Linear Models and Animal Breeding

Lawrence R. Schaeffer
Centre for Genetic Improvement of Livestock
Department of Animal and Poultry Science
University of Guelph
Guelph, ON N1G 2W1

June 2010 - Norway

Introduction to Course

1 Aim of the Course

The aim of the course is to present linear model methodology for genetic evaluation of livestock (plants) and general analysis of livestock (plant) data.

2 The Assumed Genetic Model

The **Infinitesimal Model** (1909) is assumed. There are an estimated 30,000 genes in mammals. If each gene has only two alleles (variants), then there are only 3 possible genotypes at each locus. The number of possible genotypes across all loci would be 3^{30000}, which is a number greater than the total number of animals in any livestock species. Many genes are likely to have more than two alleles, and hence the number of possibly different genotypes is even greater then 3^{30000}. Logically, one can assume a genome with essentially an infinite number of loci as being approximately correct for all practical purposes.

Infinitesimal Model: An infinite number of loci are assumed, all with an equal and small effect on a quantitative trait.

True Breeding Value: The sum of the additive effects of all loci on a quantitative trait is known as the True Breeding Value (TBV).

3 Types of Genetic Effects

In this course, only additive genetic effects are of interest. Additive effects are generally the largest of the genetic effects, and the allelic effects are passed directly to offspring while the other genetic effects are not transmitted to progeny, and are generally smaller in magnitude.

3.1 Additive Effects

Assume one locus with three alleles, A_1, A_2, and A_3. Assume also that the effect of the alleles are +3, +1, and -1, respectively. If the genetic effects are entirely **additive**, then the value of the possible genotypes would be the sum of their respective allele effects, i.e.

$$\begin{aligned}
A_1A_1 &= 3 \ + \ 3 = +6 \\
A_1A_2 &= 3 \ + \ 1 = +4 \\
A_1A_3 &= 3 \ - \ 1 = +2 \\
A_2A_2 &= 1 \ + \ 1 = +2 \\
A_2A_3 &= 1 \ - \ 1 = 0 \\
A_3A_3 &= -1 \ - \ 1 = -2.
\end{aligned}$$

3.2 Dominance Effects

Dominance genetic effects are the interactions among alleles at a given locus. This is an effect that is extra to the sum of the additive allelic effects. Each genotype would have its own dominance effect, let these be denoted as δ_{ij}, and each of them are non-zero quantities. Using the previous example, the additive and dominance effects would give

$$\begin{aligned}
A_1A_1 &= 3 \ + \ 3 \ + \ \delta_{11} = +6 + \delta_{11} \\
A_1A_2 &= 3 \ + \ 1 \ + \ \delta_{12} = +4 + \delta_{12} \\
A_1A_3 &= 3 \ - \ 1 \ + \ \delta_{13} = +2 + \delta_{13} \\
A_2A_2 &= 1 \ + \ 1 \ + \ \delta_{22} = +2 + \delta_{22} \\
A_2A_3 &= 1 \ - \ 1 \ + \ \delta_{23} = 0 + \delta_{23} \\
A_3A_3 &= -1 \ - \ 1 \ + \ \delta_{33} = -2 + \delta_{33}.
\end{aligned}$$

3.3 Epistatic Genetic Effects

Epistatic genetic effects encompass all possible interactions among the m loci (m being approximately 30,000). This includes all two way interactions, three way interactions, etc. As well, epistasis includes interactions between additive effects at different loci, interactions between additive effects at one locus with dominance effects at a second locus, and interactions between dominance effects at different loci.

4 Necessary Information

Genetic improvement of a livestock species requires four pieces of information.

4.1 Pedigrees

Animals, their sires, and their dams need to be uniquely identified in the data. Birthdates, breed composition, and genotypes for various markers or QTLs could also be stored. If animals are not uniquely identified, then genetic change of the population may not be possible. In aquaculture species, for example, individual identification may not be feasible, but family identification (sire and dam) may be known.

4.2 Data

Traits of economic importance need to be recorded accurately and completely. All animals within a production unit (herd, flock, ranch) should be recorded. Animals should not be selectively recorded. Data includes the dates of events when traits are observed, factors that could influence an animal's performance, and an identification of contemporaries that are raised and observed in the same environment under the same management regime. Observations should be objectively measured, if at all possible.

4.3 Production System

A knowledge and understanding of the production system of a livestock species is important for designing optimum selection and mating strategies. The key elements are the gestation length and the age at first breeding. The number of offspring per female per gestation will influence family structure. The use of artificial insemination and/or embryo transfer could be important. Other management practices are also useful to know. All information is used to formulate appropriate linear models for the analysis of the data and accurate estimation of breeding values of animals.

4.4 Prior Information

Read the literature. Most likely other researchers have already made analyses of the same species and traits. Their models could be useful starting points for further analyses. Their parameter estimates could predict the kinds of results that might be found. The idea is to avoid the pitfalls and problems that other researchers have already encountered. Be aware of new kinds of analyses of the same data, that may not involve linear models.

5 Tools for Genetic Evaluation

5.1 Statistical Linear Models

A model describes the factors that affect each trait in a linear fashion. That is, a factor has an additive effect on a trait. All models are simple approximations to how factors influence a trait. The goal is to find the best practical model that explains the most variation. Statistical knowledge is required.

5.2 Matrix Algebra

Matrix algebra is a notation for describing models and statistical analyses in a simplified manner. Matrix addition, multiplication, inversion, differentiation, and other operations should be mastered at an efficient level of expertise.

5.3 Computing

Animal breeders should know how to write programs, or at least to be capable of using software written by others. Available software may not be enough for some analyses, and so new programs may need to be written. The best languages for programming animal breeding work are FORTRAN (77 or 90), C, or C++. For this course, R software will be used to solve the example problems. R is an open source language available on the internet. Go to the CRAN site and download the latest version of R onto your desktop machine or laptop. R is continually being updated with one or two new versions per year. One version should suffice for at least a year. Some of the basics of R will be given in these notes.

6 EXERCISES

1. Search the literature or internet to fill in the missing values in the following table.

Species	Age at first breeding		Gestation	Pairs of
	Males(days)	Females (days)	Length (d)	Chromosomes
Cattle				
Swine				
Sheep				
Goats				
Horse				
Elk				
Deer				
Llama				
Rabbit				
Mink				
Chicken				
Turkey				
Dog				
Cat				
Mouse				

2. Describe a typical production system for one of the livestock species in the previous table.

3. Take one of the livestock species in the first question and make a file of the different breeds within that species and the number of animals in each breed today in Canada (your country). Include a picture and brief background of each breed.

4. Describe an existing data recording program for a quantitative trait in one species in Canada (your country).

7 Matrices

A matrix is a two dimensional array of numbers. The number of rows and number of columns defines the order of the matrix. Matrices are denoted by boldface capital letters.

7.1 Examples

$$\mathbf{A} = \begin{pmatrix} 7 & 18 & -2 & 22 \\ -16 & 3 & 55 & 1 \\ 9 & -4 & 0 & 31 \end{pmatrix}_{\mathbf{3\times 4}}$$

$$\mathbf{B} = \begin{pmatrix} x & y+1 & x+y+z \\ a-b & c\log d & e \\ \sqrt{x-y} & (m+n)/n & p \end{pmatrix}_{\mathbf{3\times 3}}$$

and

$$\mathbf{C} = \begin{pmatrix} \mathbf{C_{11}} & \mathbf{C_{12}} \\ \mathbf{C_{21}} & \mathbf{C_{22}} \end{pmatrix}_{\mathbf{2\times 2}}$$

7.2 Making a Matrix in R

```
A = matrix(data=c(7,18,-2,22,-16,3,55,1,9,-4,0,31),byrow=TRUE,
nrow=3,ncol=4)

# Check the dimensions
dim(A)
```

7.3 Vectors

Vectors are matrices with either one row (row vector) or one column (column vector), and are denoted by boldface small letters.

7.4 Scalar

A scalar is a matrix with just one row and one column, and is denoted by an letter or symbol.

8 Special Matrices

8.1 Square Matrix

A matrix with the same number of rows and columns.

8.2 Diagonal Matrix

Let $\{a_{ij}\}$ represent a single element in the matrix **A**. A diagonal matrix is a square matrix in which all a_{ij} are equal to zero except when i equals j.

8.3 Identity Matrix

This is a diagonal matrix with all a_{ii} equal to one (1). An identity matrix is usually written as **I**.

To make an identity matrix with r rows and columns, use

```
id = function(n) diag(c(1),nrow=n,ncol=n)

# To create an identity matrix of order 12
I12 = id(12)
```

8.4 J Matrix

A **J** matrix is a general matrix of any number of rows and columns, but in which all elements in the matrix are equal to one (1).

The following function will make a **J** matrix, given the number or rows, r, and number of columns, c.

```
jd = function(n,m) matrix(c(1),nrow=n,ncol=m)

# To make a matrix of 6 rows and 10 columns of all ones
M = jd(6,10)
```

8.5 Null Matrix

A null matrix is a **J** matrix multiplied by 0. That is, all elements of a null matrix are equal to 0.

8.6 Triangular Matrix

A lower triangular matrix is a square matrix where elements with j greater than i are equal to zero (0), $\{\ a_{ij}\}$ equal 0 for j greater than i. There is also an upper triangular matrix in which $\{a_{ij}\}$ equal 0 for i greater than j.

8.7 Tridiagonal Matrix

A tridiagonal matrix is a square matrix with all elements equal to zero except the diagonals and the elements immediately to the left and right of the diagonal. An example is shown below.

$$\mathbf{B} = \begin{pmatrix} 10 & 3 & 0 & 0 & 0 & 0 \\ 3 & 10 & 3 & 0 & 0 & 0 \\ 0 & 3 & 10 & 3 & 0 & 0 \\ 0 & 0 & 3 & 10 & 3 & 0 \\ 0 & 0 & 0 & 3 & 10 & 3 \\ 0 & 0 & 0 & 0 & 3 & 10 \end{pmatrix}.$$

9 Matrix Operations

9.1 Transposition

Let $\{a_{ij}\}$ represent a single element in the matrix $\mathbf{A}$. The transpose of $\mathbf{A}$ is defined as

$$\mathbf{A}' = \{a_{ji}\}.$$

If $\mathbf{A}$ has r rows and c columns, then $\mathbf{A}'$ has c rows and r columns.

$$\begin{aligned}
\mathbf{A} &= \begin{pmatrix} 7 & 18 & -2 & 22 \\ -16 & 3 & 55 & 1 \\ 9 & -4 & 0 & 31 \end{pmatrix} \\
\mathbf{A}' &= \begin{pmatrix} 7 & -16 & 9 \\ 18 & 3 & -4 \\ -2 & 55 & 0 \\ 22 & 1 & 31 \end{pmatrix}.
\end{aligned}$$

In R,

```
At = t(A)
# t() is the transpose function
```

9.2 Diagonals

The diagonals of matrix $\mathbf{A}$ are $\{a_{ii}\}$ for i going from 1 to the number of rows in the matrix.

Off-diagonal elements of a matrix are all other elements excluding the diagonals.

Diagonals can be extracted from a matrix in R by using the `diag()` function.

9.3 Addition of Matrices

Matrices are *conformable for addition* if they have the same order. The resulting sum is a matrix having the same number of rows and columns as the two matrices to be added. Matrices that are not of the same order cannot be added together.

$$\mathbf{A} = \{\mathbf{a_{ij}}\} \text{ and } \mathbf{B} = \{\mathbf{b_{ij}}\}$$

$$\mathbf{A} + \mathbf{B} = \{\mathbf{a_{ij}} + \mathbf{b_{ij}}\}.$$

An example is

$$\mathbf{A} = \begin{pmatrix} 4 & 5 & 3 \\ 6 & 0 & 2 \end{pmatrix} \text{ and } \mathbf{B} = \begin{pmatrix} 1 & 0 & 2 \\ 3 & 4 & 1 \end{pmatrix}$$

then

$$\mathbf{A} + \mathbf{B} = \begin{pmatrix} 4+1 & 5+0 & 3+2 \\ 6+3 & 0+4 & 2+1 \end{pmatrix}$$

$$= \begin{pmatrix} 5 & 5 & 5 \\ 9 & 4 & 3 \end{pmatrix} = \mathbf{B} + \mathbf{A}.$$

Subtraction is the addition of two matrices, one of which has all elements multiplied by a minus one (-1). That is,

$$\mathbf{A} + (-1)\mathbf{B} = \begin{pmatrix} 3 & 5 & 1 \\ 3 & -4 & 1 \end{pmatrix}.$$

R will check matrices for conformability, and will not perform the operation unless they are conformable.

9.4 Multiplication of Matrices

Two matrices are *conformable for multiplication* if the number of columns in the first matrix equals the number of rows in the second matrix.

If **C** has order $p \times q$ and **D** has order $m \times n$, then the product **CD** exists only if $q = m$. The product matrix has order $p \times n$.

In general, **CD** does not equal **DC**, and most often the product **DC** may not even exist because **D** may not be conformable for multiplication with **C**. Thus, the ordering of matrices in a product must be carefully and precisely written.

The computation of a product is defined as follows: let

$$\mathbf{C}_{p\times q} = \{c_{ij}\}$$

and

$$\mathbf{D}_{m\times n} = \{d_{ij}\}$$

and $q = m$, then

$$\mathbf{CD}_{p\times n} = \{\sum_{k=1}^{m} c_{ik}d_{kj}\}.$$

As an example, let

$$\mathbf{C} = \begin{pmatrix} 6 & 4 & -3 \\ 3 & 9 & -7 \\ 8 & 5 & -2 \end{pmatrix} \text{ and } \mathbf{D} = \begin{pmatrix} 1 & 1 \\ 2 & 0 \\ 3 & -1 \end{pmatrix},$$

then

$$\mathbf{CD} = \begin{pmatrix} 6(1)+4(2)-3(3) & 6(1)+4(0)-3(-1) \\ 3(1)+9(2)-7(3) & 3(1)+9(0)-7(-1) \\ 8(1)+5(2)-2(3) & 8(1)+5(0)-2(-1) \end{pmatrix} = \begin{pmatrix} 5 & 9 \\ 0 & 10 \\ 12 & 10 \end{pmatrix}.$$

In R,

```
# C times D - conformability is checked
CD = C %*% D
```

9.4.1 Transpose of a Product

The transpose of the product of two or more matrices is the product of the transposes of each matrix in reverse order. That is, the transpose of **CDE**, for example, is $\mathbf{E}'\mathbf{D}'\mathbf{C}'$.

9.4.2 Idempotent Matrix

A matrix, say **A**, is *idempotent* if the product of the matrix with itself equals itself, i.e. $\mathbf{AA} = \mathbf{A}$. This implies that the matrix must be square, but not necessarily symmetric.

9.4.3 Nilpotent Matrix

A matrix is *nilpotent* if the product of the matrix with itself equals a null matrix, i.e. $\mathbf{BB} = \mathbf{0}$, then **B** is nilpotent.

9.4.4 Orthogonal Matrix

A matrix is *orthogonal* if the product of the matrix with its transpose equals an identity matrix, i.e. $\mathbf{UU}' = \mathbf{I}$, which also implies that $\mathbf{U}'\mathbf{U} = \mathbf{I}$.

9.5 Traces of Square Matrices

The trace is the sum of the diagonal elements of a matrix. The sum is a scalar quantity. Let

$$\mathbf{A} = \begin{pmatrix} .51 & -.32 & -.19 \\ -.28 & .46 & -.14 \\ -.21 & -.16 & .33 \end{pmatrix},$$

then the trace is

$$tr(\mathbf{A}) = .51 + .46 + .33 = 1.30.$$

In R, the trace is achieved using the `sum()` and `diag()` functions together. The `diag()` function extracts the diagonals of the matrix, and the `sum()` function adds them together.

```
# Trace of the matrix A
trA = sum(diag(A))
```

9.5.1 Traces of Products - Rotation Rule

The trace of the product of conformable matrices has a special property known as the *rotation rule of traces*. That is,

$$tr(\mathbf{ABC}) = tr(\mathbf{BCA}) = tr(\mathbf{CAB}).$$

The traces are equal, because they are scalars, even though the dimensions of the three products might be greatly different.

9.6 Euclidean Norm

The Euclidean Norm is a matrix operation usually used to determine the degree of difference between two matrices of the same order. The norm is a scalar and is denoted as $\|\mathbf{A}\|$.

$$\|\mathbf{A}\| = [tr(\mathbf{AA}')]^{.5} = (\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij}^{2})^{.5}.$$

For example, let

$$\mathbf{A} = \begin{pmatrix} 7 & -5 & 2 \\ 4 & 6 & 3 \\ 1 & -1 & 8 \end{pmatrix},$$

then

$$
\begin{aligned}
\|\mathbf{A}\| &= (49 + 25 + 4 + \cdots + 1 + 64)^{.5} \\
&= (205)^{.5} = 14.317821.
\end{aligned}
$$

Other types of norms also exist.

In R,

```
# Euclidean Norm
EN = sqrt(sum(diag(A %*% t(A))))
```

9.7 Direct Sum of Matrices

For matrices of any dimension, say $\mathbf{H}_1$, $\mathbf{H}_2$, $\ldots \mathbf{H}_n$, the direct sum is

$$
\textstyle\sum_i^+ \mathbf{H}_i = \mathbf{H}_1 \oplus \mathbf{H}_2 \oplus \cdots \oplus \mathbf{H}_n = \begin{pmatrix} \mathbf{H}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{H}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{H}_n \end{pmatrix}.
$$

In R, the direct sum is accomplished by the `block()` function which is shown below.

```
# Direct sum operation via the block function

block <- function( ...  ) {
argv = list( ...  )
i = 0
for( a in argv ) {
m = as.matrix(a)
if(i == 0)
rmat = m
else
{
nr = dim(m)[1]
nc = dim(m)[2]
aa = cbind(matrix(0,nr,dim(rmat)[2]),m)
rmat = cbind(rmat,matrix(0,dim(rmat)[1],nc))
rmat = rbind(rmat,aa)
}
i = i+1
}
rmat
}
```

To use the function,

```
Htotal = block(H1,H2,H3,H4)
```

9.8 Kronecker Product

The Kronecker product, also known as the direct product, is where every element of the first matrix is multiplied, as a scalar, times the second matrix. Suppose that $\mathbf{B}$ is a matrix of order $m \times n$ and that $\mathbf{A}$ is of order 2×2, then the direct product of $\mathbf{A}$ times $\mathbf{B}$ is

$$\mathbf{A} \otimes \mathbf{B} \; = \; \left(\begin{array}{cc} a_{11}\mathbf{B} & a_{12}\mathbf{B} \\ a_{21}\mathbf{B} & a_{22}\mathbf{B} \end{array} \right).$$

Notice that the dimension of the example product is $2m \times 2n$.

In R, a direct product can be obtained as follows:

```
AB = A %x% B
# Note the small x between % %
```

9.9 Hadamard Product

The Hadamard product exists for two matrices that are conformable for addition. The corresponding elements of the two matrices are multiplied together. The order of the resulting product is the same as the matrices in the product. For two matrices, **A** and **B** of order 2 × 2, then the Hadamard product is

$$\mathbf{A} \odot \mathbf{B} = \begin{pmatrix} a_{11}b_{11} & a_{12}b_{12} \\ a_{21}b_{21} & a_{22}b_{22} \end{pmatrix}.$$

In R,

```
AB = A * B
```

10 Elementary Operators

Elementary operators are identity matrices that have been modified for a specific purpose.

10.1 Row or Column Multiplier

The first type of elementary operator matrix has one of the diagonal elements of an identity matrix replaced with a constant other than 1. In the following example, the {1,1} element has been set to 4. Note what happens when the elementary operator is multiplied times the matrix that follows it.

$$\begin{pmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 8 & 7 & 6 & 5 \\ 9 & 10 & 11 & 12 \end{pmatrix} = \begin{pmatrix} 4 & 8 & 12 & 16 \\ 8 & 7 & 6 & 5 \\ 9 & 10 & 11 & 12 \end{pmatrix}.$$

10.2 Interchange Rows or Columns

The second type of elementary operator matrix interchanges rows or columns of a matrix. To change rows i and j in a matrix, the identity matrix is modified by interchange rows i and j, as shown in the following example. Note the effect on the matrix that follows it.

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 8 & 7 & 6 & 5 \\ 9 & 10 & 11 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 9 & 10 & 11 & 12 \\ 8 & 7 & 6 & 5 \end{pmatrix}.$$

10.3 Combine Two Rows

The third type of elementary operator is an identity matrix that has an off-diagonal zero element changed to a non-zero constant. An example is given below, note the effect on the matrix that follows it.

$$\begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 & 2 & 3 & 4 \\ 8 & 7 & 6 & 5 \\ 9 & 10 & 11 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 7 & 5 & 3 & 1 \\ 9 & 10 & 11 & 12 \end{pmatrix}.$$

11 Matrix Inversion

An inverse of a square matrix $\mathbf{A}$ is denoted by $\mathbf{A}^{-1}$. An inverse of a matrix pre- or post-multiplied times the original matrix yields an identity matrix. That is,

$$\mathbf{A}\mathbf{A}^{-\mathbf{1}} = \mathbf{I}, \text{ and } \mathbf{A}^{-\mathbf{1}}\mathbf{A} = \mathbf{I}.$$

A matrix can be inverted if it has a nonzero determinant.

11.1 Determinant of a Matrix

The determinant of a matrix is a single scalar quantity. For a 2 × 2 matrix, say

$$\mathbf{A} = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix},$$

then the determinant is

$$|\mathbf{A}| = a_{11}a_{22} - a_{21}a_{12}.$$

For a 3 × 3 matrix, the determinant can be reduced to a series of determinants of 2 × 2 matrices. For example, let

$$\mathbf{B} = \begin{pmatrix} 6 & -1 & 2 \\ 3 & 4 & -5 \\ 1 & 0 & -2 \end{pmatrix},$$

then

$$\begin{aligned} |\mathbf{B}| &= 6\begin{vmatrix} 4 & -5 \\ 0 & -2 \end{vmatrix} - 1(-1)\begin{vmatrix} 3 & -5 \\ 1 & -2 \end{vmatrix} + 2\begin{vmatrix} 3 & 4 \\ 1 & 0 \end{vmatrix} \\ &= 6(-8) \ + 1(-1) \ + 2(-4) \\ &= -57. \end{aligned}$$

The general expression for the determinant of a matrix is

$$|\mathbf{B}| = \sum_{j=1}^{n}(-1)^{i+j}b_{ij}|\mathbf{M}_{ij}|$$

where $\mathbf{B}$ is of order n, and $\mathbf{M}_{ij}$ is a **minor** submatrix of $\mathbf{B}$ resulting from the deletion of the i^{th} row and j^{th} column of $\mathbf{B}$. Any row of $\mathbf{B}$ may be used to compute the determinant because the result should be the same for each row. Columns may also be used instead of rows.

In R, the `det()` function may be used to compute the determinant.

11.2 Matrix of Signed Minors

If the determinant is non-zero, the next step is to find the matrix of signed minors, known as the **adjoint matrix**. Using the same matrix, **B** above, the minors and their determinants are as follows:

$$\begin{aligned}
\mathbf{M}_{11} &= +1\begin{pmatrix} 4 & -5 \\ 0 & -2 \end{pmatrix}, \text{ and } |\mathbf{M}_{11}| = -8, \\
\mathbf{M}_{12} &= -1\begin{pmatrix} 3 & -5 \\ 1 & -2 \end{pmatrix}, \text{ and } |\mathbf{M}_{12}| = +1, \\
\mathbf{M}_{13} &= +1\begin{pmatrix} 3 & 4 \\ 1 & 0 \end{pmatrix}, \text{ and } |\mathbf{M}_{13}| = -4, \\
\mathbf{M}_{21} &= -1\begin{pmatrix} -1 & 2 \\ 0 & -2 \end{pmatrix}, \text{ and } |\mathbf{M}_{21}| = -2, \\
\mathbf{M}_{22} &= +1\begin{pmatrix} 6 & 2 \\ 1 & -2 \end{pmatrix}, \text{ and } |\mathbf{M}_{22}| = -14, \\
\mathbf{M}_{23} &= -1\begin{pmatrix} 6 & -1 \\ 1 & 0 \end{pmatrix}, \text{ and } |\mathbf{M}_{23}| = -1, \\
\mathbf{M}_{31} &= +1\begin{pmatrix} -1 & 2 \\ 4 & -5 \end{pmatrix}, \text{ and } |\mathbf{M}_{31}| = -3, \\
\mathbf{M}_{32} &= -1\begin{pmatrix} 6 & 2 \\ 3 & -5 \end{pmatrix}, \text{ and } |\mathbf{M}_{32}| = 36, \text{ and} \\
\mathbf{M}_{33} &= +1\begin{pmatrix} 6 & -1 \\ 3 & 4 \end{pmatrix}, \text{ and } |\mathbf{M}_{33}| = 27.
\end{aligned}$$

The adjoint matrix of signed minors, $\mathbf{M}_B$ is then

$$\mathbf{M}_B = \begin{pmatrix} -8 & 1 & -4 \\ -2 & -14 & -1 \\ -3 & 36 & 27 \end{pmatrix}.$$

11.3 The Inverse

The inverse of **B** is then

$$\begin{aligned}
\mathbf{B}^{-1} &= |\mathbf{B}|^{-1}\mathbf{M}'_B \\
&= \frac{1}{-57}\begin{pmatrix} -8 & -2 & -3 \\ 1 & -14 & 36 \\ -4 & -1 & 27 \end{pmatrix}.
\end{aligned}$$

If the determinant is zero, then the inverse is not defined or does not exist. A square matrix with a non-zero determinant is said to be **nonsingular**. Only nonsingular matrices have inverses. Matrices with zero determinants are called **singular** and do not have an inverse.

In R, there are different ways to compute an inverse.

```
BI = ginv(B) # will give generalized inverse if
# determinant is zero
```

11.4 Inverse of an Inverse

The inverse of an inverse matrix is equal to the original matrix. That is,

$$(\mathbf{A^{-1}})^{-1} = \mathbf{A}.$$

11.5 Inverse of a Product

The inverse of the product of two or more nonsingular matrices follows a rule similar to that for the transpose of a product of matrices. Let $\mathbf{A}, \mathbf{B},$ and $\mathbf{C}$ be three nonsingular matrices, then

$$(\mathbf{ABC})^{-1} = \mathbf{C^{-1}B^{-1}A^{-1}},$$

and

$$\mathbf{C^{-1}B^{-1}A^{-1}ABC} = \mathbf{I}.$$

12 Generalized Inverses

A matrix having a determinant of zero, can have a generalized inverse calculated. There are an infinite number of generalized inverses for any one singular matrix. A unique generalized inverse is the Moore-Penrose inverse which satisfies the following conditions:

1. $\mathbf{AA^{-}A} \;=\; \mathbf{A}$,
2. $\mathbf{A^{-}AA^{-}} \;=\; \mathbf{A}^{-}$,
3. $(\mathbf{A^{-}A})' \;=\; \mathbf{A^{-}A}$, and
4. $(\mathbf{AA^{-}})' \;=\; \mathbf{AA^{-}}$.

Usually, a generalized inverse that satisfies only the first condition is sufficient for practical purposes.

12.1 Linear Independence

For a square matrix with a nonzero determinant, the rows and columns of the matrix are linearly independent. If $\mathbf{A}$ is the matrix, then linear independence means that no vector, say $\mathbf{k}$, exists such that $\mathbf{Ak} = \mathbf{0}$ except for $\mathbf{k} = \mathbf{0}$.

For a square matrix with a zero determinant, at least one non-null vector, $\mathbf{k}$, exists such that $\mathbf{Ak} = \mathbf{0}$.

12.2 Rank of a Matrix

The rank of a matrix is the number of linearly independent rows and columns in the matrix. Elementary operators are used to determine the rank of a matrix. The objective is to reduce the matrix to an upper triangular form.

Pre-multiplication of a matrix by an elementary operator matrix does not change the rank of a matrix.

Reduction of a matrix to a diagonal matrix is called **reduction to canonical form**, and the reduced matrix is called the **canonical form under equivalence**.

12.2.1 Example Reduction to Find Rank

Let

$$\mathbf{A} = \begin{pmatrix} 2 & 1 & -1 & 3 & 8 \\ 4 & 7 & 5 & -2 & -1 \\ 6 & 8 & 4 & 1 & 7 \end{pmatrix}.$$

The rank of a matrix can not be greater than the minimum of the number of rows or columns, whichever is smaller. In the example above, **A** has 3 rows and 5 columns, and therefore the rank of **A** can not be greater than 3. Let $\mathbf{P_1}$ be the first elementary operator matrix.

$$\mathbf{P_1} = \begin{pmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ -3 & 0 & 1 \end{pmatrix},$$

then

$$\mathbf{P_1 A} = \begin{pmatrix} 2 & 1 & -1 & 3 & 8 \\ 0 & 5 & 7 & -8 & -17 \\ 0 & 5 & 7 & -8 & -17 \end{pmatrix}.$$

Now use $\mathbf{P_2}$ to subtract the third row from the second row, so that

$$\mathbf{P_2} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

and

$$\mathbf{P_2 P_1 A} = \begin{pmatrix} 2 & 1 & -1 & 3 & 8 \\ 0 & 5 & 7 & -8 & -17 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}.$$

The rank is the number of non-zero diagonal elements of the reduced matrix, i.e. $r(\mathbf{A}) = 2$.

Full-row rank: If **A** has order $m \times n$ with rank equal to m, then **A** has full row rank.

Full-column rank: A matrix with rank equal to the number of columns has full-column rank.

Full rank: A square matrix with rank equal to the number of rows or columns has full rank. A full rank matrix is nonsingular, has a non-zero determinant, and has an inverse.

Rank of zero: A null matrix has a rank of zero.

Rank of one: A **J** matrix has a rank of one.

Idempotent Matrix: Has rank equal to the trace of the matrix.

Rank of a Product: If **A** has rank r and **B** has rank q, and the two matrices are conformable for multiplication, then the product, **AB**, has a maximum possible rank equal to the lesser of r or q.

12.3 Consequences of Linear Dependence

A matrix with rank less than the number of rows or columns in the matrix means that the matrix can be partitioned into a square matrix of order equal to the rank (with full rank), and into other matrices of the remaining rows and columns. The other rows and columns are linearly dependent upon the rows and columns in that square matrix. That means that they can be formed from the rows and columns that are linearly independent.

Let $\mathbf{A}$ be a matrix of order $p \times q$ with rank r, and r is less than either p or q, then there are $p - r$ rows of $\mathbf{A}$ and $q - r$ columns which are not linearly independent. Partition $\mathbf{A}$ as follows:

$$\mathbf{A}_{p\times q} = \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{array} \right)$$

such that $\mathbf{A_{11}}$ has order $r \times r$ and rank of r. Re-arrangement of rows and columns of $\mathbf{A}$ may be needed to find an appropriate $\mathbf{A_{11}}$. $\mathbf{A_{12}}$ has order $r \times (q - r)$, $\mathbf{A_{21}}$ has order $(p - r) \times r$, and $\mathbf{A_{22}}$ has order $(p - r) \times (q - r)$.

There exist matrices, $\mathbf{K_1}$ and $\mathbf{K_2}$ such that

$$\left(\begin{array}{cc} \mathbf{A_{21}} & \mathbf{A_{22}} \end{array} \right) = \mathbf{K_2} \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \end{array} \right),$$

and

$$\left(\begin{array}{c} \mathbf{A_{12}} \\ \mathbf{A_{22}} \end{array} \right) = \left(\begin{array}{c} \mathbf{A_{11}} \\ \mathbf{A_{21}} \end{array} \right) \mathbf{K_1},$$

and

$$\mathbf{A} = \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{11}K_1} \\ \mathbf{K_2A_{11}} & \mathbf{K_2A_{11}K_1} \end{array} \right).$$

To illustrate, let

$$\mathbf{A} = \left(\begin{array}{rrr} 3 & -1 & 4 \\ 1 & 2 & -1 \\ -5 & 4 & -9 \\ 4 & 1 & 3 \end{array} \right),$$

and the rank of this matrix is 2. A 2×2 full rank submatrix within $\mathbf{A}$ is the upper left 2×2 matrix. Let

$$\mathbf{A_{11}} = \left(\begin{array}{rr} 3 & -1 \\ 1 & 2 \end{array} \right),$$

then

$$\mathbf{A_{12}} = \left(\begin{array}{r} 4 \\ -1 \end{array} \right) = \mathbf{A_{11}K_1},$$

where

$$\mathbf{K_1} = \left(\begin{array}{r} 1 \\ -1 \end{array} \right),$$

$$\mathbf{A_{21}} = \left(\begin{array}{rr} -5 & 4 \\ 4 & 1 \end{array} \right) = \mathbf{K_2 A_{11}},$$

where

$$\mathbf{K_2} = \left(\begin{array}{rr} -2 & 1 \\ 1 & 1 \end{array} \right),$$

and

$$\begin{aligned} \mathbf{A_{22}} &= \mathbf{K_2 A_{11} K_1} \\ &= \left(\begin{array}{rr} -2 & 1 \\ 1 & 1 \end{array} \right) \left(\begin{array}{rr} 3 & -1 \\ 1 & 2 \end{array} \right) \left(\begin{array}{r} 1 \\ -1 \end{array} \right) = \left(\begin{array}{r} -9 \\ 3 \end{array} \right). \end{aligned}$$

This kind of partitioning of matrices with less than full rank is **always** possible. In practice, we need only know that this kind of partitioning is possible, but $\mathbf{K_1}$ and $\mathbf{K_2}$ do not need to be derived explicitly.

12.4 Calculation of a Generalized Inverse

For a matrix, **A**, of less than full rank, there is an infinite number of possible generalized inverses, all of which would satisfy $\mathbf{A A_{-} A} = \mathbf{A}$. However, only one generalized inverse needs to be computed in practice. A method to derive one particular type of generalized inverse has the following steps:

1. Determine the rank of **A**.
2. Obtain $\mathbf{A_{11}}$, a square, full rank submatrix of **A** with rank equal to the rank of **A**.
3. Partition **A** as
$$\mathbf{A} = \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{array} \right).$$
4. Compute the generalized inverse as
$$\mathbf{A}^{-} = \left(\begin{array}{cc} \mathbf{A_{11}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right).$$

If **A** has order $p \times q$, then $\mathbf{A}^{-}$ must have order $q \times p$. To prove that $\mathbf{A}^{-}$ is a generalized inverse of **A**, then multiply out the expression

$$\begin{aligned} \mathbf{A A^{-} A} &= \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{array} \right) \left(\begin{array}{cc} \mathbf{A_{11}^{-1}} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right) \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{22}} \end{array} \right) \\ &= \left(\begin{array}{cc} \mathbf{A_{11}} & \mathbf{A_{12}} \\ \mathbf{A_{21}} & \mathbf{A_{21} A_{11}^{-1} A_{12}} \end{array} \right). \end{aligned}$$

From the previous section, $\mathbf{A_{21}} = \mathbf{K_2 A_{11}}$ so that

$$\mathbf{A_{21} A_{11}^{-1} A_{12}} = \mathbf{K_2 A_{11} A_{11}^{-1} A_{12}} = \mathbf{K_2 A_{12}} = \mathbf{A_{22}}.$$

12.5 Solutions to Equations Not of Full Rank

Because there are an infinite number of generalized inverses to a matrix that has less than full rank, then it logically follows that for a system of consistent equations, $\mathbf{Ax} = \mathbf{r}$, where the solutions are computed as $\mathbf{x} = \mathbf{A}^{-}\mathbf{r}$, then there would also be an infinite number of solution vectors for $\mathbf{x}$. Having computed only one generalized inverse, however, it is possible to compute many different solution vectors. If $\mathbf{A}$ has q columns and if $\mathbf{G}$ is one generalized inverse of $\mathbf{A}$, then the consistent equations $\mathbf{Ax} = \mathbf{r}$ have solution

$$\tilde{\mathbf{x}} = \mathbf{Gr} + (\mathbf{GA} - \mathbf{I})\mathbf{z},$$

where $\mathbf{z}$ is any arbitrary vector of length q. The number of linearly independent solution vectors, however, is $(q - r + 1)$.

Other generalized inverses of the same matrix can be produced from an existing generalized inverse. If $\mathbf{G}$ is a generalized inverse of $\mathbf{A}$ then so is

$$\mathbf{F} = \mathbf{GAG} + (\mathbf{I} - \mathbf{GA})\mathbf{X} + \mathbf{Y}(\mathbf{I} - \mathbf{AG})$$

for any $\mathbf{X}$ and $\mathbf{Y}$. Pre- and post- multiplication of $\mathbf{F}$ by $\mathbf{A}$ shows that this is so.

12.6 Generalized Inverses of $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$

The product $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$ occurs frequently where $\mathbf{X}$ is a matrix that usually has more rows than it has columns and has rank less than the number of columns, and $\mathbf{R}$ is a square matrix that is usually diagonal. Generalized inverses of this product matrix have special features. Let $\mathbf{X}$ be a matrix of order $N \times p$ with rank r. The product, $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$ has order $p \times p$ and is symmetric with rank r. Let $\mathbf{G}$ represent any generalized inverse of the product matrix, then the following results are true.

