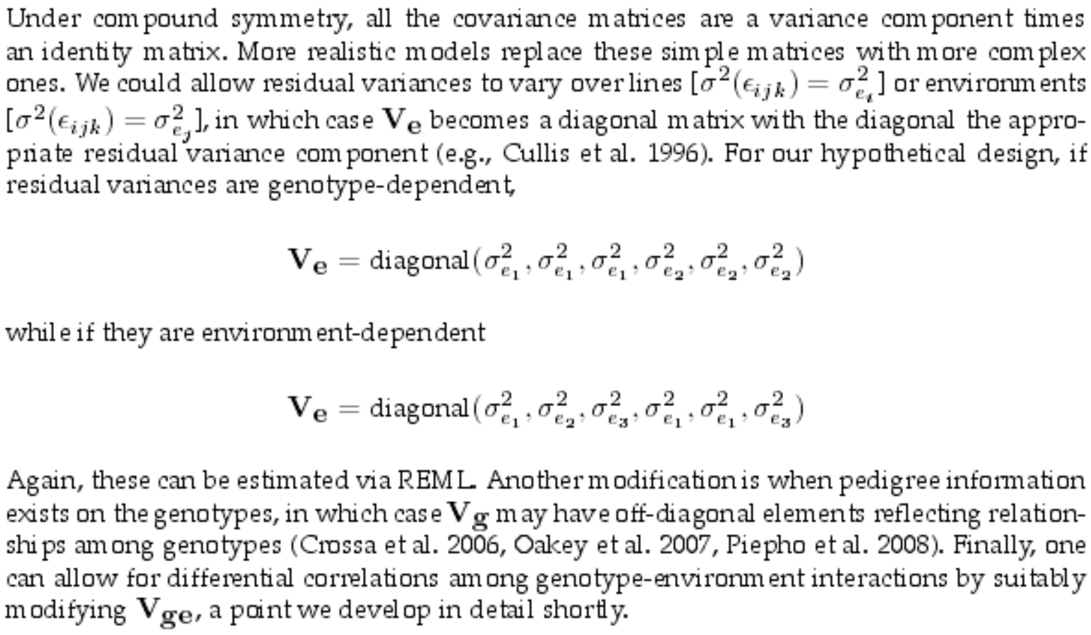


$$h^2_G = \frac{\sigma^2_{GE} + n_e\sigma^2_G}{\sigma^2_{GE} + n_e\sigma^2_G + \sigma^2_e}, \qquad h^2_{GE} = \frac{\sigma^2_{GE}}{\sigma^2_{GE} + \sigma^2_e}$$



In particular, the BLUP contribution of the genotypic effect is $h^2_G\,\widehat{G}_i$, which is a shrinkage of the BLUE estimate $\widehat{G}_i$ back to its mean (zero). The same is true for the G × E effect. The amount of shrinkage is proportional to the lack of repeatability of these two contributions. If h^2 is near one, there is very little shrinkage, while if h^2 is near zero, its contribution is shrunk back towards nearly zero. An informal (but helpful) way of thinking about shrinkage is that the coefficient of shrinkage is the ratio of signal over signal plus noise, and is a measure of the "borrowing strength" from correlated observations. If there is little such information, there is much more noise than signal, and the resulting shrinkage is considerable, while if there is a strong signal, there is little shrinkage.

Modification of the residual covariance



Under compound symmetry, all the covariance matrices are a variance component times an identity matrix. More realistic models replace these simple matrices with more complex ones. We could allow residual variances to vary over lines $[\sigma^2(\epsilon_{ijk}) = \sigma^2_{e_i}]$ or environments $[\sigma^2(\epsilon_{ijk}) = \sigma^2_{e_j}]$, in which case $\mathbf{V_e}$ becomes a diagonal matrix with the diagonal the appropriate residual variance component (e.g., Cullis et al. 1996). For our hypothetical design, if residual variances are genotype-dependent,

$$\mathbf{V_e} = \text{diagonal}(\sigma^2_{e_1}, \sigma^2_{e_1}, \sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_2}, \sigma^2_{e_2})$$

while if they are environment-dependent

$$\mathbf{V_e} = \text{diagonal}(\sigma^2_{e_1}, \sigma^2_{e_2}, \sigma^2_{e_3}, \sigma^2_{e_1}, \sigma^2_{e_1}, \sigma^2_{e_3})$$

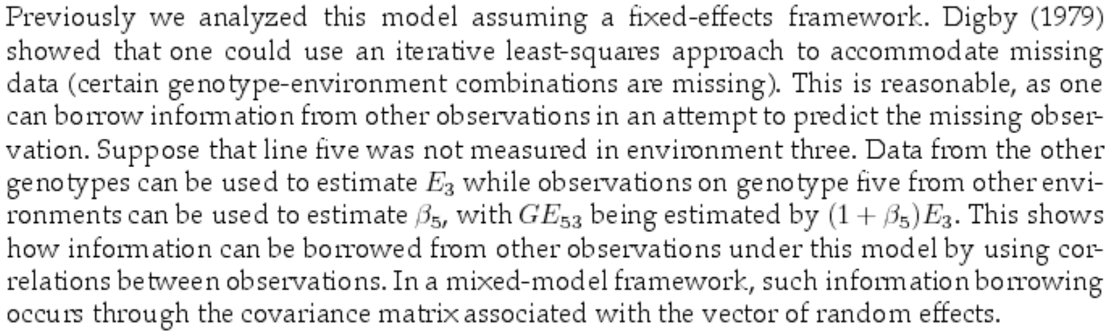
Again, these can be estimated via REML. Another modification is when pedigree information exists on the genotypes, in which case $\mathbf{V_g}$ may have off-diagonal elements reflecting relationships among genotypes (Crossa et al. 2006, Oakey et al. 2007, Piepho et al. 2008). Finally, one can allow for differential correlations among genotype-environment interactions by suitably modifying $\mathbf{V_{ge}}$, a point we develop in detail shortly.

Extending genetic covariances

- Shukla's model: starts with the compound symmetry model, but allows for different G x E variances over genotypes,
  - $GE_{ij} \sim N(0, \sigma^2_{GiE})$
  - $G_i \sim N(0, \sigma_G^2)$
  - Cov(ge) = Diagonal $(\sigma^2_{G1E}, \cdots, \sigma^2_{GnE})$
  - The covariance of a genotype across environments is still σ_G^2
- Structured covariance models allow more complicated (and more general) covariance matrices

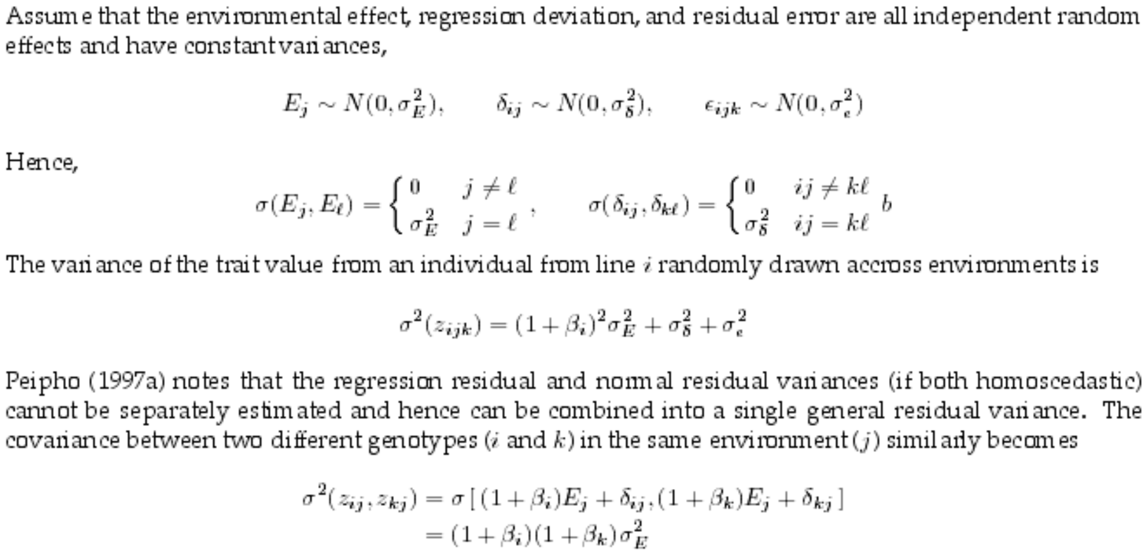
Covariances based on Finlay-Wilkinson

$$z_{ijk} = \mu + G_i + (1+\beta_i)E_j + \delta_{ij} + \epsilon_{ijk}$$



Previously we analyzed this model assuming a fixed-effects framework. Digby (1979) showed that one could use an iterative least-squares approach to accommodate missing data (certain genotype-environment combinations are missing). This is reasonable, as one can borrow information from other observations in an attempt to predict the missing observation. Suppose that line five was not measured in environment three. Data from the other genotypes can be used to estimate E_3 while observations on genotype five from other environments can be used to estimate β_5, with GE_{53} being estimated by $(1+\beta_5)E_3$. This shows how information can be borrowed from other observations under this model by using correlations between observations. In a mixed-model framework, such information borrowing occurs through the covariance matrix associated with the vector of random effects.

We treat G_i and β_i as fixed effects, δ_{ij} and e_{ijk} as random (fixed genetic effects, random environmental effects)



Assume that the environmental effect, regression deviation, and residual error are all independent random effects and have constant variances,

$$E_j \sim N(0,\sigma_E^2), \qquad \delta_{ij} \sim N(0,\sigma_\delta^2), \qquad \epsilon_{ijk} \sim N(0,\sigma_e^2)$$

Hence,

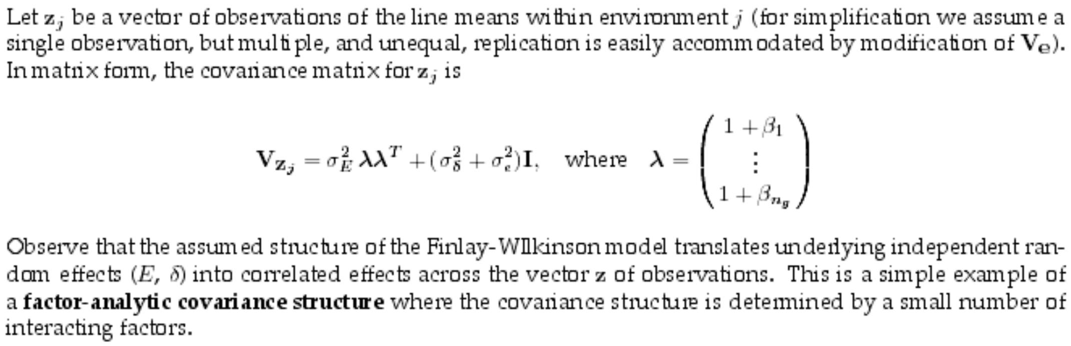
$$\sigma(E_j,E_\ell) = \begin{cases} 0 & j \neq \ell \\ \sigma_E^2 & j = \ell \end{cases}, \qquad \sigma(\delta_{ij},\delta_{k\ell}) = \begin{cases} 0 & ij \neq k\ell \\ \sigma_\delta^2 & ij = k\ell \end{cases} b$$

The variance of the trait value from an individual from line i randomly drawn accross environments is

$$\sigma^2(z_{ijk}) = (1+\beta_i)^2\sigma_E^2 + \sigma_\delta^2 + \sigma_e^2$$

Peipho (1997a) notes that the regression residual and normal residual variances (if both homoscedastic) cannot be separately estimated and hence can be combined into a single general residual variance. The covariance between two different genotypes (i and k) in the same environment (j) similarly becomes

$$\begin{aligned}\sigma^2(z_{ij},z_{kj}) &= \sigma\,[\,(1+\beta_i)E_j + \delta_{ij}, (1+\beta_k)E_j + \delta_{kj}\,]\\ &= (1+\beta_i)(1+\beta_k)\sigma_E^2\end{aligned}$$

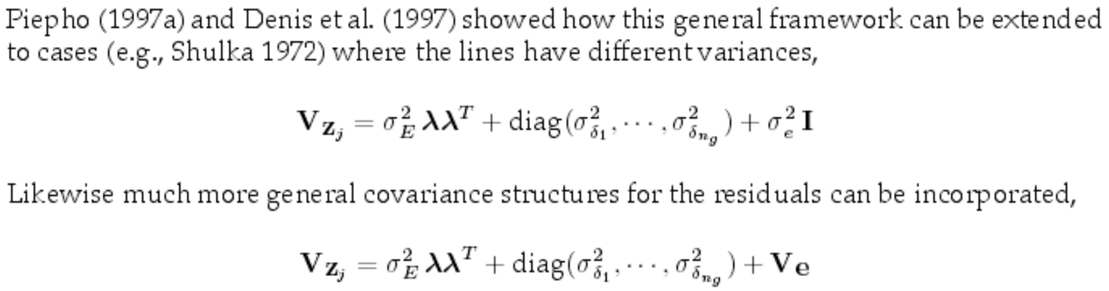


Let $\mathbf{z}_j$ be a vector of observations of the line means within environment j (for simplification we assume a single observation, but multiple, and unequal, replication is easily accommodated by modification of $\mathbf{V_e}$). In matrix form, the covariance matrix for $\mathbf{z}_j$ is

$$\mathbf{V}_{\mathbf{z}_j} = \sigma^2_E \, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + (\sigma^2_\delta + \sigma^2_e)\mathbf{I}, \quad \text{where} \quad \boldsymbol{\lambda} = \begin{pmatrix} 1+\beta_1 \\ \vdots \\ 1+\beta_{n_g} \end{pmatrix}$$

Observe that the assumed structure of the Finlay-WIlkinson model translates underlying independent random effects (E, δ) into correlated effects across the vector $\mathbf{z}$ of observations. This is a simple example of a **factor-analytic covariance structure** where the covariance structure is determined by a small number of interacting factors.

Factor-analytic covariance structures allow one to consider more general covariance structures informed by the data, rather than assumed by the investigator.



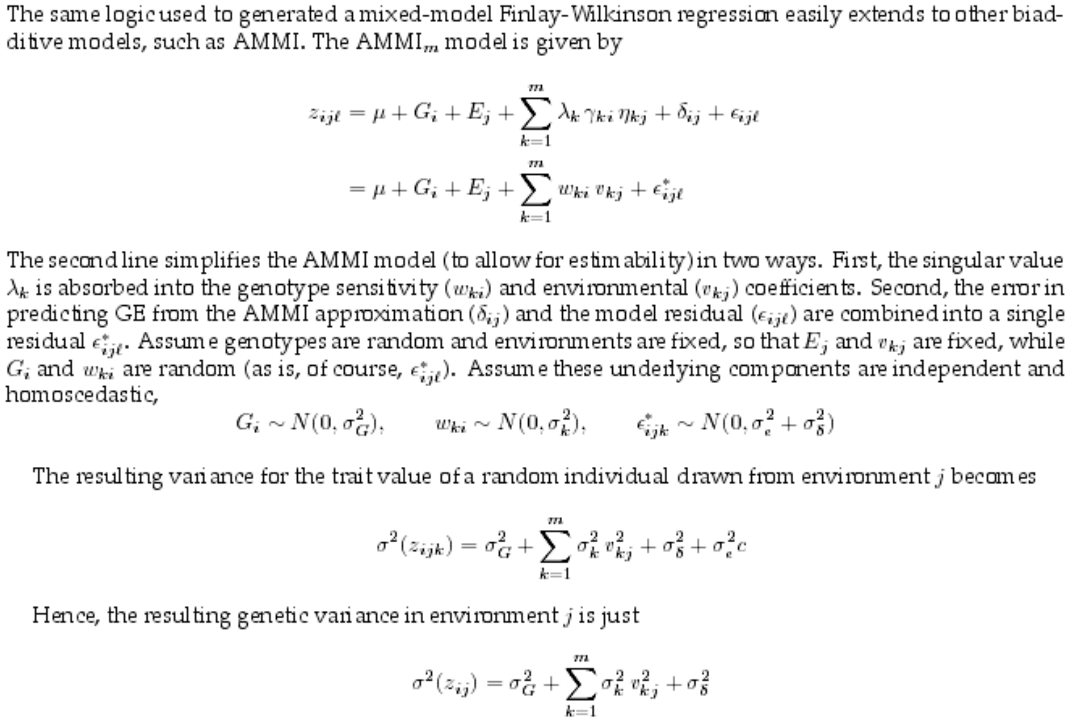
Piepho (1997a) and Denis et al. (1997) showed how this general framework can be extended to cases (e.g., Shulka 1972) where the lines have different variances,

$$\mathbf{V}_{\mathbf{z}_j} = \sigma^2_E \, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \text{diag}(\sigma^2_{\delta_1}, \cdots, \sigma^2_{\delta_{n_g}}) + \sigma^2_e \, \mathbf{I}$$

Likewise much more general covariance structures for the residuals can be incorporated,

$$\mathbf{V}_{\mathbf{z}_j} = \sigma^2_E \, \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \text{diag}(\sigma^2_{\delta_1}, \cdots, \sigma^2_{\delta_{n_g}}) + \mathbf{V_e}$$

AMMI-based structured covariance models



The same logic used to generated a mixed-model Finlay-Wilkinson regression easily extends to other biadditive models, such as AMMI. The AMMI$_m$ model is given by

$$\begin{aligned} z_{ij\ell} &= \mu + G_i + E_j + \sum_{k=1}^{m} \lambda_k \, \gamma_{ki} \, \eta_{kj} + \delta_{ij} + \epsilon_{ij\ell} \\ &= \mu + G_i + E_j + \sum_{k=1}^{m} w_{ki} \, v_{kj} + \epsilon^*_{ij\ell} \end{aligned}$$

The second line simplifies the AMMI model (to allow for estimability) in two ways. First, the singular value λ_k is absorbed into the genotype sensitivity (w_{ki}) and environmental (v_{kj}) coefficients. Second, the error in predicting GE from the AMMI approximation (δ_{ij}) and the model residual ($\epsilon_{ij\ell}$) are combined into a single residual $\epsilon^*_{ij\ell}$. Assume genotypes are random and environments are fixed, so that E_j and v_{kj} are fixed, while G_i and w_{ki} are random (as is, of course, $\epsilon^*_{ij\ell}$). Assume these underlying components are independent and homoscedastic,

$$G_i \sim N(0, \sigma^2_G), \qquad w_{ki} \sim N(0, \sigma^2_k), \qquad \epsilon^*_{ijk} \sim N(0, \sigma^{\,2}_e + \sigma^2_\delta)$$

The resulting variance for the trait value of a random individual drawn from environment j becomes

$$\sigma^2(z_{ijk}) = \sigma^2_G + \sum_{k=1}^{m} \sigma^2_k \, v^2_{kj} + \sigma^2_\delta + \sigma^{\,2}_e c$$

Hence, the resulting genetic variance in environment j is just

$$\sigma^2(z_{ij}) = \sigma^2_G + \sum_{k=1}^{m} \sigma^2_k \, v^2_{kj} + \sigma^2_\delta$$