1. $\mathbf{G}$ is not necessarily a symmetric matrix, in which case $\mathbf{G}'$ is also a generalized inverse of the product matrix.
2. $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{X}'$ or $\mathbf{X}\mathbf{G}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} = \mathbf{X}$.
3. $\mathbf{G}\mathbf{X}'\mathbf{R}^{-1}$ is a generalized inverse of $\mathbf{X}$.
4. $\mathbf{X}\mathbf{G}\mathbf{X}'$ is always symmetric and unique for all generalized inverses of the product matrix, $\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}$.
5. If $\mathbf{1}' = \mathbf{k}'\mathbf{X}$ for some $\mathbf{k}'$, then $\mathbf{1}'\mathbf{R}^{-1}\mathbf{X}\mathbf{G}\mathbf{X}' = \mathbf{1}'$.

13 Eigenvalues and Eigenvectors

There are a number of square, and sometimes symmetric, matrices involved in statistical procedures that must be *positive definite*. Suppose that $\mathbf{Q}$ is any square matrix then

- $\mathbf{Q}$ is positive definite if $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is always greater than zero for all vectors, $\mathbf{y}$.
- $\mathbf{Q}$ is *positive semi-definite* if $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is greater than or equal to zero for all vectors $\mathbf{y}$, and for at least one vector $\mathbf{y}$, then $\mathbf{y}'\mathbf{Q}\mathbf{y} = 0$.
- $\mathbf{Q}$ is *non-negative definite* if $\mathbf{Q}$ is either positive definite or positive semi-definite.

The eigenvalues (or latent roots or characteristic roots) of the matrix must be calculated. The eigenvalues are useful in that

- The product of the eigenvalues equals the determinant of the matrix.
- The number of non-zero eigenvalues equals the rank of the matrix.
- If all the eigenvalues are greater than zero, then the matrix is positive definite.
- If all the eigenvalues are greater than or equal to zero and one or more are equal to zero, then the matrix is positive semi-definite.

If $\mathbf{Q}$ is a square, symmetric matrix, then it can be represented as

$$\mathbf{Q} = \mathbf{U}\mathbf{D}\mathbf{U}'$$

where $\mathbf{D}$ is a diagonal matrix, the canonical form of $\mathbf{Q}$, containing the eigenvalues of $\mathbf{Q}$, and $\mathbf{U}$ is an orthogonal matrix of the corresponding eigenvectors. Recall that for a matrix to be orthogonal then $\mathbf{U}\mathbf{U}' = \mathbf{I} = \mathbf{U}'\mathbf{U}$, and $\mathbf{U}^{-\mathbf{1}} = \mathbf{U}'$.

The eigenvalues and eigenvectors are found by solving

$$\mid \mathbf{Q} - d\mathbf{I} \mid \; = 0,$$

and

$$\mathbf{Q}\mathbf{u} - d\mathbf{u} = \mathbf{0},$$

where d is one of the eigenvalues of $\mathbf{Q}$ and $\mathbf{u}$ is the corresponding eigenvector. There are numerous computer routines for calculating $\mathbf{D}$ and $\mathbf{U}$.

In R, the `eigen()` function is used and both $\mathbf{U}$ and $\mathbf{D}$ are returned to the user.

14 Differentiation

The differentiation of mathematical expressions involving matrices follows similar rules as for those involving scalars. Some of the basic results are shown below.

Let

$$\begin{aligned} c &= 3x_1 + 5x_2 + 9x_3 = \mathbf{b}'\mathbf{x} \\ &= \begin{pmatrix} 3 & 5 & 9 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}. \end{aligned}$$

With scalars, the derivatives are

$$\begin{aligned} \frac{\partial c}{\partial x_1} &= 3 \\ \frac{\partial c}{\partial x_2} &= 5 \\ \frac{\partial c}{\partial x_3} &= 9, \end{aligned}$$

but with vectors they are

$$\frac{\partial c}{\partial \mathbf{x}} = \mathbf{b}.$$

The general rule is

$$\frac{\partial \mathbf{A}'\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}.$$

Another function might be

$$c = 9x_1^2 + 6x_1x_2 + 4x_2^2$$

or

$$c = \begin{pmatrix} x_1 & x_2 \end{pmatrix} \begin{pmatrix} 9 & 3 \\ 3 & 4 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \mathbf{x}'\mathbf{A}\mathbf{x}.$$

With scalars the derivatives are

$$\begin{aligned} \frac{\partial c}{\partial x_1} &= 2(9)x_1 + 6x_2 \\ \frac{\partial c}{\partial x_2} &= 6x_1 + 2(4)x_2, \end{aligned}$$

and in matrix form they are,

$$\frac{\partial c}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}.$$

If **A** was not a symmetric matrix, then

$$\frac{\partial \mathbf{x}'\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{A}'\mathbf{x}.$$

15 Cholesky Decomposition

In simulation studies or applications of Gibb's sampling there is frequently a need to factor a symmetric positive definite matrix into the product of a matrix times its transpose. The Cholesky decomposition of a matrix, say $\mathbf{V}$, is a lower triangular matrix such that

$$\mathbf{V} = \mathbf{T}\mathbf{T}',$$

and $\mathbf{T}$ is lower triangular. Suppose that

$$\mathbf{V} = \left(\begin{array}{rrr} 9 & 3 & -6 \\ 3 & 5 & 0 \\ -6 & 0 & 21 \end{array} \right).$$

The problem is to derive

$$\mathbf{T} = \left(\begin{array}{rrr} t_{11} & 0 & 0 \\ t_{21} & t_{22} & 0 \\ t_{31} & t_{32} & t_{33} \end{array} \right),$$

such that

$$\begin{array}{rcl} t_{11}^2 & = & 9 \\ t_{11}t_{21} & = & 3 \\ t_{11}t_{31} & = & -6 \\ t_{21}^2 + t_{22}^2 & = & 5 \\ t_{21}t_{31} + t_{22}t_{32} & = & 0 \ \text{ and} \\ t_{31}^2 + t_{32}^2 + t_{33}^2 & = & 21 \end{array}$$

These equations give $t_{11} = 3$, then $t_{21} = 3/t_{11} = 1$, and $t_{31} = -6/t_{11} = -2$. From the fourth equation, t_{22} is the square root of $(5 - t_{21}^2)$ or $t_{22} = 2$. The fifth equation says that $(1)(-2) + (2)t_{32} = 0$ or $t_{32} = 1$. The last equation says that t_{33}^2 is $21 - (-2)^2 - (1)^2 = 16$. The end result is

$$\mathbf{T} = \left(\begin{array}{rrr} 3 & 0 & 0 \\ 1 & 2 & 0 \\ -2 & 1 & 4 \end{array} \right).$$

The derivation of the Cholesky decomposition is easily programmed for a computer. Note that in calculating the diagonals of $\mathbf{T}$ the square root of a number is needed, and consequently this number must always be positive. Hence, if the matrix is positive definite, then all necessary square roots will be of positive numbers. However, the opposite is not true. That is, if all of the square roots are of positive numbers, the matrix is not necessarily guaranteed to be positive definite. The only way to guarantee positive definiteness is to calculate the eigenvalues of a matrix and to see that they are all positive.

16 Inverse of a Lower Triangular Matrix

The inverse of a lower triangular matrix is also a lower triangular matrix, and can be easily derived. The diagonals of the inverse are the inverse of the diagonals of the original matrix. Using the matrix $\mathbf{T}$ from the previous section, then

$$\mathbf{T}^{-1} = \begin{pmatrix} t^{11} & 0 & 0 \\ t^{21} & t^{22} & 0 \\ t^{31} & t^{32} & t^{33} \end{pmatrix},$$

where

$$\begin{aligned} t^{ii} &= 1/t_{ii} \\ t_{21}t^{11} + t_{22}t^{21} &= 0 \\ t_{31}t^{11} + t_{32}t^{21} + t_{33}t^{31} &= 0 \\ t_{32}t^{22} + t_{33}t^{32} &= 0. \end{aligned}$$

These equations give

$$\begin{aligned} t^{11} &= \frac{1}{3} \\ t^{21} &= -\frac{1}{6} \\ t^{22} &= \frac{1}{2} \\ t^{31} &= \frac{5}{24} \\ t^{32} &= -\frac{1}{8} \text{ and} \\ t^{33} &= \frac{1}{4} \end{aligned}$$

Likewise the determinant of a triangular matrix is the product of all of the diagonal elements. Hence all diagonal elements need to be non-zero for the inverse to exist.

The natural logarithm of the determinant of a triangular matrix is the summation of the natural logarithms of the individual diagonal elements. This property is useful in derivative free restricted maximum likelihood.

17 EXERCISES

For each of the following exercises, first do the calculations by hand (or in your head). Then use R to obtain results, and check to make sure the results are identical.

1. Given matrices **A** and **B**, as follows.

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ -2 & 3 & -1 \\ -2 & -2 & 4 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 6 & -5 & 8 \\ -4 & 9 & -3 \\ -5 & -7 & 1 \\ 3 & 4 & -5 \end{pmatrix}.$$

   If legal to do so, do the following calculations:

   (a) $\mathbf{A}'$.

   (b) $\mathbf{A} + \mathbf{A}'$

   (c) $\mathbf{AA}$. Is $\mathbf{A}$ idempotent?

   (d) $\mathbf{AB}$.

   (e) $|\mathbf{A}|$.

   (f) $\mathbf{B}'\mathbf{A}$.

   (g) Find the rank of $\mathbf{B}$.

   (h) $\mathbf{A}^{-1}$.

   (i) $\mathbf{B} + \mathbf{B}'$.

   (j) $tr(\mathbf{BB}'$.

2. Given a new matrix, **D**.

$$\mathbf{D} = \begin{pmatrix} 8 & 3 & 5 & -1 & 7 \\ 4 & 1 & 2 & -3 & 6 \\ -2 & 6 & 7 & -5 & 2 \\ 1 & -2 & -4 & 0 & 3 \\ 1 & 6 & 14 & -3 & -4 \end{pmatrix}$$

   Find the rank and a generalized inverse of **D**.

3. Find the determinant and an inverse of the following lower triangular matrix.

$$\mathbf{L} = \begin{pmatrix} 10 & 0 & 0 & 0 & 0 \\ -2 & 20 & 0 & 0 & 0 \\ 1 & -2 & 16 & 0 & 0 \\ 4 & -1 & -2 & 5 & 0 \\ -1 & -6 & 3 & 1 & 4 \end{pmatrix}$$

4. If matrix $\mathbf{R}$ has N rows and columns with a rank of $N-2$, and matrix $\mathbf{W}$ has N rows and p columns for $p < N$ with rank of $p-3$, then what would be the rank of $\mathbf{W}'\mathbf{R}$?

5. Create a square matrix of order 5 that has rank of 2, and show that the rank of your matrix is indeed 2.

6. Obtain a generalized inverse of matrix $\mathbf{B}$ in the first question, and show that it satisfies the first Moore-Penrose condition.

Data Manipulations and R

18 Small Data Sets

Most ‘real life’ examples used in these notes to illustrate methods can be given in one table on less than a page of paper. In these cases, the student can enter the data into R manually in a few minutes.

Below are data on 10 beef calves born at a research station within one week of each other.

Beef calf data on birthweights (BW) and calving ease (CE).

Calf	Breed	Sex	CE	BW(lbs)
1	AN	M	U	55
2	CH	M	E	68
3	HE	M	U	60
4	AN	M	U	52
5	CH	F	H	65
6	HE	F	E	64
7	CH	F	H	70
8	AN	F	E	61
9	HE	F	E	63
10	CH	M	C	75

An easy way to enter the data is by columns of the table.

```
calf = c(1:10) # makes a string of numbers 1 to 10
breed = c("AN","CH","HE","AN","CH","HE","CH","AN",
"HE","CH")
sex = c("M","M","M","M","F","F","F","F","F","M")
CE = c("U","E","U","U","H","E","H","E","E","C")
BWT = c(55,68,60,52,65,64,70,61,63,75)
```

Then the columns can be put into a data frame, as follows:

```
beefdat = data.frame(calf,breed,sex,CE,BWT)
beefdat # looks at the data, exactly like the table
```

The data frame can be saved and used at other times. The saved file can not be viewed because it is stored in binary format

```
setwd(choose.dir())
save(beefdat,file="beef.RData")

# and retrieved later as
load("beef.RData")
```

18.1 Creating Design Matrices

A desgin matrix is a matrix that relates levels of a factor to the observations. The observations, in this example, are the birthweights. The factors are breed, sex, and CE. The breed factor has 3 levels, namely AN, CH, and HE. The sex factor has 2 levels, M and F, and the CE factors has 4 levels, U, E, H, and C.

A function to make a design matrix is as follows:

```
desgn <- function(v) {
if(is.numeric(v)) { vn = v  }
else
{ vn = as.numeric(factor(v)) }

mrow = length(vn)
mcol = length(levels(vn))
X = matrix(data=c(0),nrow=mrow,ncol=mcol)
for(i in 1:mrow) {
ic = vn[i]
X[i,ic] = 1 }
return(X)
}

# To use, then
B = desgn(breed)
S = desgn(sex)
C = desgn(CE)
```

These matrices are

$$\mathbf{B} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Each row of a design matrix has one 1 and the remaining elements are zero. The location of the 1 indicates the level of the factor corresponding to that observation. If you summed together the elements of each row you will always get a vector of ones, or a **J** matrix with just one column.

18.2 The `summary` Function

The data frame, `beefdat`, was created earlier. Now enter

```
summary(beefdat)
```

This gives information about each column of the data frame. If appropriate it gives the minimum and maximum value, median, and mean of numeric columns. For non-numeric columns it gives the levels of that column and number of observations for each, or the total number of levels. This is useful to check if data have been entered correctly or if there are codes in that data that were not expected.

The information may not be totally correct. For example, if missing birthweights were entered as 0, then R does not know that 0 means missing and assumes that 0 was a valid birthweight. The letters `NA`, signify a missing value, and these are skipped in the summary function.

18.3 Means and Variances

The functions to calculate the means and variances are straightforward. Let `y` be a vector of the observations on a trait of interest.

```
# The mean is
mean(y)
# The variance and standard deviation are
var(y)
sd(y)
```

18.4 Plotting

The plot function is handy for obtaining a visual appreciation of the data. There are also the hist(), boxplot(), and qqnorm() functions that plot information. Use the ?hist, for example, to find out more about a given function. This will usually show you all of the available options and examples of how the function is used. There are enough options and additional functions to make about any kind of graphical display that you like.

19 Large Data Sets

Large, in these notes, means a data set with more than 30 observations. This could be real data that exists in a file on your computer. There are too many observations and variables to enter it manually into R. The data set can be up to 100,000 records, but R is not unlimited in space, and some functions may not be efficient with a large number of records. If the data set is larger than 100,000 records, then other programming approaches like FORTRAN or C++ should be considered, and computing should be on servers with multiprocessors and gigabytes of memory.

For example, setting up a design matrix for a factor with a large data set may require too much memory. Techniques that do not require an explicit representation of the design matrix should be used. A chapter on analyses using this approach is given towards the end of the notes.

To read a file of trotting horse data, as an example, into R, use

```
zz = file.choose() # allows you to browse for file
# zz is the location or handle for the file
trot = read.table(file = zz, header=FALSE, col.names=
c("race","horse","year","month","track","dist","time"))
```

When a data frame is saved in R, it is written as a binary file, and the names of the columns form a header record in the file. Normally, data files do not have a header record, thus, `header=FALSE` was indicated in the `read.table()` function, and the `col.names` had to be provided, otherwise R provides its own generic header names like `V1, V2, ...`.

19.1 Exploring the Data

```
summary(trot) # as before with the small data sets

dim(trot) # will indicate number of records and
# number of columns in trot data frame
horsef = length(factor(trot$horse)) # number of different horses
# in the data set
yearsf = length(factor(trot$year)) # number of different years

yearsf = factor(trot$year)
levels(yearsf) # list of years represented in data

tapply(trot$times,trot$track,mean) # mean racing times by
# track location
```

20 EXERCISES

1. Enter the data from the following table into a data frame.

Trotting horses racing times.

Horse	Sex	Race	Time(sec)
1	M	1	135
2	S	1	130
3	S	1	133
4	M	1	138
5	G	1	132
2	S	2	123
4	M	2	131
5	G	2	125
6	S	2	134

   (a) Find the average time by race number and by sex of horse.

   (b) Find the mean and variance of race times in the data.

   (c) Create design matrices for horse, sex, and race.

   (d) Do a histogram of race times.

   (e) Save the data frame. Remove the data frame from your R-workspace using `rm(trot)`. Use `ls()` to determine that it is gone. Load the data frame back into the R-workspace.

   (f) Create a second data.frame by removing the ”sex” column from the first data frame.

   (g) Change the racing time of horse 6 in the second race to 136.

2. Read in any large data file, (provided by the instructor), and summarize the information as much as possible.

Writing a Linear Model

21 Parts of a Model

A linear model, in the traditional sense, is composed of three parts:

1. The equation.
2. Expectations and Variance-Covariance matrices of random variables.
3. Assumptions, restrictions, and limitations.

21.1 The Equation

The equation of the model contains the observation vector for the trait(s) of interest, the factors that explain how the observations came to be, and a residual effect that includes everything not explainable.

A matrix formulation of a general model equation is:

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

where

$\mathbf{y}$ is the vector of observed values of the trait,

$\mathbf{b}$ is a vector of factors, collectively known as fixed effects,

$\mathbf{u}$ is a vector of factors known as random effects,

$\mathbf{e}$ is a vector of residual terms, also random,

$\mathbf{X}, \mathbf{Z}$ are known matrices, commonly known as design or indicator matrices, that relate the elements of $\mathbf{b}$ and $\mathbf{u}$ to their corresponding element in $\mathbf{y}$.

21.1.1 Observation Vector

The observation vector contains elements resulting from measurements, either subjective or objective, on the experimental units (usually animals) under study. The elements in the observation vector are random variables that have a multivariate distribution, and if the form of the distribution is known, then advantage should be taken of that knowledge.

Usually **y** is assumed to have a multivariate normal distribution, but that is not always true.

The elements of **y** should represent random samples of observations from some defined population. If the elements are not randomly sampled, then bias in the estimates of **b** and **u** can occur, which would lead to errors in ranking animals or conclusions to hypothesis tests.

21.1.2 Factors

Discrete or Continuous A continuous factor is one that has an infinite-like range of possible values. For example, if the observation is the distance a rock can be thrown, then a continuous factor would be the weight of the rock. If the observation is the rate of growth, then a continuous factor would be the amount of feed eaten.

Discrete factors usually have *classes* or *levels* such as age at calving might have four levels (e.g. 20 to 24 months, 25 to 28 months, 29 to 32 months, and 33 months or greater). An analysis of milk yields of cows would depend on the age levels of the cows.

Fixed Factors In the traditional "frequentist" approach, fixed and random factors need to be distinguished.

If the number of *levels* of a factor is small or limited to a fixed number, then that factor is usually *fixed*.

If inferences about a factor are going to be limited to that set of *levels*, and to no others, then that factor is usually *fixed*.

If a new sample of observations were made (a new experiment), and the same *levels* of a factor are in both samples, then the factor is usually *fixed*.

If the *levels* of a factor were determined as a result of selection among possible available levels, then that factor should probably be a fixed factor.

Regressions of a continuous factor are usually a fixed factor (but not always).

Random Factors If the number of *levels* of a factor is large, then that factor can be a *random* factor.

If the inferences about a factor are going to be made to an entire population of conceptual *levels*, then that factor can be a *random* factor.

If the *levels* of a factor are a sample from an infinitely large population, then that factor is usually *random*.

If a new sample of observations were made (a new experiment), and the *levels* were completely different between the two samples, then the factors if usually *random*.

Examples of fixed and random factors.

Fixed	Random
Diets	Animals
Breeds	Contemporary Groups
Sex	Herd-Year-Seasons
Age levels	Permanent Environment
Cages, Tanks	Maternal Effects
Seasons	Litters

21.2 Expectations and VCV Matrices

In general terms, the expectations are

$$E\begin{pmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{Xb}\\ \mathbf{0}\\ \mathbf{0}\end{pmatrix},$$

and the variance-covariance matrices are

$$V\begin{pmatrix}\mathbf{u}\\ \mathbf{e}\end{pmatrix} = \begin{pmatrix}\mathbf{G} & \mathbf{0}\\ \mathbf{0} & \mathbf{R}\end{pmatrix},$$

where $\mathbf{G}$ and $\mathbf{R}$ are general square matrices assumed to be nonsingular and positive definite. Also,

$$Var(\mathbf{y}) \;=\; \mathbf{ZGZ}' + \mathbf{R} \;=\; \mathbf{V}.$$

21.3 Assumptions and Limitations

The third part of a model includes items that are not apparent in parts 1 and 2. For example, information about the manner in which data were sampled or collected. Were the animals randomly selected or did they have to meet some minimum standards? Did the data arise from many environments, at random, or were the environments specially chosen? Examples will follow.

A linear model is not complete unless all three parts of the model are present. Statistical procedures and strategies for data analysis are determined only after a complete model is in place.

22 Example 1. Beef Calf Weights

Weights on beef calves taken at 200 days of age are shown in the table below.

Males	Females
198	187
211	194
220	202
	185

22.1 Equation of the Model

$$y_{ij} = s_i + c_j + e_{ij},$$

where y_{ij} is one of the 200-day weights, s_i is an effect due to the sex of the calf (fixed factor), c_j is an effect of the calf (random factor), and e_{ij} is a residual effect or unexplained variation (random factor).

22.2 Expectations and Variances

$$\begin{aligned} E(c_j) &= 0 \\ E(e_{ij}) &= 0 \\ Var(c_j) &= \sigma^2_c \\ Var(e_{ij}) &= \sigma^2_{ei} \end{aligned}$$

Additionally, $Cov(c_j, c_{j'}) = 0$, which says that all of the calves are independent of each other, i.e. unrelated. Note that σ^2_{ei} implies that the residual variance is different for each sex of calf, because of the subscript i. Also, $Cov(e_{ij}, e_{ij'}) = 0$ and $Cov(e_{ij}, e_{i'j'}) = 0$ says that all residual effects are independent of each other within and between sexes.

22.3 Assumptions and Limitations

1. All calves are assumed to be of the same breed.
2. All calves were reared in the same environment and time period.
3. All calves were from dams of the same age (e.g. 3 yr olds).
4. Maternal effects are ignored. Pedigree information is missing and maternal effects can not be estimated.
5. Calf effects contain all genetic effects, direct and maternal.
6. All weights were accurately recorded (i.e. not guessed) at age 200 days.

22.4 Matrix Representation

Ordering the observations by males, then females, the matrix representation of the model would be

$$\mathbf{y} = \begin{pmatrix} 198 \\ 211 \\ 220 \\ 187 \\ 194 \\ 202 \\ 185 \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix},$$

and $\mathbf{Z} = \mathbf{I}$ of order 7. Also,

$$\begin{aligned} \mathbf{G} &= \mathbf{I}\sigma_c^2 = diag\{\sigma_c^2\} \\ \mathbf{R} &= diag\{\sigma_{e1}^2 \ \sigma_{e1}^2 \ \sigma_{e1}^2 \ \sigma_{e2}^2 \ \sigma_{e2}^2 \ \sigma_{e2}^2 \ \sigma_{e2}^2\} \end{aligned}$$

23 Example 2. Temperament Scores of Dairy Cows

Below are progeny data of three sires on temperament scores (on a scale of 1(easy to handle) to 40(very cantankerous)) taken at milking time.

CG	Age	Sire	Score
1	1	1	17
1	2	2	29
1	1	2	34
1	2	3	16
2	2	3	20
2	1	3	24
2	2	1	13
2	1	1	18
2	2	2	25
2	1	2	31

23.1 Equation of the Model

$$y_{ijkl} = c_i + a_j + s_k + e_{ijkl},$$

where y_{ijkl} is a temperament score, c_i is a contemporary group effect (CG) which identifies animals that are typically reared and treated alike together; a_j is an age group effect, in this case just two age groups; s_k is a sire effect; and e_{ijkl} is a residual effect. Contemporary groups and age groups are often taken to be fixed factors, and sires are generally random factors. Age group 1 was for daughters between 18 and 24 mo of age, and age group 2 was for daughters between 25 and 32 mo of age.

23.2 Expectations and Variances

$$\begin{aligned}
E(y_{ijkl}) &= c_i + a_j \\
E(s_k) &= 0 \\
E(e_{ijkl}) &= 0 \\
Var(s_k) &= \sigma_s^2 \\
Cov(s_k, s_{k'}) &= 0 \\
Var(e_{ijkl}) &= \sigma_{ei}^2
\end{aligned}$$

Thus, the residual variance differs between contemporary groups. The sire variance represents one quarter of the additive genetic variance because all progeny are assumed to be half-sibs (i.e. from different dams). The sires are assumed to be unrelated.

23.3 Assumptions and Limitations

1. Daughters were approximately in the same stage of lactation when temperament scores were taken.

2. The same person assigned temperament scores for all daughters.

3. The age groupings were appropriate.

4. Sires were unrelated to each other.

5. Sires were mated randomly to dams (with respect to milking temperament or any correlated traits).

6. Only one offspring per dam.

7. Only one score per daughter.

8. No preferential treatment towards particular daughters.

24 Example 3. Feed Intake in Pigs

24.1 Equation of the Model

$$y_{ijkmn} = (HYM)_i + S_j + L_k + a_{km} + p_{km} + e_{ijkmn},$$

where y_{ijkmn} is a feed intake measurement at a specified moment in time, n, on the m^{th} pig from litter k, whose sow was in age group j, within the i^{th} herd-year-month of birth subclass; HYM is a herd-year-month of birth or contemporary group effect; S_j is an age of sow effect identified by parity number of the sow; L_k is a litter effect which identifies a group of pigs with the same genetic and environmental background; a_{km} is an additive genetic animal effect; p_{km} is an animal permanent environmental effect common to all measurements on an animal; and e_{ijkmn} is a residual effect specific to each measurement.

24.2 Expectations and Variances

$$\begin{aligned}
E(L_k) &= 0 \\
Var(L_k) &= \sigma^2_L \\
E(a_{km}) &= 0 \\
Var(\mathbf{a}) &= \mathbf{A}\sigma^2_a \\
E(p_{km}) &= 0 \\
Var(\mathbf{p}) &= \mathbf{I}\sigma^2_p \\
E(e_{ijkmn}) &= 0 \\
Var(\mathbf{e}) &= \mathbf{R} = \mathbf{I}\sigma^2_e \\
\mathbf{G} &= \left(\begin{array}{ccc} \mathbf{I}\sigma^2_L & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}\sigma^2_a & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_p \end{array} \right).
\end{aligned}$$

All pigs were purebred Landrace. Two males and two females were taken randomly from each litter for feed intake measurements.

24.3 Assumptions and Limitations

1. There are no sex differences in feed intake.
2. There are no maternal effects on feed intake.
3. All measurements were taken at approximately the same age of the pigs.

4. All measurements were taken within a controlled environment at one location.

5. Feed and handling of pigs was uniform for all pigs within a herd-year-month subclass.

6. Litters were related through the use of boars from artificial insemination.

7. Feed intake was the average of 3 daily intakes during the week, and weekly averages were available for 5 consecutive weeks.

8. Growth was assumed to be linear during the test period.

25 Comments on Models

1. Read the literature first to find out what should be in the model. Know what has already been researched.

2. Explain your model to other people in a workshop. Get ideas from other people.

3. Talk to industry people to know how data are collected.

4. Test your models. Do not be afraid to change the model as evidence accumulates.

5. Not everyone will agree to the same model.

6. Make sure you can justify all three parts.

7. Consider non linear or other types of models if appropriate.

26 EXERCISES

Write models for each of the following situations.

1. Dogs compete at tracking at several levels of skill. The lowest level is called a TD (Tracking Dog) trial. A track layer will set a track that is 600 to 800 meters in length with at least two right angle turns in it and with two scent articles for the dog to indicate. The dog must pick up or lie down next to an article. The track must be at least one hour old before the dog runs it. The tracks are designed by a judge based on the fields in which the tracks are set and weather conditions. Dogs of many breed compete. Someone decided to analyze data from two years of trials in Canada. The results of 50 trialswere collected. Dogs either passed or failed the test, so the observation is either 1 (if they pass) or 0 (if they fail). Write a model to analyze these data.

2. Cows have been challenged with a foreign substance injected into their blood to induce an immune response. Cows' blood is sampled at 2 hr, 6 hr, 12 hr, 24 hr, and 48 hr. Levels of immune response are measured in each sample. Four different levels of the foreign substance have been used, and a fifth group of cows were given a placebo injection. Each group contained 3 cows. All cows were between 60 to 120 days in milk.

3. Beef bulls undergo a 112 day growth test at stations located in 3 places in Ontario. The traits measured are the amount of growth on test and amount of feed eaten during the test. Data are from several years, and bulls are known to be related to each other across years. Several breeds and crossbreds are involved in the tests. Age at start of the test is not the same for each bull, but between 150 to 250 days.

4. Weights of individually identified rainbow trout were collected over five years at two years of age. Fish are reared in tanks in a research facility with the capability of controlling water temperature and hours of daylight. Tanks differ in size and number of fish. Pedigree information is available on all fish. Sex and maturity are known. When a trout matures they stop growing.

5. Rabbit rearing for meat consumption is concerned with raising the most rabbits per litter. Average litter size at birth may be 9, but the average number at weaning is only 7. Litter size at weaning is the economically important trait measured on each doe.

6. Describe your own research project and write models for your data or experiment.

Estimation

27 Description of Problem

Below are data on gross margins (GM) of dairy cows. For each cow there are also observations on protein yield, type score, non-return rate, milking speed, and somatic cell score. The problem is to regress the observed traits onto gross margins to derive a prediction equation. Given protein yield, type score, non-return rate, milking speed, and somatic cell score, predict the gross margins of the cow.

Gross Margins (GM) of dairy cows.

Cow	Prot	Type	Non-return	Milking	Somatic	Gross
	kg	score	rate	speed	Cell score	Margins($)
1	246	75	66	3	3.5	-284
2	226	80	63	4	3.3	-402
3	302	82	60	2	3.1	-207
4	347	77	58	3	4.3	267
5	267	71	66	5	3.7	-201
6	315	86	71	4	3.5	283
7	241	90	68	3	3.6	-45
8	290	83	70	2	3.9	246
9	271	78	67	1	4.1	70
10	386	80	64	3	3.4	280

28 Theory Background

28.1 General Model

A general fixed effects model is

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

with $E(\mathbf{y}) = \mathbf{Xb}$, and $Var(\mathbf{y}) = \mathbf{V} = V(\mathbf{e})$. $\mathbf{V}$ is assumed to be positive definite, i.e. all eigenvalues are positive.

28.2 Function to be Estimated

Let a function of $\mathbf{b}$ be $\mathbf{K}'\mathbf{b}$, for some matrix $\mathbf{K}'$.

28.3 Estimator

The estimator is a linear function of the observation vector, $\mathbf{L}'\mathbf{y}$, where $\mathbf{L}'$ is to be determined.

28.4 Error of Estimation

The error of estimation is given by $\mathbf{L}'\mathbf{y} - \mathbf{K}'\mathbf{b}$, and the variance-covariance matrix of the error vector is

$$Var(\mathbf{L}'\mathbf{y} - \mathbf{K}'\mathbf{b}) = Var(\mathbf{L}'\mathbf{y}) = \mathbf{L}'\mathbf{V}\mathbf{L}.$$

28.5 Properties of the Estimator

The following criteria are used to derive $\mathbf{L}'$.

1. $\mathbf{L}'\mathbf{y}$ should have the same expectation as $\mathbf{K}'\mathbf{b}$.

   $$E(\mathbf{L}'\mathbf{y}) = \mathbf{L}'E(\mathbf{y}) = \mathbf{L}'\mathbf{X}\mathbf{b},$$

   and therefore, $\mathbf{L}'\mathbf{y}$ is an unbiased estimator of $\mathbf{K}'\mathbf{b}$ if $\mathbf{L}'\mathbf{X} = \mathbf{K}'$.

2. The variance-covariance matrix of the error vector, $\mathbf{L}'\mathbf{V}\mathbf{L}$, should have diagonal elements that are as small as possible. Minimization of the diagonal elements of $\mathbf{L}'\mathbf{V}\mathbf{L}$ results in an estimator that is called *best*.

28.6 Function to be Minimized

$$\mathbf{F} = \mathbf{L}'\mathbf{V}\mathbf{L} + (\mathbf{L}'\mathbf{X} - \mathbf{K}')\boldsymbol{\Phi}$$

where $\boldsymbol{\Phi}$ is a *LaGrange Multiplier* that imposes the restriction $\mathbf{L}'\mathbf{X} = \mathbf{K}'$.

The function, $\mathbf{F}$, is differentiated with respect to the unknowns, $\mathbf{L}$ and $\boldsymbol{\Phi}$, to give

$$\frac{\partial \mathbf{F}}{\partial \mathbf{L}} = 2\mathbf{V}\mathbf{L} + \mathbf{X}\boldsymbol{\Phi},$$

and

$$\frac{\partial \mathbf{F}}{\partial \boldsymbol{\Phi}} = \mathbf{X}'\mathbf{L} - \mathbf{K}.$$

28.7 Solving for $\mathbf{L}'$

The derivatives are equated to null matrices and rewritten as

$$\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{L} \\ \theta \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{K} \end{pmatrix}$$

where $\theta = .5\mathbf{\Phi}$. Note that $\mathbf{VL} + \mathbf{X}\theta = \mathbf{0}$ from the first equation. Further,

$$\mathbf{L} = -\mathbf{V}^{-1}\mathbf{X}\theta$$

Substitution into the second row gives

$$\mathbf{X}'\mathbf{L} = -\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\theta = \mathbf{K},$$

so that

$$\theta = -(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{K}.$$

Putting θ into the equation for $\mathbf{L}$, then

$$\mathbf{L} = \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{K}.$$

The estimator of $\mathbf{K}'\mathbf{b}$ is then

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

where

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

BLUE stands for **B**est **L**inear **U**nbiased **E**stimator. $\mathbf{K}'\hat{\mathbf{b}}$ is BLUE of $\mathbf{K}'\mathbf{b}$.

GLS stands for **G**eneralized **L**east **S**quares, which are

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

and $\hat{\mathbf{b}}$ is the GLS solution.

Weighted LS is equivalent to GLS except that $\mathbf{V}$ is assumed to be a diagonal matrix whose diagonal elements could be different from each other.

Ordinary LS is equivalent to GLS except that $\mathbf{V}$ is assumed to be an identity matrix times a scalar. That is, all of the diagonals of $\mathbf{V}$ are the same value.

29 BLUE for Example Data

The model is

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

with

$$Var(\mathbf{y}) \;=\; \mathbf{V} \;=\; \mathbf{I}\sigma_e^2,$$

$$\mathbf{X} \;=\; \begin{pmatrix} 1 & 246 & 75 & 66 & 3 & 3.5 \\ 1 & 226 & 80 & 63 & 4 & 3.3 \\ 1 & 302 & 82 & 60 & 2 & 3.1 \\ 1 & 347 & 77 & 58 & 3 & 4.3 \\ 1 & 267 & 71 & 66 & 5 & 3.7 \\ 1 & 315 & 86 & 71 & 4 & 3.5 \\ 1 & 241 & 90 & 68 & 3 & 3.6 \\ 1 & 290 & 83 & 70 & 2 & 3.9 \\ 1 & 271 & 78 & 67 & 1 & 4.1 \\ 1 & 386 & 80 & 64 & 3 & 3.4 \end{pmatrix}, \;\; \mathbf{y} \;=\; \begin{pmatrix} -284 \\ -402 \\ -207 \\ 267 \\ -201 \\ 283 \\ -45 \\ 246 \\ 70 \\ 280 \end{pmatrix}.$$

Ordinary LS equations are sufficient for this problem.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 2891 & 802 & 653 & 30 & 36.40 \\ 2891 & 858337 & 231838 & 188256 & 8614 & 10547.60 \\ 802 & 231838 & 64588 & 52444 & 2393 & 2915.00 \\ 653 & 188256 & 52444 & 42795 & 1961 & 2377.10 \\ 30 & 8614 & 2393 & 1961 & 102 & 108.20 \\ 36.4 & 10547.60 & 2915 & 2377.10 & 108.2 & 133.72 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} \;=\; \begin{pmatrix} 7 \\ 92442 \\ 4420 \\ 2593 \\ -679 \\ 469 \end{pmatrix},$$

$$\mathbf{y}'\mathbf{y} \;=\; 622,729.$$

The solutions are given by

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

$$\hat{\mathbf{b}} = \begin{pmatrix} -4909.611560 \\ 4.158675 \\ 14.335441 \\ 20.833125 \\ 1.493961 \\ 327.871209 \end{pmatrix}.$$

If the `solve()` function in R is used, then the $\hat{\mathbf{b}}$ above is obtained. If the solution is obtained by taking the inverse of the coefficient matrix, $\mathbf{X}'\mathbf{X}$, then a different solution vector is found. This solution vector is invalid due to rounding errors in the calculation of the inverse. To avoid rounding errors, subtract a number close to the mean of each $x-$variable. For example, subtract 289 from all protein values, 80 from all type values, 65 from NRR, 3 from milking speed values, and 3.6 from somatic cell scores. Then

$$\mathbf{X} = \begin{pmatrix} 1 & -43 & -5 & 1 & 0 & -0.1 \\ 1 & -63 & 0 & -2 & 1 & -0.3 \\ 1 & 13 & 2 & -5 & -1 & -0.5 \\ 1 & 58 & -3 & -7 & 0 & 0.7 \\ 1 & -22 & -9 & 1 & 2 & 0.1 \\ 1 & 26 & 6 & 6 & 1 & -0.1 \\ 1 & -48 & 10 & 3 & 0 & 0.0 \\ 1 & 1 & 3 & 5 & -1 & 0.3 \\ 1 & -18 & -2 & 2 & -2 & 0.5 \\ 1 & 97 & 0 & -1 & 0 & -0.2 \end{pmatrix},$$

and

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 10 & 1 & 2 & 3 & 0 & 0.40 \\ 1 & 22549 & -20 & -526 & -59 & 24.40 \\ 2 & -20 & 268 & 74 & -13 & -4.20 \\ 3 & -526 & 74 & 155 & 2 & 0.30 \\ 0 & -59 & -13 & 2 & 12 & -1.00 \\ 0.4 & 24.40 & -4.20 & 0.30 & -1.00 & 1.24 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 7 \\ 90419 \\ 3860 \\ 2138 \\ -700 \\ 443 \end{pmatrix},$$

Notice that the elements in the coefficient matrix are much smaller, which leads to less rounding error. The solutions are the same except for the intercept, which is now -21.947742, which is related to the first solution in that

$$-4909.611560 = -21.947742 - \hat{b}_1(289) - \hat{b}_2(80) - \hat{b}_3(65) - \hat{b}_4(3) - \hat{b}_5(3.6).$$

30 Analysis of Variance

30.1 Basic Table Format

All analysis of variance tables have a basic format.

Source	Degrees of Freedom	Sum of Squares	Formula
Total	N	SST	$\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}$
Mean	1	SSM	$\mathbf{y}'\mathbf{V}^{-1}\mathbf{1}(\mathbf{1}'\mathbf{V}^{-1}\mathbf{1})^{-1}\mathbf{1}'\mathbf{V}^{-1}\mathbf{y}$
Model	r($\mathbf{X}$)	SSR	$\hat{\mathbf{b}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = \mathbf{y}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$
Residual	N-r($\mathbf{X}$)	SSE	SST - SSR

For the example problem the table is as follows.

Analysis of Variance Table.

Source	df	SS
Total	10	622,729.00
Mean	1	4.90
Model	6	620,209.10
Residual	4	2519.90

Alternative tables can be found from different software packages. A common one is shown below, with corrections for the mean.

Alternative Analysis of Variance Table.

Source	df	SS
Total-Mean	9	622,724.10
Model-Mean	5	620,204.20
Residual	4	2519.90

30.2 Requirements for Valid Tests of Hypothesis

The distribution of $\mathbf{y}$ should be multivariate normal. An F-statistic assumes that the numerator sum of squares has a central Chi-square distribution, and the denominator sum of squares has a central Chi-square distribution, and the numerator and denominator are independent.

A sum of squares, say $\mathbf{y}'\mathbf{Q}\mathbf{y}$, has a Chi-square distribution if $\mathbf{QV}$ is idempotent.

30.2.1 Distribution of SSR

$$\text{SSR} = \mathbf{y}'\mathbf{Q_R}\mathbf{y}$$

$$\mathbf{Q_R} = \mathbf{V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}}.$$

SSR has a chi-square distribution if $\mathbf{Q_R V}$ is idempotent.

Proof:

$$\begin{aligned}
\mathbf{Q_R V Q_R V} &= [\mathbf{V^{-1}X(X'V^{-1}X)^{-}X'}][\mathbf{V^{-1}X(X'V^{-1}X)^{-}X'}] \\
&= [\mathbf{V^{-1}X(X'V^{-1}X)^{-}}][\mathbf{X'V^{-1}X(X'V^{-1}X)^{-}X'}] \\
&= [\mathbf{V^{-1}X(X'V^{-1}X)^{-}}][\mathbf{X'}] \\
&= \mathbf{Q_R V}.
\end{aligned}$$

30.2.2 Distribution of SSE

$$\text{SSE} = \mathbf{y}'\mathbf{Q_E}\mathbf{y},$$

$$\mathbf{Q_E} = \mathbf{V^{-1}} - \mathbf{Q_R}.$$

SSE has a chi-square distribution if $\mathbf{Q_E V}$ is idempotent.

Proof:

$$\begin{aligned}
\mathbf{Q_E V Q_E V} &= (\mathbf{V^{-1}} - \mathbf{Q_R})\mathbf{V}(\mathbf{V^{-1}} - \mathbf{Q_R})\mathbf{V} \\
&= (\mathbf{I} - \mathbf{Q_R V})(\mathbf{I} - \mathbf{Q_R V}) \\
&= \mathbf{I} - \mathbf{Q_R V} - \mathbf{Q_R V} + \mathbf{Q_R V Q_R V} \\
&= \mathbf{I} - 2\mathbf{Q_R V} + \mathbf{Q_R V} \\
&= \mathbf{I} - \mathbf{Q_R V} \\
&= \mathbf{Q_E V}
\end{aligned}$$

30.2.3 Independence of SSR and SSE

SSR and SSE are independent chi-square variables if $\mathbf{Q_R V Q_E} = \mathbf{0}$.

Proof:

$$\begin{aligned}
\mathbf{Q_R V Q_E} &= \mathbf{Q_R V (V^{-1} - Q_R)} \\
&= \mathbf{Q_R - Q_R V Q_R} \\
&= \mathbf{Q_R - Q_R V Q_R V V^{-1}} \\
&= \mathbf{Q_R - Q_R V V^{-1}} \\
&= \mathbf{Q_R - Q_R} \\
&= \mathbf{0}
\end{aligned}$$

30.2.4 Noncentrality parameters for SSR and SSE

The denominator of an F-statistic must be a central chi-square variable, while the numerator of the F-statistic will have a central chi-square distribution only if the null hypothesis is true. The noncentrality parameter of SSR, if the null hypothesis is false, is

$$\lambda_R = .5\mathbf{b}'\mathbf{X}'\mathbf{Q_R}\mathbf{X}\mathbf{b}.$$

The noncentrality parameter of SSE is always zero.

Proof:

$$\begin{aligned}
\mathbf{Q_E X} &= \mathbf{(V^{-1} - Q_R)X} \\
&= \mathbf{V^{-1}X - V^{-1}X(X'V^{-1}X)^{-}X'V^{-1}X} \\
&= \mathbf{V^{-1}X - V^{-1}X} \\
&= \mathbf{0}
\end{aligned}$$

30.3 Testing the Model

To test the adequacy of the model,

$$F_M = \frac{SSR/r(\mathbf{X})}{SSE/(N - r(\mathbf{X}))}.$$

SSR could have a noncentral chi-square distribution. The noncentrality parameter is non-null except when $\mathbf{b} = 0$. If $\mathbf{b} = 0$, then the model is non-significant.

A significant F_M is one that differs from the table F-values, and indicates that $\mathbf{b}$ is not a null vector and that the model does explain some of the major sources of variation. The model should usually be significant because $\mathbf{b}$ includes the mean of the observations which is usually different from zero.

The test of the model for the example data is

$$F_M = \frac{620,209.1/6}{2519.9/4} = 164.083,$$

which is highly significant.

30.4 R^2 Values

The multiple correlation coefficient, R^2, is another measure of the adequacy of a model to fit the data.

$$R^2 \ = \ \frac{SSR - SSM}{SST - SSM}.$$

The value goes from 0 to 1, and the higher the R^2, the better is the fit of the model to the data.

If the number of observations is small, then there is an adjustment for this situation. Let N be the number of observations and r be the number of regression coefficients (including intercept), then the adjusted R^2 value is

$$R^{2*} \ = \ 1 - \frac{(N-1)(1-R^2)}{(N-r)}.$$

For low R^2, the adjusted value could be negative, which means it is actually 0, i.e. no fit at all.

31 General Linear Hypothesis Testing

The general linear hypothesis test partitions SSR into sub-hypotheses about functions of $\mathbf{b}$. An hypothesis test consists of

1. a null hypothesis,
2. an alternative hypothesis,
3. a test statistic, and
4. a probability level or rejection region.

The alternative hypothesis is usually unrestricted. For hypothesis tests with viable alternative hypotheses, the reader is refered to Henderson (1984) or Searle (1971).

The null hypothesis is written as

$$\mathbf{H}'\mathbf{b} = \mathbf{c}$$

or as

$$\mathbf{H}'\mathbf{b} - \mathbf{c} = \mathbf{0}$$

where

1. $\mathbf{H}'$ must have full row rank, and
2. $\mathbf{H}'\mathbf{b}$ must be an estimable function.

If these conditions are met, then $\mathbf{H}'\mathbf{b}$ is *testable.*

The test statistic is

$$F = \frac{s/r(\mathbf{H}')}{SSE/(N - r(\mathbf{X}))}$$

where

$$s = (\mathbf{H}'\hat{\mathbf{b}} - \mathbf{c})'(\mathbf{H}'\mathbf{C}\mathbf{H})^{-1}(\mathbf{H}'\hat{\mathbf{b}} - \mathbf{c}),$$

and

$$\mathbf{C} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}.$$

The statistic, s, is always independent of SSE and has a central Chi-square distribution if the null hypothesis is true.

31.1 Alternative Computing Formula for s

The following method for computing s shows that the test is comparing the solutions for $\hat{\mathbf{b}}$ which were obtained from GLS equations with another set of solutions obtained from a restricted set of GLS equations that were restricted assuming the null hypothesis was true. If the sum of squares of differences are significant, then the null hypothesis can be rejected (or fail to be accepted). The alternative computing form is

$$s = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} - (\hat{\mathbf{b}}_{\mathbf{o}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \hat{\theta}_{\mathbf{o}}'\mathbf{c})$$

where

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

and

$$\begin{pmatrix} \hat{\mathbf{b}}_{\mathbf{o}} \\ \hat{\theta}_{\mathbf{o}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{H} \\ \mathbf{H}' & \mathbf{0} \end{pmatrix}^{-} \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{c} \end{pmatrix}.$$

The equivalence of the alternative formula to that in the previous section is as follows. From the first row of the above equation,

$$\mathbf{X'V^{-1}X\hat{b}_o} + \mathbf{H'}\hat{\theta}_\mathbf{o} = \mathbf{X'V^{-1}y}$$

which can be re-arranged and solved for $\mathbf{\hat{b}_o}$ as

$$\begin{aligned}\mathbf{\hat{b}_o} &= \mathbf{C(X'V^{-1}y - H}\hat{\theta}_\mathbf{o}) \\ &= \mathbf{CX'V^{-1}y - CH}\hat{\theta}_\mathbf{o} \\ &= \mathbf{\hat{b} - CH}\hat{\theta}_\mathbf{o}\end{aligned}$$

and consequently from the second row of the restricted equations,

$$\begin{aligned}\mathbf{H'\hat{b}_o} &= \mathbf{c} \\ &= \mathbf{H'\hat{b} - H'CH}\hat{\theta}_\mathbf{o}\end{aligned}$$

Solving for $\hat{\theta}_\mathbf{o}$ gives

$$\hat{\theta}_\mathbf{o} = (\mathbf{H'CH})^{-1}(\mathbf{H'\hat{b} - c}).$$

Now substitute this result into that for $\mathbf{\hat{b}_o}$ to obtain

$$\mathbf{\hat{b}_o} = \mathbf{\hat{b} - CH(H'CH)^{-1}(H'\hat{b} - c)}.$$

Taking the alternative form of s and putting in the new solutions for $\mathbf{\hat{b}_o}$ and $\hat{\theta}_\mathbf{o}$ the original linear hypothesis formula is obtained, i.e.

$$\begin{aligned}s &= SSR - (\mathbf{\hat{b}'_oX'V^{-1}y} + \hat{\theta}'_\mathbf{o}\mathbf{c}) \\ &= SSR - [\mathbf{\hat{b}' - (H'\hat{b} - c)'(H'CH)^{-1}H'C]X'V^{-1}y} \\ &\quad + \mathbf{(H'\hat{b} - c)'(H'CH)}^{-1}\mathbf{c} \\ &= SSR - SSR + \mathbf{(H'\hat{b} - c)'(H'CH)^{-1}(H'\hat{b} - c)} \\ &= \mathbf{(H'\hat{b} - c)'(H'CH)^{-1}(H'\hat{b} - c)}\end{aligned}$$

31.2 Hypotheses having rank of X

Suppose there are two null hypotheses such that

$$\mathbf{H'_1b} = \mathbf{0} \text{ with } r(\mathbf{H'_1}) = r(\mathbf{X}),$$

$$\mathbf{H'_2b} = \mathbf{0} \text{ with } r(\mathbf{H'_2}) = r(\mathbf{X}),$$

but

$$\mathbf{H'_1} \neq \mathbf{H'_2},$$

then s_1 is equal to s_2 and both of these are equal to SSR. To simplify the proof, but not necessary to the proof, let $\mathbf{X}$ have full column rank. Both null hypotheses represent estimable functions and therefore, each $\mathbf{H_i'}$ may be written as $\mathbf{T'X}$ for some $\mathbf{T'}$, and $\mathbf{T'X}$ has order and rank equal to $r(\mathbf{X})$. Consequently, $\mathbf{T'X}$ can be inverted. Then

$$\begin{aligned} s_i &= \hat{\mathbf{b}}'\mathbf{H_i}(\mathbf{H_i'CH_i})^{-1}\mathbf{H_i'}\hat{\mathbf{b}} \\ &= \hat{\mathbf{b}}'\mathbf{X'T}(\mathbf{T'XCX'T})^{-1}\mathbf{T'X}\hat{\mathbf{b}} \\ &= \hat{\mathbf{b}}'(\mathbf{X'T})[(\mathbf{X'T})^{-1}\mathbf{C}^{-}(\mathbf{T'X})^{-1}](\mathbf{T'X})\hat{\mathbf{b}} \\ &= \hat{\mathbf{b}}'\mathbf{C}^{-}\hat{\mathbf{b}} \\ &= \mathbf{y'V^{-1}XC(C^{-})CX'V^{-1}y} \\ &= \mathbf{y'V^{-1}XCX'V^{-1}y} \\ &= SSR \end{aligned}$$

31.3 Orthogonal Hypotheses

Let $\mathbf{H'b} = \mathbf{0}$ be a null hypothesis such that $r(\mathbf{H'}) = r(\mathbf{X}) = r$, and therefore,

$$s = (\mathbf{H'}\hat{\mathbf{b}})'(\mathbf{H'CH})^{-1}\mathbf{H'}\hat{\mathbf{b}} = SSR.$$

Now partition $\mathbf{H'}$ into r rows as

$$\mathbf{H'} = \begin{pmatrix} \mathbf{h_1'} \\ \mathbf{h_2'} \\ \vdots \\ \mathbf{h_r'} \end{pmatrix}$$

where each $\mathbf{h_i'}$ has rank of one, then the rows of $\mathbf{H'}$ are *orthogonal* if

$$\mathbf{h_i'Ch_j} = 0 \text{ for all i} \neq \text{ j}.$$

If the rows are orthogonal to each other then

$$\mathbf{H'CH} = \begin{pmatrix} \mathbf{h_1'Ch_1} & 0 & \dots & 0 \\ 0 & \mathbf{h_2'Ch_2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{h_r'Ch_r} \end{pmatrix}$$

which means that

$$s = SSR = \sum_{i=1}^{r}(\mathbf{h_i'}\hat{\mathbf{b}})^2/(\mathbf{h_i'Ch_i}).$$

When data are unbalanced, orthogonal contrasts are difficult to obtain and depend on the number of observations in each subclass. Orthogonal contrasts are not essential for hypothesis testing. The subpartitions of SSR do not necessarily have to sum to SSR.

31.4 Power of the Test

Every hypothesis test results in one of four possible outcomes depending on the true state of nature and the outcome of the test, as shown below:

True State	Test Result	
of Nature	Reject	Accept
Null hypothesis true	Type I error	No error
Null hypothesis false	No error	Type II error

Type I errors occur when the null hypothesis is rejected even though it is true, and Type II errors occur when the null hypothesis is not rejected when it is false. Depending on the risks associated with each type of error, then either Type I or Type II errors can be minimized. For example, if the hypothesis is that an animal does not have a disease and the animal does not have the disease, then rejection of the hypothesis (Type I error) results in the animal being treated for the disease. This could be very costly and might involve surgery, or the treatment could be minimal in cost and without harmful effects on the animal. On the other hand, if the animal really does have the disease and the hypothesis is not rejected (Type II error), then the animal would not be treated. The cost of no treatment might be death of the animal.

The *Power of the Test* is 1 minus the probability of a Type II error. If a Type II error results in death loss or heavy financial loss, then the researcher should use a very high Power of the Test. The Power of the Test is important in clinical studies, but not as critical in animal breeding research on field data.

32 Reduction Notation

Another computing technique for hypothesis testing is the use of reduction notation. The equivalence to the general linear hypothesis method will be demonstrated. Let the general fixed effects model be re-written as

$$\mathbf{y} = \sum_{i=1}^{p} \mathbf{X_i b_i} + \mathbf{e}$$

where p is the number of fixed factors in the model. Then the reduction due to fitting the full model is denoted as

$$R(\mathbf{b_1}, \mathbf{b_2}, \ldots, \mathbf{b_p}) = \hat{\mathbf{b}}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} = SSR.$$

To test the importance of factor $\mathbf{b_1}$, for example, the null hypothesis would be

$$\mathbf{H}_1'\mathbf{b} = \mathbf{0} \text{ or } \mathbf{b_1} = \mathbf{0}.$$

To obtain the appropriate reduction to test this hypothesis, construct a set of equations with $\mathbf{X_1 b_1}$ omitted from the model. Let

$$\mathbf{W} = \left(\begin{array}{cccc} \mathbf{X_2} & \mathbf{X_3} & \ldots & \mathbf{X_p} \end{array} \right),$$

then the reduction due to fitting the submodel with factor 1 omitted is

$$R(\mathbf{b_2}, \mathbf{b_3}, \ldots \mathbf{b_p}) \ = \ \mathbf{y'V^{-1}W(W'V^{-1}W)^{-}W'V^{-1}y}$$

and the test statistic is computed as

$$\begin{aligned} s &= R(\mathbf{b_1}, \mathbf{b_2}, \mathbf{b_3}, \ldots \mathbf{b_p}) - R(\mathbf{b_2}, \mathbf{b_3}, \ldots \mathbf{b_p}) \\ &= R(\mathbf{b_1} \mid \mathbf{b_2}, \mathbf{b_3}, \ldots \mathbf{b_p}) \end{aligned}$$

with $r(\mathbf{X}) - r(\mathbf{W})$ degrees of freedom. The above s is equivalent to the general linear hypothesis form because

$$R(\mathbf{b_2}, \mathbf{b_3}, \ldots \mathbf{b_p}) \ = \ \hat{\mathbf{b}}'_{\mathbf{o}}\mathbf{X'V^{-1}y}$$

where $\hat{\mathbf{b}}_{\mathbf{o}}$ is a solution to

$$\left(\begin{array}{cc} \mathbf{X'V^{-1}X} & \mathbf{H_1} \\ \mathbf{H'_1} & \mathbf{0} \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}}_{\mathbf{o}} \\ \hat{\theta}_{\mathbf{o}} \end{array} \right) \ = \ \left(\begin{array}{c} \mathbf{X'V^{-1}y} \\ \mathbf{0} \end{array} \right).$$

The addition of $\mathbf{H'_1}$ as a LaGrange Multiplier to $\mathbf{X'V^{-1}X}$ accomplishes the same purpose as omitting the first factor from the model and solving the reduced equations $\mathbf{W'V^{-1}W}\hat{\mathbf{b}}_{\mathbf{s}} = \mathbf{W'V^{-1}y}$, where $\hat{\mathbf{b}}_{\mathbf{s}}$ is $\hat{\mathbf{b}}$ excluding $\hat{\mathbf{b}}_{\mathbf{1}}$.

As long as the reductions due to fitting submodels are subtracted from SSR, the same s values will be calculated. The reduction notation method, however, assumes the null hypothesis that $\mathbf{H'b} = \mathbf{0}$ while the general linear hypothesis procedure allows the more general null hypothesis, namely, $\mathbf{H'b} = \mathbf{c}$.

33 Example in R

```
# X is a matrix with N rows and r columns
# y is a vector of the N observations
# Ordinary LS equations are

XX = t(X) %*% X
Xy = t(X) %*% y

bhat = solve(XX,Xy)
# OR
C = ginv(XX)
bhat = C %*% Xy
# The two vectors should be identical.
# If they are not equal, then rounding errors are
# occurring
```

```
# AOV Table
SST = t(y) %*% y
SSR = t(bhat) %*% Xy
SSE = SST - SSR
SSM = sum(y)*mean(y)
sors = c("Total","Mean","Model","Residual")
df = c(N, 1, r, (N-r) )
SS = c(SST, SSM, SSR, SSE)
FF = c(0,0,((SSR/r)/(SSE/(N-r))) , 0)
AOV = cbind(sors,df,SS,FF)
AOV

# R-squared
R2 = (SSR - SSM)/(SST - SSM)
R2S = 1 - ( (N-1)*(1-R2)/(N-r) )
R2
R2S
```

```
# General Linear Hypothesis (assuming r=4, e.g.)
H0 = matrix(data=c(1, 0, 0, 0, 0, 1, 0, 0),byrow=TRUE,ncol=4)
c0 = matrix(data=c(250, 13),ncol=1)
w = (H0 %*% bhat) -c0
HCH = ginv(H0 %*% C %*% t(H0))
s = t(w) %*% HCH %*% w
df = 2 # two rows of H0
F = (s/df)/(SSE/(N-r)) # with df and (N-r) degrees of freedom
```

34 EXERCISES

1. Prove that s in the general linear hypothesis is Chi-square and is independent of SSE. This is complicated and difficult.

2. Using the data from the example problem, test the hypothesis that

$$b_1 = 5.$$

3. Using the data from the example problem, test the following hypothesis:

$$\begin{aligned} b_2 &= 0 \\ b_4 - b_5 &= 0 \end{aligned}$$

4. Derive a regression equation to predict the weight of rainbow trout at 9 months of age from their fork lengths and body circumference measures.

Data on Rainbow Trout

Fish	Fork L	Circum.	Weight
	cm	cm	g
1	14.5	6.77	24.7
2	12.6	6.15	24.3
3	16.0	7.51	25.2
4	14.2	8.24	23.4
5	14.7	8.06	24.3
6	14.9	7.88	26.1
7	15.6	7.87	24.4
8	16.0	9.66	24.6
9	13.6	6.14	23.5
10	14.8	7.39	23.0
11	15.0	8.22	23.6
12	11.5	6.13	22.1
13	17.5	8.62	26.6

(a) Construct the AOV table.

(b) Test each regression for significance from zero.

(c) Subtract 14.5 from fork lengths, and 7.5 from circumferences, and re-analyze the data.

5. In a spline regression analysis inclusion of one of the x-variables depends on the value of another x-variable. In this problem, x_2 is included only if x_1 is greater than 7, then $x_2 = X_1 - 7$, otherwise $x_2 = 0$. The number 7 is called a knot, a point in the curve where the shape changes direction. Below are the data and model.

Spline Regression Data.

x_1	x_1^2	x_2	x_2^2	y_i
8	64	1	1	147
5	25	0	0	88
13	169	6	36	202
6	36	0	0	135
9	81	2	4	151
11	121	4	16	198
18	324	11	121	94
12	144	5	25	173
7	49	0	0	169
2	4	0	0	122

$$y_i = a \ + \ b_1 x_1 \ + \ b_2 x_1^2 \ + \ b_3 x_2 \ + \ b_4 x_2^2 \ + \ e_i,$$

with $Var(\mathbf{y}) = \mathbf{I}\sigma_e^2$.

(a) Give the AOV table and R^2 value.

(b) Test the hypothesis that

$$a = 250.$$

(c) Test the hypothesis that

$$\begin{aligned} b_1 &= -5 \\ b_2 &= 0 \\ b_2 - b_4 &= 3 \end{aligned}$$

6. Below are data on protein yields of a cow during the course of 365 days lactation.

Daily Protein Yields, kg for Agathe.

Days in Milk(d)	Yield(kg)
6	1.20
14	1.25
25	1.32
38	1.41
52	1.49
73	1.35
117	1.13
150	0.95
179	0.88
214	0.74
246	0.65
271	0.50
305	0.40
338	0.25
360	0.20

$$y_i = b_0 \ + \ b_1 s \ + \ b_2 s^2 \ + \ b_3 t \ + \ b_4 t^2 \ + \ e_i,$$

and assume for simplicity that $Var(\mathbf{y}) = \mathbf{I}\sigma_e^2$. Also,

$$\begin{aligned} s &= d/365 \\ t &= log(400/d) \end{aligned}$$

(a) Analyze the data and give the AOV table.

(b) Estimate the protein on day 200 and the standard error of that estimate.

(c) Estimate the total protein yield from day 5 to day 305, and a standard error of that estimate.

(d) Could this prediction equation be used on another cow? Could it be used on a clone of Agathe?

Estimability

35 Introduction

Consider models where the rank of $\mathbf{X}$ is less than the number of columns in $\mathbf{X}$. An example is a two-way cross classified model without interaction where

$$y_{ijk} = \mu + A_i + B_j + e_{ijk},$$

and

y_{ijk} is an observation on the postweaning gain (165 days) of male beef calves,

μ is an overall mean,

A_i is an effect due to the age of the dam of the calf,

B_j is an effect due to the breed of the calf, and

e_{ijk} is a residual effect specific to each observation.

There are four age of dam groups, namely 2-yr-olds, 3-yr-olds, 4-yr-olds, and 5-yr-olds or greater, and three breeds, i.e. Angus(AN), Hereford(HE), and Simmental(SM). The following assumptions are made,

1. There are no breed by age of dam interactions.
2. Diet and management effects were the same for all calves.
3. All calves were raised in the same environment in the same months and year and at the same age of development.
4. Calves resulted from random mating of sires to dams, and there is only one progeny per sire and per dam.

Also,

$$\begin{aligned} E(y_{ijk}) &= \mu + A_i + B_j \quad \text{and} \\ Var(y_{ijk}) &= \sigma_e^2, \end{aligned}$$

that is, the same residual variance was assumed for all calves regardless of breed group or age of dam group. The data are given in the table below.

Growth Data on Beef Calves

Calf Tattoo	Age of Dam (yr)	Breed of calf	PWG (kg)
16K	2	AN	346
18K	3	AN	355
22K	4	AN	363
101L	5+	HE	388
121L	5+	HE	384
98L	2	HE	366
115L	3	HE	371
117L	3	HE	375
52J	4	SM	412
49J	5+	SM	429
63J	2	SM	396
70J	2	SM	404

Determine the effects of breed of calf and age of dam on postweaning gains of male calves at a fixed age. The model in matrix notation is $\mathbf{y} = \mathbf{Xb} + \mathbf{e}$, with

$$E(\mathbf{y}) = \mathbf{Xb} \text{ and } Var(\mathbf{y}) = \mathbf{I}\sigma_e^2,$$

where

$$\mathbf{Xb} = \left(\begin{array}{ccccccccc} 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \left(\begin{array}{c} \mu \\ A_2 \\ A_3 \\ A_4 \\ A_{5+} \\ B_{AN} \\ B_{HE} \\ B_{SM} \end{array} \right).$$

36 The Rank of X

Procedures for determining the rank of a matrix using elementary operators were given in previous notes. In practice, the matrix $\mathbf{X}$ is too large to apply elementary operators. Also, in the use of classification models, $\mathbf{X}$ has mostly just zeros and ones in it. For one factor in a model, say diets, in a row of $\mathbf{X}$ there will be a single one, indicating the particular diet, and the remaining elements in the row will be zero. If there were five diets, then there would be five columns in $\mathbf{X}$, and if those five columns were added together, the result would be one column with all values equal to one.

If there were two factors, diets and breeds, in $\mathbf{X}$, then the columns for the diet effects would always sum to give one, and the columns for the breed effects would also sum to give one. Thus, there would be a dependency between diet effects and breed effects. The dependencies need to be identified and removed. For example, if the column for diet 1 was removed from $\mathbf{X}$, then there would not be a dependency between breed effects and diet effects any longer. Any diet or any breed column could have been removed to eliminate the dependency.

In the example problem of the previous section, the model equation is

$$y_{ijk} = \mu + A_i + B_j + e_{ijk},$$

where age of dam has 4 levels, and breed of calf has 3 levels. The total number of columns in $\mathbf{X}$ is 8. The columns of age of dam effects will have a dependency with the columns of breed of calf, and therefore, one of those columns has to be removed. Suppose the column for 2-yr-old age of dam is removed. The columns for breed of calf still sum to a column of ones, which is the same as the column for μ. Thus, another column (one of the breed of calf columns OR the μ column) needs to be removed. There are no further dependencies after removing the μ column. There were two restrictions imposed, so the rank of $\mathbf{X}$ is therefore, 6.

Given the model, and number of levels of each factor, it is possible to determine, before analyzing the data, the rank of $\mathbf{X}$ without using elementary operators. This takes practice, and problems are given in the Exercise section.

Another example is as follows. Let

$$y_{ijk} = \mu + A_i + B_{ij} + e_{ijk},$$

where factor A has 5 levels and factor B has 3 levels for every level of factor A. The total number of columns in $\mathbf{X}$ is $1 + 5 + 5 * (3)$ or 21. Factors A and B both have i as a subscript, and therefore, there is a dependency for every i, or 5 dependencies. All columns for factor A need to be removed. Also, summing all 15 columns of factor B gives a column of ones, equivalent to the column for μ. Remove the column for μ and all dependencies are removed. The rank of $\mathbf{X}$ is 15.

37 Estimable Functions

If the rank of $\mathbf{X}$ is less than the number of columns in $\mathbf{X}$, then

- the inverse of $\mathbf{X}'\mathbf{X}$ does not exist because its determinant is zero.

- a generalized inverse must be used, let

  $$\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-}$$

  such that

  $$\tilde{\mathbf{b}} = \mathbf{CX}'\mathbf{y}.$$

- there are an infinite number of generalized inverses of $\mathbf{X}'\mathbf{X}$, and therefore, an infinite number of possible solution vectors, $\tilde{\mathbf{b}}$.

Recall that,

$$\begin{aligned} \mathbf{X}'\mathbf{XCX}'\mathbf{X} &= \mathbf{X}'\mathbf{X} \\ \mathbf{XCX}'\mathbf{X} &= \mathbf{X} \end{aligned}$$

and $\mathbf{XCX}'$ is invariant to $\mathbf{C}$.

37.1 Solution vector

If

$$\tilde{\mathbf{b}} = \mathbf{CX}'\mathbf{y},$$

then

$$E(\tilde{\mathbf{b}}) = \mathbf{CX}'\mathbf{Xb}$$

is the first set of estimable functions.

37.2 Other Estimable Functions

Let $\mathbf{K}'\mathbf{b}$ represent a set of functions of the elements of $\mathbf{b}$, to determine if that function is estimable in terms of the working linear model, then show that

$$\mathbf{K}'\mathbf{CX}'\mathbf{X} = \mathbf{K}'.$$

If this equality does not hold, then the function $\mathbf{K}'\mathbf{b}$ is not estimable.

Estimable functions are unique regardless of the solution vector or generalized inverse of $\mathbf{X}'\mathbf{X}$ that is derived.

38 Generalized Least Squares Equations

38.1 Equations

Because $Var(\mathbf{y}) = \mathbf{I}\sigma_e^2$, the GLS equations reduce to ordinary LS equations, i.e. $\mathbf{X}'\mathbf{X}\tilde{\mathbf{b}} = \mathbf{X}'\mathbf{y}$. For the example data,

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 4 & 3 & 2 & 3 & 3 & 5 & 4 \\ 4 & 4 & 0 & 0 & 0 & 1 & 1 & 2 \\ 3 & 0 & 3 & 0 & 0 & 1 & 2 & 0 \\ 2 & 0 & 0 & 2 & 0 & 1 & 0 & 1 \\ 3 & 0 & 0 & 0 & 3 & 0 & 2 & 1 \\ 3 & 1 & 1 & 1 & 0 & 3 & 0 & 0 \\ 5 & 1 & 2 & 0 & 2 & 0 & 5 & 0 \\ 4 & 2 & 0 & 1 & 1 & 0 & 0 & 4 \end{pmatrix}, \mathbf{X}'\mathbf{y} = \begin{pmatrix} 4589 \\ 1512 \\ 1101 \\ 775 \\ 1201 \\ 1064 \\ 1884 \\ 1641 \end{pmatrix},$$

and $\mathbf{y}'\mathbf{y} = 1,761,549$.

38.2 Generalized Inverse and Solution Vector

The rank of $\mathbf{X}$ in this example is 6. The dependencies are: columns 2, 3, 4, and 5 sum to give column 1 as do columns 6, 7, and 8. Two constraints on the solutions are needed. Let $\hat{\mu}$ and $\hat{A}_2$ be set equal to zero, then a generalized inverse of $(\mathbf{X}'\mathbf{X})$ is $\mathbf{C}$ where $\mathbf{C}$ is equal to

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & .71366 & .22467 & .32159 & -.31278 & -.41410 & -.13656 \\ 0 & 0 & .22467 & .79295 & .19383 & -.33921 & -.16740 & -.24670 \\ 0 & 0 & .32159 & .19383 & .66960 & -.17181 & -.39648 & -.21586 \\ 0 & 0 & -.31278 & -.33921 & -.17181 & .55066 & .19383 & .12775 \\ 0 & 0 & -.41410 & -.16740 & -.39648 & .19383 & .52423 & .14097 \\ 0 & 0 & -.13656 & -.24670 & -.21586 & .12775 & .14097 & .36564 \end{pmatrix}.$$

The corresponding solution vector is $\hat{\mathbf{b}} = \mathbf{C}\mathbf{X}'\mathbf{y}$, or

$$\begin{pmatrix} \tilde{\mu} \\ \tilde{A}_2 \\ \tilde{A}_3 \\ \tilde{A}_4 \\ \tilde{A}_{5+} \\ \tilde{B}_{AN} \\ \tilde{B}_{HE} \\ \tilde{B}_{SM} \end{pmatrix} = \begin{pmatrix} 0. \\ 0. \\ 9.0264 \\ 13.5639 \\ 24.4934 \\ 347.1366 \\ 363.3921 \\ 400.7357 \end{pmatrix}.$$

38.3 Expectation of Solution Vector

The expectation of $\hat{\mathbf{b}}$ is

$$E(\hat{\mathbf{b}}) = E(\mathbf{CX'y}) = \mathbf{CX'Xb}.$$

For the example data, the expectations are

$$E \begin{pmatrix} \hat{\mu} \\ \hat{A}_2 \\ \hat{A}_3 \\ \hat{A}_4 \\ \hat{A}_{5+} \\ \hat{B}_{AN} \\ \hat{B}_{HE} \\ \hat{B}_{SM} \end{pmatrix} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \mu \\ A_2 \\ A_3 \\ A_4 \\ A_{5+} \\ B_{AN} \\ B_{HE} \\ B_{SM} \end{pmatrix}$$

$$= \begin{pmatrix} 0 \\ 0 \\ A_3 - A_2 \\ A_4 - A_2 \\ A_{5+} - A_2 \\ \mu + A_2 + B_{AN} \\ \mu + A_2 + B_{HE} \\ \mu + A_2 + B_{SM} \end{pmatrix}.$$

Hence, $\hat{A}_3 = 9.0264$ is not an estimate of A_3, but is rather an estimate of the difference of $A_3 - A_2$. Also, $\hat{B}_{AN} = 347.1366$ is not an estimate of B_{AN}, but of $\mu + A_2 + B_{AN}$, an estimate of the mean of Angus calves from 2-yr-old dams.

By computing the expected value of the solution vector, the functions of the true parameters that have been estimated by a particular generalized inverse can be determined. These functions are estimable because the solution vector is a linear function of $\mathbf{y}$, which is always estimable.