Likewise, the covariance between the same random genotype over different environments (j and ℓ) becomes

$$\begin{aligned}\sigma^2(z_{ij}, z_{i\ell}) &= \sigma\left(G_i + \sum_{k=1}^{m} w_{ki}\, v_{kj}, G_i + \sum_{k=1}^{m} w_{ki}\, v_{k\ell} \right) \\ &= \sigma_G^2 + \sum_{k=1}^{m} \sigma(\, w_{ki}\, v_{kj}, w_{ki}\, v_{k\ell}) = \sigma_G^2 + \sum_{k=1}^{m} v_{kj}\, v_{k\ell}\, \sigma(\, w_{ki}, w_{ki}) \\ &= \sigma_G^2 + \sum_{k=1}^{m} \sigma_k^2\, v_{kj}\, v_{k\ell}\end{aligned}$$

The resulting covariance matrix for the vector $\mathbf{z}_i$ of observations of genotypes over the n_e environments can be written in the form

$$\mathbf{V}_{\mathbf{z}_i} = \sigma_G^2\,\mathbf{J} + \boldsymbol{\lambda}\boldsymbol{\lambda}^T + \sigma_e^2\mathbf{I}$$

where $\mathbf{J}$ is a matrix of ones, and $\boldsymbol{\lambda}$ is the $n_e \times m$ matrix,

$$\boldsymbol{\lambda} = (\,\boldsymbol{\lambda}_1 \quad \cdots \quad \boldsymbol{\lambda}_m\,)\,, \quad \text{where} \quad \boldsymbol{\lambda}_k = \sigma_k^2 \begin{pmatrix} v_{k1} \\ \vdots \\ v_{kn_e} \end{pmatrix} = \begin{pmatrix} \lambda_{k1} \\ \vdots \\ \lambda_{kn_e} \end{pmatrix}$$

Summary: Structured Covariance models

	G random, E Fixed		E random, G Fixed	
	$\sigma^2(z_{ij})$	$\sigma(z_{ij}, z_{ik})$	$\sigma^2(z_{ij})$	$\sigma(z_{ij}, z_{kj})$
Index range	$1 \leq j \leq n_e$	$1 \leq k, j \leq n_e$	$1 \leq i \leq n_g$	$1 \leq i, k \leq n_g$
C	$\sigma_G^2 + \sigma_{GE}^2$	σ_G^2	$\sigma_E^2 + \sigma_{GE}^2$	σ_E^2
S	$\sigma_G^2 + \sigma_{GE_j}^2$	σ_G^2	$\sigma_E^2 + \sigma_{G_iE}^2$	σ_E^2
FW	$\sigma_G^2 + E_j^2\sigma_\beta^2 + \sigma_{\delta_j}^2$	$\sigma_G^2 + E_jE_k\sigma_\beta^2$	$\alpha_i^2\sigma_E^2 + \sigma_{\delta_i}^2$	$\alpha_i\alpha_k\sigma_E^2$
FA(m)	$\sigma_G^2 + \sum_\ell^m \lambda_{\ell j}^2$	$\sigma_G^2 + \sum_\ell^m \lambda_{\ell j}\lambda_{\ell k}$	$\sigma_E^2 + \sum_\ell^m \lambda_{\ell i}^2$	$\sigma_E^2 + \sum_\ell^m \lambda_{\ell i}\lambda_{\ell k}$
U	σ_j^2	σ_{jk}	σ_i^2	σ_{ik}

Summary the covariance structures for various mixed-models for G×E . When G is taken as random with E fixed, $\sigma^2(z_{ij})$ is genetic variance in environment j, while $\sigma(z_{ij}, z_{ik})$ is the covariance between a random genotype (i) measured in environments j and k. When E taken as random, $\sigma^2(z_{ij})$ corresponds to the variance for an individual from genotype i drawn from a random environment, while $\sigma(z_{ij}, z_{kj})$ is the covariance between genotypes i and k when measured in across a random environment (j). C corresponds to the Compound Symmetry model, S is Shukla's extension, FW is Finlay-Wilkinson where $\alpha_i = 1+\beta_i$, FA(m) is factor-analytic model (i.e., a mixed AMMI-type model) with m factors, U is the completely Unstructured model.

Lecture 10:

Infinite-dimensional/Function-valued Traits: Covariance Functions and Random Regressions

Bruce Walsh lecture notes
Summer Institute in Statistical Genetics
Seattle, 20 – 22 July 2016

Longitudinal traits

- Many classic quantitative traits are longitudinal -- measured at multiple time points --- milk yield, body size, etc.
- We have already examined the repeated-measures design wherein an identical trait (assumed to be unchanging) is measured multiple times.
- For most longitudinal traits, we expect the trait to change over time, such as a growth curve.
- These are function-valued traits, also called infinite-dimensional traits.
- One critical feature of such traits is that their additive variances change with t, and trait values from different time points have different correlations.

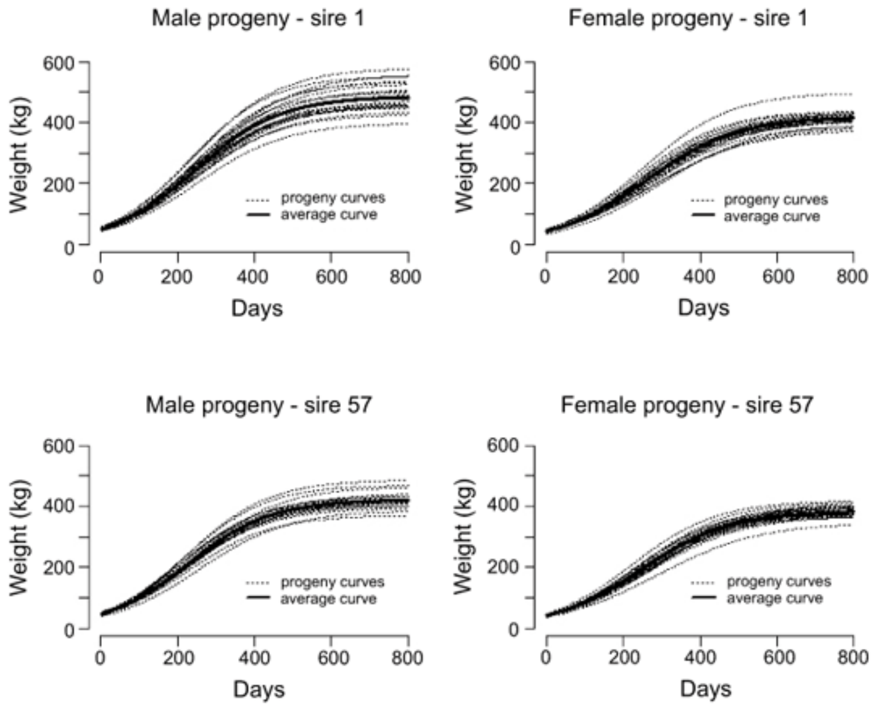


Figure 3 - Mixed logistic growth curves (---) fitted for all progeny of sire 1 (24 males and 32 females) and all progeny of sire 57 (20 males and 59 females) and associated average growth curves (—).

Sci Agric. 66: 85-89

Norms of reaction

- The other type of function-valued trait is one indexed by some continuous environmental variable (as opposed to time), such as adult body weight as a function of temperature or grain yield as a function of total rainfall.
- The measurement of such traits generally requires replication of individuals over environments (versus the sequential evaluation of a single individual with longitudinal traits). As with G x E, this can be done
  - Using clones/pure lines
  - Using family members
- Such curves are common in ecology & evolution and are called norms of reaction, and are measures of G x E
  - Norms of reaction measure phenotypic plasticity --- variation that can be expressed from a fixed genotype, which is often an important adaptation in changing environments.

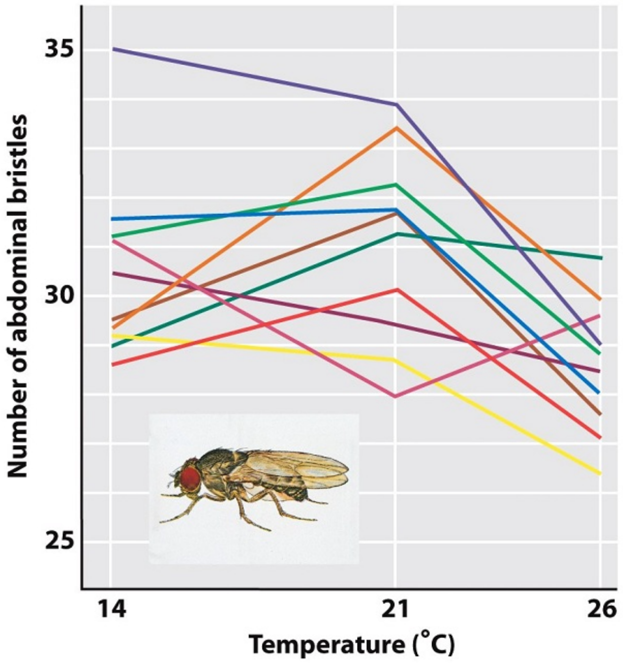


Figure 18-6
Introduction to Genetic Analysis, Ninth Edition
© 2008 W. H. Freeman and Company

How to model such traits?

- One obvious approach is to treat the trait measured at discrete time points as a series of correlated traits.
  - Makes sense to do this for something like parity (litter number), as individuals are all measured at the same event, i.e., parity one, parity two, etc.
  - However, with a trait like a growth or some performance curve, we often expect to have different time measurements for different individuals.
    - We could either lump these into groups (reducing precision) or treat each different time/tuning variable value as a different trait (much missing data).
  - Better solution: estimate the trait covariance function, where $C(t_1,t_2) = Cov[z(t_1),z(t_2)]$ or $Cov[A(t_1),A(t_2)]$

Covariance function approach

- Kirkpatrick popularized the use of covariance functions (largely in evolutionary biology) in the mid-late 1980's.
- He noted that traits measured with respect to some continuous indexing variable (such as time or temperature) have effectively infinite dimensions, as one could (in theory) always consider finer and finer time scales.
  - Thus, rather than treat them as a (potentially) every-expanding set of discrete correlated traits, better to simply consider the covariance $C(t_1,t_2)$ between any two time points within the range of the sampled data. Note that $C(t_1,t_1)$ is the trait variance at time t_1.
  - $C(t_1,t_2)$ is the covariance function, the logical extension of the covariance matrix C(i,j) used for correlated traits, using continuous, rather than integer, indexes.

Covariance functions (cont)

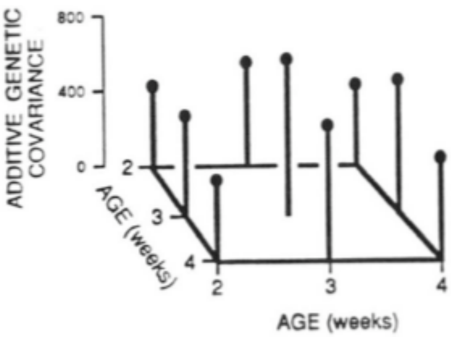
- As with any quantitative trait, the covariance between the values at two time points can be decomposed into an additive-genetic (breeding value) covariance function and a residual (or environmental) covariance function,
  - $C_z(t_1,t_2) = C_A(t_1,t_2) + C_E(t_1,t_2)$
- The issue in the estimation of the additive covariance function is how one proceeds from an additive-covariance matrix estimate **G** from discrete time points to a continuous function covering all possible values with the span of time sampled to estimate **G**.
  - Basic (initial) idea: Use curve-fitting based on low-degree polynomials to use **G** to fit a covariance function
  - This is typically done by using Legendre polynomials as the basis function.

Riska et al. (1984) data on breeding values for log(body weight)

The basic idea was illustrated by Kirkpatrick with a data set on mouse body weight measured at ages 2, 3, and 4 weeks. Riska et al. estimated the G matrix as

$$\begin{array}{cccc} & 2 & 3 & 4 \\ \hat{G} = & 436 & 522 & 424 \\ & 522 & 808 & 665 \\ & 424 & 665 & 558 \end{array}$$

Plotting these values on a lattice at these discrete time points gives



Ideally, would like some sort of smooth curve for this data.

Towards the covariance function

- Suppose we assume the breeding value at time t (for $2 \leq t \leq 4$ weeks) is in the form of a quadratic, so that individual's i breeding value is given by
  - $A_i(t) = a_{io} + a_{i1}\, t + a_{i2}\, t^2$.
  - Here the a_{ij} (for $0 \leq j \leq 2$) are regression coefficients unique to individual i, and are unchanging over time.
- A different individual (j) also has a quadratic regression, but with different coefficients
  - $A_j(t) = a_{jo} + a_{j1}\, t + a_{j2}\, t^2$.
  - the a_{ij} are referred to as random regression coefficients, as they are random (drawn from some distribution) OVER individuals, but constant over time WITHIN an individual.

Towards the covariance function (cont)

We can think of these random regression coefficients as being drawn from a distribution:

$$\begin{pmatrix} a_0 \\ a_1 \\ a_2 \end{pmatrix} \sim \mathbf{0}, \mathbf{C_G}, \quad \text{where} \quad \mathbf{C_G} = \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix}$$

Ideally, we would like to use our estimate of G to make inferences on the elements in C_G.

We can write the additive value in time t for individual i as $\mathbf{a}_i^T{*}\mathbf{t}$, where = $\mathbf{a}_i^T = (a_{i0}, a_{i1}, a_{i2})$ and $\mathbf{t}^T = (1, t, t^2)$

Towards the covariance function

The regression $A_i(t) = a_{io} + a_{i1}t + a_{i2}t^2 = \mathbf{a}_i^T\mathbf{t}$ yields the covariance function, as the value of the vector t for different times are constants, giving

$$\begin{aligned} \text{Cov}[A_i(t_1), A_i(t_2)] &= \text{Cov}[\mathbf{a}_i^T\mathbf{t}_1, \mathbf{a}_i^T\mathbf{t}_2] \\ &= t_1^T \,\text{Cov}(\mathbf{a}_{i,},\mathbf{a}_i)\, t_2 \\ &= t_1^T\, C_G\, t_2 \end{aligned}$$

This is a bilinear form (the generalization of a quadratic form).

$$\mathrm{Cov}[A(t_1), A(t_2)] = \mathbf{t}_1^T \mathbf{C}_\mathbf{G}\, \mathbf{t}_2$$

$$= \begin{pmatrix} 1 & t_1 & t_1^2 \end{pmatrix} \begin{pmatrix} \sigma_0^2 & \sigma_{01} & \sigma_{02} \\ \sigma_{01} & \sigma_1^2 & \sigma_{12} \\ \sigma_{02} & \sigma_{12} & \sigma_2^2 \end{pmatrix} \begin{pmatrix} 1 \\ t_2 \\ t_2^2 \end{pmatrix}$$

Expanding gives

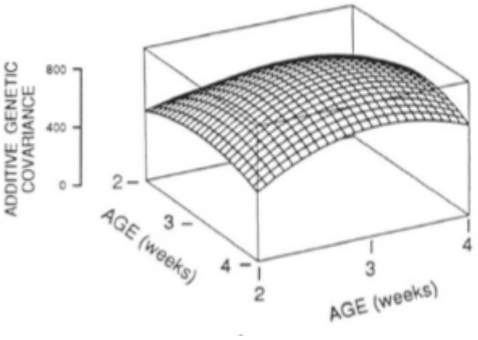
$$\mathrm{Cov}[A(t_1), A(t_2)] = \sigma_0^2 + \sigma_{01}(t_1 + t_2) + \sigma_{02}(t_1^2 + t_2^2) \\ + \sigma_1^2 t_1\, t_2 + \sigma_{12}(t_1^2\, t_2 + t_1\, t_2^2) + \sigma_2^2\, t_1^2\, t_2^2$$

More generally, fitting an m-th degree polynomial for A gives
the product of two m-degree polynomials for the covariance function

$$A_i(t) = \sum_{j=0}^{m} a_{ij}\, t^j$$

$$\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \sum_{j=0}^{m} \sum_{k=0}^{m} a_{jk}\, t_1^j\, t_2^k$$

Kirkpatrick estimated to covariance function for the Riska data by assuming an individual's breeding value over time can be modeled by 2nd degree polynomial. The resulting covariance function gives the following surface:



Estimated additive-genetic covariance function

Details

- Before building on these basic ideas to estimate the covariance function, some background on Legendre polynominals is required, as these are used as the basis functions (building blocks) for curve-fitting instead of the set $(1, t, t^2, \ldots t^k)$
  - Specifically, we could approximate a function $f(t)$ by the k-th degree polynomial $f(t) = \Sigma^k a_i t^i$.
  - Instead, we approximate it by a weighted sum of the functions $\phi_0(t), \phi_1(t), \ldots, \phi_k(t)$, where $\phi_j(t)$ is a polynomial of degree j (the Legendre polynomial of order j, for $0 \leq j \leq k$), using $f(t) = \Sigma^k b_i \phi_i(t)$.

Legendre Polynomials

For curve-fitting, orthogonal polynomials are often used, where $\phi_k(t)$ denotes a k-th degree polynomial. The set of these building blocks $\phi_o(t), \phi_1(t), \ldots \phi_k(t)$ .. are defined to be orthogonal in the sense that the integral of $\phi_i(t)\, \phi_j(t) = 0$ when i and j are not equal. We also assume they are scaled to have unit length, with the integral $\phi_i^2(t) = 1$.

For $-1 \leq t \leq 1$, the first five scaled Legendre polynomials are given by

$$\phi_0(t) = 0.7071$$
$$\phi_1(t) = 1.2247\, t$$
$$\phi_2(t) = -0.7906 + 2.3717\, t^2$$
$$\phi_3(t) = -2.8062\, t + 4.6771\, t^3$$
$$\phi_4(t) = 0.7955 - 7.9550\, t^2 + 9.2808\, t^4$$
$$\phi_5(t) = 4.2973\, t - 20.5205\, t^3 + 18.4685\, t^5$$

For example, the curve $y = a + b\,t$ can be written as
$y = a/(0.7071)\, \phi_0(t) + b/(1.2247)\, \phi_1(t)$ for $-1 \leq t \leq 1$.
More generally, any k-th degree polynomial can be written as
$\Sigma^\kappa a_i\, \phi_i(t)$

$\phi_0(t) = 0.7071$
$\phi_1(t) = 1.2247\ t$
$\phi_2(t) = -0.7906 + 2.3717\ t^2$
$\phi_3(t) = -2.8062\ t + 4.6771\ t^3$
$\phi_4(t) = 0.7955 - 7.9550\ t^2 + 9.2808\ t^4$
$\phi_5(t) = 4.2973\ t - 20.5205\ t^3 + 18.4685\ t^5$

In matrix form, $\phi = \mathbf{M}\mathbf{t}, \quad \text{where} \quad \phi = \begin{pmatrix} \phi_0(t) \\ \phi_1(t) \\ \phi_2(t) \\ \phi_3(t) \\ \phi_4(t) \\ \phi_5(t) \end{pmatrix}, \quad \mathbf{t} = \begin{pmatrix} 1 \\ t \\ t^2 \\ t^3 \\ t^4 \\ t^5 \end{pmatrix}$

j-th row of **M** are the coefficients for the jth Legendre polynomial

Row 4 = coefficients for ϕ_4.