38.4 Other Solution Vectors

Other possible solution vectors are given by the formula,

$$\mathbf{b^o} = \mathbf{CX'y} + (\mathbf{I} - \mathbf{CX'X})\mathbf{z},$$

where $\mathbf{z}$ is any vector of arbitrary constants. For example, let

$$\mathbf{z}' = \begin{pmatrix} 1 & -2 & 6 & 0 & 1 & -5 & 1 & 3 \end{pmatrix},$$

then another solution vector is

$$\mathbf{b}^o = \hat{\mathbf{b}} + \left(\begin{array}{rrrrrrrr} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \left(\begin{array}{r} 1 \\ -2 \\ 6 \\ 0 \\ 1 \\ -5 \\ 1 \\ 3 \end{array}\right) = \left(\begin{array}{r} 1 \\ -2 \\ 7.0264 \\ 11.5639 \\ 22.4934 \\ 348.1366 \\ 364.3921 \\ 401.7357 \end{array}\right).$$

In order to check that $\mathbf{b^o}$ is a valid solution vector, multiply $\mathbf{b^o}$ by $\mathbf{X'X}$ which gives

$$\begin{aligned} \mathbf{X'Xb^o} &= \mathbf{X'XCX'y} + \mathbf{X'X}(\mathbf{I} - \mathbf{CX'X})\mathbf{z} \\ &= \mathbf{X'y} + (\mathbf{X'X} - \mathbf{X'XCX'X})\mathbf{z} \\ &= \mathbf{X'y} \end{aligned}$$

Hence, $\mathbf{b^o}$ regenerates $\mathbf{X'y}$ and is therefore, a valid solution vector. Because $\mathbf{z}$ is an arbitrary vector, there are an infinite number of valid solution vectors to these equations.

38.5 Estimable Functions

The solution for A_3 from $\tilde{\mathbf{b}}$ was 9.0264 which was an estimate of $A_3 - A_2$. More generally,

$$\mathbf{k}'\tilde{\mathbf{b}} = \left(\begin{array}{cccccccc} 0 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \end{array}\right) \tilde{\mathbf{b}} = \tilde{A}_3 - \tilde{A}_2.$$

The solution for A_3 from $\mathbf{b}^o$ was 7.0264, and the expectation is not known. However,

$$\mathbf{k'b^o} = A_3^o - A_2^o = 7.0264 - (-2) = 9.0264,$$

and the expected value of $\mathbf{k'b^o}$ is $A_3 - A_2$, the same as the expected value of $\mathbf{k}'\tilde{\mathbf{b}}$.

Suppose the following two functions are of interest.

$$\mathbf{k_1'} = \left(\begin{array}{cccccccc} 1 & 1 & -1 & 0 & 0 & 1 & 0 & 0 \end{array}\right),$$

and

$$\mathbf{k_2'} = \left(\begin{array}{cccccccc} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{array}\right).$$

One way to quickly determine if these functions are not estimable is to multiply each of them times $\tilde{\mathbf{b}}$ and $\mathbf{b^o}$ and if the results are different, then that function is not estimable. For the example data the results are

$$\mathbf{k}_1'\tilde{\mathbf{b}} = 338.1102 \text{ and } \mathbf{k_1'b^o} = 340.1102,$$

and

$$\mathbf{k}_2'\tilde{\mathbf{b}} = 414.2996 \text{ and } \mathbf{k_2'b^o} = 414.2996.$$

Thus, $\mathbf{k_1'b}$ would definitely not be an estimable function because different results were obtained with different solution vectors. The function, $\mathbf{k_2'b}$ gave the same results for these two solution vectors, and therefore **might** be an estimable function.

An exact method of determining estimability would be to check if

$$\mathbf{k'CX'X} - \mathbf{k'} = \mathbf{0} \text{ or } \mathbf{k'CX'X} = \mathbf{k'}.$$

For the two functions in the previous paragraph,

$$\mathbf{k_1'CX'X} = \begin{pmatrix} 1 & 2 & -1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \neq \mathbf{k_1'},$$

and

$$\mathbf{k_2'CX'X} = \mathbf{k_2'}.$$

Hence, $\mathbf{k_2'b}$ is an estimable function.

38.6 Comments

Some basic results on estimability for classification models are

- μ is generally not estimable by itself.
- Differences among levels of main factors are estimable provided that factor has no interactions with other factors.
- Only estimable functions have any meaning or value in an analysis.
- Only estimable functions can be tested.

39 Sum of Squares Due to Model

The infinite number of possible solution vectors to the GLS equations yield the same sum of squares due to the model. That is,

$$\begin{aligned} SSR &= \mathbf{b^{o'}X'y} \\ &= \mathbf{y'XCX'y} + \mathbf{z'}(\mathbf{I} - \mathbf{X'XC})\mathbf{X'y} \\ &= \mathbf{y'XCX'y} \end{aligned}$$

because $(\mathbf{I} - \mathbf{X'XC})\mathbf{X'} = \mathbf{0}$, and note that $\mathbf{XCX'} = \mathbf{XC'X'}$ is unique for all possible generalized inverses of $\mathbf{X'X}$. For the example data,

$$\begin{aligned} SSR &= \hat{\mathbf{b}}'\mathbf{X'y} \\ &= 1,761,457.9 \\ &= \mathbf{b^{o'}X'y} \end{aligned}$$

This means that the test of the significance of the model is always unique, i.e. the test does not depend on the particular solution vector that has been calculated.

40 Least Squares Means

Least squares means are frequently reported in the scientific literature and are provided by several statistical packages. Least squares means are estimable functions and are therefore unique. For the example data, the least squares means for age of dam effects would be

Least Squares Means
For Age of Dam

Age of Dam	Least Square Mean
2	370.4215
3	379.4479
4	383.9854
5+	394.9149

The functions that these numbers are estimating are not equal to $\mu + A_i$ because $\mu + A_i$ is not an estimable function. Instead, these means are estimating $\mu + A_i + \frac{1}{3}(B_{AN} + B_{HE} + B_{SM})$.

Similarly, the least square means for the breed effects are estimating the function

$$\mu + \frac{1}{4}(A_2 + A_3 + A_4 + A_{5+}) + B_j.$$

The values were 358.9075, 375.1630, and 412.5066 for AN, HE, and SM, respectively.

Least square means are not well understood by all researchers because estimability is not understood.

41 Variance of Estimable Functions

If $\mathbf{k}'\mathbf{b}$ is an estimable function, then

$$\mathbf{k}'\mathbf{C}\mathbf{X}'\mathbf{X} = \mathbf{k}'$$

or $\mathbf{k}' = \mathbf{t}'\mathbf{X}$ with

$$\mathbf{t}' = \mathbf{k}'\mathbf{C}\mathbf{X}'.$$

Then

$$\begin{aligned}
Var(\mathbf{k}'\tilde{\mathbf{b}}) &= \mathbf{k}'Var(\tilde{\mathbf{b}})\mathbf{k} \\
&= \mathbf{k}'Var(\mathbf{CX}'\mathbf{y})\mathbf{k} \\
&= \mathbf{k}'\mathbf{CX}'Var(\mathbf{y})\mathbf{XC}'\mathbf{k} \\
&= \mathbf{k}'\mathbf{CX}'\mathbf{XC}'\mathbf{k}\sigma_e^2 \\
&= \mathbf{t}'\mathbf{XC}'\mathbf{k}\sigma_e^2 \\
&= \mathbf{k}'\mathbf{C}'\mathbf{k}\sigma_e^2 \\
&= \mathbf{k}'\mathbf{C}\mathbf{k}\sigma_e^2
\end{aligned}$$

If $\mathbf{k}' = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}$ and $\mathbf{k}'\tilde{\mathbf{b}} = 425.2291$, then

$$\begin{aligned}
Var(\mathbf{k}'\hat{\mathbf{b}}) &= (.79295 + .36564 + 2(-.24670))\sigma_e^2 \\
&= .66519\sigma_e^2.
\end{aligned}$$

Similarly, if

$$\mathbf{K}' = \begin{pmatrix} 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & -1 & 1 \end{pmatrix},$$

then

$$\begin{aligned}
Var(\mathbf{K}'\hat{\mathbf{b}}) &= \mathbf{K}'\mathbf{CK}\sigma_e^2 \\
&= \begin{pmatrix} .66960 & -.18062 \\ -.18062 & .60793 \end{pmatrix}\sigma_e^2
\end{aligned}$$

42 Analysis of Variance

AOV for Calf Postweaning Gain Data.

Source	Degrees of Freedom	Sum of Squares	Notation
Total	12	1,761,549	SST
Mean	1	1,754,910.08	SSM
Model	6	1,761,457.9	SSR
Residual	6	91.1	SSE

An estimate of σ_e^2 is

$$\hat{\sigma}_e^2 = \frac{SSE}{(N - r(\mathbf{X}))} = 15.1833.$$

42.1 The Model

A test of the adequacy of the model is

$$F_M = \frac{SSR/r(\mathbf{X})}{SSE/(N-r(\mathbf{X}))} = \frac{293,576.32}{15.1833} = 19,335.5.$$

Therefore, the model is highly significant. Another criterion would be the R^2 where

$$R^2 = \frac{SSR - SSM}{SST - SSM} = .9863.$$

Adjusting for the small sample size gives

$$R^{2*} = .9748.$$

Thus, the model is very adequate at explaining variation in calf postweaning gains.

42.2 Partitioning SSR

Two tests of interest for the example data would be

1. Age of dam effects, and
2. Breed of calf effects.

42.2.1 Age of Dam Effects

The null hypothesis would be $\mathbf{H}_1'\mathbf{b} = \mathbf{0}$ where

$$\mathbf{H}_1' = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & -1 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{H}_1'\mathbf{b} = \begin{pmatrix} A_2 - A_3 \\ A_2 - A_4 \\ A_2 - A_{5+} \end{pmatrix}.$$

$\mathbf{H}_1'\mathbf{b}$ is an estimable function, and therefore, the test is testable.

The sum of squares is

$$s_1 = (\mathbf{H}_1'\tilde{\mathbf{b}})'(\mathbf{H}_1'\mathbf{G}\mathbf{H}_1)^{-1}\mathbf{H}_1'\tilde{\mathbf{b}}$$

where

$$\mathbf{H}_1'\tilde{\mathbf{b}} = \begin{pmatrix} -9.0264 \\ -13.5639 \\ -24.4934 \end{pmatrix},$$

$$\mathbf{H_1'GH_1} = \begin{pmatrix} .71366 & .22467 & .32159 \\ .22467 & .79295 & .19383 \\ .32159 & .19383 & .66960 \end{pmatrix},$$

$$(\mathbf{H_1'GH_1})^{-1} = \begin{pmatrix} 1.86667 & -.33333 & -.80000 \\ -.33333 & 1.41667 & -.25000 \\ -.80000 & -.25000 & 1.95000 \end{pmatrix},$$

and $s_1 = 981.10653$ with 3 degrees of freedom. The F-test is

$$F_1 = \frac{s_1/r(\mathbf{H}_1')}{15.185022} = \frac{327.0355}{15.185022} = 21.5.$$

Age of dam effects, i.e. differences among age of dam groups, are significantly different from zero.

Suppose the null hypothesis for age of dam effects had been written as $\mathbf{H_2'b} = \mathbf{0}$ where

$$\mathbf{H}_2' = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 & 0 \end{pmatrix},$$

and

$$\mathbf{H_2'b} = \begin{pmatrix} A_2 - A_3 \\ A_3 - A_4 \\ A_4 - A_{5+} \end{pmatrix},$$

then it can be shown that $s_2 = s_1$, and the same conclusions would be drawn with respect to the age of dam differences. To prove that $s_2 = s_1$, note that

$$\mathbf{H}_2' = \mathbf{PH}_1'$$

where $\mathbf{P}$ is an elementary operator matrix,

$$\mathbf{P} = \begin{pmatrix} 1 & 0 & 0 \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix},$$

then

$$\begin{aligned} s_2 &= \hat{\mathbf{b}}'\mathbf{H_1P'}(\mathbf{PH_1'CH_1P'})^{-1}\mathbf{PH_1'}\hat{\mathbf{b}} \\ &= \hat{\mathbf{b}}'\mathbf{H_1P'P'^{-1}}(\mathbf{H_1'CH_1})^{-1}\mathbf{P^{-1}PH_1'}\hat{\mathbf{b}} \\ &= \hat{\mathbf{b}}'\mathbf{H_1}(\mathbf{H_1'CH_1})^{-1}\mathbf{H_1'}\hat{\mathbf{b}} \\ &= s_1. \end{aligned}$$

Thus, the differences that are used in the null hypothesis only need to represent one set of linearly independent differences. All other differences can be generated by use of the appropriate elementary operator matrices.

42.2.2 Breed of Calf Effects

The hypothesis is

$$\mathbf{H}_3' = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{pmatrix},$$

$$\mathbf{H}_3'\tilde{\mathbf{b}} = \begin{pmatrix} -16.2555 \\ -53.5991 \end{pmatrix}, \mathbf{H}_3'\mathbf{C}\mathbf{H}_3 = \begin{pmatrix} .68722 & .37004 \\ .37004 & .66079 \end{pmatrix},$$

and $s_3 = 4742.0565$ with 2 degrees of freedom.

42.2.3 Summary

The tests are summarized in the table below.

Partitions of SSR for Calf PWG Data

Source	D.F.	Sum of Squares
Model	6	1,761,457.9**
Age of Dam	3	981.1065**
Breed of Calf	2	4742.0565**

$**$Significance at .05 level

43 Reduction Notation

Hypothesis testing may also be conducted using reduction notation as in the regression models. For the example data, the model and submodels can be written as

Full Model $\mathbf{y} = \mu\mathbf{1} + \mathbf{X_1a} + \mathbf{X_2b} + \mathbf{e}$ where $\mathbf{a}$ refers to the vector of four age of dam effects, and $\mathbf{b}$ refers to the vector of three breed of calf effects.

Model 1 $\mathbf{y} = \mu\mathbf{1} + \mathbf{X_2b} + \mathbf{e}$

Model 2 $\mathbf{y} = \mu\mathbf{1} + \mathbf{X_1a} + \mathbf{e}$

Model 3 $\mathbf{y} = \mu\mathbf{1} + \mathbf{e}$

The corresponding reductions for these models were obtained by constructing the ordinary LS equations for each model, solving, and multiplying the solutions times their corresponding right hand sides. The results gave

$R(\mu,\mathbf{a},\mathbf{b}) = SSR =$	1,761,457.9	6 df
$R(\mu,\mathbf{b}) =$	1,760,476.7935	3 df
$R(\mu,\mathbf{a}) =$	1,756,715.8435	4 df
$R(\mu) = SSM$	1,754,910.08	1 df

To test the age of dam differences for significance from zero,

$$\begin{aligned} s_1 &= R(\mu,\mathbf{a},\mathbf{b}) - R(\mu,\mathbf{b}) \\ &= R(\mathbf{a} \mid \mu,\mathbf{b}) \\ &= 981.1065 \end{aligned}$$

with 3 degrees of freedom (6-3).

For breed of calf differences,

$$\begin{aligned} s_3 &= R(\mu,\mathbf{a},\mathbf{b}) - R(\mu,\mathbf{a}) \\ &= R(\mathbf{b} \mid \mu,\mathbf{a}) \\ &= 4742.0565 \end{aligned}$$

with 2 (6-4) degrees of freedom.

Lastly,

$$\begin{aligned} s &= R(\mu,\mathbf{a},\mathbf{b}) - R(\mu) \\ &= R(\mathbf{a},\mathbf{b} \mid \mu) \\ &= SSR - SSM \\ &= 6,547.82 \end{aligned}$$

provides a test of the model excluding the mean with 5 (6-1) degrees of freedom.

All tests using reduction notation involve SSR in order to be equivalent to tests from the general linear hypothesis approach. All tests are for differences being not zero.

44 Connectedness

A procedure to determine connectedness was suggested by Weeks and Williams (1964). If all subclasses of the fixed effects are full, i.e. contain at least one observation, then the data are completely connected and there are no problems of estimability.

When several subclasses are empty, i.e. do not contain any observations, then some estimable functions of **b** may no longer be estimable.

Following Weeks and Williams (1964): An N-tuple is a collection of numbers that identify a particular subclass. Suppose there are three fixed factors, then a general 3-tuple that identifies a subclass is (i, j, k), where i indicates the level number of the first factor, j indicates the level number of the second factor, and k indicates the level number of the third factor. Two N-tuples are said to be *nearly identical* if all of the level indicators are the same except for one factor. The steps to determine connectedness are

1. Construct a table of all N-tuples that contain one or more observations.
2. Select any N-tuple from the table. The subclass with the largest number of observations would be a logical choice.
3. Find all N-tuples in the table that are nearly identical to the selected N-tuple, then any N-tuples that are nearly identical to these, and so on until no more nearly identical N-tuples can be found.

The resulting subset of N-tuples obtained in this manner form a connected set of subclasses. If any other N=tuples remain in the table, then repeat the procedure on the remaining N-tuples, and continue until all N-tuples have been allocated to a connected subset.

The following list of 2-tuples is used to illustrate the procedure.

(1,1)	(1,2)			(1,5)	(1,6)		
	(2,2)		(2,4)		(2,6)		
		(3,3)					(3,8)
(4,1)			(4,4)	(4,5)		(4,7)	
		(5,3)					(5,8)

Pick one of the points from the table, say (1,1). The 2-tuples that are nearly identical to (1,1) are

(1,2), (1,5), (1,6), and (4,1).

Now find the 2-tuples that are nearly identical to each of these four subclasses, and so on. The set of subclasses forms a connected subset consisting of

(1,1), (1,2), (1,5), (1,6), (4,1)
(2,2), (4,5), (2,6), (4,4), (4,7)
and (2,4).

The remaining points are (3,3), (3,8), (5,3), and (5,8) which form a second connected subset, but this subset is *disconnected* from the first.

There are a few options for handling disconnected subclasses.

1. Discard all of the small subsets and analyze only the largest group of connected subclasses.
2. Analyze each connected subset separately.
3. Collect more data for those subclasses that would improve connectedness between the disconnected groups. In the previous example, the addition of observations to subclasses (1,3), (2,3), (1,8), (2,8), and (3,5) would make all subclasses in the table connected.

The choice depends on the goals of the analysis and the costs of collecting more data versus analyzing the available data.

The procedure of Weeks and Williams (1964) applies to a completely fixed effects model without interactions or covariates. Connectedness may also be a problem in some mixed model situations and in the estimation of variance components. Computer programs for checking connectedness are not trivial for large data sets. A simple, efficient procedure for 2-tuples has been described by Fernando et al. (1983).

45 Classification Model with Interactions

Below are data on pheasants raised in two locations on three diets.

Example Data on Pheasants

Location	Diet	Males	Females
1	1	3, 6, 4	2, 5, 1
	2	1, 4, 5	1, 3
	3	2, 4	1, 3, 2
2	1	5, 6, 4	2, 2, 4
	2	2, 3, 6	3
	3	4, 6, 5	2, 3, 3

Assume the model,

$$y_{ijkm} = \mu + L_i + D_j + S_k + LD_{ij} + LS_{ik} + DS_{jk} + LDS_{ijk} + e_{ijkm}$$

where

μ is the overall mean,

L_i is a location effect,

D_j is a diet effect,

S_k is a sex effect, and

$LD_{ij}, LS_{ik}, DS_{jk}$ are the two-way interaction effects among the main factors, and

LDS_{ijk} is the three-way interaction of the main effects.

All factors are assumed to be fixed effects. The $\mathbf{X}$ matrix for this model is of order 32 by 36 with a rank of 12, which is the number of three-way interaction subclasses.

An alternative (equivalent) form of the above model is

$$y_{ijkm} = \mu_{ijk} + e_{ijkm}$$

which has a corresponding $\mathbf{X}$ matrix of order 32 by 12 with a rank of 12. Assuming that $Var(\mathbf{e}) = \mathbf{I}\sigma_e^2$, then

$$\mathbf{X}'\mathbf{X} = \text{diag}\begin{pmatrix} 3 & 3 & 3 & 2 & 2 & 3 & 3 & 3 & 3 & 1 & 3 & 3 \end{pmatrix},$$

and

$$(\mathbf{X}'\mathbf{y})' = \begin{pmatrix} 13 & 8 & 10 & 4 & 6 & 6 & 15 & 8 & 11 & 3 & 15 & 8 \end{pmatrix},$$

where subclasses are ordered as in the table.

45.1 Solution Vector

The solutions to the LS equations are the subclass means in this case. For example, the solution for location 1, diet 2, females was

$$\tilde{\mu}_{122} = 4/2 = 2,$$

with expectation equal to

$$E(\tilde{\mu}_{122}) = \mu + L_1 + D_2 + S_2 + LD_{12} + LS_{12} + DS_{22} + LDS_{122}.$$

The expectation of all subclass means has a similar format to the above example. Notice that all of the main effects and all of the interaction terms are involved. Consequently, there is no function of the $\tilde{\mu}_{ijk}$ that can be formulated that would be totally free of the interaction effects. For example,

$$\begin{aligned} E(\tilde{\mu}_{211} - \tilde{\mu}_{111}) &= \mu + L_2 + D_1 + S_1 + LD_{21} + LS_{21} + DS_{11} \\ &\quad +LDS_{211} - \mu - L_1 - D_1 - S_1 - LD_{11} \\ &\quad -LS_{11} - DS_{11} - LDS_{111} \\ &= L_2 - L_1 + LD_{21} - LD_{11} + LS_{21} - LS_{11} \\ &\quad +LDS_{211} - LDS_{111} \end{aligned}$$

The consequence is that tests of hypotheses for the main effects can not be conducted unless the tests for all interaction effects are not significantly different from zero.

The least squares mean for location 1 is

$$\begin{aligned}
&(\hat{\mu}_{111} + \hat{\mu}_{112} + \hat{\mu}_{121} + \hat{\mu}_{122} + \hat{\mu}_{131} + \hat{\mu}_{132})/6 \\
&= (\frac{13}{3} + \frac{8}{3} + \frac{10}{3} + \frac{4}{2} + \frac{6}{2} + \frac{6}{3})/6 \\
&= \frac{17.33333}{6} = 2.88889,
\end{aligned}$$

which is an estimate of

$$\begin{aligned}
\mu &+ L_1 + (D_1 + D_2 + D_3)/3 + (S_1 + S_2)/2 \\
&+ (LD_{11} + LD_{12} + LD_{13})/3 + (LS_{11} + LS_{12})/2 \\
&+ (DS_{11} + DS_{12} + DS_{21} + DS_{22} + DS_{31} + DS_{32})/6 \\
&+ (LDS_{111} + LDS_{112} + LDS_{121} + LDS_{122} + LDS_{131} + LDS_{132})/6
\end{aligned}$$

45.2 AOV Table

Example Three-Way Classification
With Interaction Model

Source	D. F.	Sum of Squares
Total	32	435.00000
Mean	1	357.78125
Model	12	391.00000
Residual	20	44.00000

The adequacy of the model is given by

$$F_M = \frac{391/12}{44/20} = 14.81,$$

which is highly significant with

$$R^2 = \frac{391. - 357.78125}{435. - 357.78125} = .43.$$

45.3 Partitioning SSR

Begin by constructing the null hypothesis matrices for the main effects using the simple μ_{ijk} model. Below are the hypothesis matrices for locations (L), diets (D), and sexes (S).

$$\mathbf{H}'_L = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & -1 & -1 & -1 & -1 & -1 & -1 \end{pmatrix},$$

$$\mathbf{H}'_D = \left(\begin{array}{cccccccccccc} 1 & 1 & -1 & -1 & 0 & 0 & 1 & 1 & -1 & -1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & -1 & 1 & 1 & 0 & 0 & -1 & -1 \end{array} \right),$$

and

$$\mathbf{H}'_S = \left(\begin{array}{cccccccccccc} 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 & 1 & -1 \end{array} \right).$$

The matrices for testing the interaction effects are obtained from the matrices for the main effects by computing the product of each row of one matrix times each row of the second matrix, element by element. The matrix for testing location by diet interaction would be formed by the product of rows of $\mathbf{H}'_L$ times $\mathbf{H}'_D$. Because the matrix for locations has one row and that for diets has two rows, the resulting matrix will have $2(= 1 \times 2)$ rows, as follows:

$$\mathbf{H}'_{LD} = \left(\begin{array}{cccccccccccc} 1 & 1 & -1 & -1 & 0 & 0 & -1 & -1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & -1 & -1 & -1 & -1 & 0 & 0 & 1 & 1 \end{array} \right).$$

Similarly, for location by sex and diet by sex interactions,

$$\mathbf{H}'_{LS} = \left(\begin{array}{cccccccccccc} 1 & -1 & 1 & -1 & 1 & -1 & -1 & 1 & -1 & 1 & -1 & 1 \end{array} \right),$$

and

$$\mathbf{H}'_{DS} = \left(\begin{array}{cccccccccccc} 1 & -1 & -1 & 1 & 0 & 0 & 1 & -1 & -1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 & -1 & 1 \end{array} \right).$$

The matrix for testing the three-way interaction can be formed in a number of ways, one of which is to multiply $\mathbf{H}'_{LD}$ times $\mathbf{H}'_S$ giving

$$\mathbf{H}'_{LDS} = \left(\begin{array}{cccccccccccc} 1 & -1 & -1 & 1 & 0 & 0 & -1 & 1 & 1 & -1 & 0 & 0 \\ 1 & -1 & 0 & 0 & -1 & 1 & -1 & 1 & 0 & 0 & -1 & 1 \end{array} \right).$$

Partitions of SSR

Three-way model with interaction example.

Source	D.F.	Sum of Squares	F-values
L,locations	1	4.3556	1.98
D,diets	2	2.4482	0.56
S,sexes	1	17.4222	7.92*
LD	2	1.4380	0.33
LS	1	0.3556	0.16
DS	2	1.1468	0.26
LDS	2	1.1013	0.25

None of the interaction effects were significantly different from zero. Tests of the main effects may be interpretted ignoring the interactions.

If the three-way interaction was significant, then all of the other tests would not be needed because they could not be interpretted without reference to the three-way interactions.

45.4 Missing Subclasses

If one or more subclasses had been empty, i.e. no observations present, then the formulation of the hypothesis test matrices and the interpretation of results becomes more complicated, and in some cases tests may not be possible at all.

As the number of missing subclasses increases, one must ”customize” the contrasts to be those of the most interest. Another possibility is to remove data until all remaining subclasses are full. This could have the effect of eliminating one or more levels of several factors which may not be desirable. If too many subclasses are missing, then a model that ignores interactions could be analyzed. Interpretation of results would need to consider the possibility that interactions might exist.

46 Classification Model with Covariates

Consider a nutrition trial of dairy heifers designed to study growth and development of the mammary gland relative to the plane of nutrition. The amount of adipose tissue in the mammary gland is one of the traits of interest which is dependent upon the amount of growth hormone in the blood as well as the diet. A model might be as follows:

$$y_{ij} = \mu + T_i + b(X_{ij}) + e_{ij}$$

where

y_{ij} is the amount of adipose tissue,

μ is the overall mean,

T_i is the effect of diet, $i = 1, 2, 3$,

b is the regression coefficient of y_{ij} on X_{ij}, and

e_{ij} is a residual effect.

Assume that $Var(e_{ij}) = \sigma_e^2$, and that all residual effects are uncorrelated. The levels of growth hormone in the blood are assumed to be unaffected by the diets in the study. If there were an effect, then the inclusion of growth hormone levels as a covariate could reduce the effects of diets in the analysis. A way to counteract this problem would be to estimate a separate regression coefficient for each diet.

Example data on dairy heifers.

Heifer Number	Diet	Adipose Tissue	Growth Hormone
1	1	147	55
2	1	150	47
3	1	145	56
4	1	166	50
5	2	114	53
6	2	140	48
7	2	105	49
8	2	130	54
9	2	133	58
10	3	101	42
11	3	97	45
12	3	112	51
13	3	90	57

46.1 Equations and Solution Vector

$$\left(\begin{array}{rrrrr} 13 & 4 & 5 & 4 & 665 \\ 4 & 4 & 0 & 0 & 208 \\ 5 & 0 & 5 & 0 & 262 \\ 4 & 0 & 0 & 4 & 195 \\ 665 & 208 & 262 & 195 & 34303 \end{array}\right) \left(\begin{array}{c} \hat{\mu} \\ \hat{T}_1 \\ \hat{T}_2 \\ \hat{T}_3 \\ \hat{b} \end{array}\right) = \left(\begin{array}{r} 1,630 \\ 608 \\ 622 \\ 400 \\ 83,645 \end{array}\right)$$

with solutions

$$\left(\begin{array}{c} \hat{\mu} \\ \hat{T}_1 \\ \hat{T}_2 \\ \hat{T}_3 \\ \hat{b} \end{array}\right) = \left(\begin{array}{r} 0.0000 \\ 165.1677 \\ 137.6690 \\ 112.3447 \\ -.2532 \end{array}\right).$$

The expectation of this particular solution vector is $E(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{X}\mathbf{b}$ or

$$E \left(\begin{array}{c} \hat{\mu} \\ \hat{T}_1 \\ \hat{T}_2 \\ \hat{T}_3 \\ \hat{b} \end{array}\right) = \left(\begin{array}{ccccc} 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right) \left(\begin{array}{c} \mu \\ T_1 \\ T_2 \\ T_3 \\ b \end{array}\right).$$

46.2 Least Squares Means

The least squares means for diets include the covariate as follows:

Diet 1

$$\hat{\mu} + \hat{T}_1 + \hat{b}\overline{X}$$

$$= 0 + 165.1677 + (-.2532)(51.1538)$$

$$= 152.2155$$

where $\overline{X}$ is the average value of the covariate (growth hormone) in the data.

Diet 2

$$\hat{\mu} + \hat{T}_2 + \hat{b}\overline{X} = 124.7168.$$

Diet 3

$$\hat{\mu} + \hat{T}_3 + \hat{b}\overline{X} = 99.3925.$$

46.3 Analysis of Variance

AOV Table for Example
Data on Dairy Heifers.

Source	D.F.	Sum of Squares	F-value
Total,SST	13	211,154.0000	
Mean,SSM	1	204,376.9231	
Model,SSR	4	209,808.9557	350.97**
Diets	2	5,187.6512	17.36**
Regression	1	16.1557	0.11
Residual,SSE	9	1,345.0443	

The R^2 for this model was 80.15%. The regression coefficient was not significantly different from zero. However, diets were highly significant.

47 EXERCISES

1. **Rank of the Model**

   Determine the number of columns in $\mathbf{X}$ and the rank of $\mathbf{X}$ for the following model equations where all factors are fixed factors:

   (a) $y_{ijkl} = \mu + A_i + B_j + C_k + e_{ijkl}$ where $i = 1,5,\ j = 1,4,\ k = 1,8$.

   (b) $y_{ijkl} = \mu + A_i + B_j + C_k + (AC)_{ik} + e_{ijkl}$ where $i = 1,5,\ j = 1,3,\ k = 1,6$.

   (c) $y_{ijklm} = \mu + (AB)_{ij} + (CD)_{kl} + (BC)_{jk} + e_{ijklm}$ where $i = 1,3,\ j = 1,4,\ k = 1,6,$ $l = 1,5$.

   (d) $y_{ijklm} = A_i + B_{ij} + C_{ijk} + D_l + e_{ijklm}$ where $i = 1,2,\ j = 1,3,\ k = 1,4,$ and $l = 1,9$.

   (e) $y_{ijkl} = \mu + A_i + BC(jk) + b_1 x_{1ijkl} + b_2 x_{2ijkl} + e_{ijkl}$ where $i = 1,6,\ j = 1,2,$ and $k = 1,4$.

   (f) $y_{ijkl} = \mu + T_i + H_j + HY_{jk} + e_{ijkl}$ where $i = 1,10,\ j = 1,20,$ and $k = 1,15$.

2. **Disconnectedness**

   Determine the connected groups within each of the following sets of filled subclass numbers

(a)

(1,1,1)	(2,2,2)	(3,3,4)	(4,3,1)
(1,3,2)	(2,1,4)	(3,1,1)	(4,4,1)
(1,3,1)	(2,1,1)	(3,2,4)	(4,2,3)
(1,1,4)	(2,4,1)	(3,3,3)	(4,4,4)
(1,2,3)	(2,3,1)	(3,1,4)	(4,3,2)
(1,4,1)	(2,2,3)	(3,3,1)	(4,4,3)
(1,2,1)	(2,3,2)	(3,4,2)	(4,1,1)

(b)

(A,1,6)	(A,2,5)	(B,3,4)	(B,1,5)	(C,2,6)
(C,3,4)	(D,2,1)	(D,1,6)	(A,3,3)	(B,2,6)
(C,1,5)	(D,3,1)	(A,2,4)	(B,1,6)	(C,1,4)
(D,1,1)	(A,3,1)	(B,4,4)	(C,3,3)	(D,4,2)

(c)

(1,1,1,1)	(1,2,1,2)	(2,1,3,1)	(2,3,3,1)	(3,2,2,1)
(1,3,3,3)	(1,2,1,3)	(3,2,1,2)	(1,3,3,1)	(2,3,1,1)
(2,2,3,1)	(2,2,2,2)	(1,2,3,1)	(2,1,2,2)	(2,2,3,1)
(3,2,3,1)	(2,3,1,3)	(1,3,2,1)	(3,3,3,2)	(1,1,1,2)

3. **Classification Model**

(a) Guinea pigs were given different types of water, and small amounts of one of three varieties of nuts (as a snack). They were housed in separate cages within one of four rooms that were kept at different temperatures. Below are records on the amount of regular feed eaten per day. The factors in the model are Room (R), Water Type (W)(either tap water or Perrier), and Nut Variety (V) (either almonds, walnuts, or cashews). Nuts and water are offered through the day - equal amounts to each guinea pig.

The objective is to estimate and test differences in regular feed intake (g) due to room, water type, and nut variety.

Guinea Pig Data on Daily Feed Intake

Room	Water	Nuts	Obs.	Room	Water	Nuts	Obs.
1	1	1	117	2	2	2	178
1	2	2	110	2	2	3	82
1	1	3	71	2	1	2	41
1	2	2	170	2	1	3	91
1	1	3	150	2	1	1	116
1	2	3	130	2	2	1	61
1	1	3	64	2	2	2	115
1	2	2	89	2	1	2	87
1	1	1	141				
3	2	2	48	4	2	3	59
3	2	1	75	4	1	1	90
3	1	3	139	4	2	2	138
3	1	3	55	4	2	3	77
3	2	2	53	4	1	1	129
3	2	2	84	4	2	1	82
3	1	3	62	4	2	3	45
3	1	1	125	4	1	1	67
3	2	1	103	4	2	3	180
				4	2	2	123

Let the model equation be:

$$y_{ijkl} = R_i \; + \; W_j \; + \; V_k \; + \; e_{ijkl}$$

Assume the residual variance is constant for all subclasses (i.e. $\mathbf{I}\sigma_e^2$).

i. Construct the OLS equations and obtain a solution.

ii. Test for differences among rooms.

iii. Test for differences among water types.

iv. Test for difference among nut varieties.

v. Write a complete AOV table, and compute the multiple R^2 coefficient.

vi. Write a “Conclusions” paragraph about this analysis.

(b) Below are scores of warmblood mares from Inspection Tests given in three locations, (L_j). Age of the horse at test, (A_i), is another important variable.

Mare	Age	Location	Score
1	2	Alberta	71
2	3	Alberta	83
3	4	Alberta	92
4	2	Alberta	74
5	3	Sask.	68
6	4	Sask.	78
7	2	Sask.	80
8	3	Sask.	94
9	4	Man.	89
10	2	Man.	77
11	3	Man.	85
12	4	Man.	86

Let the model equation be

$$y_{ijk} = \mu \ + \ A_i \ + \ L_j \ + \ e_{ijk}.$$

The total sum of squares is 80,285.

i. Set up $\mathbf{X}'\mathbf{X}$ and $\mathbf{X}'\mathbf{y}$.

ii. Assume that the solution vector was as follows:

$$\begin{pmatrix} \hat{\mu} \\ \hat{A}_1 \\ \hat{A}_2 \\ \hat{A}_3 \\ \hat{L}_{AL} \\ \hat{L}_{SK} \\ \hat{L}_{MB} \end{pmatrix} = \begin{pmatrix} 0.0 \\ 0.0 \\ 5.4 \\ 9.3 \\ 71.3 \\ 72.4 \\ 75.5 \end{pmatrix},$$

Construct the basic AOV table.

iii. Calculate R^2.

iv. Write the null hypothesis, $(\mathbf{H}_0')$, for testing the effect of location.

v. Calculate the Least Squares Mean for L_{AL}.

(c) Every time that I go to the cottage I have to clear the cottage of mice. Every mouse that is killed, I measure its tail length and record its sex. Below is a table of my findings over the last four years. Numbers in parentheses are the number of mice, and the first number is the sum of the tail lengths of those mice.

Sex	2004	2005	2006	2007	Row Totals
Females	54(12)	42(9)	51(10)	47(9)	194(40)
Males	34(8)	50(11)	46(10)	41(11)	171(40)
Column Totals	88(20)	92(20)	97(20)	88(20)	365(80)

Construct Ordinary Least Squares equations from the numbers in the table for a model with sex and year fixed effects.

(d) Below are quantities calculated for an analysis of data on squirrels. Use these quantities to complete the Analysis of Variance Table below.

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= 54,079 \\
N &= 22 \\
\hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} &= 43,266.84 \\
r(\mathbf{X}) &= 8 \\
\mathbf{y}'\mathbf{1}\mathbf{1}'\mathbf{y}/N &= 32,417.26
\end{aligned}$$

Analysis of Variance Table

Source	degrees of freedom	Sum of Squares	Mean Squares	F-value
Total				
Model				
Residual				

What is the R^2 value for this analysis?