$$\mathbf{M} = \begin{array}{c} \begin{pmatrix} 0.7071 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2247 & 0 & 0 & 0 & 0 \\ -0.7906 & 0 & 2.3717 & 0 & 0 & 0 \\ 0 & -2.8062 & 0 & 4.5777 & 0 & 0 \\ 0.7944 & 0 & -7.9950 & 0 & 9.2808 & 0 \\ 0 & 4.2973 & 0 & -20.5205 & 0 & 18.4685 \end{pmatrix} \\ \begin{array}{cccccc} 1 & t & t^2 & t^3 & t^4 & t^5 \end{array} \end{array}$$

How do we write the following 5th order polynomial in terms of Legendre polynomials?

$$y = 4 - 6x + 14x^2 + 26x^3 + 50x^4 - 110x^5$$

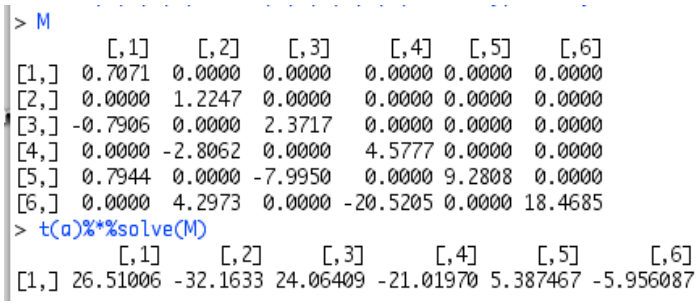
Note that y = $\mathbf{a}^T\mathbf{x}$, where $\mathbf{a} = \begin{pmatrix} 4 \\ -6 \\ 14 \\ 26 \\ 50 \\ -110 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix}$

$$\begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \phi_5(x) \end{pmatrix} = \mathbf{M} \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix} \quad \text{implies} \quad \begin{pmatrix} 1 \\ x \\ x^2 \\ x^3 \\ x^4 \\ x^5 \end{pmatrix} = \mathbf{M}^{-1} \begin{pmatrix} \phi_0(x) \\ \phi_1(x) \\ \phi_2(x) \\ \phi_3(x) \\ \phi_4(x) \\ \phi_5(x) \end{pmatrix}$$

Giving $\mathbf{x} = \mathbf{M}^{-1}\phi$. Since $\mathbf{y} = \mathbf{a}^T\mathbf{x} = \mathbf{a}^T\mathbf{M}^{-1}\phi$, weights on Legendre polynomials are given by $\mathbf{a}^T\mathbf{M}^{-1}$

Weights are given by $\mathbf{a}^T\mathbf{M}^{-1}$

R returns



```
> M
        [,1]    [,2]    [,3]     [,4]   [,5]    [,6]
[1,]  0.7071  0.0000  0.0000   0.0000 0.0000  0.0000
[2,]  0.0000  1.2247  0.0000   0.0000 0.0000  0.0000
[3,] -0.7906  0.0000  2.3717   0.0000 0.0000  0.0000
[4,]  0.0000 -2.8062  0.0000   4.5777 0.0000  0.0000
[5,]  0.7944  0.0000 -7.9950   0.0000 9.2808  0.0000
[6,]  0.0000  4.2973  0.0000 -20.5205 0.0000 18.4685
> t(a)%*%solve(M)
         [,1]     [,2]     [,3]      [,4]     [,5]      [,6]
[1,] 26.51006 -32.1633 24.06409 -21.01970 5.387467 -5.956087
```

Giving y = 26.51006*ϕ_0(x) -32.1633 *ϕ_1(x) + 24.06409 *ϕ_2(x) -21.01970 *ϕ_3(x) + 5.387467 *ϕ_4(x) -5.956087 *ϕ_5(x)

More generally, any k-degree polynomial $\mathbf{y} = \mathbf{a}^T\mathbf{x}_k$ can be expressed as a weighted series of the first k+1 Legendre polynomials ϕ_0, .., ϕ_k, where the weights are $\mathbf{a}^T\mathbf{M}^{-1}$. **M** is (k+1) x (k+1), with the jth row being the coefficients on x for the j-th order Legendre polynomial.

The Covariance function in terms of Legendre polynomials

- Express the trait breeding value for individual i at time t_j by an m-th order polynomial,
  - $A_i(t_j) = \Sigma_k{}^m\, a_{ik}\, \phi_k(t_j)$, where $\mathbf{a}_i \sim 0, \mathbf{C}_G$
  - Define the vectors
    - $\boldsymbol{\phi}_m(t) = (\phi_0(t), \phi_1(t), \ldots, \phi_m(t))^T$, which we often write as just $\boldsymbol{\phi}_m$ or $\boldsymbol{\phi}$ for brevity
    - $\mathbf{a}_i = (a_{i0}, a_{i1}, \ldots, a_{im})^T$.
- Hence $A_i(t_j) = \boldsymbol{\phi}_m(t)^T\mathbf{a}_i = \mathbf{a}_i^T\boldsymbol{\phi}_m(t)$.
    - $\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \mathrm{Cov}[\mathbf{a}_i^T \boldsymbol{\phi}_m(t_1), \mathbf{a}_i^T \boldsymbol{\phi}_m(t_2)]$
    - $\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \boldsymbol{\phi}_m(t_1)^T\, \mathbf{C}_G\, \boldsymbol{\phi}_m(t_2)$

Covariance function (cont)

- $\mathrm{Cov}[A_i(t_1), A_i(t_2)] = \phi_m(t_1)^T \, C_G \, \phi_m(t_2)$
- Recall for $\mathbf{t}_m = (1, t, t^2, \ldots, t^m)^T$ that
  - $\phi_m(t) = \mathbf{M}\mathbf{t}_m$, where M is the (m+1) x (m+1) matrix of coefficients for the first (m+1) Legendre polynomials
- Substituting in $\phi(t) = \mathbf{Mt}$ yields
  - $\mathrm{Cov}[A_i(t_1), A_i(t_2)\,] = \mathbf{t}_1^T \, \mathbf{M}^T C_G \mathbf{M} \, \mathbf{t}_2$, or
  - $\mathrm{Cov}[A_i(t_1), A_i(t_2)\,] = \mathbf{t}_1^T \, \mathbf{H} \, \mathbf{t}_2$, with $\mathbf{H} = \mathbf{M}^T C_G \mathbf{M}$
    - This allows us to express the covariance function in terms t_1 and t_2 directly

From G to C_G

- The key component to the covariance function is the covariance matrix C_G for the additive genetic random regression coefficients. How do we obtain this?
- We start with what Kirkpatrick called the "full estimate"
  - Given an estimated **G** matrix of the trait measured at m time points, we can describe trait breeding value as an m-1 degree polynomial
  - This is done as a weighted combination of the first m Legendre polynomials, ϕ_0, ϕ_1, … ϕ_{m-1}.
  - $G_{ij} = \mathrm{Cov}[A\,(t_i), A\,(t_j)\,] = \phi_m(t_i) \, C_G \, \phi_m(t_j)^T$

The full estimate does an element-by-element matching of G to functions of $\phi_m(t_i)$ (which are known constants) and C_G.

$$\mathbf{G} = \begin{pmatrix} G_{11} & \cdots & G_{1m} \\ \vdots & \ddots & \vdots \\ G_{m1} & \cdots & G_{mm} \end{pmatrix}, \quad \text{where} \quad G_{ij} = \boldsymbol{\phi}^T(t_i)\mathbf{G}_{\mathbf{C}}\boldsymbol{\phi}(t_j)$$

$$= \begin{pmatrix} \boldsymbol{\phi}^T(t_1)\mathbf{G}_{\mathbf{C}}\boldsymbol{\phi}(t_1) & \cdots & \boldsymbol{\phi}^T(t_1)\mathbf{G}_{\mathbf{C}}\boldsymbol{\phi}(t_m) \\ \vdots & \ddots & \vdots \\ \boldsymbol{\phi}^T(t_m)\mathbf{G}_{\mathbf{C}}\boldsymbol{\phi}(t_1) & \cdots & \boldsymbol{\phi}^T(t_m)\mathbf{G}_{\mathbf{C}}\boldsymbol{\phi}(t_m) \end{pmatrix}$$

$$= \begin{pmatrix} \boldsymbol{\phi}^T(t_1) \\ \vdots \\ \boldsymbol{\phi}^T(t_m) \end{pmatrix} \mathbf{G}_{\mathbf{C}} \begin{pmatrix} \boldsymbol{\phi}(t_1) \\ \vdots \\ \boldsymbol{\phi}(t_m) \end{pmatrix} = \boldsymbol{\Phi}^T\mathbf{G}_{\mathbf{C}}\boldsymbol{\Phi}$$

$$\mathbf{G} = \boldsymbol{\Phi}^T\mathbf{G}_{\mathbf{C}}\boldsymbol{\Phi} \quad \text{implies} \quad \mathbf{G}_{\mathbf{C}} = \left(\boldsymbol{\Phi}^T\right)^{-1}\mathbf{G}\boldsymbol{\Phi}^{-1}$$

where

$$\boldsymbol{\Phi}^T = \begin{pmatrix} \boldsymbol{\phi}^T(t_1) \\ \boldsymbol{\phi}^T(t_2) \\ \vdots \\ \boldsymbol{\phi}^T(t_m) \end{pmatrix} = \begin{pmatrix} \phi_0(t_1) & \phi_1(t_1) & \cdots & \phi_{m-1}(t_1) \\ \phi_0(t_2) & \phi_1(t_2) & \cdots & \phi_{m-1}(t_2) \\ \vdots & \vdots & \cdots & \vdots \\ \phi_0(t_m) & \phi_1(t_m) & \cdots & \phi_{m-1}(t_m) \end{pmatrix}$$

Note that $\boldsymbol{\Phi}$ is a matrix of constants --- the Legendre polynomials evaluated at the sample time points. Note that time points are scaled to be within (-1, 1), so ordering time on the original scale as $T_1 < \ldots < T_m$, scaled values are given by $t_i = 2(T_i - T_1)/(T_m - T_1) - 1$

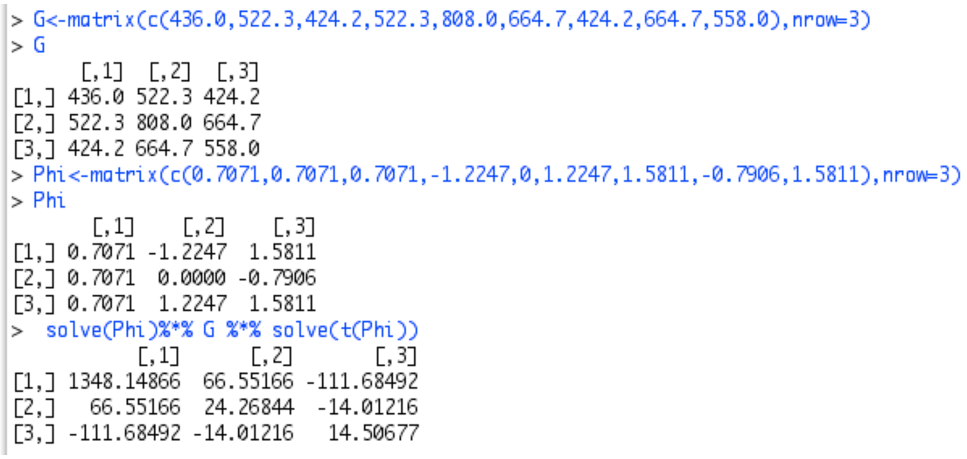
Example: Riska's data

$$\mathbf{G} = \begin{pmatrix} 436.0 & 522.3 & 424.2 \\ 522.3 & 808.0 & 664.7 \\ 424.2 & 664.7 & 558.0 \end{pmatrix}$$

$$\begin{array}{ll} \boldsymbol{\Phi}^T = \begin{pmatrix} \phi_0(-1) & \phi_1(-1) & \phi_2(-1) \\ \phi_0(0) & \phi_1(0) & \phi_2(0) \\ \phi_0(1) & \phi_1(1) & \phi_2(1) \end{pmatrix} & \begin{array}{l} \leftarrow\cdots\text{ 2 weeks, t = -1} \\ \leftarrow\cdots\text{ 3 weeks, t = 0} \\ \leftarrow\cdots\text{ 4 weeks, t = 1} \end{array} \end{array}$$

$$= \begin{pmatrix} 0.7071 & -1.2247 & 1.5811 \\ 0.7071 & 0 & -0.7906 \\ 0.7071 & 1.2247 & 1.5811 \end{pmatrix}$$

$$\mathbf{G_C} = \left(\boldsymbol{\Phi}^T\right)^{-1}\mathbf{G}\boldsymbol{\Phi}^{-1} = \begin{pmatrix} 1348.1 & 66.6 & -111.7 \\ 66.6 & 24.2 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{pmatrix}$$



```
> G<-matrix(c(436.0,522.3,424.2,522.3,808.0,664.7,424.2,664.7,558.0),nrow=3)
> G
      [,1]  [,2]  [,3]
[1,] 436.0 522.3 424.2
[2,] 522.3 808.0 664.7
[3,] 424.2 664.7 558.0
> Phi<-matrix(c(0.7071,0.7071,0.7071,-1.2247,0,1.2247,1.5811,-0.7906,1.5811),nrow=3)
> Phi
       [,1]    [,2]    [,3]
[1,] 0.7071 -1.2247  1.5811
[2,] 0.7071  0.0000 -0.7906
[3,] 0.7071  1.2247  1.5811
>  solve(Phi)%*% G %*% solve(t(Phi))
           [,1]      [,2]       [,3]
[1,] 1348.14866  66.55166 -111.68492
[2,]   66.55166  24.26844  -14.01216
[3,] -111.68492 -14.01216   14.50677
```

The resulting covariance function becomes

$$\mathrm{Cov}(t_1,t_2) = \boldsymbol{\phi}^T(t_1)\mathbf{G_C}\boldsymbol{\phi}(t_2)$$

$$= \begin{pmatrix}\phi_0(t_1) & \phi_1(t_1) & \phi_2(t_1)\end{pmatrix}\begin{pmatrix}1348.1 & 66.6 & -111.7\\ 66.6 & 24.2 & -14.0\\ -111.7 & -14.0 & 14.5\end{pmatrix}\begin{pmatrix}\phi_0(t_1)\\ \phi_1(t_1)\\ \phi_2(t_1)\end{pmatrix}$$

This bilinear form expresses the covariance function in terms of the Legendre polynomials. Usually we would like to express this as a polynomial in t_1 & t_2:

One could do this by first substituting in the polynomial form for $\phi_i(t)$, expanding and collecting terms. However, much easier to do this in matrix form. Recall the coefficient matrix **M** from earlier in the notes, where $\phi = Mt$. Writing the covariance function as $\phi_1^T G_C \phi_2 = (Mt_1)^T G_C(Mt_2) = t_1^T M^T G_C M t_2 = t_1^T H t_2$, where $H = M^T C_G M$.

The covariance function becomes $t_1^T H t_2$, with $H = M^T C_G M$

Since the first three Legendre polynomials are used, **M** is 3 x 3

$$\mathbf{M} = \begin{pmatrix}0.7071 & 0 & 0\\ 0 & 1.2247 & 0\\ -0.7906 & 0 & 2.3717\end{pmatrix}$$

$H = M^T C_G M$ gives

$$\mathbf{H} = \begin{pmatrix}808.0 & 71.2 & -214.5\\ 71.2 & 36.4 & -40.7\\ -214.5 & -40.7 & 81.6\end{pmatrix}$$

Expanding this out gives

$$\begin{aligned}\mathrm{Cov}(A_1,A_2) = {} & 808 + 71.2(t_1 + t_2) + 36.4\, t_1 t_2\\ & - 40.7(t_1^2 t_2 + t_1 t_2^2) - 215.0(t_1^2 + t_2^2)\\ & + 81.6 t_1^2 t_2^2\end{aligned}$$

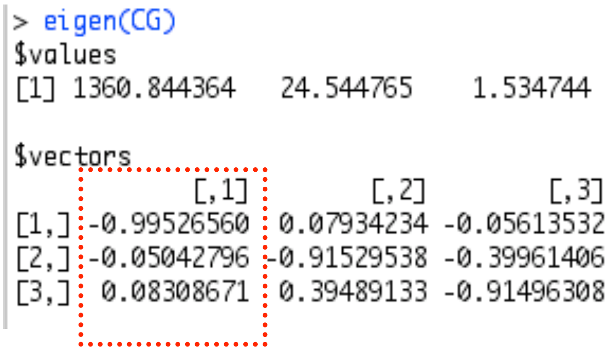
More generally, the coefficient on $t_1^{i-1} t_2^{j-1}$ in the covariance expansion is given by H_{ij}. -- the (i,j)-th element of **H**.

The Eigenstructure of C_G

- The variance-covariance matrix C_G of the random regression coefficients is extremely information on the nature of variation for the function-valued trait.
- The function-valued analogue of the eigenvector is the eigenfunction, which also has an associated eigenvalue. Akin to the eigvenvector associated with the largest eigenvalue accounting for the largest single direction of variation, the eigenfunction associated with the largest eigenvalue is the functional curve associated with the most variation.
- The eigenvalues of C_G are the same as those for the covariance function, while the associated eigenvectors of C_G give the weights on the orthogonal polynomials that recover the eigenfunctions of the covariance function.

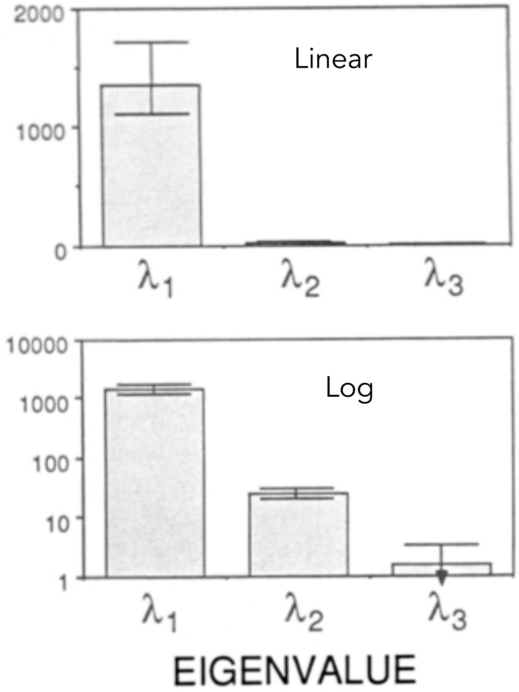
Back to Riska's data

$$\mathbf{G_C} = \left(\boldsymbol{\Phi}^T\right)^{-1} \mathbf{G}\boldsymbol{\Phi}^{-1} = \begin{pmatrix} 1348.1 & 66.6 & -111.7 \\ 66.6 & 24.2 & -14.0 \\ -111.7 & -14.0 & 14.5 \end{pmatrix}$$