(e) We know that

$$E(\hat{\mathbf{b}}) = \mathbf{C}\mathbf{X}'\mathbf{X}\mathbf{b},$$

for $\mathbf{C} = (\mathbf{X}'\mathbf{X})^{-}$, and for a given analysis it was found that

$$\mathbf{C}\mathbf{X}'\mathbf{X} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

Determine if $\mathbf{K}'\mathbf{b}$ is estimable if

$$\mathbf{K}' = \begin{pmatrix} 0 & 1 & -1 & -2 & 1 & 1 & 0 & 0 \\ 2 & 0 & 2 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

(f) Please use the data in Table AA. The first number is the number of observations, and the second number in parentheses is the SUM of the observations in that subclass.

Table AA

Farm	Breeds			Totals
	1	2	3	
1	5(105)	5(155)	30(780)	40(1040)
2	5(45)	15(285)	10(140)	30(470)
Totals	10(150)	20(440)	40(920)	70(1510)

i. Construct the OLS equations (order 6) for the model

$$y_{ijk} = (BF)_{ij} \ + \ e_{ijk},$$

where both factors are fixed.

ii. Assume the total sum of squares is 37842, and a solution vector is equal to the subclass means (and $\mu = 0$), and its expectation are

$$\hat{\mathbf{b}} = \begin{pmatrix} 21 \\ 9 \\ 31 \\ 19 \\ 26 \\ 14 \end{pmatrix} = \begin{pmatrix} BF_{11} \\ BF_{12} \\ BF_{21} \\ BF_{22} \\ BF_{31} \\ BF_{32} \end{pmatrix}.$$

Construct the basic ANOVA table from this information.

iii. Construct the null hypothesis matrices for the two main effects (breed and farm), and the one for the interaction between breed and farm.

iv. What are the R^2 and estimate of residual variance for this analysis?

Prediction Theory

48 Introduction

Best Linear Unbiased Prediction (BLUP) was developed for animal breeding by Dr. Charles Roy Henderson around 1949. The methods were first applied to genetic evaluation of dairy sires in the northeastern United States in 1970. BLUP is used in nearly all countries and all livestock species for genetic evaluation of individual animals.

DEFINITION: Prediction is the estimation of the realized value of a random variable (from data) that has been sampled from a population with a known variance-covariance structure.

49 General Linear Mixed Model

49.1 Equation of the Model

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

where

$\mathbf{y}$ is an $N \times 1$ vector of observations,

$\mathbf{b}$ is a $p \times 1$ vector of unknown constants,

$\mathbf{u}$ is a $q \times 1$ vector of unknown effects of random variables,

$\mathbf{e}$ is an $N \times 1$ vector of unknown residual effects,

$\mathbf{X}, \mathbf{Z}$ are known matrices of order $N \times p$ and $N \times q$ respectively, that relate elements of $\mathbf{b}$ and $\mathbf{u}$ to elements of $\mathbf{y}$.

The elements of $\mathbf{b}$ are considered to be fixed effects while the elements of $\mathbf{u}$ are the random effects from populations of random effects with known variance-covariance structures. Both $\mathbf{b}$ and $\mathbf{u}$ may be partitioned into one or more factors depending on the situation.

49.2 Expectations and Covariance Structure

The expectations of the random variables are

$$\begin{aligned} E(\mathbf{u}) &= \mathbf{0} \\ E(\mathbf{e}) &= \mathbf{0} \\ E(\mathbf{y}) &= E(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \\ &= E(\mathbf{Xb}) + E(\mathbf{Zu}) + E(\mathbf{e}) \\ &= \mathbf{X}E(\mathbf{b}) + \mathbf{Z}E(\mathbf{u}) + E(\mathbf{e}) \\ &= \mathbf{Xb} + \mathbf{Z0} + \mathbf{0} \\ &= \mathbf{Xb} \end{aligned}$$

and the variance-covariance structure is typically represented as

$$V \begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{R} \end{pmatrix},$$

Both G and R known
(not estimating var comps)

where $\mathbf{G}$ and $\mathbf{R}$ are known, positive definite matrices. Consequently,

$$\begin{aligned} Var(\mathbf{y}) &= Var(\mathbf{Xb} + \mathbf{Zu} + \mathbf{e}) \\ &= Var(\mathbf{Zu} + \mathbf{e}) \\ &= \mathbf{Z}Var(\mathbf{u})\mathbf{Z}' + Var(\mathbf{e}) + \mathbf{Z}Cov(\mathbf{u}, \mathbf{e}) + Cov(\mathbf{e}, \mathbf{u})\mathbf{Z}' \\ &= \mathbf{ZGZ}' + \mathbf{R}, \text{ and} \\ Cov(\mathbf{y}, \mathbf{u}) &= \mathbf{ZG} \\ Cov(\mathbf{y}, \mathbf{e}) &= \mathbf{R} \end{aligned}$$

TRUE values, not estimated

If $\mathbf{u}$ is partitioned into s factors as

$$\mathbf{u}' = \begin{pmatrix} \mathbf{u}_1' & \mathbf{u}_2' & \cdots & \mathbf{u}_s' \end{pmatrix},$$

then

$$Var(\mathbf{u}) = Var \begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_s \end{pmatrix} = \begin{pmatrix} \mathbf{G}_{11} & \mathbf{G}_{12} & \cdots & \mathbf{G}_{1s} \\ \mathbf{G}_{12}' & \mathbf{G}_{22} & \cdots & \mathbf{G}_{2s} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{G}_{1s}' & \mathbf{G}_{2s}' & \cdots & \mathbf{G}_{ss} \end{pmatrix}.$$

Each $\mathbf{G}_{ij}$ is assumed to be known.

50 Predictors

The problem is to predict the function

$$\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u},$$

provided that $\mathbf{K}'\mathbf{b}$ is an estimable function.

50.1 Best Predictor

The **best predictor**, for any type of model, requires knowledge of the distribution of the random variables as well as the moments of that distribution. Then, the best predictor is the conditional mean of the predictor given the data vector, i.e.

$$E(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}|\mathbf{y})$$

which is unbiased and has the smallest mean squared error of all predictors (Cochran 1951). The computational form of the predictor depends on the distribution of $\mathbf{y}$. The computational form could be linear or nonlinear. The word **best** means that the predictor has the smallest mean squared error of all predictors of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$.

50.2 Best Linear Predictor

The best predictor may be linear OR nonlinear. Nonlinear predictors are often difficult to manipulate or to derive a feasible solution. The predictor could be restricted to class of linear functions of $\mathbf{y}$. Then, the distributional form of $\mathbf{y}$ does not need to be known, and only the first (means) and second (variances) moments of $\mathbf{y}$ must be known. If the first moment is $\mathbf{Xb}$ and the second moment is $Var(\mathbf{y}) = \mathbf{V}$, then the **best linear predictor** is

$$E(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}) = \mathbf{K}'\mathbf{b} + \mathbf{C}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})$$

where

$$\mathbf{C}' = Cov(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}, \mathbf{y}).$$

When $\mathbf{y}$ has a multivariate normal distribution, then the best linear predictor (BLP) is the same as the best predictor (BP). The BLP has the smallest mean squared error of all **linear** predictors of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$.

50.3 Best Linear Unbiased Predictor

In general, the first moment of $\mathbf{y}$, namely $\mathbf{Xb}$, is not known, but $\mathbf{V}$, the second moment, is commonly assumed to be known. Then predictors can be restricted further to those that are linear and also unbiased. The **best linear unbiased predictor** is

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

where

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

and $\mathbf{C}$ and $\mathbf{V}$ are as before.

This predictor is the same as the BLP except that $\hat{\mathbf{b}}$ has replaced $\mathbf{b}$ in the formula. Note that $\hat{\mathbf{b}}$ is the GLS estimate of $\mathbf{b}$. Of all *linear, unbiased* predictors, BLUP has the smallest mean squared error. However, if $\mathbf{y}$ is not normally distributed, then nonlinear predictors of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$ could potentially exist that have smaller mean squared error than BLUP.

51 Derivation of BLUP

51.1 Predictand and Predictor

DEFINITION: The predictand is the function to be predicted, in this case

$$\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}.$$

DEFINITION: The predictor is the function to predict the *predictand*, a linear function of $\mathbf{y}$, i.e. $\mathbf{L}'\mathbf{y}$, for some $\mathbf{L}$.

51.2 Requiring Unbiasedness

Equate the expectations of the predictor and the predictand to determine what needs to be true in order for unbiasedness to hold. That is,

$$\begin{aligned} E(\mathbf{L}'\mathbf{y}) &= \mathbf{L}'\mathbf{X}\mathbf{b} \\ E(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}) &= \mathbf{K}'\mathbf{b} \end{aligned}$$

then to be unbiased for all possible vectors $\mathbf{b}$,

$$\mathbf{L}'\mathbf{X} = \mathbf{K}'$$

or

$$\mathbf{L}'\mathbf{X} - \mathbf{K}' = \mathbf{0}.$$

51.3 Variance of Prediction Error

The prediction error is the difference between the predictor and the predictand. The covariance matrix of the prediction errors is

$$\begin{aligned} Var(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u} - \mathbf{L}'\mathbf{y}) &= Var(\mathbf{M}'\mathbf{u} - \mathbf{L}'\mathbf{y}) \\ &= \mathbf{M}'Var(\mathbf{u})\mathbf{M} + \mathbf{L}'Var(\mathbf{y})\mathbf{L} \\ &\quad -\mathbf{M}'Cov(\mathbf{u},\mathbf{y})\mathbf{L} - \mathbf{L}'Cov(\mathbf{y},\mathbf{u})\mathbf{M} \\ &= \mathbf{M}'\mathbf{G}\mathbf{M} + \mathbf{L}'\mathbf{V}\mathbf{L} - \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{L} - \mathbf{L}'\mathbf{Z}\mathbf{G}\mathbf{M} \\ &= Var(PE) \end{aligned}$$

51.4 Function to be Minimized

Because the predictor is required to be unbiased, then the mean squared error is equivalent to the variance of prediction error. Combine the variance of prediction error with a LaGrange Multiplier to force unbiasedness to obtain the matrix **F**, where

$$\mathbf{F} = Var(PE) + (\mathbf{L}'\mathbf{X} - \mathbf{K}')\mathbf{\Phi}.$$

Minimization of the diagonals of **F** is achieved by differentiating **F** with respect to the unknowns, **L** and $\mathbf{\Phi}$, and equating the partial derivatives to null matrices.

$$\begin{aligned} \frac{\partial \mathbf{F}}{\partial \mathbf{L}} &= 2\mathbf{VL} - 2\mathbf{ZGM} + \mathbf{X\Phi} = \mathbf{0} \\ \frac{\partial \mathbf{F}}{\partial \mathbf{\Phi}} &= \mathbf{X}'\mathbf{L} - \mathbf{K} = \mathbf{0} \end{aligned}$$

Let $\theta = .5\mathbf{\Phi}$, then the first derivative can be written as

$$\mathbf{VL} = \mathbf{ZGM} - \mathbf{X}\theta$$

then solve for **L** as

$$\begin{aligned} \mathbf{V}^{-1}\mathbf{VL} &= \mathbf{L} \\ &= \mathbf{V}^{-1}\mathbf{ZGM} - \mathbf{V}^{-1}\mathbf{X}\theta. \end{aligned}$$

Substituting the above for **L** into the second derivative, then we can solve for θ as

$$\begin{aligned} \mathbf{X}'\mathbf{L} - \mathbf{K} &= \mathbf{0} \\ \mathbf{X}'(\mathbf{V}^{-1}\mathbf{ZGM} - \mathbf{V}^{-1}\mathbf{X}\theta) - \mathbf{K} &= \mathbf{0} \\ \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\theta &= \mathbf{X}'\mathbf{V}^{-1}\mathbf{ZGM} - \mathbf{K} \\ \theta &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{ZGM} - \mathbf{K}) \end{aligned}$$

Substituting this solution for θ into the equation for **L** gives

$$\begin{aligned} \mathbf{L}' = & \ \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1} + \mathbf{K}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1} \\ & -\mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}. \end{aligned}$$

Let

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

GLS Estimate of b(hat)

then the predictor becomes

$$\mathbf{L}'\mathbf{y} = \mathbf{K}'\hat{\mathbf{b}} + \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

which is the BLUP of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$, and $\hat{\mathbf{b}}$ is a GLS solution for **b**. A special case for this predictor would be to let $\mathbf{K}' = \mathbf{0}$ and $\mathbf{M}' = \mathbf{I}$, then the predictand is $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u} \ = \ \mathbf{u}$, and

$$\mathbf{L}'\mathbf{y} = \hat{\mathbf{u}} = \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}).$$

Hence the predictor of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$ is $\begin{pmatrix} \mathbf{K}' & \mathbf{M}' \end{pmatrix}$ times the predictor of $\begin{pmatrix} \mathbf{b}' & \mathbf{u}' \end{pmatrix}'$ which is $\begin{pmatrix} \hat{\mathbf{b}}' & \hat{\mathbf{u}}' \end{pmatrix}'$.

52 Variances of Predictors

Let

$$\begin{aligned}\mathbf{P} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-} \\ \hat{\mathbf{b}} &= \mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}\end{aligned}$$

then

$$\begin{aligned}\hat{\mathbf{u}} &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}) \\ &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}\mathbf{y}\end{aligned}$$

for $\mathbf{W} = (\mathbf{I} - \mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1})$. From the results on generalized inverses of $\mathbf{X}$,

$$\mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} = \mathbf{X},$$

and therefore,

$$\begin{aligned}\mathbf{W}\mathbf{X} &= (\mathbf{I} - \mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1})\mathbf{X} \\ &= \mathbf{X} - \mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \\ &= \mathbf{X} - \mathbf{X} = \mathbf{0}.\end{aligned}$$

The variance of the predictor is,

$$\begin{aligned}Var(\hat{\mathbf{u}}) &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}(Var(\mathbf{y}))\mathbf{W}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \\ &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{W}\mathbf{V}\mathbf{W}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \\ &= \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} - \mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}.\end{aligned}$$

The covariance between $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ is

$$\begin{aligned}Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}}) &= \mathbf{P}\mathbf{X}'\mathbf{V}^{-1}Var(\mathbf{y})\mathbf{W}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \\ &= \mathbf{P}\mathbf{X}'\mathbf{W}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G} \\ &= \mathbf{0} \text{ because } \mathbf{X}'\mathbf{W} = \mathbf{0}\end{aligned}$$

Therefore, the total variance of the predictor is

$$\begin{aligned}Var(\mathbf{K}'\hat{\mathbf{b}} + \mathbf{M}'\hat{\mathbf{u}}) &= \mathbf{K}'\mathbf{P}\mathbf{K} + \mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{M} \\ &\quad -\mathbf{M}'\mathbf{G}\mathbf{Z}'\mathbf{V}^{-1}\mathbf{X}\mathbf{P}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Z}\mathbf{G}\mathbf{M}.\end{aligned}$$

53 Variance of Prediction Error

The main results are

$$\begin{aligned}Var(\hat{\mathbf{b}} - \mathbf{b}) &= Var(\hat{\mathbf{b}}) + Var(\mathbf{b}) - Cov(\hat{\mathbf{b}}, \mathbf{b}) - Cov(\mathbf{b}, \hat{\mathbf{b}}) \\ &= Var(\hat{\mathbf{b}}) \\ &= \mathbf{P}. \\ Var(\hat{\mathbf{u}} - \mathbf{u}) &= Var(\hat{\mathbf{u}}) + Var(\mathbf{u}) - Cov(\hat{\mathbf{u}}, \mathbf{u}) - Cov(\mathbf{u}, \hat{\mathbf{u}}),\end{aligned}$$

where

$$
\begin{aligned}
Cov(\hat{\mathbf{u}}, \mathbf{u}) &= \mathbf{GZ'V^{-1}W}Cov(\mathbf{y}, \mathbf{u}) \\
&= \mathbf{GZ'V^{-1}WZG} \\
&= \mathbf{GZ'(V^{-1} - V^{-1}XPX'V^{-1})ZG} \\
&= Var(\hat{\mathbf{u}})
\end{aligned}
$$

so that

$$
\begin{aligned}
Var(\hat{\mathbf{u}} - \mathbf{u}) &= Var(\hat{\mathbf{u}}) + \mathbf{G} - 2Var(\hat{\mathbf{u}}) \\
&= \mathbf{G} - Var(\hat{\mathbf{u}}).
\end{aligned}
$$

Also,

$$
\begin{aligned}
Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}} - \mathbf{u}) &= Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}}) - Cov(\hat{\mathbf{b}}, \mathbf{u}) \\
&= \mathbf{0} - \mathbf{PX'V^{-1}ZG}.
\end{aligned}
$$

54 Mixed Model Equations

The covariance matrix of $\mathbf{y}$ is $\mathbf{V}$ which is of order N. N is usually too large to allow $\mathbf{V}$ to be inverted. The BLUP predictor has the inverse of $\mathbf{V}$ in the formula, and therefore, would not be practical when N is large. Henderson(1949) developed the mixed model equations for computing BLUP of $\mathbf{u}$ and the GLS of $\mathbf{b}$. However, Henderson did not publish a proof of these properties until 1963 with the help of S. R. Searle, which was one year after Goldberger (1962).

Take the first and second partial derivatives of $\mathbf{F}$,

$$
\begin{pmatrix} \mathbf{V} & \mathbf{X} \\ \mathbf{X}' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{L} \\ \theta \end{pmatrix} = \begin{pmatrix} \mathbf{ZGM} \\ \mathbf{K} \end{pmatrix}
$$

Recall that $\mathbf{V} = Var(\mathbf{y}) = \mathbf{ZGZ'} + \mathbf{R}$, and let

$$
\mathbf{S} = \mathbf{G(Z'L - M)}
$$

which when re-arranged gives

$$
\mathbf{M} = \mathbf{Z'L} - \mathbf{G}^{-1}\mathbf{S},
$$

then the previous equations can be re-written as

$$
\begin{pmatrix} \mathbf{R} & \mathbf{X} & \mathbf{Z} \\ \mathbf{X}' & \mathbf{0} & \mathbf{0} \\ \mathbf{Z}' & \mathbf{0} & -\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{L} \\ \theta \\ \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathbf{K} \\ \mathbf{M} \end{pmatrix}.
$$

Take the first row of these equations and solve for **L**, then substitute the solution for **L** into the other two equations.

$$\mathbf{L} \ = \ -\mathbf{R}^{-1}\mathbf{X}\theta - \mathbf{R}^{-1}\mathbf{Z}\mathbf{S}$$

and

$$-\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \theta \\ \mathbf{S} \end{pmatrix} = \begin{pmatrix} \mathbf{K} \\ \mathbf{M} \end{pmatrix}.$$

Let a solution to these equations be obtained by computing a generalized inverse of

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix}$$

denoted as

$$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xz} \\ \mathbf{C}_{zx} & \mathbf{C}_{zz} \end{pmatrix},$$

then the solutions are

$$\begin{pmatrix} \theta \\ \mathbf{S} \end{pmatrix} = -\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xz} \\ \mathbf{C}_{zx} & \mathbf{C}_{zz} \end{pmatrix} \begin{pmatrix} \mathbf{K} \\ \mathbf{M} \end{pmatrix}.$$

Therefore, the predictor is

$$\begin{aligned} \mathbf{L}'\mathbf{y} &= \begin{pmatrix} \mathbf{K}' & \mathbf{M}' \end{pmatrix} \begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xz} \\ \mathbf{C}_{zx} & \mathbf{C}_{zz} \end{pmatrix} \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix} \\ &= \begin{pmatrix} \mathbf{K}' & \mathbf{M}' \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix}, \end{aligned}$$

where $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ are solutions to

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}+\mathbf{G}^{-1} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

The equations are known as *Henderson's Mixed Model Equations* or MME. The equations are of order equal to the number of elements in **b** and **u**, which is usually much less than the number of elements in **y**, and therefore, are more practical to solve. Also, these equations require the inverse of **R** rather than **V**, both of which are of the same order, but **R** is usually diagonal or has a more simple structure than **V**. Also, the inverse of **G** is needed, which is of order equal to the number of elements in **u**. The ability to compute the inverse of **G** depends on the model and the definition of **u**.

The MME are a useful computing algorithm for obtaining BLUP of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$. Please keep in mind that BLUP is a statistical procedure such that if the conditions for BLUP are met, then the predictor has the smallest mean squared error of all linear, unbiased predictors. The conditions are that the model is the true model and the variance-covariance matrices of the random variables are known without error.

In the strictest sense, all models approximate an unknown true model, and the variance-covariance parameters are usually guessed, so that there is never a truly BLUP analysis of data, except possibly in simulation studies.

55 Equivalence Proofs

The equivalence of the BLUP predictor to the solution from the MME was published by Henderson in 1963. In 1961 Henderson was in New Zealand (on sabbatical leave) visiting Shayle Searle learning matrix algebra and trying to derive the proofs in this section. Henderson needed to prove that

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1}$$

where

$$\mathbf{T} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}$$

and

$$\mathbf{V} = \mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R}.$$

Henderson says he took his coffee break one day and left the problem on Searle's desk, and when he returned from his coffee break the proof was on his desk.

$$\begin{aligned}
\mathbf{V}\mathbf{V}^{-1} &= (\mathbf{Z}\mathbf{G}\mathbf{Z}' + \mathbf{R})(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1}) \\
&= \mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1} + \mathbf{I} - \mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&\quad -\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&= \mathbf{I} + (\mathbf{Z}\mathbf{G}\mathbf{T}^{-1} - \mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}) \\
&\quad \mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&= \mathbf{I} + (\mathbf{Z}\mathbf{G}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}) \\
&\quad -\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z})\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&= \mathbf{I} + (\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{Z} \\
&\quad -\mathbf{Z}\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z})\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&= \mathbf{I} + (\mathbf{0})\mathbf{T}\mathbf{Z}'\mathbf{R}^{-1} \\
&= \mathbf{I}.
\end{aligned}$$

Now take the equation for $\hat{\mathbf{u}}$ from the MME

$$\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X}\hat{\mathbf{b}} + (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y}$$

which can be re-arranged as

$$(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})\hat{\mathbf{u}} = \mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})$$

or

$$\hat{\mathbf{u}} = \mathbf{T}\mathbf{Z}'\mathbf{R}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}).$$

The BLUP formula was

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

Then

$$\begin{aligned}
\mathbf{GZ'V^{-1}} &= \mathbf{GZ'(R^{-1} - R^{-1}ZTZ'R^{-1})} \\
&= \mathbf{(GZ'R^{-1} - GZ'R^{-1}ZTZ'R^{-1})} \\
&= \mathbf{(GT^{-1} - GZ'R^{-1}Z)TZ'R^{-1}} \\
&= \mathbf{(G(Z'R^{-1}Z + G^{-1}) - GZ'R^{-1}Z)TZ'R^{-1}} \\
&= \mathbf{TZ'R^{-1}}.
\end{aligned}$$

Similarly, the MME solution for $\hat{\mathbf{u}}$ and substituting it into the first equation in the MME gives

$$\mathbf{X'R^{-1}X\hat{b} + X'R^{-1}Z(TZ'R^{-1}(y - X\hat{b}))} = \mathbf{X'R^{-1}y}.$$

Combine the terms in $\hat{\mathbf{b}}$ and $\mathbf{y}$ to give

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

which are the same as the GLS equations,

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

Goldberger (1962) published these results before Henderson (1963), but Henderson knew of these equivalences back in 1949 through numerical examples. After he discovered Goldberger's paper (sometime after his retirement) Henderson insisted on citing it along with his work. Most people in animal breeding, however, refer to Henderson as the originator of this work and its primary proponent.

56 Variances of Predictors and Prediction Errors From MME

The covariance matrices of the predictors and prediction errors can be expressed in terms of the generalized inverse of the coefficient matrix of the MME, $\mathbf{C}$. Recall that

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{C_{xx}} & \mathbf{C_{xz}} \\ \mathbf{C_{zx}} & \mathbf{C_{zz}} \end{pmatrix} \begin{pmatrix} \mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}} \end{pmatrix} \mathbf{y},$$

or as

$$\hat{\mathbf{b}} = \mathbf{C'_b y},$$

and

$$\hat{\mathbf{u}} = \mathbf{C'_u y}.$$

If the coefficient matrix of the MME is full rank (or a full rank subset) (to simplify the presentation of results), then

$$\left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \\ \mathbf{C_{zx}} & \mathbf{C_{zz}} \end{array}\right)\left(\begin{array}{cc} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z+G^{-1}} \end{array}\right) = \left(\begin{array}{cc} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{array}\right),$$

which gives the result that

$$\begin{array}{c} \left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \\ \mathbf{C_{zx}} & \mathbf{C_{zz}} \end{array}\right)\left(\begin{array}{cc} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} \end{array}\right) \\ = \left(\begin{array}{cc} \mathbf{I} & -\mathbf{C_{xz}G^{-1}} \\ \mathbf{0} & \mathbf{I}-\mathbf{C_{zz}G^{-1}} \end{array}\right). \end{array}$$

This last result is used over and over in deriving the remaining results.

Now,

$$\begin{aligned} Var(\hat{\mathbf{b}}) &= Var(\mathbf{C_b'y}) \\ &= \mathbf{C_b'}Var(\mathbf{y})\mathbf{C_b} \\ &= \mathbf{C_b'(ZGZ'+R)C_b} \\ &= \left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \end{array}\right)\left(\begin{array}{c} \mathbf{X'R^{-1}} \\ \mathbf{Z'R^{-1}} \end{array}\right)\mathbf{(ZGZ'+R)C_b} \\ &= \left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \end{array}\right)\left(\begin{array}{c} \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}Z} \end{array}\right)\mathbf{GZ'C_b} \\ &\quad +\left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \end{array}\right)\left(\begin{array}{c} \mathbf{X'} \\ \mathbf{Z'} \end{array}\right)\mathbf{C_b} \\ &= -\mathbf{C_{xz}G^{-1}GZ'C_b} \\ &\quad +\left(\begin{array}{cc} \mathbf{C_{xx}} & \mathbf{C_{xz}} \end{array}\right)\left(\begin{array}{cc} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} \end{array}\right)\left(\begin{array}{c} \mathbf{C_{xx}} \\ \mathbf{C_{zx}} \end{array}\right) \\ &= \mathbf{C_{xz}G^{-1}C_{zx}} \\ &\quad +\left(\begin{array}{cc} \mathbf{I} & -\mathbf{C_{xz}G^{-1}} \end{array}\right)\left(\begin{array}{c} \mathbf{C_{xx}} \\ \mathbf{C_{zx}} \end{array}\right) \\ &= \mathbf{C_{xz}G^{-1}C_{zx}+C_{xx}-C_{xz}G^{-1}C_{zx}} \\ &= \mathbf{C_{xx}}. \end{aligned}$$

The remaining results are derived in a similar manner. These give

$$\begin{aligned} Var(\hat{\mathbf{u}}) &= \mathbf{C_u}'Var(\mathbf{y})\mathbf{C_u} \\ &= \mathbf{G}-\mathbf{C_{zz}} \end{aligned}$$

$$Cov(\hat{\mathbf{b}},\hat{\mathbf{u}}) = \mathbf{0}$$

$$\begin{aligned} Var(\hat{\mathbf{u}}-\mathbf{u}) &= Var(\hat{\mathbf{u}}) + Var(\mathbf{u}) - Cov(\hat{\mathbf{u}},\mathbf{u}) - Cov(\mathbf{u},\hat{\mathbf{u}}) \\ &= Var(\mathbf{u}) - Var(\hat{\mathbf{u}}) \\ &= \mathbf{G} - (\mathbf{G} - \mathbf{C_{zz}}) \\ &= \mathbf{C_{zz}} \end{aligned}$$

$$\begin{aligned} Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}}-\mathbf{u}) &= Cov(\hat{\mathbf{b}}, \mathbf{u}) \\ &= \mathbf{C_{xz}} \end{aligned}$$

In matrix form, the variance-covariance matrix of the predictors is

$$Var\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{C_{xx}} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} - \mathbf{C_{zz}} \end{pmatrix},$$

and the variance-covariance matrix of prediction errors is

$$Var\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}-\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{C_{xx}} & \mathbf{C_{xz}} \\ \mathbf{C_{zx}} & \mathbf{C_{zz}} \end{pmatrix}.$$

As the number of observations in the analysis increases, two things can be noted from these results:

1. $Var(\hat{\mathbf{u}})$ increases in magnitude towards a maximum of $\mathbf{G}$, and
2. $Var(\hat{\mathbf{u}}-\mathbf{u})$ decreases in magnitude towards a minimum of $\mathbf{0}$.

57 Hypothesis Testing

When $\mathbf{G}$ and $\mathbf{R}$ are assumed known, as in BLUP, then the solutions for $\hat{\mathbf{b}}$ from the MME are BLUE and tests of hypotheses that use these solutions are best. Tests involving $\hat{\mathbf{u}}$ are unnecessary because when $\mathbf{G}$ and $\mathbf{R}$ have been assumed to be known, then the variation due to the random factors has already been assumed to be different from zero. The general linear hypothesis procedures are employed as in the fixed effects model. The null hypothesis is

$$\begin{pmatrix} \mathbf{H}'_o & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{b} \\ \mathbf{u} \end{pmatrix} = \mathbf{c}$$

or

$$\mathbf{H'_o b} = \mathbf{c},$$

where $\mathbf{H'_o b}$ must be an estimable function of $\mathbf{b}$ and $\mathbf{H'_o}$ must have full row rank. $\mathbf{H'_o b}$ is estimable if

$$\mathbf{H'_o} \begin{pmatrix} \mathbf{C_{xx}} & \mathbf{C_{xz}} \end{pmatrix} \begin{pmatrix} \mathbf{X'R^{-1}X} \\ \mathbf{Z'R^{-1}X} \end{pmatrix} = \mathbf{H'_o}.$$

The test statistic is

$$s = (\mathbf{H_o'\hat{b}} - \mathbf{c})'(\mathbf{H_o'C_{xx}H_o})^{-1}(\mathbf{H_o'\hat{b}} - \mathbf{c})$$

with $r(\mathbf{H_o'})$ degrees of freedom, and the test is

$$F = (s/r(\mathbf{H_o'}))/\hat{\sigma}_e^2,$$

where

$$\hat{\sigma}_e^2 = (\mathbf{y'R^{-1}y} - \mathbf{\hat{b}'X'R^{-1}y} - \mathbf{\hat{u}'Z'R^{-1}y})/(N - r(\mathbf{X})).$$

The degrees of freedom for F are $r(\mathbf{H_o'})$ and $(N - r(\mathbf{X}))$. Note that

$$\mathbf{y'R^{-1}y} - \mathbf{\hat{b}'X'R^{-1}y} - \mathbf{\hat{u}'Z'R^{-1}y} = \mathbf{y'V^{-1}y} - \mathbf{\hat{b}'X'V^{-1}y}.$$

If **G** and **R** are not known, then there is no best test because BLUE of **b** is not possible. Valid tests exist only under certain circumstances. If estimates of **G** and **R** are used to construct the MME, then the solution for $\mathbf{\hat{b}}$ is not BLUE and the resulting tests are only approximate.

If the estimate of **G** is considered to be inappropriate, then a test of $\mathbf{H_o'b} = \mathbf{c}$ can be constructed by treating **u** as a fixed factor, assuming that $\mathbf{H_o'b}$ is estimable in the model with **u** as fixed. That is,

$$\begin{pmatrix} \mathbf{\hat{b}} \\ \mathbf{\hat{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{Z'R^{-1}X} & \mathbf{Z'R^{-1}Z} \end{pmatrix}^{-} \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{pmatrix},$$

$$= \begin{pmatrix} \mathbf{P_{xx}} & \mathbf{P_{xz}} \\ \mathbf{P_{zx}} & \mathbf{P_{zz}} \end{pmatrix} \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{Z'R^{-1}y} \end{pmatrix},$$

and

$$\begin{aligned} \hat{\sigma}_e^2 &= (\mathbf{y'R^{-1}y} - \mathbf{\hat{b}'X'R^{-1}y} - \mathbf{\hat{u}'Z'R^{-1}y})/(N - r\begin{pmatrix} \mathbf{X} & \mathbf{Z} \end{pmatrix}), \\ s &= (\mathbf{H_o'\hat{b}} - \mathbf{c})'(\mathbf{H_o'P_{xx}H_o})^{-1}(\mathbf{H_o'\hat{b}} - \mathbf{c}), \\ F &= (s/r(\mathbf{H_o'}))/\hat{\sigma}_e^2. \end{aligned}$$

58 Restrictions on Fixed Effects

There may be functions of **b** that are known and this knowledge should be incorporated into the estimation process. For example, in beef cattle, male calves of a particular breed are known to weigh 25 kg more than female calves of the same breed at 200 days of age. By incorporating a difference of 25 kg between the sexes in an analysis then all other estimates of fixed and random effects would be changed accordingly and also their variances.

Let $\mathbf{B}'\mathbf{b} = \mathbf{d}$ be the restriction to be placed on $\mathbf{b}$, then the appropriate equations would be

$$\left(\begin{array}{lll} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{B} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} & \mathbf{0} \\ \mathbf{B}' & \mathbf{0} & \mathbf{0} \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \phi \end{array} \right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{d} \end{array} \right).$$

Because $\mathbf{B}'\mathbf{b} = \mathbf{d}$ is any general function, then there are three possible effects of this function on the estimability of $\mathbf{K}'\mathbf{b}$ in the model. The conditions on $\mathbf{B}'$ are that it

1. must have full row rank, and
2. must not have more than $r(\mathbf{X})$ rows.

58.1 B′b is an estimable function

If $\mathbf{B}'\mathbf{b}$ represents a set of estimable functions of $\mathbf{b}$ in the original model, then

1. the estimability of $\mathbf{b}$ is unchanged, and
2. the modified equations above do not have an inverse.

58.2 B′b is not an estimable function

If $\mathbf{B}'\mathbf{b}$ represents a set of non-estimable functions of $\mathbf{b}$ with $(p - r(\mathbf{X}))$ rows, where p is the number of columns of $\mathbf{X}$, then

1. $\mathbf{b}$ is estimable as if $\mathbf{X}$ was full column rank, and
2. the modified equations above have a unique inverse.

58.3 B′b is not an estimable function

If $\mathbf{B}'\mathbf{b}$ represents a set of non-estimable functions of $\mathbf{b}$ with fewer than $(p - r(\mathbf{X}))$ rows, and if we let

$$\left(\begin{array}{cc} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{array} \right) = \left(\begin{array}{cc} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{B} \\ \mathbf{B}' & \mathbf{0} \end{array} \right)^{-}$$

then $\mathbf{K}'\mathbf{b}$ is estimable if

$$\left(\begin{array}{cc} \mathbf{K}' & \mathbf{0} \end{array} \right) \left(\begin{array}{cc} \mathbf{P}_{11} & \mathbf{P}_{12} \\ \mathbf{P}_{21} & \mathbf{P}_{22} \end{array} \right) \left(\begin{array}{cc} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{B} \\ \mathbf{B}' & \mathbf{0} \end{array} \right)$$

$$= \left(\begin{array}{cc} \mathbf{K}' & \mathbf{0} \end{array} \right).$$

The modified MME do not have a unique inverse in this situation.