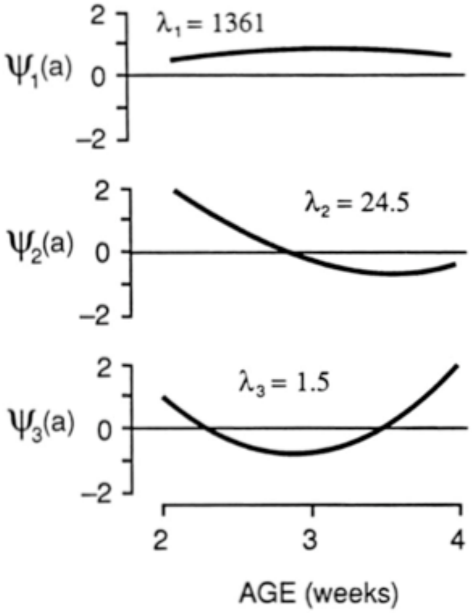
```
> eigen(CG)
$values
[1] 1360.844364   24.544765    1.534744

$vectors
            [,1]        [,2]        [,3]
[1,] -0.99526560  0.07934234 -0.05613532
[2,] -0.05042796 -0.91529538 -0.39961406
[3,]  0.08308671  0.39489133 -0.91496308
```

First eigenvector



CG has a dominant eigenvalue --- most of the variation in the breeding value for growth is along one curve

Associated eigenfunctions for C_G for the Riska dataset



Eigenfunctions of C_G

- If $\mathbf{e}_i$ denotes the eigenvector associated with the ith eigenvalue λ_i of C_G, then for the covariance function
  - λ_i is the ith eigenvalue
  - associated eigenfunction is $\underline{\phi}_m(t)^T \mathbf{e}_i$
  - $= e_{i1}\phi_0(t) + e_{i2}\phi_1(t) + \cdots + e_{im}\phi_{m-1}(t)$
  - Since $\phi = \mathbf{Mt}$, we have $(\mathbf{Mt})^T \mathbf{e}_i = t^T (M^T \mathbf{e}_i)$, giving the weights on $(1, t, t^2, .. ,t^{m-1})$ as $M^T \mathbf{e}_i$
  - For Riska's data, the leading eigenfunction is
  - $\psi_1(t) = 0.7693 - 0.0617\ t - 0.1971\ t^2$

Eigenfunctions: $\psi_i(t) = t^T (M^T \mathbf{e}_i)$

$$\mathbf{M} = \begin{pmatrix} 0.7071 & 0 & 0 \\ 0 & 1.2247 & 0 \\ -0.7906 & 0 & 2.3717 \end{pmatrix}$$

$$\mathbf{e}_1 = \begin{pmatrix} 0.995 \\ 0.050 \\ -0.083 \end{pmatrix}, \qquad \mathbf{e}_2 = \begin{pmatrix} -0.079 \\ 0.915 \\ -0.395 \end{pmatrix}, \qquad \mathbf{e}_3 = \begin{pmatrix} 0.056 \\ 0.400 \\ 0.915 \end{pmatrix}$$

$$\mathbf{M}^T\mathbf{e}_1 = \begin{pmatrix} 0.769 \\ 0.062 \\ -0.197 \end{pmatrix}, \quad \mathbf{M}^T\mathbf{e}_2 = \begin{pmatrix} 0.256 \\ 1.121 \\ -0.937 \end{pmatrix}, \quad \mathbf{M}^T\mathbf{e}_3 = \begin{pmatrix} -0.684 \\ 0.490 \\ 2.170 \end{pmatrix}$$

$\psi_2(t) = 0.256 + 1.121*t - 0.937*t^2$

$\psi_3(t) = -0.684 + 0.490*t +2.170*t^2$

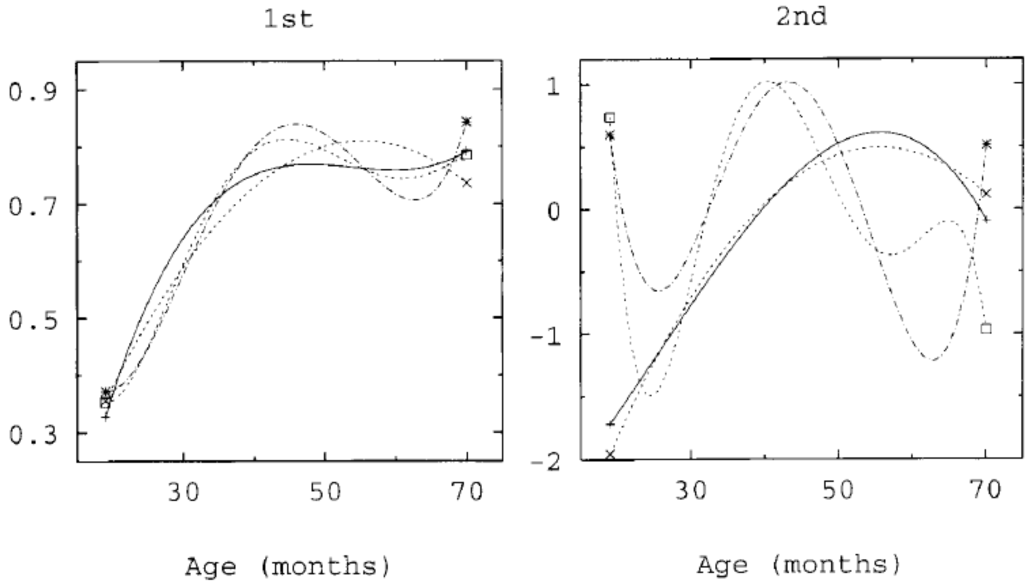


Figure 3. Estimated first and second eigenfunction of the genetic covariance function, for orders of polynomial fit of 3 (×), 4 (+), 5 (∗) and 6 (□), respectively (rank 3 estimates of the coefficient matrices).

Meyer's data on Cattle Weight

Over-fitting G_C?

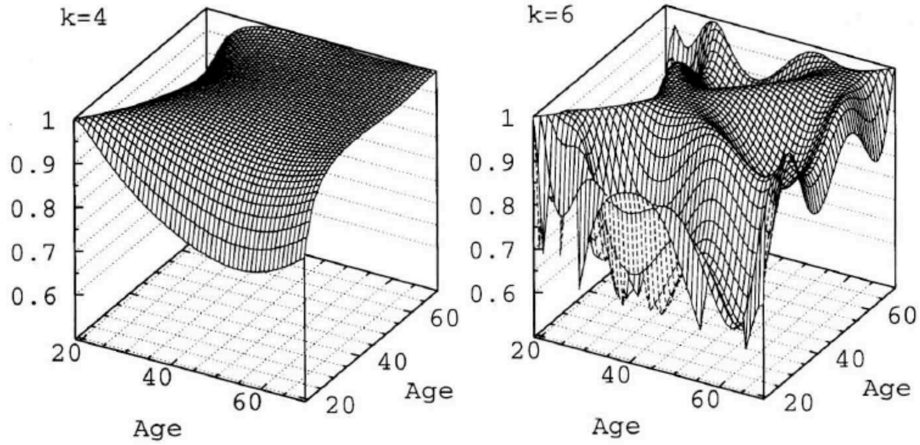


Figure 5. Estimates of genetic correlations for orders of polynomial fit (k) of 4 and 6.

Meyer's data showing how increasing the degree of polynomial used results in over-fitting. In her words: "surfaces become 'wiggly' "

Reduced estimation of C_G

- While the full estimate (rank $\mathbf{C_G}$ = rank of observed $\mathbf{G}$) is (relatively) straightforward, this likely results in an overfit of the data, as the covariance function is forced to exactly fit the observed values for all t_1, t_2, some of which are sampling noise
  - Results in a less smooth covariance function than one based on using a reduced dimension.
  - Kirkpatrick originally suggested a least-squares approach, while Meyer & Hill suggested a REML-based approach
  - Key breakthrough, first noticed by Goddard, and fully developed by Meyer, is the connection between covariance functions and random regressions.
  - This should not be surprising given that we started with random regressions to motivate covariance functions.
  - The key is that standard BLUP approaches (for multivariate traits) can be used for random regressions.

Mixed-Models (BLUPs) for Longitudinal traits

- Simplest setting is the repeatability model, the trait breeding and residual (permanent environmental) values are assumed constant over time. The jth observation on i is
  - $y_{ij} = u + a_i + pe_i + e_{ij}$
  - $\mathbf{a} \sim \mathbf{0}$, Var(A)$\mathbf{A}$
- At the other extreme is the multiple-trait approach, where each sampled time point is considered as a separate, but correlated, trait. Here y_{ij} is the jth "trait" (sampled time point) for individual i.
  - $y_{ij} = u + a_{ij} + e_{ij}$
  - $\mathbf{a} \sim 0$, $\mathbf{G} \text{ X } \mathbf{A}$
- In the middle are random-regressions, where for the jth observation (time t_j) on individual i is
  - $y_{ij} = u + \Sigma_k^n\, a_{ik}\phi_k(t_j) + \Sigma_k^m\, pe_{ik}\phi_k(t_j) + e_{ij}$
  - $\mathbf{a}_i \sim \mathbf{0}$, $\mathbf{C_G}$ and $\mathbf{p}_i \sim \mathbf{0}$, $\mathbf{C_E}$

The repeatability model

- The repeatability model assumes that the trait is unchanging between observations, but multiple observations (records) are taken over time to smooth out sampling noise (e)
- 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 $\sigma_e^2 = (1-r)\, \sigma_z^2$

Mixed-model $\quad \mathbf{y = X\beta + Za + Zp + e}$

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

Mixed-model equations

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

where

$$\lambda_A = \frac{\sigma_e^2}{\sigma_A^2} = \frac{1-r}{h^2} \quad \text{and} \quad \lambda_u = \frac{\sigma_e^2}{\sigma_p^2} = \frac{1-r}{r-h^2}$$

The multiple-trait model

- With a clearly discrete number of stages (say k), a longitudinal trait could be modeled as k correlated traits, so that individual i has values y_{i1}, y_{i2}, .., y_{ik}.
- In this case, there is no need for permanent environmental effects, as these now appear in correlations among the residuals, the within-individual environmental correlations (which are estimated by REML).
- This can be put into standard Mixed Model equations by simply "stacking" the vectors for each trait to create one vector for each random effect.