59 Restricted BLUP

BLUP is commonly applied to models to evaluate the genetic merit of livestock in order to make decisions on culling and breeding of animals. In these cases, an objective of selection might be to improve the performance of animals for one trait while leaving another trait unchanged. In matrix notation, we might have two functions,

$$\mathbf{K}_1'\mathbf{b} + \mathbf{M}_1'\mathbf{u} \text{ and } \mathbf{K}_2'\mathbf{b} + \mathbf{M}_2'\mathbf{u},$$

representing the vectors of elements upon which selection decisions are to be made. One technique of achieving the objective is to force the covariance between the predictor of one function with the predictand of the other function to be zero. A zero covariance would result in no correlated response in $\mathbf{K}_2'\mathbf{b} + \mathbf{M}_2'\mathbf{u}$ as a consequence of selecting on $\mathbf{L}_1'\mathbf{y}$, provided $\mathbf{y}$ has a multivariate normal distribution. The covariance matrix of concern is

$$Cov(\mathbf{L}_1'\mathbf{y}, \mathbf{K}_2'\mathbf{b} + \mathbf{M}_2'\mathbf{u}) = \mathbf{L}_1'\mathbf{Z}\mathbf{G}\mathbf{M}_2.$$

Therefore, in deriving $\mathbf{L}_1'$ we must add another LaGrange Multiplier to $\mathbf{F}$ to give

$$\mathbf{F} = Var(\mathbf{L}_1'\mathbf{y} - \mathbf{K}_1'\mathbf{b} - \mathbf{M}_1'\mathbf{u}) + (\mathbf{L}_1'\mathbf{X} - \mathbf{K}_1')\mathbf{\Phi} + \mathbf{L}_1'\mathbf{Z}\mathbf{G}\mathbf{M}_2\varphi.$$

Minimize the diagonals of $\mathbf{F}$ with respect to $\mathbf{L}_1, \mathbf{\Phi},$ and φ, and equate the partial derivatives to null matrices. The resulting modified MME would be

$$\left(\begin{array}{lll} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{G}\mathbf{M}_2 \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{G}\mathbf{M}_2 \\ \mathbf{M}_2'\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{M}_2'\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{M}_2'\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z}\mathbf{G}\mathbf{M}_2 \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \\ \hat{\mathbf{t}} \end{array} \right)$$

$$= \left(\begin{array}{l} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{M}_2'\mathbf{G}\mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{array} \right).$$

Let a generalized inverse of the coefficient matrix be

$$\left(\begin{array}{ccc} \mathbf{C}_{11} & \mathbf{C}_{12} & \mathbf{C}_{13} \\ \mathbf{C}_{21} & \mathbf{C}_{22} & \mathbf{C}_{23} \\ \mathbf{C}_{31} & \mathbf{C}_{32} & \mathbf{C}_{33} \end{array} \right),$$

then the following results may be derived:

$$Var(\hat{\mathbf{b}}) = \mathbf{C}_{11},$$

$$\begin{aligned}
Var(\hat{\mathbf{u}}-\mathbf{u}) &= \mathbf{C_{22}}, \\
Var(\hat{\mathbf{u}}) &= \mathbf{G}-\mathbf{C_{22}}, \\
Cov(\hat{\mathbf{b}},\hat{\mathbf{u}}) &= \mathbf{0}, \\
Cov(\hat{\mathbf{u}},\mathbf{M_2'u}) &= \mathbf{0}, \\
Cov(\hat{\mathbf{b}},\mathbf{M_2'u}) &= \mathbf{0}, \\
Cov(\hat{\mathbf{b}},\hat{\mathbf{u}}-\mathbf{u}) &= \mathbf{C_{12}}, \\
\mathbf{M_2'}\hat{\mathbf{u}} &= \mathbf{0}.
\end{aligned}$$

Another technique of obtaining the same result is to compute $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ in the usual manner from the MME, then derive the appropriate weights to apply, in $\mathbf{K_1}$ and $\mathbf{M_1}$, such that

$$Cov(\mathbf{K_1'}\hat{\mathbf{b}}+\mathbf{M_1'}\hat{\mathbf{u}},\mathbf{M_2'u}) = \mathbf{0},$$

for a given $\mathbf{M_2}$.

60 Singular G

By definition, variance-covariance matrices should always be nonsingular. In particular, **G** and **R** should be nonsingular because the MME utilize the inverse of these matrices to obtain BLUP. The matrix **V** must always be nonsingular, but there may be cases when either **G** or **R** may be singular.

Consider the case where **G** is singular, and therefore **G** does not have an inverse. The BLUP of **u** is unaffected since the inverse of **G** is not needed, but in the MME there is a problem. Harville (1976) and Henderson(1973) suggest pre-multiplying the last equation of the MME by **G** to give

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{GZ'R^{-1}X} & \mathbf{GZ'R^{-1}Z+I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{GZ'R^{-1}y} \end{pmatrix}.$$

A disadvantage of these equations is that the coefficient matrix is no longer symmetric, and solving the equations by Gauss-Seidel iteration may be slow to achieve convergence, if the solutions converge at all. Also, the variance-covariance matrix of prediction errors has to be obtained as follows:

$$Var\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}-\mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}Z} \\ \mathbf{GZ'R^{-1}X} & \mathbf{GZ'R^{-1}Z+I} \end{pmatrix}^{-} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix}.$$

The equations could be made symmetric as follows:

$$\begin{pmatrix} \mathbf{X'R^{-1}X} & \mathbf{X'R^{-1}ZG} \\ \mathbf{GZ'R^{-1}X} & \mathbf{GZ'R^{-1}ZG+G} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} \mathbf{X'R^{-1}y} \\ \mathbf{GZ'R^{-1}y} \end{pmatrix},$$

where

$$\hat{\mathbf{u}} = \mathbf{G}\hat{\alpha},$$

and the variance-covariance matrix of prediction errors is calculated as

$$Var\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} - \mathbf{u} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix} \mathbf{C} \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{G} \end{pmatrix},$$

where $\mathbf{C}$ represents a generalized inverse of the coefficient matrix in the symmetric set of equations.

61 Singular R

When $\mathbf{R}$ is singular, the MME can not be used to compute BLUP. However, the calculation of $\mathbf{L}'$ can still be used and the results given earlier on variances of predictors and prediction errors still holds. The disadvantage is that the inverse of $\mathbf{V}$ is needed and may be too large to solve.

Another alternative might be to partition $\mathbf{R}$ and $\mathbf{y}$ into a full rank subset and analyze that part ignoring the linearly dependent subset. However, the solutions for $\hat{\mathbf{b}}$ and $\hat{\mathbf{u}}$ may be dependent on the subsets that are chosen, unless $\mathbf{X}$ and $\mathbf{Z}$ may be partitioned in the same manner as $\mathbf{R}$.

Singular $\mathbf{R}$ matrices do not occur frequently with continuously distributed observations, but do occur with categorical data where the probabilities of observations belonging to each category must sum to one.

62 When u and e are correlated

Nearly all applications of BLUP have been conducted assuming that $Cov(\mathbf{u}, \mathbf{e}) = \mathbf{0}$, but suppose that $Cov(\mathbf{u}, \mathbf{e}) = \mathbf{T}$ so that

$$Var(\mathbf{y}) = \mathbf{ZGZ}' + \mathbf{R} + \mathbf{ZT}' + \mathbf{TZ}'.$$

A solution to this problem is to use an equivalent model where

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wu} + \epsilon$$

for

$$\mathbf{W} = \mathbf{Z} + \mathbf{TG}^{-1}$$

and

$$Var\begin{pmatrix} \mathbf{u} \\ \epsilon \end{pmatrix} = \begin{pmatrix} \mathbf{G} & \mathbf{0} \\ \mathbf{0} & \mathbf{B} \end{pmatrix}$$

where $\mathbf{B} = \mathbf{R} - \mathbf{TG}^{-1}\mathbf{T}'$, and consequently,

$$\begin{aligned} Var(\mathbf{y}) &= \mathbf{WGW}' + \mathbf{B} \\ &= (\mathbf{Z} + \mathbf{TG}^{-1})\mathbf{G}(\mathbf{Z}' + \mathbf{G}^{-1}\mathbf{T}') + (\mathbf{R} - \mathbf{TG}^{-1}\mathbf{T}') \\ &= \mathbf{ZGZ}' + \mathbf{ZT}' + \mathbf{TZ}' + \mathbf{R} \end{aligned}$$

The appropriate MME for the equivalent model are

$$\left(\begin{array}{ll} \mathbf{X}'\mathbf{B}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{B}^{-1}\mathbf{W} \\ \mathbf{W}'\mathbf{B}^{-1}\mathbf{X} & \mathbf{W}'\mathbf{B}^{-1}\mathbf{W} + \mathbf{G}^{-1} \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{array} \right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{B}^{-1}\mathbf{y} \\ \mathbf{W}'\mathbf{B}^{-1}\mathbf{y} \end{array} \right).$$

The inverse of $\mathbf{B}$ can be written as

$$\mathbf{B}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{T}(\mathbf{G} - \mathbf{T}'\mathbf{R}^{-1}\mathbf{T})^{-1}\mathbf{T}'\mathbf{R}^{-1},$$

but this form may not be readily computable.

The biggest difficulty with this type of problem is to define $\mathbf{T} = Cov(\mathbf{u}, \mathbf{e})$, and then to estimate the values that should go into $\mathbf{T}$. A model with a non-zero variance-covariance matrix between $\mathbf{u}$ and $\mathbf{e}$ can be re-parameterized into an equivalent model containing $\mathbf{u}$ and ϵ which are uncorrelated.

63 G and R Unknown

For BLUP an assumption is that $\mathbf{G}$ and $\mathbf{R}$ are known without error. In practice this assumption almost never holds. Usually the proportional relationships among parameters in these matrices (i.e. such as heritabilities and genetic correlations) are known. In some cases, however, both $\mathbf{G}$ and $\mathbf{R}$ may be unknown, then linear unbiased estimators of $\mathbf{b}$ and $\mathbf{u}$ may exist, but these may not necessarily be best.

Unbiased estimators of $\mathbf{b}$ exist even if $\mathbf{G}$ and $\mathbf{R}$ are unknown. Let $\mathbf{H}$ be any nonsingular, positive definite matrix, then

$$\mathbf{K}'\mathbf{b}^o = \mathbf{K}'(\mathbf{X}'\mathbf{H}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y} = \mathbf{K}'\mathbf{C}\mathbf{X}'\mathbf{H}^{-1}\mathbf{y}$$

represents an unbiased estimator of $\mathbf{K}'\mathbf{b}$, if estimable, and

$$Var(\mathbf{K}'\mathbf{b}^o) = \mathbf{K}'\mathbf{C}\mathbf{X}'\mathbf{H}^{-1}\mathbf{V}\mathbf{H}^{-1}\mathbf{X}\mathbf{C}\mathbf{K}.$$

This estimator is *best* when $\mathbf{H} = \mathbf{V}$. Some possible matrices for $\mathbf{H}$ are $\mathbf{I}$, diagonals of $\mathbf{V}$, diagonals of $\mathbf{R}$, or $\mathbf{R}$ itself.

The $\mathbf{u}$ part of the model has been ignored in the above. Unbiased estimators of $\mathbf{K}'\mathbf{b}$ can also be obtained from

$$\left(\begin{array}{ll} \mathbf{X}'\mathbf{H}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{H}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{H}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{H}^{-1}\mathbf{Z} \end{array} \right) \left(\begin{array}{c} \mathbf{b}^o \\ \mathbf{u}^o \end{array} \right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{H}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{H}^{-1}\mathbf{y} \end{array} \right)$$

provided that $\mathbf{K}'\mathbf{b}$ is estimable in a model with $\mathbf{u}$ assumed to be fixed. Often the inclusion of $\mathbf{u}$ as fixed changes the estimability of $\mathbf{b}$.

If $\mathbf{G}$ and $\mathbf{R}$ are replaced by estimates obtained by one of the usual variance component estimation methods, then use of those estimates in the MME yield unbiased estimators of $\mathbf{b}$ and unbiased predictors of $\mathbf{u}$, provided that $\mathbf{y}$ is normally distributed (Kackar and Harville, 1981). Today, Bayesian methods are applied using Gibbs sampling to simultaneously estimate $\mathbf{G}$ and $\mathbf{R}$, and to estimate $\mathbf{b}$ and $\mathbf{u}$.

64 Example 1

Below are data on progeny of three sires distributed in two contemporary groups. The first number is the number of progeny, and the second number in parentheses is the sum of the progeny observations.

Sire	Contemporary Group	
	1	2
A	3(11)	6(19)
B	4(16)	3(18)
C	5(14)	

64.1 Operational Models

Let

$$y_{ijk} \ = \ \mu \ + \ C_i \ + \ S_j \ + \ e_{ijk},$$

where y_{ijk} are the observations on the trait of interest of individual progeny, assumed to be one record per progeny only, μ is an overall mean, C_i is a random contemporary group effect, S_j is a random sire effect, and e_{ijk} is a random residual error term associated with each observation.

$$\begin{aligned} E(y_{ijk}) &= \mu, \\ Var(e_{ijk}) &= \sigma^2_e \\ Var(C_i) &= \sigma^2_c \ = \ \sigma^2_e/6.0 \\ Var(S_j) &= \sigma^2_s \ = \ \sigma^2_e/11.5 \end{aligned}$$

The ratio of four times the sire variance to total phenotypic variance (i.e. $(\sigma^2_c + \sigma^2_s + \sigma^2_e)$), is known as the heritability of the trait, and in this case is 0.2775. The ratio of the contemporary group variance to the total phenotypic variance is 0.1329. The important ratios are

$$\begin{aligned}\sigma_e^2/\sigma_c^2 &= 6.0 \\ \sigma_e^2/\sigma_s^2 &= 11.5\end{aligned}$$

There are a total of 21 observations, but only five filled subclasses. The individual observations are not available, only the totals for each subclass. Therefore, an equivalent model is the "means" model.

$$\bar{y}_{ij} \; = \; \mu \; + \; C_i \; + \; S_j \; + \; \bar{e}_{ij},$$

where $\bar{y}_{ij}$ is the mean of the progeny of the j^{th} sire in the i^{th} contemporary group, and $\bar{e}_{ij}$ is the mean of the residuals for the $(ij)^{th}$ subclass.

The model assumes that

- Sires were mated randomly to dams within each contemporary group.
- Each dam had only one progeny.
- Sires were not related to each other.
- Progeny were all observed at the same age (or observations are perfectly adjusted for age effects).
- The contemporary groups were independent from each other.

64.2 Mixed Model Equations

The process is to define **y**, **X**, and **Z**, and also **G** and **R**. After that, the calculations are straightforward. The "means" model will be used for this example.

64.2.1 Observations

The observation vector for the "means" model is

$$\mathbf{y} \; = \; \left(\begin{array}{c} 11/3 \\ 16/4 \\ 14/5 \\ 19/6 \\ 18/3 \end{array} \right).$$

64.2.2 Xb and Zu

$$\mathbf{Xb} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \mu.$$

The overall mean is the only column in **X** for this model.

There are two random factors and each one has its own design matrix.

$$\mathbf{Zu} = \begin{pmatrix} \mathbf{Z}_c & \mathbf{Z}_s \end{pmatrix} \begin{pmatrix} \mathbf{c} \\ \mathbf{s} \end{pmatrix},$$

where

$$\mathbf{Z}_c\mathbf{c} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix}, \ \mathbf{Z}_s\mathbf{s} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} S_A \\ S_B \\ S_C \end{pmatrix},$$

so that, together,

$$\mathbf{Zu} = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ S_A \\ S_B \\ S_C \end{pmatrix}.$$

64.2.3 G and R

The covariance matrix of the means of residuals is **R**. The variance of a mean of random variables is the variance of individual variables divided by the number of variables in the mean. Let n_{ij} equal the number of progeny in a sire by contemporary group subclass, then the variance of the subclass mean is σ_e^2/n_{ij}. Thus,

$$\mathbf{R} = \begin{pmatrix} \sigma_e^2/3 & 0 & 0 & 0 & 0 \\ 0 & \sigma_e^2/4 & 0 & 0 & 0 \\ 0 & 0 & \sigma_e^2/5 & 0 & 0 \\ 0 & 0 & 0 & \sigma_e^2/6 & 0 \\ 0 & 0 & 0 & 0 & \sigma_e^2/3 \end{pmatrix}.$$

The matrix **G** is similarly partitioned into two submatrices, one for contemporary groups and one for sires.

$$\mathbf{G} \; = \; \left(\begin{array}{cc} \mathbf{G}_c & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_s \end{array} \right),$$

where

$$\mathbf{G}_c \; = \; \left(\begin{array}{cc} \sigma^2_c & 0 \\ 0 & \sigma^2_c \end{array} \right) \; = \; \mathbf{I}\sigma^2_c \; = \; \mathbf{I}\frac{\sigma^2_e}{6.0},$$

and

$$\mathbf{G}_s \; = \; \left(\begin{array}{ccc} \sigma^2_s & 0 & 0 \\ 0 & \sigma^2_s & 0 \\ 0 & 0 & \sigma^2_s \end{array} \right) \; = \; \mathbf{I}\sigma^2_s \; = \; \mathbf{I}\frac{\sigma^2_e}{11.5}.$$

The inverses of **G** and **R** are needed for the MME.

$$\mathbf{R}^{-1} \; = \; \left(\begin{array}{ccccc} 3 & 0 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 6 & 0 \\ 0 & 0 & 0 & 0 & 3 \end{array} \right) \frac{1}{\sigma^2_e},$$

and

$$\mathbf{G}^{-1} \; = \; \left(\begin{array}{ccccc} 6 & 0 & 0 & 0 & 0 \\ 0 & 6 & 0 & 0 & 0 \\ 0 & 0 & 11.5 & 0 & 0 \\ 0 & 0 & 0 & 11.5 & 0 \\ 0 & 0 & 0 & 0 & 11.5 \end{array} \right) \frac{1}{\sigma^2_e}.$$

Because both are expressed in terms of the inverse of σ^2_e, then that constant can be ignored. The relative values between **G** and **R** are sufficient to get solutions to the MME.

64.2.4 MME and Inverse Coefficient Matrix

The left hand side of the MME (LHS) is

$$\left(\begin{array}{ll} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{array} \right),$$

and the right hand side of the MME (RHS) is

$$\left(\begin{array}{c} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{array} \right).$$

Numerically,

$$\text{LHS} = \begin{pmatrix} 21 & 12 & 9 & 9 & 7 & 5 \\ 12 & 18 & 0 & 3 & 4 & 5 \\ 9 & 0 & 15 & 6 & 3 & 0 \\ 9 & 3 & 6 & 20.5 & 0 & 0 \\ 7 & 4 & 3 & 0 & 18.5 & 0 \\ 5 & 5 & 0 & 0 & 0 & 16.5 \end{pmatrix} \begin{pmatrix} \hat{\mu} \\ \hat{C}_1 \\ \hat{C}_2 \\ \hat{S}_A \\ \hat{S}_B \\ \hat{S}_C \end{pmatrix},$$

and

$$\text{RHS} = \begin{pmatrix} 78 \\ 41 \\ 37 \\ 30 \\ 34 \\ 14 \end{pmatrix}.$$

The inverse of LHS coefficient matrix is

$$\mathbf{C} = \begin{pmatrix} .1621 & -.0895 & -.0772 & -.0355 & -.0295 & -.0220 \\ -.0895 & .1161 & .0506 & .0075 & .0006 & -.0081 \\ -.0772 & .0506 & .1161 & -.0075 & -.0006 & .0081 \\ -.0355 & .0075 & -.0075 & .0655 & .0130 & .0085 \\ -.0295 & .0006 & -.0006 & .0130 & .0652 & .0088 \\ -.0220 & -.0081 & .0081 & .0085 & .0088 & .0697 \end{pmatrix}.$$

C has some interesting properties.

- Add elements (1,2) and (1,3) = -.1667, which is the negative of the ratio of σ^2_c/σ^2_e.
- Add elements (1,4), (1,5), and (1,6) = -.08696, which is the negative of the ratio of σ^2_s/σ^2_e.
- Add elements (2,2) and (2,3), or (3,2) plus (3,3) = .1667, ratio of contemporary group variance to residual variance.
- Add elements (4,4) plus (4,5) plus (4,6) = .08696, ratio of sire variance to residual variance. Also, ((5,4)+(5,5)+(5,6) = (6,4)+(6,5)+(6,6)).
- The sum of ((4,2)+(5,2)+(6,2)) = ((4,3)+(5,3)+(6,3)) = 0.

64.2.5 Solutions and Variance of Prediction Error

Let SOL represent the vector of solutions to the MME, then

$$\text{SOL} \ = \ \mathbf{C} \ * \ \text{RHS} \ = \ \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} \ = \ \begin{pmatrix} \hat{\mu} \\ \hat{C}_1 \\ \hat{C}_2 \\ \hat{S}_A \\ \hat{S}_B \\ \hat{S}_C \end{pmatrix} \ = \ \begin{pmatrix} 3.7448 \\ -.2183 \\ .2183 \\ -.2126 \\ .4327 \\ -.2201 \end{pmatrix}.$$

The two contemporary group solutions add to zero, and the three sire solutions add to zero.

The variances of prediction error are derived from the diagonals of **C** corresponding to the random effect solutions multiplied times the residual variance. Hence, the variance of prediction error for contemporary group 1 is .1161 σ_e^2. An estimate of the residual variance is needed. An estimate of the residual variance is given by

$$\hat{\sigma}_e^2 \ = \ (SST - SSR)/(N - r(\mathbf{X})).$$

SST was not available from these data because individual observations were not available. Suppose $SST = 322$, then

$$\hat{\sigma}_e^2 \ = \ (322 - 296.4704)/(21 - 1) \ = \ 1.2765.$$

SSR is computed by multiply the solution vector times the RHS of the MME. That is,

$$SSR = 3.7448(78) - .2183(41) + .2183(37) - .2126(30) + .4327(34) - .2201(14) \ = \ 296.4704.$$

The variance of prediction error for contemporary group 1 is

$$Var(PE) \ = \ .1161(1.2765) \ = \ .1482.$$

The standard error of prediction, or SEP, is the square root of the variance of prediction error, giving .3850. Thus, the solution for contemporary group 1 is -.2183 plus or minus .3850.

Variances of prediction error are calculated in the same way for all solutions of random effects.

Effect	Solution	SEP
C_1	-.2183	.3850
C_2	.2183	.3850
S_A	-.2126	.2892
S_B	.4327	.2885
S_C	-.2201	.2983

Sire A has 9 progeny while sire B has 7 progeny, but sire B has a slightly smaller SEP. The reason is due to the distribution of progeny of each sire in the two contemporary groups. Sire C, of course, has the larger SEP because it has only 5 progeny and all of these are in contemporary group 1. The differences in SEP in this small example are not large.

64.2.6 Repeatability or Reliability

Variances of prediction error are often expressed as a number going from 0 to 100 % , known as repeatability or reliability (REL) (depending on the species). The general formula is

$$\text{REL} \ = \ (Var(\text{True Values}) - Var(PE))/(Var(\text{True Values}),$$

times 100. Thus, for contemporary group 1, the reliability would be

$$\text{REL} \ = \ 100(.1667 - .1482)/(.1667) \ = \ 11.10.$$

For Sire A, the REL would be

$$\text{REL} \ = \ 100(.08696 - (.0655 * 1.2765))/(.08696) \ = \ 3.85.$$

Thus, sires have smaller reliabilities than contemporary groups, but SEP for sires is smaller than for contemporary groups. This is because contemporary groups have more progeny in them than sires have, and because the variance of contemporary groups is larger than the variance of sire transmitting abilities.

64.3 R Methods for MME

Given the matrices $\mathbf{X}$, $\mathbf{Z}$, $\mathbf{G}^{-1}$, $\mathbf{R}^{-1}$, and $\mathbf{y}$, then an R-function can be written to set up the MME, solve, and compute SSR. This function works for small examples, as given in these notes. For large problems, other methods can be used to solve the equations by iterating on the data. The function for small examples is given here.

```
MME = function(X,Z,GI,RI,y) {
XX = t(X) %*% RI %*% X
XZ = t(X) %*% RI %*% Z
ZZ = (t(Z) %*% RI %*% Z) + GI
Xy = t(X) %*% RI %*% y
Zy = t(Z) %*% RI %*% y

# Combine the pieces into LHS and RHS
piece1 = cbind(XX,XZ)
piece2 = cbind(t(XZ),ZZ)
LHS = rbind(piece1,piece2)
RHS = rbind(Xy,Zy)

# Invert LHS and solve
C = ginv(LHS)
SOL = C %*% RHS
SSR = t(SOL) %*% RHS
SOLNS = cbind(SOL,sqrt(diag(C)))

return(list(LHS=LHS,RHS=RHS,C=C,SSR=SSR,SOLNS=SOLNS))
}
```

To use the function,

```
Exampl = MME(X1,Z1,GI,RI,y)
str(Exampl)

# To view the results
Exampl$LHS
Exampl$RHS
Exampl$C
Exampl$SOLNS
Exampl$SSR
```

65 EXERCISES

1. Below are data on progeny of 6 rams used in 5 sheep flocks (for some trait). The rams were unrelated to each other and to any of the ewes to which they were mated. The first number is the number of progeny in the herd, and the second (within parentheses) is the sum of the observations.

Ram	Flocks				
ID	1	2	3	4	5
1	6(638)	8(611)	6(546)	5(472)	0(0)
2	5(497)	5(405)	5(510)	0(0)	4(378)
3	15(1641)	6(598)	5(614)	6(639)	5(443)
4	6(871)	11(1355)	0(0)	3(412)	3(367)
5	2(235)	4(414)	8(874)	4(454)	6(830)
6	0(0)	0(0)	4(460)	12(1312)	5(558)

Let the model equation be

$$y_{ijk} = \mu + F_i + R_j + e_{ijk}$$

where F_i is a flock effect, R_j is a ram effect, and e_{ijk} is a residual effect. There are a total of 149 observations and the total sum of squares was equal to 1,793,791. Assume that $\sigma_e^2 = 7\sigma_f^2 = 1.5\sigma_r^2$ when doing the problems below.

(a) Set up the mixed model equations and solve. Calculate the SEPs and reliabilities of the ram solutions.

(b) Repeat the above analysis, but assume that flocks are a fixed factor (i.e. do not add any variance ratio to the diagonals of the flock equations). How do the evaluations, SEP, and reliabilities change from the previous model?

(c) Assume that rams are a fixed factor, and flocks are random. Do the rams rank similarly to the previous two models?

2. When a model has one random factor and its covariance matrix is an identity matrix times a scalar constant, then prove the the solutions for that factor from the MME will sum to zero. Try to make the proof as general as possible.

Genetic Relationships

66 Pedigree Preparation

Pedigrees of animals need to be arranged in chronological order. Parents should appear in a list before (ahead of) their progeny. Ordering a pedigree is most easily accomplished by sorting animals by birthdate. Birthdates can be incorrectly recorded or entered, or for many individuals may not be available. One approach is to assume that all birthdates are incorrect. Animals can be arranged by assigning generation numbers to animals, then iterate through the pedigrees modifying the generation numbers of the sire and dam to be at least one greater than the generation number of the offspring. The number of iterations depends on the number of generations of animals in the list. Probably 20 or less iterations are needed for most situations.

If the number of iterations reaches 50 or more, then there is an increased likelihood that there is a loop in the pedigrees. That means that an animal is its own ancestor, somewhere back in the pedigree. For example, A might be the parent of B, and B is the parent of C, and C is the parent of A. In this case the generation numbers will keep increasing in each iteration. Thus, if more than 50 iterations are used, then look at the animals with the highest generation numbers and try to find the loop. A loop is an error in the pedigrees and must be repaired. Either correct the parentage, or remove the parent of the older animal.

66.1 Example Pedigree to Sort

Animal	Sire	Dam	Generation Number
BF	DD	HE	1
DD	GA	EC	1
GA			1
EC	GA	FB	1
FB			1
AG	BF	EC	1
HE	DD	FB	1

All animals begin with generation number 1. Proceed through the pedigrees one animal at a time.

1. Take the current generation number of the animal and increase it by one (1), call it m. The first animal is BF, for example, and its generation number is 1, increased by 1 becomes m=2.

2. Compare m to the generation numbers of the animal's sire and dam. In the case of BF, the sire is DD and DD's generation number is 1. That is less than 2 so DD's generation number has to be changed to 2 (m). The dam is HE, and HE's generation number is also changed to 2.

Repeat for each animal in the pedigree list. Keep modifying the generation numbers until no more need to be changed. The animal with the highest generation number is the oldest animal.

The end result after four iterations of the example pedigree is shown below.

Animal	Sire	Dam	Generation Number
BF	DD	HE	2
DD	GA	EC	4
GA			6
EC	GA	FB	5
FB			6
AG	BF	EC	1
HE	DD	FB	3

Now sort the list by decreasing order of the generation number.

Animal	Sire	Dam	Generation Number
GA			6
FB			6
EC	GA	FB	5
DD	GA	EC	4
HE	DD	FB	3
BF	DD	HE	2
AG	BF	EC	1

The order of animals GA or FB is not important. The order of animals with the same generation number is not critical.

Once the pedigree is sorted, then the birthdates can be checked. Errors can be spotted more readily. Once the errors are found and corrected, then the generation numbers could be checked again. Animals should then be numbered consecutively according to the last list from 1 to the total number of animals in the list. That means that parent numbers should always be smaller than progeny ID numbers. Having animals in this order facilitates calculation of inbreeding coefficients, assignment of animals with unknown parents to groups, and utilization of the inverse of the relationship matrix in the solution of mixed model equations.

67 Genomic Relationships

Genomic relationships are constructed by identifying the genomic sources for each animal. One half of the alleles, genomic effects, are from the male parent and the other half of alleles are from the female parent. Let $\mathbf{g}$ be a vector of the genomic effects for all animals, of length $2N$ where N is the number of animals, then

$$Var(\mathbf{g}) = \mathbf{G}\sigma_g^2.$$

The genomic relationship matrix, $\mathbf{G}$, can be viewed as the average of an infinite number of gametic relationship matrices, $\mathbf{G}_i$, for the i^{th} gene. The genomic relationship matrix can be constructed using simple rules.

67.1 Example Pedigree

Parentage of five animals are given below.

Example Pedigree.

Animal	Sire	Dam
A	-	-
B	-	-
C	A	B
D	A	C
E	D	B

Expand this table to identify the genomic structure. Parent1 and Parent2 indicate the genomic sources for the male or female parents of the sire of an animal, respectively. For example, for animal C, the male source of alleles is C1, and the source of alleles for C1 comes from animal A's genes. The source of alleles for C2 is from the female parent, B.

Example Genomic Pedigree.

Animal	Genome	Parent1	Parent2
A	A1	-	-
A	A2	-	-
B	B1	-	-
B	B2	-	-
C	C1	A1	A2
C	C2	B1	B2
D	D1	A1	A2
D	D2	C1	C2
E	E1	D1	D2
E	E2	B1	B2

This genomic relationship matrix will be of order 10. The diagonals of all genomic relationship matrices are always equal to 1. The quantities in the off-diagonals of the matrix are probabilities of genes being identical by descent (an average probability across all genes).

		A		B		C		D		E	
		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
A	A1	1	0	0	0						
	A2	0	1	0	0						
B	B1	0	0	1	0						
	B2	0	0	0	1						
C	C1					1					
	C2						1				
D	D1							1			
	D2								1		
E	E1									1	
	E2										1

Because the parents of A and B are unknown, then they are assumed to be randomly drawn from a large random mating population and assumed to have no genes identical by descent between them.

Let (A1,C1) indicate an element in the above table between the A1 male parent contribution of animal A and the C1 male parent contribution of animal C, then the value that goes into that location is

```
(A1,C1) = 0.5 * [ (A1,A1) + (A1,A2) ] = 0.5.
```

Similarly, for the rest of the A1 row,

```
(A1,C2) = 0.5 * [ (A1,B1) + (A1,B2) ] = 0,
(A1,D1) = 0.5 * [ (A1,A1) + (A1,A2) ] = 0.5,
(A1,D2) = 0.5 * [ (A1,C1) + (A1,C2) ] = 0.25,
(A1,E1) = 0.5 * [ (A1,D1) + (A1,D2) ] = 0.375,
(A1,E2) = 0.5 * [ (A1,B1) + (A1,B2) ] = 0.
```

This recursive pattern follows through the entire table. Relationships should be determined row-wise, and when a row is completed, the values are transcribed down the corresponding column. Thus, if (X,Y) corresponds to the relationship between two genomic contributions, then X should always chronologically precede Y. If this is not the case, then errors in relationship calculations can result. The completed table is shown below.

		A		B		C		D		E	
		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
A	A1	1	0	0	0	.5	0	.5	.25	.375	0
	A2	0	1	0	0	.5	0	.5	.25	.375	0
B	B1	0	0	1	0	0	.5	0	.25	.125	.5
	B2	0	0	0	1	0	.5	0	.25	.125	.5
C	C1	.5	.5	0	0	1	0	.5	.5	.5	0
	C2	0	0	.5	.5	0	1	0	.5	.25	.5
D	D1	.5	.5	0	0	.5	0	1	.25	.625	0
	D2	.25	.25	.25	.25	.5	.5	.25	1	.625	.25
E	E1	.375	.375	.125	.125	.5	.25	.625	.625	1	.125
	E2	0	0	.5	.5	0	.5	0	.25	.125	1

Animals D and E are inbred and the offdiagonals between D1 and D2 and between E1 and E2 show the inbreeding coefficient.

67.2 Additive Genetic Relationships

Additive and dominance relationships may be obtained from this genomic relationship table. The **additive relationship** between animals A and C is given by

```
0.5 * [ (A1,C1) + (A1,C2) + (A2,C1) + (A2,C2) ] = 0.5.
```

Add the four numbers in each square of the table and divide by 2. Then the matrix of additive relationships is

$$\mathbf{A} = \left(\begin{array}{lllll} 1 & 0 & .5 & .75 & .375 \\ 0 & 1 & .5 & .25 & .625 \\ .5 & .5 & 1 & .75 & .625 \\ .75 & .25 & .75 & 1.25 & .75 \\ .375 & .625 & .625 & .75 & 1.125 \end{array} \right).$$

67.3 Dominance Genetic Relationships

The **dominance genetic relationship** between animals X and Y, in general, is given by

```
(X1,Y1)*(X2,Y2) + (X1,Y2)*(X2,Y1).
```

The complete dominance relationship matrix is

$$\mathbf{D} = \begin{pmatrix} 1 & 0 & 0 & .25 & 0 \\ 0 & 1 & 0 & 0 & .125 \\ 0 & 0 & 1 & .25 & .25 \\ .25 & 0 & .25 & 1.0625 & .15625 \\ 0 & .125 & .25 & .15625 & 1.015625 \end{pmatrix}.$$

68 Example Genomic Model

Assume the five animals (A through E) had records equal to 5, 7, 9, 2, and 4, respectively. The process is to define **y**, **X**, **Z**, **G**, and **R**.

$$\mathbf{y} \ = \begin{pmatrix} 50 \\ 70 \\ 90 \\ 20 \\ 40 \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix},$$

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

$$\mathbf{G} \ = \ \frac{1}{8}\begin{pmatrix} 8 & 0 & 0 & 0 & 4 & 0 & 4 & 2 & 3 & 0 \\ 0 & 8 & 0 & 0 & 4 & 0 & 4 & 2 & 3 & 0 \\ 0 & 0 & 8 & 0 & 0 & 4 & 0 & 2 & 1 & 4 \\ 0 & 0 & 0 & 8 & 0 & 4 & 0 & 2 & 1 & 4 \\ 4 & 4 & 0 & 0 & 8 & 0 & 4 & 4 & 4 & 0 \\ 0 & 0 & 4 & 4 & 0 & 8 & 0 & 4 & 2 & 4 \\ 4 & 4 & 0 & 0 & 4 & 0 & 8 & 2 & 5 & 0 \\ 2 & 2 & 2 & 2 & 4 & 4 & 2 & 8 & 5 & 2 \\ 3 & 3 & 1 & 1 & 4 & 2 & 5 & 5 & 8 & 1 \\ 0 & 0 & 4 & 4 & 0 & 4 & 0 & 2 & 1 & 8 \end{pmatrix},$$

and

$$\mathbf{R} \;=\; \mathbf{I}.$$

The variances are $\sigma^2_g = \sigma^2_e$, so that the ratio of residual to genomic variances is equal to 1.

Solving the MME for this model can be done using the function given in the notes on prediction theory.

R Statements

```
GI = ginv(G)
RI = ginv(R)
genom = MME(X,Z,GI,RI,y)
```

The solutions to the equations are

$$\left(\begin{array}{c} \hat{\mu} \\ \hat{A}1 \\ \hat{A}2 \\ \hat{B}1 \\ \hat{B}2 \\ \hat{C}1 \\ \hat{C}2 \\ \hat{D}1 \\ \hat{D}2 \\ \hat{E}1 \\ \hat{E}2 \end{array}\right) = \left(\begin{array}{r} 56.4194 \\ -3.0015 \\ -3.0015 \\ 3.0015 \\ 3.0015 \\ -.7295 \\ 5.2736 \\ -8.1364 \\ -2.8628 \\ -6.8468 \\ 1.2053 \end{array}\right).$$

The total additive genetic merit of an animal is equal to the sum of the two genomic contributions. Thus, for animal E, an estimate of the total additive genetic merit is called an *Estimated Breeding Value.*

$$\mathrm{EBV}_E \;=\; -6.8468 + 1.2053 \;=\; -5.6415.$$

Animal E received the more favourable alleles from its female parent.

69 Inverse of the Genomic Matrix

The methods of Henderson(1975), Quaas(1976), and Meuwissen and Luo (1992) were combined to find a fast way of inverting the genomic relationship matrix.

Partition the genomic relationship matrix as

$$\mathbf{G} = \mathbf{TDT}'$$

where **T** is a lower triangular matrix and **D** is a diagonal matrix. The diagonals of **D** are obtained while forming a row of **T**. Animal genomes are processed in order from oldest to youngest (i.e. parents before progeny).

For animal genomes with unknown parent genomes, the diagonals of **D** are equal to 1. Therefore, the diagonals of **D** for A1, A2, B1, and B2 are equal to 1.

Begin with C1, the parent genomes are A1 and A2. Form a table as follows:

Genome	t	D
C1	1	x
A1	.5	1
A2	.5	1

The diagonal element for (C1,C1) in **G** is equal to 1, which is equal to $\mathbf{t}'\mathbf{Dt}$, which is

$$(1)^2 x \ + \ (.5)^2(1) \ + \ (.5)^2(1) \ = \ 1,$$

which can be re-arranged and solved for x,

$$x = 1 \ - \ .25 \ - \ .25 \ = \ .5.$$

A similar table and calculations can be made for C2, D1, and E2. Thus, the diagonal elements of **D** for these genomic contributions are also equal to .5.