For trait j (1 ≤ j ≤ k), the mixed model becomes

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta}_j + \mathbf{Z}_j\mathbf{a}_i + \mathbf{e}_j$$

$$\begin{pmatrix} \mathbf{a}_j \\ \mathbf{e}_j \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma^2_{A_j}\,\mathbf{A} & \mathbf{0} \\ \mathbf{0} & \sigma^2_{e_j}\,\mathbf{I} \end{pmatrix}$$

We can write this as **y** = **Xβ** + **Za** + **e**, where

$$\begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_k \end{pmatrix} = \begin{pmatrix} \mathbf{X}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{X}_k \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_k \end{pmatrix} + \begin{pmatrix} \mathbf{Z}_1 & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{Z}_k \end{pmatrix} \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_k \end{pmatrix} + \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_k \end{pmatrix}$$

Again, the BLUP for the vector of all EBVs is given by

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

With **V** the covariance structure for this model

Covariance structure for EBVS

The resulting covariance structure for the stacked vector of breeding values is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\end{pmatrix}=\begin{pmatrix}\sigma^2(A_1)\mathbf{A} & \cdots & \sigma(A_1,A_k)\mathbf{A}\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1)\mathbf{A} & \cdots & \sigma^2(A_k)\mathbf{A}\end{pmatrix}=\mathbf{G}\otimes\mathbf{A}$$

where ⊗ denotes the Kronecker (or direct) product (LW Chapter 26) and

$$\mathbf{G}=\begin{pmatrix}\sigma^2(A_1) & \cdots & \sigma(A_1,A_k)\\ \vdots & \ddots & \vdots\\ \sigma(A_k,A_1) & \cdots & \sigma^2(A_k)\end{pmatrix}$$

is the matrix of genetic covariances of interest.

The genetic variance-covariance matrix G accounts for the genetic covariances among traits. G has k variances and k(k-1)/2 covariances, which must be estimated (REML) from the data.

Covariance structure for residuals

Similarly, the covariance structure for the stacked vectors of residuals is

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{e}_1\\ \vdots\\ \mathbf{e}_k\end{pmatrix}=\mathbf{E}\otimes\mathbf{I},\quad\text{where}\quad\mathbf{E}=\begin{pmatrix}\sigma^2(e_1) & \cdots & \sigma(e_1,e_k)\\ \vdots & \ddots & \vdots\\ \sigma(e_k,e_1) & \cdots & \sigma^2(e_k)\end{pmatrix}$$

Finally, we need to specify any covariances between **a** and **e**. By construction $\sigma(a_z,e_z)=\sigma(a_w,e_w)=0$, while the standard assumption is $\sigma(A_z,e_w)=\sigma(A_w,e_z)=0$, giving the covariance structure as

$$\boldsymbol{\sigma}\begin{pmatrix}\mathbf{a}_1\\ \vdots\\ \mathbf{a}_k\\ \mathbf{e}_1\\ \vdots\\ \mathbf{e}_k\end{pmatrix}=\begin{pmatrix}\mathbf{G}\otimes\mathbf{A} & \mathbf{0}\\ \mathbf{0} & \mathbf{E}\otimes\mathbf{I}\end{pmatrix}$$

Here the matrix E accounts for within-individual correlations in the environmental (or residual) values.

Random regressions

- Random regression models are basically a hybrid between repeated records models and multiple-trait models.
  - The basic structure of the model is that the trait at time t is the sum of potentially time-dependent fixed effects $\mu(t)$, a time-dependent breeding value $a(t)$, a time-dependent permanent environmental effect $p(t)$, and a residual error e. These last three are random effects
  - $y(t) = \mu(t) + a(t) + p(t) + e$
  - $a(t)$ and $p(t)$ are both approximated by random regressions, of order n and m, respectively (usually $n = m$)
  - $a_i(t_j) = \Sigma_k^n a_{ik}\phi_k(t_j)$ and $p_i(t_j) = \Sigma_k^m b_{ik}\phi_k(t_j)$
  - The vectors $\mathbf{a}_i$ and $\mathbf{b}_i$ for individual i are handled in a multiple-trait framework, with covariance matrices $\mathbf{C}_G$ and $\mathbf{C}_E$ for the within-individual vectors of additive and permanent environmental effects.

To build up the random regression model, consider the q_i observations from different times for individual i

$$\mathbf{y}_i = \begin{pmatrix} y(t_{i1}) \\ \vdots \\ y(t_{iq_i}) \end{pmatrix} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{a}_i = \begin{pmatrix} a_{i0} \\ \vdots \\ a_{im} \end{pmatrix}, \quad \mathbf{p}_i = \begin{pmatrix} p_{i0} \\ \vdots \\ p_{im} \end{pmatrix}, \quad \mathbf{e}_i = \begin{pmatrix} e_{i0} \\ \vdots \\ e_{im} \end{pmatrix}$$

Here are fitting m-degree polynomials ($m < q_i$) for both the breeding value and permanent environmental value regressions. We also assume that any fixed-effects are not time dependent. Both of these assumptions are easily relaxed.

Model & covariance structure for vector **y**$_i$ of observations from individual i

$$\mathbf{y}_i = \begin{pmatrix} y(t_{i1}) \\ \vdots \\ y(t_{iq_i}) \end{pmatrix} = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{a}_i = \begin{pmatrix} a_{i0} \\ \vdots \\ a_{im} \end{pmatrix}, \quad \mathbf{p}_i = \begin{pmatrix} p_{i0} \\ \vdots \\ p_{im} \end{pmatrix}, \quad \mathbf{e}_i = \begin{pmatrix} e_{i0} \\ \vdots \\ e_{im} \end{pmatrix}$$

Covariance structure

$$\begin{pmatrix} \mathbf{a}_i \\ \mathbf{p}_i \\ \mathbf{e}_i \end{pmatrix} \sim \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \mathbf{C_G} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C_E} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \sigma_e^2\,\mathbf{I} \end{pmatrix}$$

The design matrix for the regression coefficients on the breeding values is very information

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \mathbf{Z}_{i1}\mathbf{a}_i + \mathbf{Z}_{i2}\mathbf{p}_i + \mathbf{e}_i$$

$$\mathbf{Z}_{i1} = \begin{pmatrix} \phi_0(t_{i1}) & \cdots & \phi_m(t_{i1}) \\ \phi_0(t_{i2}) & \cdots & \phi_m(t_{i2}) \\ \vdots & \ddots & \vdots \\ \phi_0(t_{iq_i}) & \cdots & \phi_m(t_{iq_i}) \end{pmatrix}$$

$\mathbf{Z}_{i1}$ is a q_i x (m+1) matrix of fixed constants that depend on the values of order zero through m Legendre polynomials, where the jth row represents these evaluated at time t_{ij}.
A KEY FEATURE is that this set of times could be different for each individual, yet the mixed model does all the bookkeeping to fully account for this.

As with the multiple trait model, stacking the individual vectors allows us to put this model in standard form. Note that while the vectors stacked for the multiple trait model represented the vectors for each trait separately, here the stacked vectors are the observations for each individual.

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}$$

Z_1, Z_2 Block diagonal

$$\mathbf{Z}_1 = \begin{pmatrix} \mathbf{Z}_{11} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_{12} & \cdots & \mathbf{0} \\ \vdots & & \ddots & \mathbf{0} \\ \mathbf{0} & \cdots & \cdots & \mathbf{Z}_{1n} \end{pmatrix}$$

Full Model & covariance structure

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{pmatrix}, \quad \mathbf{a} = \begin{pmatrix} \mathbf{a}_1 \\ \vdots \\ \mathbf{a}_n \end{pmatrix}, \quad \mathbf{p} = \begin{pmatrix} \mathbf{p}_1 \\ \vdots \\ \mathbf{p}_n \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} \mathbf{e}_1 \\ \vdots \\ \mathbf{e}_n \end{pmatrix}$$

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e}$$

Covariance structure

$$\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} \sim \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \mathbf{A} \otimes \mathbf{C}_{\mathbf{G}} & 0 & 0 \\ 0 & \mathbf{I} \otimes \mathbf{C}_{\mathbf{E}} & 0 \\ 0 & 0 & \sigma_e^2\,\mathbf{I} \end{pmatrix}$$

More generally, we can replace σ_e^2 I by **R**.

Mixed-model equations (slightly more generalized covariance structure)

$$\mathbf{H}\begin{pmatrix}\widehat{\mathbf{b}}\\ \widehat{\mathbf{a}}\\ \widehat{\mathbf{p}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{y}\end{pmatrix}\quad\begin{pmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{pmatrix}\sim\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}\mathbf{A}\otimes\mathbf{C}_{\mathbf{G}} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\otimes\mathbf{C}_{\mathbf{E}} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{pmatrix}$$

where

$$\mathbf{H}=\begin{pmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_1+\mathbf{A}^{-1}\otimes\mathbf{C}_{\mathbf{G}}^{-1} & \mathbf{Z}_1^T\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2^T\mathbf{R}^{-1}\mathbf{Z}_2+\mathbf{I}\otimes\mathbf{C}_{\mathbf{E}}^{-1}\end{pmatrix}$$

Model-fitting issues

- A central issue is what degree m of polynomials to use.
- Standard likelihood tests can be used (compare m = k with m = k + 1).
- Meyer suggests that tests should be comparing k with k + 2, as often going from odd to even does not improve fit, but going from even to even (k+2) does, and vice-versa.

Response to selection

- Standard BLUP selection can be used, based on some criteria for an optimal functional value (curve) in the offspring.
- The expected response in the offspring is simply obtained by substituting the average of the parental breeding values into the polynomial regression for the breeding value to generate an expected offspring curve.