The table for D2 is a little longer. Start with parent genomes C1 and C2

Genome	t	D
D2	1	x
C1	.5	.5
C2	.5	.5

Now add the parent genomes of C1 and C2, as follows:

Genome	t	D
D2	1	x
C1	.5	.5
C2	.5	.5
A1	.25	1
A2	.25	1
B1	.25	1
B2	.25	1

The next step would be to add the 'parents' of A1 and A2, then B1 and B2, but these 'parents' are unknown, and so no further additions to the table are made. Now compute $\mathbf{t}'\mathbf{D}\mathbf{t}$ as

$$x + (.5)^2(.5) + (.5)^2(.5) + 4(.25)^2(1) = 1,$$

or

$$x = 1 - .125 - .125 - 4(.0625) \ = \ .5.$$

The table of E1 is more complex. The parent genomes are D1 and D2. As the animals become younger, the length of these tables can become greater, and with n generations there can be up to $2^n + 1$ elements in a table.

Genome	t	D
E1	1	x
D1	.5	.5
D2	.5	.5
A1	.25	1
A2	.25	1
C1	.25	.5
C2	.25	.5
A1	.125	1
A2	.125	1
B1	.125	1
B2	.125	1

Note that A1 and A2 appear twice in the table. Their coefficients in $\mathbf{t}$ must be added together before computing $\mathbf{t}'\mathbf{D}\mathbf{t}$. The new table, after adding coefficents is

Genome	t	D
E1	1	x
D1	.5	.5
D2	.5	.5
A1	.375	1
A2	.375	1
C1	.25	.5
C2	.25	.5
B1	.125	1
B2	.125	1

Then

$$x = 1 - 2(.5)^2(.5) - 2(.375)^2(1) - 2(.25)^2(.5) - 2(.125)^2(1) = .375.$$

The complete results for the diagonals of $\mathbf{D}$ are given in the next table.

Diagonals of **D**

Animal	Genome	Parent1	Parent2	**D**
A	A1	-	-	1
A	A2	-	-	1
B	B1	-	-	1
B	B2	-	-	1
C	C1	A1	A2	.5
C	C2	B1	B2	.5
D	D1	A1	A2	.5
D	D2	C1	C2	.5
E	E1	D1	D2	.375
E	E2	B1	B2	.5

The inverse of **G** is

$$\mathbf{G}^{-1} = \mathbf{T}^{-T}\mathbf{D}^{-1}\mathbf{T}^{-1},$$

and as Henderson (1975) discovered, the elements in $\mathbf{T}^{-1}$ are all 1's on the diagonals, and each row has a -.5 in the columns corresponding to the two parent genomes. All other elements are equal to 0. This structure leads to a simple set of rules for creating the inverse of **G**, which can be accomplished by going through the pedigrees, one genome at a time.

Let d^i be equal to one over the diagonal of **D** for the i^{th} genome, and let $p1$ and $p2$ be the parent genomes, then the contributions to the inverse of **G** from this genome would be to add the following values:

	i	$p1$	$p2$
i	d^i	$.5d^i$	$.5d^i$
$p1$	$.5d^i$	$.25d^i$	$.25d^i$
$p2$	$.5d^i$	$.25d^i$	$.25d^i$

Applying these rules, then the complete inverse is shown in the table below.

		A		B		C		D		E	
		A1	A2	B1	B2	C1	C2	D1	D2	E1	E2
A	A1	2	1	0	0	-1	0	-1	0	0	0
	A2	1	2	0	0	-1	0	-1	0	0	0
B	B1	0	0	2	1	0	-1	0	0	0	-1
	B2	0	0	1	2	0	-1	0	0	0	-1
C	C1	-1	-1	0	0	2.5	.5	0	-1	0	0
	C2	0	0	-1	-1	.5	2.5	0	-1	0	0
D	D1	-1	-1	0	0	0	0	2.6667	.6667	-1.3333	0
	D2	0	0	0	0	-1	-1	.6667	2.6667	-1.3333	0
E	E1	0	0	0	0	0	0	-1.3333	-1.3333	2.6667	0
	E2	0	0	-1	-1	0	0	0	0	0	2

70 Additive Relationship Matrix

The additive genetic relationships between animals were obtained from the genomic relationship matrix. The order of the additive genetic relationship matrix, **A**, equals the number of animals (N) in the pedigree. However, elements of **A** can be determined by the tabular method directly, and its inverse can be derived directly using the methods of Henderson (1975) and Meuwissen and Luo (1992).

Sewell Wright, in his work on genetic relationships and inbreeding, defined the relationship between two animals to be a correlation coefficient. That is, the genetic covariance between two animals divided by the square root of the product of the genetic variances of each animal. The genetic variance of an animal was equal to $(1 + F_i)\sigma_a^2$, where F_i is the inbreeding coefficient of that animal, and σ_a^2 is the population additive genetic variance. Correlations range from -1 to +1, and therefore, represented a percentage relationship between two individuals, usually positive only.

The elements of the additive relationship matrix are the numerators of Wright's correlation coefficients. Consequently, the diagonals of **A** can be as high as 2, and relationships between two individuals can be greater than 1. The **A** is a matrix that represents the relative genetic variances and covariances among individuals.

70.1 The Tabular Method

Additive genetic relationships among animals may be calculated using a recursive procedure called the Tabular Method (attributable to Henderson and perhaps to Wright before him). To begin, make a list of all animals that have observations in your data, and for each of these determine their parents (called the sire and dam). An example list is shown below.

Animal	Sire	Dam
A	-	-
B	-	-
C	-	-
D	A	B
E	A	C
F	E	D

The list should be in chronological order so that parents appear before progeny. The sire and dam of animals A, B, and C are assumed to be unknown, and consequently animals A, B, and C are assumed to be genetically unrelated. In some instances the parentage of animals may be traced for several generations, and for each animal the parentage should be traced to a common *base* generation.

Using the completed list of animals and pedigrees, form a two-way table with n rows and columns, where n is the number of animals in the list, in this case $n = 6$. Label the rows and columns with the corresponding animal identification and above each animal ID write the ID of its parents as shown below.

Tabular Method Example,
Starting Values.

	-,-	-,-	-,-	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0			
B	0	1	0			
C	0	0	1			
D						
E						
F						

For each animal whose parents were unknown a one was written on the diagonal of the table (i.e for animals A, B, and C), and zeros were written in the off-diagonals between these three animals, assuming they were unrelated. Let the elements of this table (refered to as matrix **A**) be denoted as a_{ij}. Thus, by putting a 1 on the diagonals for animals with unknown parents, the additive genetic relationship of an animal with itself is one. The additive genetic relationship to animals without common parents or whose parents are unknown is assumed to be zero.

The next step is to compute relationships between animal A and animals D, E, and F. The relationship of any animal to another is equal to the average of the relationships of that animal with the parents of another animal. For example, the relationship between A and D is the average of the relationships between A and the parents of D, who are A and B. Thus,

$$\begin{array}{lll} a_{AD} & = .5 \ (\ a_{AA} + a_{AB} \ ) & = .5(1 + 0) = .5 \\ a_{AE} & = .5 \ (\ a_{AA} + a_{AC} \ ) & = .5(1 + 0) = .5 \\ a_{AF} & = .5 \ (\ a_{AE} + a_{AD} \ ) & = .5(.5 + .5) = .5 \end{array}$$

The relationship table, or **A** matrix, is symmetric, so that $a_{AD} = a_{DA}$, $a_{AE} = a_{EA}$, and $a_{AF} = a_{FA}$. Continue calculating the relationships for animals B and C to give the following table.

Tabular Method Example,
Partially Completed.

	-,-	-,-	-,-	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0	.5	.5	.5
B	0	1	0	.5	0	.25
C	0	0	1	0	.5	.25
D	.5	.5	0			
E	.5	0	.5			
F	.5	.25	.25			

Next, compute the diagonal element for animal D. By definition this is one plus the inbreeding coefficient, i.e.

$$a_{DD} \ = \ 1 + F_D.$$

The inbreeding coefficient, F_D is equal to one-half the additive genetic relationship between the parents of animal D, namely,

$$F_D \; = \; .5a_{AB} = 0.$$

When parents are unknown, the inbreeding coefficient is zero assuming the parents of the individual were unrelated. After computing the diagonal element for an animal, like D, then the remaining relationships to other animals in that row are calculated as before. The completed matrix is given below. Note that only animal F is inbred in this example. The inbreeding coefficient is a measure of the percentage of loci in the genome of an animal that has become homogeneous, that is, the two alleles at a locus are the same (identical by descent). Sometimes these alleles may be lethal and therefore, inbreeding is generally avoided.

Tabular Method Example
Completed Table.

	-,-	-,-	-,-	A,B	A,C	E,D
	A	B	C	D	E	F
A	1	0	0	.5	.5	.5
B	0	1	0	.5	0	.25
C	0	0	1	0	.5	.25
D	.5	.5	0	1	.25	.625
E	.5	0	.5	.25	1	.625
F	.5	.25	.25	.625	.625	1.125

Generally, the matrix **A** is nonsingular, but if the matrix includes two animals that are identical twins, then two rows and columns of **A** for these animals would be identical, and therefore, **A** would be singular. In this situation assume that the twins are genetically equal and treat them as one animal (by giving them the same registration number or identification) (see Kennedy and Schaeffer, 1989).

71 Inbreeding Calculations

The inbreeding coefficients and the inverse of **A** for inbred animals are generally required for BLUP analyses of animal models. Thus, fast methods of doing both of these calculations, and for very large populations of animals are necessary.

$$\mathbf{A} = \mathbf{TBT}',$$

where $\mathbf{T}$ is a lower triangular matrix and $\mathbf{B}$ is a diagonal matrix. Quaas (1976) showed that the diagonals of $\mathbf{B}$, say b_{ii} were

$$b_{ii} \ = \ (.5 - .25(F_s + F_d)),$$

where F_s and F_d are the inbreeding coefficients of the sire and dam, respectively, of the i^{th} individual. If one parent is unknown, then

$$b_{ii} \ = \ (.75 - .25F_p),$$

where F_p is the inbreeding coefficient of the parent that is known. Lastly, if neither parent is known then $b_{ii} = 1$.

One of the more efficient algorithms for calculating inbreeding coefficients is that of Meuwissen and Luo (1992). Animals should be in chronological order, as for the Tabular Method. To illustrate consider the example given in the Tabular Method section. The corresponding elements of $\mathbf{B}$ for animals A to F would be

$$\left(\begin{array}{cccccc} 1 & 1 & 1 & .5 & .5 & .5 \end{array} \right).$$

Now consider a new animal, G, with parents F and B. The first step is to set up three vectors, where the first vector contains the identification of animals in the pedigree of animal G, the second vector will contain the elements of a row of matrix $\mathbf{T}$, and the third vector will contain the corresponding b_{ii} for each animal.

Step 1 Add animal G to the ID vector, a 1 to the T-vector, and

$$b_{GG} \ = \ .5 - .25(.125 + 0) = 15/32$$

to the B-vector, giving

ID vector	T-vector	B-vector
G	1	15/32

Step 2 Add the parents of G to the ID vector, and because they are one generation back, add .5 to the T-vector for each parent. In the D-vector, animal B has $b_{BB} = 1$, and animal F has $b_{FF} = .5$. The vectors now appear as

ID vector	T-vector	B-vector
G	1	15/32
F	.5	.5
B	.5	1

Step 3 Add the parents of F and B to the ID vector, add .25 (.5 times the T-vector value of the individual (F or B)) to the T-vector, and their corresponding b_{ii} values. The parents of F were E and D, and the parents of B were unknown. These give

ID vector	T-vector	B-vector
G	1	15/32
F	.5	.5
B	.5	1
E	.25	.5
D	.25	.5

Step 4 Add the parents of E and D to the ID vector, .125 to the T-vector, and the appropriate values to the B-vector. The parents of E were A and C, and the parents of D were A and B.

ID vector	T-vector	B-vector
G	1	15/32
F	.5	.5
B	.5	1
E	.25	.5
D	.25	.5
A	.125	1
C	.125	1
A	.125	1
B	.125	1

The vectors are complete because the parents of A, B, and C are unknown and no further ancestors can be added to the pedigree of animal G.

Step 5 Accumulate the values in the T-vector for each animal ID. For example, animals A and B appear twice in the ID vector. Accumulating their T-vector values gives

ID vector	T-vector	B-vector
G	1	15/32
F	.5	.5
B	.5+.125=.625	1
E	.25	.5
D	.25	.5
A	.125+.125=.25	1
C	.125	1

Do not accumulate quantities until all pathways in the pedigree have been processed, otherwise a coefficient may be missed and the wrong inbreeding coefficient could be calculated.

Step 6 The diagonal of the **A** matrix for animal G is calculated as the sum of the squares of the values in the T-vector times the corresponding value in the B-vector, hence

$$\begin{aligned} a_{GG} &= (1)^2(15/32) + (.5)^2(.5) + (.625)^2 \\ &\quad +(.25)^2(.5) + (.25)^2(.5) + (.25)^2 + (.125)^2 \\ &= 72/64 \\ &= 1\frac{1}{8} \end{aligned}$$

The inbreeding coefficient for animal G is one-eighth.

The efficiency of this algorithm depends on the number of generations in each pedigree. If each pedigree is 10 generations deep, then each of the vectors above could have over 1000 elements for a single animal. To obtain greater efficiency, animals with the same parents could be processed together, and each would receive the same inbreeding coefficient, so that it only needs to be calculated once. For situations with only 3 or 4 generation pedigrees, this algorithm would be very fast and the amount of computer memory required would be low compared to other algorithms (Golden et al. (1991), Tier(1990)).

71.1 Example Additive Matrix

Consider the pedigrees in the table below:

Animal	Sire	Dam
1	-	-
2	-	-
3	1	-
4	1	2
5	3	4
6	1	4
7	5	6

Animals with unknown parents may or may not be selected individuals, but their parents (which are unknown) are assumed to belong to a em base generation of animals, i.e. a large, random mating population of unrelated individuals. Animal 3 has one parent known and one parent unknown. Animal 3 and its sire do not belong to the base generation, but its unknown dam is assumed to belong to the base generation. If these assumptions are not valid, then the concept of phantom parent groups needs to be utilized (covered later in these notes). Using the tabular method, the **A** matrix for the above seven animals is given below.

	-,-	-,-	1,-	1,2	3,4	1,4	5,6
	1	2	3	4	5	6	7
1	1	0	.5	.5	.5	.75	.625
2	0	1	0	.5	.25	.25	.25
3	.5	0	1	.25	.625	.375	.5
4	.5	.5	.25	1	.625	.75	.6875
5	.5	.25	.625	.625	1.125	.5625	.84375
6	.75	.25	.375	.75	.5625	1.25	.90625
7	.625	.25	.5	.6875	.84375	.90625	1.28125

Now partition **A** into **T** and **B** giving:

Sire	Dam	Animal	1	2	3	4	5	6	7	**B**
		1	1	0	0	0	0	0	0	1.0
		2	0	1	0	0	0	0	0	1.0
1		3	.5	0	1	0	0	0	0	.75
1	2	4	.5	.5	0	1	0	0	0	.50
3	4	5	.5	.25	.5	.5	1	0	0	.50
1	4	6	.75	.25	0	.5	0	1	0	.50
5	6	7	.625	.25	.25	.5	.5	.5	1	.40625

Note that the rows of **T** account for the direct relationships, that is, the direct transfer of genes from parents to offspring.

71.2 The Inverse of Additive Relationship Matrix

The inverse of the relationship matrix can be constructed similarly to the inverse of the genomic relationship matrix by a set of rules. Recall the previous example of seven animals with the following values for b_{ii}.

Animal	Sire	Dam	b_{ii}	b_{ii}^{-1}
1	-	-	1.00	1.00
2	-	-	1.00	1.00
3	1	-	0.75	1.33333
4	1	2	0.50	2.00
5	3	4	0.50	2.00
6	1	4	0.50	2.00
7	5	6	0.40625	2.4615385

Let $\delta = b_{ii}^{-1}$, then if both parents are known the following constants are added to the appropriate elements in the inverse matrix:

	animal	sire	dam
animal	δ	$-.5\delta$	$-.5\delta$
sire	$-.5\delta$	$.25\delta$	$.25\delta$
dam	$-.5\delta$	$.25\delta$	$.25\delta$

If one parent is unknown, then delete the appropriate row and column from the rules above, and if both parents are unknown then just add δ to the animal's diagonal element of the inverse.

Each animal in the pedigree is processed one at a time, but any order can be taken. Let's start with animal 6 as an example. The sire is animal 1 and the dam is animal 4. In this case, $\delta = 2.0$. Following the rules and starting with an inverse matrix that is empty, after handling animal 6 the inverse matrix should appear as follows:

	1	2	3	4	5	6	7
1	.5			.5		-1	
2							
3							
4	.5			.5		-1	
5							
6	-1			-1		2	
7							

After processing all of the animals, then the inverse of the relationship matrix for these seven animals should be as follows:

	1	2	3	4	5	6	7
1	2.33333	.5	-.66667	-.5	0	-1	0
2	.5	1.5	0	-1.00000	0	0	0
3	-.66667	0	1.83333	.5	-1	0	0
4	-.5	-1	.5	3.0000	-1	-1	0
5	0	0	-1	-1	2.61538	.61538	-1.23077
6	-1	0	0	-1	.61538	2.61538	-1.23077
7	0	0	0	0	-1.23077	-1.23077	2.46154

The product of the above matrix with the original relationship matrix, **A**, gives an identity matrix.

71.3 R function To Construct A-inverse

Given a pedigree list and the corresponding b_{ii} values for each animal in the pedigree, then the inverse of **A** can be written. Below is a function in R that will do those calculations. Animals should be numbered consecutively from 1 to N.

R Statements

```
# sid is a list of the sire IDs
# did is a list of the dam IDs
AINV = function(sid,did,bi) {
nam = length(sid)
np = nam + 1
roul = matrix(data=c(1,-0.5,-0.5,
-0.5,0.25,0.25,-0.5,0.25,0.25),ncol=3)
ss = sid + 1
dd = did + 1
LAI = matrix(data=c(0),nrow=np,ncol=np)
for(i in 1:nam) {
ip = i + 1
k = cbind(ip,ss[i],dd[i])
x = 1/bi[i]
LAI[k,k] = LAI[k,k] + roul*x
}
k = c(2:np)
AI = LAI[k,k]
return(AI)  }
```

71.4 Phantom Parent Groups

Westell (1984) and Robinson (1986) assigned *phantom* parents to animals with unknown parents. Each phantom parent was assumed to have only one progeny. Phantom parents were assumed to be unrelated to all other real or phantom animals.

Phantom parents whose first progeny were born in a particular time period probably underwent the same degree of selection intensity to become a breeding animal. However, male phantom parents versus female phantom parents might have been selected differently. Phantom parents were assigned to phantom parent *groups* depending on whether they were sires or dams and on the year of birth of their first progeny.

Genetic groups may also be formed depending on breed composition and/or regions within a country. The basis for further groups depends on the existence of different selection intensities involved in arriving at particular phantom parents.

Phantom parent groups are best handled by considering them as additional animals in the pedigree. Then the inverse of the relationship matrix can be constructed using the same rules as before. These results are due to Quaas (1984). To illustrate, use the same seven animals as before. Assign the unknown sires of animals 1 and 2 to phantom group 1 ($P1$) and the unknown dams to phantom group 2 ($P2$). Assign the unknown dam of animal 3 to phantom group 3 ($P3$). The resulting matrix will be of order 10 by 10 :

$$\mathbf{A}_*^{-1} = \left(\begin{array}{cc} \mathbf{A}^{rr} & \mathbf{A}^{rp} \\ \mathbf{A}^{pr} & \mathbf{A}^{pp} \end{array} \right),$$

where $\mathbf{A}^{rr}$ is a 7 by 7 matrix corresponding to the elements among the real animals; $\mathbf{A}^{rp}$ and its transpose are of order 7 by 3 and 3 by 7, respectively, corresponding to elements of the inverse between real animals and phantom groups, and $\mathbf{A}^{pp}$ is of order 3 by 3 and contains inverse elements corresponding to phantom groups. $\mathbf{A}^{rr}$ will be exactly the same as $\mathbf{A}^{-1}$ given in the previous section. The other matrices are

$$\mathbf{A}^{rp} = \left(\begin{array}{rrr} -.5 & -.5 & .33333 \\ -.5 & -.5 & 0 \\ 0 & 0 & -.66667 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{array} \right)$$

$$\mathbf{A}^{pp} = \left(\begin{array}{rrr} .5 & .5 & 0 \\ .5 & .5 & 0 \\ 0 & 0 & .33333 \end{array} \right)$$

In this formulation, phantom groups (according to Quaas (1984)) are additional fixed factors and there is a dependency between phantom groups 1 and 2. This singularity can cause problems in deriving solutions for genetic evaluation. The dependency can be removed by adding an identity matrix to $\mathbf{A}^{pp}$. When genetic groups have many animals assigned to them, then adding the identity matrix to $\mathbf{A}^{pp}$ does not result in any significant re-ranking of animals in genetic evaluation and aids in getting faster convergence of the iterative system of equations.

Phantom groups are used in many genetic evaluation systems today. The phantom parents assigned to a genetic group are assumed to be the outcome of non random mating and similar selection differentials on their parents. This assumption, while limiting, is not as severe as assuming that all phantom parents belong to one base population.

72 Identical Genotypes

Occasionally genetically identical twins are born, arising from a single embryo. These individuals share all of the same genetic material, both nuclear and non-nuclear DNA. In an additive genetic relationship matrix the rows for those two animals will be identical, and therefore, a dependency exists in the relationship matrix and an inverse is not possible.

Clones are individuals that tend to share only the nuclear DNA, and the assumption is that the non-nuclear DNA can cause genetic and phenotypic differences between the animals in their development. The additive genetic portion, which is passed on to progeny, is in the nuclear DNA, and therefore, the additive relationship matrix will have identical rows of numbers for clones from the same animal. The additive relationship matrix would be singular.

Kennedy and Schaeffer (1989) suggested that the relationship matrix be constructed for "genotypes" rather than for individuals in the case of these identical animals. If there were five clones of one animal, then the animal and its clones would represent one genotype, and so there would be only one row of the additive relationship matrix for all six animals. If the animals were measured for some trait, then that "genotype" would have repeated observations (without permanent environmental effects). They would all receive the same estimated breeding value.

One could also treat them as full-sibs, all having the same parents, but not sharing the exact same DNA. If there were many of these in the data file, then it could cause an overestimation of the additive genetic variance. Therefore, this approach would not be suitable.

73 Unknown Sires

In some situations a female is exposed to more than one possible mate. For example, a group of cows in a beef herd may have access to 3 or 4 males during the breeding season. Another example occurs in mink breeding where the success of having conception requires mating a female three times with different males at each mating. Progeny born from one female are a mixture of progeny of those three males. That is, different eggs could be fertilized by different males. Through genetic tests, the probabilities that a progeny is from either the first, second, or third male are known.

The additive relationship matrix can be constructed using the probabilities that a specific male is the sire of a given progeny. An example is as follows: Animals A and B are males, C is a female, and D is a progeny of C with 0.25 probability that the sire was A and 0.75 probability that the sire was B. Construct the additive genetic relationship matrix for this pedigree.

	A	B	C	D
A	1	0	0	.125
B	0	1	0	.375
C	0	0	1	.5
D	.125	.375	.5	1

Normally the relationship of the sire to its progeny (if unrelated to the dam) is 0.5, but in this case, for the relationship between A and D, the value has to be multiplied times the probability of A being the sire of D. Between B and D, the relationship of .5 has to be multipled times .75.

The inverse of this matrix is derived in a similar manner as the regular additive relationship matrix. The b_i values for animals A, B, and C are equal to 1. Because D has two possible sires, its b_i value needs to be calculated differently.

ID vector	T-vector	B-vector
D	1	x
C	.5	1
B	.375	1
A	.125	1

The diagonal of **A** for animal D has to be assumed to be known, as one plus one half the relationship between the sire(s) and dam. In this case, A, B, and C are all unrelated, and therefore, D will not be inbred, so that the diagonal of **A** will be 1.

$$a_{DD} \ = \ 1 \ = \ x \ + \ (.5)^2(1) \ + \ (.375)^2(1) \ + \ (.125)^2(1)$$

Solving for x gives $.59375 \ = \ b_D$. The inverse elements added to **A** for animal D are given by

$$\begin{pmatrix} -.125 \\ -.375 \\ -.5 \\ 1 \end{pmatrix} \frac{1}{.59375} \begin{pmatrix} -.125 & -.375 & -.5 & 1 \end{pmatrix} = \begin{pmatrix} .0263 & .0789 & .1053 & -.2105 \\ .0789 & .2368 & .3158 & -.6316 \\ .1053 & .3158 & .4211 & -.8421 \\ -.2105 & -.6316 & -.8421 & 1.6842 \end{pmatrix}.$$

To complete the inverse, add 1 to the diagonals for animals A, B, and C.

74 EXERCISES

1. Use the following data for this problem.

Treatment	Animal	Sire	Dam	Observations
1	1			15
2	2			73
1	3			44
2	4	1	3	56
1	5	2	4	55
2	6	1	5	61
1	7	6	4	32
2	8	7	5	47

Let the model equation be

$$y_{ij} = T_i + p_j + m_j + e_{ij}$$

where T_i is a fixed treatment effect, p_j is a random, paternal gamete effect of animal j, m_j is a random, maternal gamete effect of animal j, and e_{ijk} is a random, residual effect. Assume that

$$\sigma^2_e = 3.2\sigma^2_G.$$

(a) Complete the genomic relationship matrix for the eight animals, and the inverse of it.

(b) Construct the MME and solve.

(c) Predict the breeding values of each animal and obtain the standard errors of prediction.

(d) Test the difference between the treatments.

(e) Form the additive relationship matrix for this pedigree.

(f) Calculate the inverse of **A**.

(g) Assume the model

$$y_{ij} = T_i + a_j + e_{ij}$$

where a_j is the animal additive genetic effect with covariance matrix $\mathbf{A}\sigma^2_a$, and where

$$\sigma^2_e \;=\; 1.6\sigma^2_a.$$

Construct the MME for this model and solve. Compare EBVs from this model and the previous model.

2. Use the following data for this problem.

Animal	Sire	Dam	CG	Observations
1				
2				
3				
4	1	3	1	6
5	2	4	1	15
6	1		1	9
7	2	6	2	11
8	1	3	2	8
9	2	4	2	13
10	5	8	3	7
11	5	3	3	10
12	2	7	3	5

Let the model equation be

$$y_{ij} = CG_i + a_j + e_{ij}$$

where CG_i is a fixed contemporary group effect, a_j is a random, animal additive genetic effect, and e_{ijk} is a random residual effect. Assume that

$$\sigma_e^2 = 1.2\ \sigma_a^2.$$

(a) Complete the additive relationship matrix and the inverse of it.

(b) Construct the MME and solve for this model.

(c) Compute SEP for the EBV and also reliabilities.

3. You are given the following pedigree information and values of b_i. Determine the b_i value and inbreeding coefficient of animal H which is a progeny of animals G and F.

Animal	Sire	Dam	F_i	b_i
A			0	1
B	A		0	3/4
C	A	B	1/4	1/2
D	C	B	3/8	7/16
E	A	D	5/16	13/32
F	C	D	1/2	11/32
G	E	B	11/32	27/64
H	G	F		

Write out $\mathbf{A}^{-1}$ for this pedigree using Henderson's rules.

4. The following boxes are from a larger genomic relationship matrix.

					L	K
	Km	Kf	Lm	Lf	Mm	Mf
Gm	$\frac{1}{2}$	0	$\frac{1}{4}$	$\frac{3}{8}$	w	x
Gf	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{8}$	y	z

(a) Calculate w, x, y, and z.

(b) Calculate a_{GL} and d_{GL}

5. Assign phantom groups to replace missing sire and dam identifications in the following pedigree.

Animal	Sire	Dam	Sex	Year of birth
1			M	1970
2			F	1970
3			M	1971
4			F	1971
5			F	1971
6			M	1970
7	1		F	1972
8		2	M	1972
9			F	1972
10			M	1972

Form the inverse of the relationship matrix, including the phantom groups.

6. For the example pedigree in the section on **Unknown Sires**, compute the inverse of the **A** matrix assuming the probability that the sire is animal A of 0.3, and for animal B is 0.7.

7. In the section on **Unknown Sires**, let three unrelated sires be animals G, H, and K. Female M was exposed to all three sires (as in mink). The probabilities of the three sires being the sire of a progeny are .2, .5, and .3, for animals G, H, and K, respectively. Also assume that animal M is related to animal H by .25. Construct the **A** matrix and derive the inverse (without inverting **A** directly).

Phantom Parent Groups

75 Background History

Pedigree information on each animal may not be traceable back to the same base generation due to lack of recording and/or movement of animals from one owner to another (in the same or different countries). Therefore, a pedigree file may have many animals with missing parent identification. Animals with missing parents should not be assumed to be from a homogeneous base population.

One technique to deal with missing parent information is to assign a parent group to an animal based upon the year of birth of the animal and the pathway of selection. If the animal is a female, for example, and the dam information is missing, then the parent group would be in the Dams of Females pathway (DF). There are also Dams of Males (DM), Sires of Females (SF), and Sires of Males (SM). Four pathways and various years of birth nested within each pathway. These have become known as phantom parent groups.

Genetic group effects are added to the model.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za} + \mathbf{e},$$

where

$\mathbf{a}$ is the vector of animal additive genetic effects,

$\mathbf{Z}$ is the matrix that relates animals to their observations,

$\mathbf{g}$ is the vector of genetic group effects, and

$\mathbf{Q}$ is the matrix that relates animals to their genetic groups, and

$\mathbf{y}, \mathbf{Xb}, \mathbf{e}$ are as described in earlier notes.

The *Estimated Breeding Value*, EBV, of an animal is equal to the sum of the group and animal solutions from MME, i.e.

$$\text{Vector of EBVs} = \mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{a}}.$$

The grouping strategy described above was developed by Westell (1988), Quaas(1988), and Robinson(1986). To illustrate, phantom parents have been added to the following pedigrees, indicated by P1 to P6.

Animal	Sire	Dam
A	P1	P4
B	P2	P5
C	P3	P6
D	A	B
E	A	C
F	E	D

Now assign P1, P2, and P3 to genetic group 1 and P4, P5, and P6 to genetic group 2. The pedigree list becomes

Animal	Sire	Dam
A	G1	G2
B	G1	G2
C	G1	G2
D	A	B
E	A	C
F	E	D

75.1 Simplifying the MME

The advantage of phantom parent grouping is that the mixed model equations simplify significantly. Using the previous model, the MME are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}\mathbf{Q} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{X} & \mathbf{Q}'\mathbf{Z}'\mathbf{Z}\mathbf{Q} & \mathbf{Q}'\mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}\mathbf{Q} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}\alpha \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Q}'\mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}.$$

Notice that $\mathbf{Q}'$ times the third row subtracted from the second row gives

$$\mathbf{Q}'\mathbf{A}^{-1}\hat{\mathbf{a}}\alpha = \mathbf{0}.$$

Quaas and Pollak (1981) showed that the above equations could be transformed so that $\mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{a}}$ can be computed directly. The derivation is as follows. Note that

$$\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{a}} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Q} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{Q} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \hat{\mathbf{a}} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{Q} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{a}} \end{pmatrix}.$$

Substituting this equality into the left hand side (LHS) of the MME gives

$$\left(\begin{array}{ccc} \mathbf{X'X} & \mathbf{0} & \mathbf{X'Z} \\ \mathbf{Q'Z'X} & \mathbf{0} & \mathbf{Q'Z'Z} \\ \mathbf{Z'X} & -\mathbf{A^{-1}Q}\alpha & \mathbf{Z'Z}+\mathbf{A^{-1}}\alpha \end{array}\right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \mathbf{Q}\hat{\mathbf{g}}+\hat{\mathbf{a}} \end{array}\right) = \left(\begin{array}{c} \mathbf{X'y} \\ \mathbf{Q'Z'y} \\ \mathbf{Z'y} \end{array}\right).$$

To make the equations symmetric again, both sides of the above equations must be premultiplied by

$$\left(\begin{array}{ccc} \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} & -\mathbf{Q'} \\ \mathbf{0} & \mathbf{0} & \mathbf{I} \end{array}\right).$$

This gives the following system of equations as

$$\left(\begin{array}{ccc} \mathbf{X'X} & \mathbf{0} & \mathbf{X'Z} \\ \mathbf{0} & \mathbf{Q'A^{-1}Q}\alpha & -\mathbf{Q'A^{-1}}\alpha \\ \mathbf{Z'X} & -\mathbf{A^{-1}Q}\alpha & \mathbf{Z'Z}+\mathbf{A^{-1}}\alpha \end{array}\right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{g}} \\ \mathbf{Q}\hat{\mathbf{g}}+\hat{\mathbf{a}} \end{array}\right) = \left(\begin{array}{c} \mathbf{X'y} \\ \mathbf{0} \\ \mathbf{Z'y} \end{array}\right).$$

Quaas (1988) examined the structure of $\mathbf{Q}$ and the inverse of $\mathbf{A}$ under phantom parent grouping and noticed that $\mathbf{Q'A^{-1}Q}$ and $-\mathbf{Q'A^{-1}}$ had properties that followed the rules of Henderson (1976) for forming the elements of the inverse of $\mathbf{A}$. Thus, the elements of $\mathbf{A^{-1}}$ and $\mathbf{Q'A^{-1}Q}$ and $-\mathbf{Q'A^{-1}}$ can be created by a simple modification of Henderson's rules. Use δ_i as computed earlier, (i.e. $\delta_i = B_{ii}^{-1}$), and let i refer to the individual animal, let s and d refer to either the parent or the phantom parent group if either is missing, then the rules are

Constant to Add	Location in Matrix
δ_i	(i, i)
$-\delta_i/2$	(i, s),(s, i),(i, d), and (d, i)
$\delta_i/4$	(s, s),(d, d),(s, d), and (d, s)

Thus, $\mathbf{Q'A}^{-1}\mathbf{Q}$ and $\mathbf{Q'A}^{-1}$ can be created directly without explicitly forming $\mathbf{Q}$ and without performing the multiplications times $\mathbf{A^{-1}}$.

75.2 Typical Computational Problems

The solution to the modified equations has a problem in that the genetic group effects are still fixed effects and some restriction on their solutions may be needed to reach convergence. For animals with both parents unknown, then there can be complete confounding between the sire parent group and the dam parent group which causes a problem in getting a solution. Kennedy proposed to break confounding by defining genetic groups such that sire parent groups might be 70-71, 72-73, etc., while dam parent groups might be 69-70, 71-72, 73-74, etc. giving an overlap between sire and dam groups. Schaeffer (personal practice) suggested adding α to the diagonals of the genetic group effect equations in the modified equations. This automatically removes confounding, and the genetic group solutions sum to zero.

76 Estimation of Variances

A simulation study was conducted to look at the effects of genetic grouping methods, in general. A base population of 100 sires and 2000 dams was generated. Matings within a generation were random. Each mating resulted in one offspring, hence 2000 offspring per generation. Two correlated traits were simulated for each animal with a genetic correlation of 0.10, and a residual correlation of 0.30. Selection of males and females for the next generation was on the basis of trait 2 phenotypes (available for males and females). Animals were selected across generations, and so generations were overlapping. Trait 1 was analyzed in this study and only females after generation 0 were assumed to have single records for trait 1. The residual and genetic variances for both traits were the same. Total phenotypic variance was 100, and heritability in the base population was 0.25. Eight generations of selection and matings were made after the base population. Complete pedigrees were known in this population, denoted as P0. No phantom grouping was necessary with this population because all animals traced back to the same base population parents.

Two subpopulations were created from P0 by randomly eliminating parents of animals. Two different percentages were utilized, being 10 and 25%, denoted as P10 and P25, respectively. Different rates were used to test the impact on EBVs and genetic parameters. Phantom parent groups were formed according to the four pathways of selection and by generation of birth, i.e. 32 groups. There were no attempts to ensure a particular number of animals per group, but on average there would be 50 per group at 10% and 125 at 25%. The unknown parents, however, have an effect on the inbreeding coefficients, and therefore, on the elements in the relationship matrix inverses.

77 Models of Analyses

Several different models were applied to each data set.

77.1 Model 1

A simple animal model was assumed where

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{a} + \mathbf{e},$$

where $\mathbf{y}$ are the trait 1 single records on females, μ is the overall mean, $\mathbf{a}$ is the vector of animal additive genetic effects for all males and females in the population, and $\mathbf{e}$ is the vector of residual effects. Also,

$$Var(\mathbf{a}) = \mathbf{A}\sigma_a^2,$$

where $\mathbf{A}^{-1}$ was created from the pedigree that was known in a given population. This model assumes all animals can be traced back to the same base population animals.

77.2 Model 2

Due to selection there are usually time trends that need to be taken into account. This is accomplished by adding a generation effect to the model replacing the overall mean effect. This is also a form of genetic grouping where $\mathbf{Q}$ denotes to which group each animal with a record belongs, and $\mathbf{X} = \mathbf{ZQ}$.

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{b}$ are the 8 generation effects. Phantom parent groups are not utilized with this model. The EBVs of animals with records is the sum of the generation group solution plus the additive genetic solution.

77.3 Model 3

The model is

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{ZQg} + \mathbf{Za} + \mathbf{e},$$

where $\mathbf{Q}$ describes the phantom parent grouping definitions, and $\mathbf{g}$ are the phantom group effects. With this model the phantom group effects can be either fixed or random. This model and model 4 were only applied to populations P10 and P25. EBVs were obtained directly from this model ($\mathbf{Q}\hat{\mathbf{g}} + \hat{\mathbf{a}}$).

77.4 Model 4

The model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{ZQg} + \mathbf{Za} + \mathbf{e},$$

which has generation effects in $\mathbf{b}$ and also uses phantom grouping in $\mathbf{g}$. The definition of an EBV is not clear in this model. There is double counting for genetic trend in $\mathbf{b}$ and in $\mathbf{g}$. The EBV was the sum of the generation group solution and the animal solution.

78 Analyses

Fifty (50) replicates were made of each population structure and model of analysis. Variance components were estimated using Bayesian methods and Gibbs sampling (30,000 total samples and 3,000 burn-in period).

EBVs were calculated in the next step using the estimated components, using a set convergence criteria or a maximum of 3000 iterations, whichever came first. No restrictions were placed on any solutions. Iteration on data methods were used to solve the MME. Average inbreeding coefficients by generation were summarized. Correlations of EBVs with true breeding values were computed for all animals in generations 1 to 8. True genetic trends in males and females by generation were averaged over replicates. The difference between true generation average and estimated average in generation 8 was used to determine if trends were under or over estimated.

Note that when phantom groups were fixed, there were severe confounding problems with the MME, and therefore, only results for when the variance ratio (residual to genetic) was added to the diagonals of the phantom group equations are reported here(i.e. when phantom groups were random). The correlations between true and estimated breeding values for fixed phantom groups were less than 0.2 with Model 3, for example.

79 Results

Table 1 gives the correlations between true and estimated breeding values of all animals from generations 1 to 8. The best model appears to be one without any generation effects or phantom groups, for all pedigree structures. As the percentage of animals with unknown parents is increased the correlation decreases. Differences between models and pedigree structures were significantly different from Model 1 and P0. Although the differences are significant statistically, they may not be practically significant.

Table 1
Correlations of true breeding values with
estimated breeding values of all animals
from generations 1 to 8.
Standard error of estimates was 0.0025.

Data	Model 1	Model 2	Model 3	Model 4
Set	μ	$\mathbf{Xb}$	$\mu + \mathbf{ZQg}$	$\mathbf{Xb} + \mathbf{ZQg}$
P0	.588	.579		
P10	.575	.567	.567	.568
P25	.555	.550	.544	.547

Average inbreeding coefficients of animals born in generation 8 are given in Table 2. When full pedigree information was available back to the base population, then the average inbreeding coefficient was 0.00872. When 10% of the parents were assumed to be unknown the average inbreeding coefficient dropped by about 0.0027 (which was statistically significant), and at 25% a drop of 0.0054 occurred. Thus, inbreeding rate is significantly underestimated when pedigrees are missing. The differences are small on a practical basis, but over time these differences would become greater.

Table 2
Average inbreeding coefficients computed from known pedigree information of animals in generation 8.
Standard error of estimates was 0.000001.

Data Set	Model 1 μ	Model 2 $\mathbf{Xb}$	Model 3 $\mu + \mathbf{ZQg}$	Model 4 $\mathbf{Xb} + \mathbf{ZQg}$
P0	.00872	.00872		
P10	.00596	.00594	.00585	.00587
P25	.00330	.00328	.00325	.00322

Estimates of genetic and residual variances and heritability are given in Table 3. Model 1 was unbiased for all three pedigree structures, as was Model 3 for P10 and P25. Thus, phantom parent groups may be ignored when estimating genetic parameters under any pedigree structure. Including a fixed generation effect in the model tended to bias the genetic component downwards and resulted in underestimated heritability. The residual component was not affected.

Table 3
Estimates of genetic (G) and residual (R) variances and heritability (h^2).
Standard errors of estimates were 0.60, 0.71, and 0.006 for R, G, and h^2, respectively.

Data Set		Model 1 μ	Model 2 $\mathbf{Xb}$	Model 3 $\mu + \mathbf{ZQg}$	Model 4 $\mathbf{Xb} + \mathbf{ZQg}$
P0	R	74.8	75.5		
	G	25.1	24.2		
	h^2	.251	.242		
P10	R	74.8	75.7	74.9	75.4
	G	25.3	23.8	25.4	24.2
	h^2	.252	.239	.252	.243
P25	R	74.9	75.9	74.9	75.4
	G	25.2	23.5	25.3	24.2
	h^2	.251	.236	.252	.242

Genetic trends were also affected by the model and pedigree structure. The average differences between true genetic averages and estimated genetic averages in generation 8 are shown in Table 4. Model 1 (with no generation effects or phantom groups) always underestimated the true genetic change in the population. The bias increased as the percentage of unknown parents increased. Model 3 (with phantom parent groups) was also underestimated, but not as much as Model 1. When generation effects were in the model (Models 2 and 4), then genetic trends were estimated unbiasedly.

Table 4
Average difference of true minus estimated genetic averages of animals in generation 8, expressed in genetic standard deviation units. Standard error of estimates was 0.008.

Data Set	Model 1 μ	Model 2 $\mathbf{Xb}$	Model 3 $\mu + \mathbf{ZQg}$	Model 4 $\mathbf{Xb} + \mathbf{ZQg}$
P0	.187	-.004		
P10	.274	-.005	.233	.007
P25	.361	-.005	.246	.010

Table 5 shows trends in phantom group solutions for Model 3 as fixed or random, averaged over replicates, and also shows true genetic trends for comparison and estimates of generation effects from Model 2. Phantom group solutions did not follow true genetic trends (either as fixed or as random), and there is no theoretical requirement anywhere that the phantom group solutions should be smooth or linear. When phantom groups are fixed, the estimates had much higher standard errors, and were too variable from replicate to replicate. The generation effect estimates from Model 2 were linear and smooth, but underestimated the true genetic trend. However, when the generation effect solutions are added to the animal solutions to give EBV, the trend in average EBV gives an unbiased estimate of the true genetic trend.

Table 5
Two sets of phantom group solutions for Model 3, P10, for fixed and random analysis, average estimates of generation effects from Model 2, and average true breeding values by generation.

Gen.	True	$\hat{b}$	SB	SC	DB	DC	
1	-.044	.006	.093	-1.220	.730	-.580	R
			2.526	-1.448	2.408	-.590	F
2	.661	.530	.276	.447	.714	.079	R
			-5.268	.780	6.311	.175	F
3	1.041	.930	.762	-.450	1.202	.407	R
			3.295	-.472	-.469	.589	F
4	1.426	1.219	.968	1.948	1.300	1.302	R
			1.892	2.743	3.443	1.682	F
5	1.841	1.625	1.072	1.558	1.172	1.086	R
			-1.974	2.206	3.455	1.401	F
6	2.220	1.996	1.518	1.457	1.704	1.467	R
			-2.589	2.116	4.278	1.885	F
7	2.665	2.301	.855	1.875	.647	1.592	R
			-1.505	2.690	-7.416	2.021	F
8	3.006	2.668	.003	2.150	.005	1.921	R
			.000	3.087	.000	2.492	F
SE		.099	.373	.373	.373	.373	R
			2.366	2.366	2.366	2.366	F

80 Discussion and Conclusions

Limitations of this study were

1. Missing parents were determined randomly for lack of knowledge about how missing parents are actually sampled. Thus, if parents are missing based on some selective strategy, then the results could be much worse than shown here for P10 and P25.

2. Selection in this simulation on phenotypes of trait 2, while analyzing only trait 1 in females was a simplification as might occur when analyzing fertility traits, for example, after selection for many generations on production.

3. Phenotypic selection would keep inbreeding from accumulating too quickly in this small population.

4. Assortative mating of parents would be another complication to this simulation that would likely increase inbreeding, as well as genetic change.

5. Results may also differ slightly with heritability of the trait.

The choice of model seems to depend on the objective of the analysis. If the objective is to estimate the genetic parameters unbiasedly, then Model 1 should be used. However, Model 1 significantly underestimates genetic trends. The best model for estimation of genetic trend was Model 2. The degree of underestimation by Model 1 depends on the percentage of missing parent information. Thus, the best course of action seems to be to use Model 1 for estimating parameters, and with those parameter estimates use Model 2 to get unbiased estimates of genetic trend and presumably EBVs. Hence, phantom parent groups are not necessary for either objective. Models 3 and 4 gave lower correlations of true BV with EBV than Models 1 or 2.

If phantom parent groups are used, then do not expect the solutions to be smooth and linear, or to follow the true genetic trend. Forcing phantom group solutions to be smooth would be incorrect and could introduce bias to EBVs. Phantom groups as random should be better than phantom groups as fixed, but there have been other studies that have shown no differences between fixed or random if the number of animals per phantom group is large.

Animal Model

81 Introduction

In animal genetics, measurements are taken on individual animals, and thus, the model of analysis should include the animal additive genetic effect. The remaining items in the model are factors that have an effect on the trait of interest. The infinitesimal genetic model is assumed, and the animals that have been measured should represent a random sample (or the entire sample) of progeny of the breeding males and females in the population. That means measurements should not be taken only from the biggest progeny, or the best progeny, or the poorest progeny, but a random sample of progeny.

A simple animal model with one record per animal is

$$y_i \;=\; \mu \;+\; a_i \;+\; e_i,$$

where y_i is the phenotypic record or observation, μ is the overall mean of the observations, a_i are the additive genetic values of animals, and e_i are the residual or environmental effects associated with the i^{th} observation. Assumptions for this model are

1. The population is large (infinite) in size.
2. The association of alleles from the two parents is assumed to be at random.
3. No selection has been imposed on individual records.
4. Only additive genetic effects exist.
5. The influence of other factors on the observations is absent.

The expectations (for the example data to follow) are

$$E(a_i) \;=\; 0, \;\; \text{and} \;\; E(e_i) \;=\; 0,$$

and the variances are

$$\begin{aligned} Var(a_i) &= \sigma^2_a, \\ Var(e_i) &= \sigma^2_e. \end{aligned}$$

The variances are assumed known, at least to proportionality. The covariance matrix of the vector of animal additive genetic effects is

$$Var(\mathbf{a}) \;=\; \mathbf{A}\sigma^2_a.$$

The relationship matrix is assumed to be complete back to the base population (a large, randomly mating population).

The heritability of the trait is

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_e^2},$$

and the ratio of variances is

$$\alpha = \frac{\sigma_e^2}{\sigma_a^2}.$$

82 Example Problem

Below is the pedigree information and data on 16 animals. The first four animals were base generation animals without records.

82.1 Data

Animal	Sire	Dam	Record	Animal	Sire	Dam	Record
1	0	0		9	5	6	36.0
2	0	0		10	7	8	66.4
3	0	0		11	5	8	28.9
4	0	0		12	7	6	73.0
5	1	2	38.5	13	1	6	44.2
6	3	4	48.9	14	3	8	53.4
7	3	2	64.3	15	5	4	33.6
8	1	4	50.5	16	7	8	49.5

The number of observations is $N = 12$, and the total sum of squares is $30,811.78$. Let $\sigma_a^2 = 36$, and $\sigma_e^2 = 64$, so that $h^2 = 0.36$, and $\alpha = 1.777778$.

82.2 Additive Genetic Relationships

The matrix of additive genetic relationships among the sixteen individuals is **A** (times 16) given below:

$$\begin{pmatrix}
16 & 0 & 0 & 0 & 8 & 0 & 0 & 8 & 4 & 4 & 8 & 0 & 8 & 4 & 4 & 4 \\
0 & 16 & 0 & 0 & 8 & 0 & 8 & 0 & 4 & 4 & 4 & 4 & 0 & 0 & 4 & 4 \\
0 & 0 & 16 & 0 & 0 & 8 & 8 & 0 & 4 & 4 & 0 & 8 & 4 & 8 & 0 & 4 \\
0 & 0 & 0 & 16 & 0 & 8 & 0 & 8 & 4 & 4 & 4 & 4 & 4 & 4 & 8 & 4 \\
8 & 8 & 0 & 0 & 16 & 0 & 4 & 4 & 8 & 4 & 10 & 2 & 4 & 2 & 8 & 4 \\
0 & 0 & 8 & 8 & 0 & 16 & 4 & 4 & 8 & 4 & 2 & 10 & 8 & 6 & 4 & 4 \\
0 & 8 & 8 & 0 & 4 & 4 & 16 & 0 & 4 & 8 & 2 & 10 & 2 & 4 & 2 & 8 \\
8 & 0 & 0 & 8 & 4 & 4 & 0 & 16 & 4 & 8 & 10 & 2 & 6 & 8 & 6 & 8 \\
4 & 4 & 4 & 4 & 8 & 8 & 4 & 4 & 16 & 4 & 6 & 6 & 6 & 4 & 6 & 4 \\
4 & 4 & 4 & 4 & 4 & 4 & 8 & 8 & 4 & 16 & 6 & 6 & 4 & 6 & 4 & 8 \\
8 & 4 & 0 & 4 & 10 & 2 & 2 & 10 & 6 & 6 & 18 & 2 & 5 & 5 & 7 & 6 \\
0 & 4 & 8 & 4 & 2 & 10 & 10 & 2 & 6 & 6 & 2 & 18 & 5 & 5 & 3 & 6 \\
8 & 0 & 4 & 4 & 4 & 8 & 2 & 6 & 6 & 4 & 5 & 5 & 16 & 5 & 4 & 4 \\
4 & 0 & 8 & 4 & 2 & 6 & 4 & 8 & 4 & 6 & 5 & 5 & 5 & 16 & 3 & 6 \\
4 & 4 & 0 & 8 & 8 & 4 & 2 & 6 & 6 & 4 & 7 & 3 & 4 & 3 & 16 & 4 \\
4 & 4 & 4 & 4 & 4 & 4 & 8 & 8 & 4 & 8 & 6 & 6 & 4 & 6 & 4 & 16
\end{pmatrix}.$$

82.3 Application of BLUP with R

The construction of mixed model equations begins by defining **X**, **Z**, **G** and **R**.

X is a vector of ones of order 12 by 1.

Z is a 12 by 16 matrix, in which the first four columns are all zero, and the last 12 columns are an identity matrix.

G is $\mathbf{A}\sigma_a^2$ of order 16, and

R is $\mathbf{I}\sigma_e^2$ of order 12.

To set up **X** in R, use the `jd` function below.

```
# function to form a J matrix - all ones
jd <- function(n,m) matrix(c(1),nrow=n,ncol=m)

X = jd(12,1)
```

There are at least two ways to create **Z**.

```
# First Method
# function to make an identity matrix
id <- function(n) diag(c(1),nrow=n,ncol=n)

Z = cbind((jd(12,4)*0),id(12))

# Second Method
# function to make a design matrix
desgn <- function(vnum) {
mrow = length(vnum)
mcol = max(vnum)
W = matrix(data=c(0),nrow=mrow,ncol=mcol)
for(i in 1:mrow) {
ic = vnum[i]
W[i,ic] = 1  }
return(W)  }

anim = c(5:16) # animals with records
Z = desgn(anim)
```

The desgn function can be used generally for creating design matrices from a list of the levels of a factor.

The inverse of the relationship matrix is needed. A list of the sires and dams of the 16 animals are needed, with zeros for missing parents, and a list of the b_i values for each animal are needed. In this example, the b_is for animals 1 to 4 are 1, and for animals 5 through 16 are 0.5.

```
y = matrix(data=c(38.5,48.9,64.3,50.5,36.0,66.4,28.9,
73.0,44.2,53.4,33.6,49.5)
bii = c(1,1,1,1,.5,.5,.5,.5,.5,.5,.5,.5,.5,.5,.5,.5)
sid = c(0,0,0,0,1,3,3,1,5,7,5,7,1,3,5,7)
did = c(0,0,0,0,2,4,2,4,6,8,8,6,6,8,4,8)

AI = AINV(sid,did,bii)

# multiply AI by the variance ratio
GI = AI * alpha

RI = id(12)
```

Alpha is the ratio of residual to additive genetic variances. In this example, the ratio is 64 divided by 36, or 1.778. The last line above creates the inverse of **R** which is an identity matrix. Finally, the mixed model equations are created and solved using the `MME` function given in an earlier section of the notes.

```
Exmp = MME(X,Z,GI,RI,y)

Exmp$LHS
Exmp$RHS
Exmp$C
Exmp$SOLNS
```

82.4 Solutions and SEP

The solutions for the animals are given in the next table. The estimate of the overall mean was 48.8955.

EBV for animals from MME and diagonals of the inverse
of the LHS.

Animal	EBV	Diagonals
1	-4.7341	.4866
2	.3728	.4959
3	5.8579	.4786
4	-1.4967	.4975
5	-7.1305	.4198
6	2.1335	.4260
7	8.4380	.4137
8	-2.1543	.4274
9	-4.7807	.4592
10	6.2947	.4582
11	-8.0126	.4795
12	9.4167	.4751
13	-2.0456	.4507
14	2.4341	.4501
15	-6.7242	.4459
16	2.5849	.4582

An estimate of the residual variance is obtained by subtracting the total sum of squares minus the sum of squares due to the model, and dividing that by $N - r(\mathbf{X})$.

$$\begin{aligned} SST &= 30,811.78 \\ SSR &= 29,618.3158 \\ SST - SSR &= 1193.464 \\ \hat{\sigma}_e^2 &= 108.4967 \end{aligned}$$

The Standard Error of Prediction (SEP) of an Estimated Breeding Value (EBV) is the square root of the product of the diagonal of the inverse of the LHS times the estimate of residual variance. Thus for animal 1 in the table above, the SEP is

$$SEP = (.4866 \times 108.4967)^{.5} \ = \ 7.2662.$$

82.5 Reliability

The reliability (REL) of an EBV is another measure of accuracy of the estimate.

$$REL_i = (a_{ii} - c_{ii}\alpha)/a_{ii},$$

where a_{ii} is the diagonal of the **A** matrix for animal i, and c_{ii} is the diagonal of the inverse of the LHS (in the table above). For animal 1,

$$REL = (1 \ - \ .4866(1.778))/1 \ = \ .1349.$$

The reliability is only 13.49 %, which is low.

Publication of "official" EBV often requires a minimum reliability, such as 75%, plus minimums on number of progeny and number of contemporary groups for those progeny.

83 Simulation of Records

The sampling processes involved in generating a set of data are better understood through simulation.

83.1 Generating Breeding Values of Animals

Let the heritability of the trait be 0.25, and the variance of phenotypic records to be 100. Then, $\sigma^2_a = 25$, and $\sigma^2_e = 75$. Breeding values of base population animals are created by multiplying a pseudo-random normal deviate, RND, times the genetic standard deviation, $\sigma_a = 5$. Let there be two base animals,

$$\begin{aligned} a_1 &= -1.8014 * 5 \ = \ -9.0070 \\ a_2 &= 0.6556 * 5 \ = \ 3.2780 \end{aligned}$$

Progeny can be generated from animals previously generated as the average of the parent breeding values plus a Mendelian sampling effect. The Mendelian sampling effect is the product of another random normal deviate times the genetic standard deviation of the Mendelian sampling effect. The variance of the Mendelian sampling effect is b_i times the genetic variance, where b_i is obtained during the calculation of inbreeding coefficients (given in previous notes). That is,

$$b_i \ = \ (0.5 - 0.25 * (F_s + F_d)),$$

where F_s and F_d are the inbreeding coefficients of the sire and dam, respectively.

Let Animal 3 be a progeny of animals 1 and 2,

$$\begin{aligned} a_3 &= 0.5 * (a_1 \ + \ a_2) \ + \ m_3 \\ b_3 &= 0.5, \\ m_3 &= (0.5)^{.5}\sigma_a \text{ RND} \\ &= 0.7071 * 5 * (-.4045), \\ &= -1.4301 \end{aligned}$$

$$\begin{aligned} a_3 &= 0.5(-9.0070 + 3.2780) - 1.4301 \\ &= -4.2946. \end{aligned}$$

83.2 Phenotypic Records

To create phenotypic records, the equation of the model must be specified. Assume that

$$y_{ij} \;=\; \mu \;+\; d_j \;+\; a_i \;+\; e_{ij},$$

where μ is an overall mean (assume a value of 50); d_j is a diet effect for one of 3 diets, a_i is the animal breeding value, and e_{ij} is a residual effect.

Animals must be assigned to a diet. One way is to have the experimental design known in advance, such that an animal is created for a specific diet. Another way is to pick one of the three diets randomly and assign it to that animal. The differences among the diets should be set before generating records. Assume that

$$\begin{aligned} d_1 &= -10.0, \\ d_2 &= 0, \\ d_3 &= 10.0. \end{aligned}$$

Diet differences would be constant for all animals generated. If the simulation were repeated, then the same differences would probably be used. This would be the traditionalist's view of a fixed factor.

Residual effects are random and specific to each observation. A residual effect is generated by multiplying a random normal deviate times the residual standard deviation.

A record for animal 3, in diet 3, would be

$$y_{33} \;=\; 50 \;+\; 10.0 \;-4.2946 + .3939(8.6602) \;=\; 59.1167.$$

Usually observations will be to the nearest whole number, so that $y_{33} = 59$ would be the final result for animal 3.

If animals are generated in chronological order, then inbreeding coefficients need to be calculated as new animals are generated, so that b_i values can be obtained for generating Mendelian sampling effects. This is an efficient algorithm for simulating data for large populations. A later set of notes will consider using R to generate large populations of animals for single or multiple traits.

84 Measuring Genetic Change

To measure genetic change, the population needs to be defined precisely. In dairy cattle, one population would be all females born, by year of birth. Then the average EBVs of cows by year of birth would estimate the genetic change in new females coming into the population.

Another population could be all females calving in a particular year. This is different from the first population because cows can have several calvings over their lifetime, and secondly not all females that are born would necessarily calve even once. A cow's EBV could appear in the calculations in more than one year of calving. This would represent the trend in genetic merit of actively lactating cows.

In beef cattle, the averages for females being bred in a given year could be an interesting trend to monitor versus the average of females that produce a calf. In all cases, EBVs are used in the calculations.

Genetic trends can be overestimated if the heritability used in the calculation of EBV is too large. Similarly, if the heritability is too low, then genetic trend can be underestimated. Thus, having good estimates of the variance parameters is crucial to good estimates of genetic trend.

85 Reduced Animal Models

85.1 Usual Animal Model Approach

Consider an animal model with periods as a fixed factor and one observation per animal, as in the table below.

Animal Model Example Data

Animal	Sire	Dam	Period	Observation
5	1	3	2	250
6	1	3	2	198
7	2	4	2	245
8	2	4	2	260
9	2	4	2	235
4	-	-	1	255
3	-	-	1	200
2	-	-	1	225

Assume that the ratio of residual to additive genetic variances is 2. The MME for this data would be of order 11 (nine animals and two periods).

$$\begin{pmatrix} 3 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 4 & 0 & 2 & 0 & -2 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 6 & 0 & 3 & 0 & 0 & -2 & -2 & -2 \\ 1 & 0 & 2 & 0 & 5 & 0 & -2 & -2 & 0 & 0 & 0 \\ 1 & 0 & 0 & 3 & 0 & 6 & 0 & 0 & -2 & -2 & -2 \\ 0 & 1 & -2 & 0 & -2 & 0 & 5 & 0 & 0 & 0 & 0 \\ 0 & 1 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 5 & 0 \\ 0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 5 \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \hat{a}_8 \\ \hat{a}_9 \end{pmatrix} = \begin{pmatrix} 680 \\ 1188 \\ 0 \\ 225 \\ 200 \\ 255 \\ 250 \\ 198 \\ 245 \\ 260 \\ 235 \end{pmatrix}$$

and the solutions to these equations are

$$\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \hat{a}_8 \\ \hat{a}_9 \end{pmatrix} = \begin{pmatrix} 225.8641 \\ 236.3366 \\ -2.4078 \\ 1.3172 \\ -10.2265 \\ 11.3172 \\ -2.3210 \\ -12.7210 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{pmatrix}.$$

Some species, such as swine or rainbow trout, have many offspring per mating, and subsequently fewer animals are needed for breeding purposes. The MME would contain many equations (one per animal), even though a large percentage of animals do not have any progeny. Reducing the order of the MME would be an advantage, to save computing time. Quaas and Pollak (1980) proposed the *reduced animal model.* There is one model for animals that are not parents, and a separate model for animals that are parents. The total number of animal genetic equations in MME is equal to the number of animals that are parents. The solutions for the animals in the reduced animal model are exactly the same as solutions from the regular animal model.

85.2 Theoretical Development

The vector of additive genetic values of animals in the animal model is $\mathbf{a}$. Denote animals with progeny as $\mathbf{a_p}$, and those without progeny as $\mathbf{a_o}$, so that

$$\mathbf{a}' = \begin{pmatrix} \mathbf{a_p}' & \mathbf{a_o}' \end{pmatrix}.$$

In terms of the example data,

$$\begin{aligned}
\mathbf{a_p}' &= \begin{pmatrix} a_1 & a_2 & a_3 & a_4 \end{pmatrix}, \\
\mathbf{a_o}' &= \begin{pmatrix} a_5 & a_6 & a_7 & a_8 & a_9 \end{pmatrix}.
\end{aligned}$$

The additive genetic value of an animal may be written as the average of the additive genetic values of the parents plus a Mendelian sampling effect, which is the animal's specific deviation from the parent average, i.e.

$$a_i = .5(a_s + a_d) + m_i.$$

Therefore,

$$\mathbf{a}_o = \mathbf{P}\mathbf{a}_p + \mathbf{m},$$

where $\mathbf{P}$ is a matrix that indicates the parents of each animal in $\mathbf{a}_o$, with elements equal to 0.5, and $\mathbf{m}$ is the vector of Mendelian sampling effects. Then

$$\begin{aligned}
\mathbf{a} &= \begin{pmatrix} \mathbf{a}_p \\ \mathbf{a}_o \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I} \\ \mathbf{P} \end{pmatrix} \mathbf{a}_p + \begin{pmatrix} \mathbf{0} \\ \mathbf{m} \end{pmatrix},
\end{aligned}$$

and

$$\begin{aligned}
Var(\mathbf{a}) &= \mathbf{A}\sigma_a^2 \\
&= \begin{pmatrix} \mathbf{I} \\ \mathbf{P} \end{pmatrix} \mathbf{A}_{pp} \begin{pmatrix} \mathbf{I} & \mathbf{P}' \end{pmatrix} \sigma_a^2 + \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D} \end{pmatrix} \sigma_a^2
\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements equal to b_i, and b_i is $0.5 - 0.25 * (F_s + F_d)$ for the i^{th} animal, and

$$Var(\mathbf{a}_p) = \mathbf{A}_{pp}\sigma_a^2.$$

The animal model can now be written as

$$\begin{pmatrix} \mathbf{y}_p \\ \mathbf{y}_o \end{pmatrix} = \begin{pmatrix} \mathbf{X}_p \\ \mathbf{X}_o \end{pmatrix} \mathbf{b} + \begin{pmatrix} \mathbf{Z}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_o \end{pmatrix} \begin{pmatrix} \mathbf{I} \\ \mathbf{P} \end{pmatrix} \mathbf{a}_p + \begin{pmatrix} \mathbf{e}_p \\ \mathbf{e}_o + \mathbf{Z}_o\mathbf{m} \end{pmatrix}.$$

The residual vector has two types of residuals and the additive genetic values of animals without progeny have been replaced with $\mathbf{P}\mathbf{a}_p$. Because every individual has only one record, then $\mathbf{Z}_o = \mathbf{I}$, but $\mathbf{Z}_p$ may have fewer rows than there are elements of $\mathbf{a}_p$ because not all parents may have observations themselves. In the example data, animal 1 does not have an observation, therefore,

$$\mathbf{Z}_p = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}.$$

Consequently,

$$\begin{aligned}
\mathbf{R} &= Var\begin{pmatrix} \mathbf{e}_p \\ \mathbf{e}_o + \mathbf{m} \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 + \mathbf{D}\sigma_a^2 \end{pmatrix} \\
&= \begin{pmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_o \end{pmatrix}\sigma_e^2
\end{aligned}$$

The mixed model equations for the reduced animal model are

$$\begin{pmatrix} \mathbf{X}_p'\mathbf{X}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{X}_p'\mathbf{Z}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{P} \\ \mathbf{Z}_p'\mathbf{X}_p + \mathbf{P}'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{Z}_p'\mathbf{Z}_p + \mathbf{P}'\mathbf{R}_o^{-1}\mathbf{P} + \mathbf{A}_{pp}^{-1}\alpha \end{pmatrix}\begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_p \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}_p'\mathbf{y}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{y}_o \\ \mathbf{Z}_p'\mathbf{y}_p + \mathbf{P}'\mathbf{R}_o^{-1}\mathbf{y}_o \end{pmatrix}.$$

Solutions for $\hat{\mathbf{a}}_o$ are derived from the following formulas:

$$\hat{\mathbf{a}}_o = \mathbf{P}\hat{\mathbf{a}}_p + \hat{\mathbf{m}},$$

where

$$\hat{\mathbf{m}} = (\mathbf{Z}_o'\mathbf{Z}_o + \mathbf{D}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\hat{\mathbf{b}} - \mathbf{P}\hat{\mathbf{a}}_p).$$

85.3 Analysis of Example Data

Using the example data,

$$\mathbf{P} = \begin{pmatrix} .5 & 0 & .5 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \end{pmatrix},$$

and

$$\mathbf{D} = \text{diag}\begin{pmatrix} .5 & .5 & .5 & .5 & .5 \end{pmatrix},$$

then the MME with $\alpha = 2$ are

$$\begin{pmatrix} 3 & 0 & 0 & 1 & 1 & 1 \\ 0 & 4 & .8 & 1.2 & .8 & 1.2 \\ 0 & .8 & 2.4 & 0 & .4 & 0 \\ 1 & 1.2 & 0 & 3.6 & 0 & .6 \\ 1 & .8 & .4 & 0 & 3.4 & 0 \\ 1 & 1.2 & 0 & .6 & 0 & 3.6 \end{pmatrix}\begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \end{pmatrix} = \begin{pmatrix} 680. \\ 950.4 \\ 179.2 \\ 521. \\ 379.2 \\ 551. \end{pmatrix}$$

The solutions are as before, i.e.

$$\begin{array}{lll} \hat{b}_1 = 225.8641, & \hat{a}_1 = \text{-}2.4078, & \hat{a}_3 = \text{-}10.2265, \\ \hat{b}_2 = 236.3366, & \hat{a}_2 = 1.3172, & \hat{a}_4 = 11.3172. \end{array}$$

To compute $\hat{\mathbf{a}}_o$, first calculate $\hat{\mathbf{m}}$ as:

$$\begin{aligned}
(\mathbf{I} + \mathbf{D}^{-1}\alpha) &= \left(\begin{array}{ccccc} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{array}\right) \\
\mathbf{y}_o &= \left(\begin{array}{c} 250 \\ 198 \\ 245 \\ 260 \\ 235 \end{array}\right) \\
\mathbf{X}_o\hat{\mathbf{b}} &= \left(\begin{array}{cc} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right) \left(\begin{array}{c} 225.8641 \\ 236.3366 \end{array}\right) \\
\mathbf{P}\hat{\mathbf{a}}_p &= \left(\begin{array}{r} -6.3172 \\ -6.3172 \\ 6.3172 \\ 6.3172 \\ 6.3172 \end{array}\right) \\
\hat{\mathbf{m}} &= (\mathbf{I} + \mathbf{D}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\hat{\mathbf{b}} - \mathbf{P}\hat{\mathbf{a}}_p) \\
&= \left(\begin{array}{r} 3.9961 \\ -6.4039 \\ .4692 \\ 3.4692 \\ -1.5308 \end{array}\right)
\end{aligned}$$

and

$$\mathbf{P}\hat{\mathbf{a}}_p + \hat{\mathbf{m}} = \left(\begin{array}{r} -2.3211 \\ -12.7211 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{array}\right).$$

Generally, with today's computer power, there is little need to use reduced animal models. However, routine genetic evaluation schemes for species with high reproductive rates may benefit from using a reduced animal model.

86 Sire and Dam Models

Another type of reduced animal model is called a sire and dam model. The assumptions to use this model are

1. Animals have only one record each,
2. Animals that have records are not parents of other animals, and
3. None of the parents have records of their own.
4. Sires and dams are mated randomly.
5. Sires and dams are random samples of parents, and not the result of intense selection.
6. Progeny have only one record each.

Parents have only progeny, and do not have records themselves. The animal model equation is re-written from

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

to

$$\mathbf{y} = \mathbf{Xb} + .5(\mathbf{Z}_s\mathbf{a}_s + \mathbf{Z}_d\mathbf{a}_d) + \mathbf{m} + \mathbf{e}$$

where the subscripts s and d refer to sire and dam, respectively. The analysis is conducted by combining $(\mathbf{m} + \mathbf{e})$ into one residual term, say ϵ. Also,

$$Var\begin{pmatrix} \mathbf{a}_s \\ \mathbf{a}_d \\ \mathbf{m}+\mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{I}\sigma_s^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_d^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}(\sigma_m^2 + \sigma_e^2) \end{pmatrix},$$

which implies that sires are unrelated to each other or to any of the dams, and dams are unrelated. The sire variance is usually assumed to be equal to one quarter of the additive genetic variance. The dam variance generally equals one quarter of the additive genetic variance plus any common environmental effects, dominance genetic effects, and maternal genetic effects. Heritability is usually estimated from the sire variance for that reason.

In a sire and dam model, there is no interest in obtaining predictions for the progeny. However, the Mendelian sampling effects can be obtained by backsolving.

Sires and dams have usually been highly selected, especially sires, and so are not random samples of males or females amongst all males and females that are born in the population. They should not be random factors in the model, but more likely should be fixed factors.

87 Sire-Maternal Grandsire Models

The sire and dam model may be further simplified, when appropriate, to a sire-maternal grandsire model. The assumptions for this model are the same as those for the sire-dam model plus:

1. Each dam has only one progeny, and the dams of the dams (i.e. maternal granddams) have only one daughter, and
2. The daughters of a maternal grandsire (MGS) represented in the data are a random sample of all daughters of that MGS.

The last assumption is the most critical, and probably not valid. In dairy cattle, usually dams of cows are selected, and only good cows are allowed to have progeny. Poor cows are either culled or not allowed to produce replacement cows. For this reason, Sire-Maternal Grandsire models are not recommended in dairy cattle, and possibly other species too.

The $\mathbf{a}_d$ vector in the sire and dam model may be further partitioned as

$$\mathbf{a}_d = .5(\mathbf{a}_{mgs} + \mathbf{a}_{mgd}) + \mathbf{m}_d$$

so that if $\mathbf{Z}_{mgd} = \mathbf{I}$, and $\mathbf{Z}_d = \mathbf{I}$, then

$$\begin{aligned} \mathbf{y} &= \mathbf{Xb} + .5\mathbf{Z}_s\mathbf{a}_s \\ &\quad +.5(.5\mathbf{Z}_{mgs}\mathbf{a}_{mgs} + .5\mathbf{a}_{mgd} + \mathbf{m}_d) + \epsilon \\ &= \mathbf{Xb} + \mathbf{Z}_s(.5\mathbf{a}_s) \\ &\quad +.5\mathbf{Z}_s(.5\mathbf{a}_{mgs}) + (.25\mathbf{a}_{mgd} + \mathbf{m}_d + \epsilon). \end{aligned}$$

The vectors $.5\mathbf{a}_s$ and $.5\mathbf{a}_{mgs}$ contain many of the same sires, and so the two can be combined into one vector, say $\mathbf{s}$, and the two incidence matrices can also be combined into one matrix, say $\mathbf{Z}_{smgs}$, which contains a 1 in each row for the sire of an animal, and a .5 for the MGS of that animal. A computational requirement at this point is that the sire and MGS are not the same individual for any one animal. The combined model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_{smgs}\mathbf{s} + \xi.$$

The solutions for $\mathbf{s}$ are in terms of estimated transmitting abilities and must be doubled to give estimated breeding values.

88 Sire Models

The next progression away from the animal model is a sire model. More assumptions are needed than the previous models. The sire model was one of the first models used in animal breeding for the evaluation of dairy bulls for the milk producing ability of their daughters. The additional assumptions of the sire model are

1. Sires are mated randomly to dams, and
2. Dams are mated to only one sire and have just one progeny.

The sire model is relatively simple to apply. Relationships among sires are usually formed based on the sire and MGS of each bull, rather than sire and dam.

Because of the many assumptions that are implied with a sire model (when compared to an animal model), the use of a sire model in the genetic evaluation of animals should be a last resort. If at all possible, an animal model should be employed.

89 EXERCISES

1. Below are pedigrees and data on 20 animals in three contemporary groups.

Animal	Sire	Dam	Group	Record
1	-	-		
2	-	-		
3	-	-		
4	-	-		
5	1	2	1	40
6	1	2	1	28
7	3	4	1	34
8	1	2	1	35
9	1	4	2	17
10	3	2	2	41
11	1	4	2	25
12	3	2	2	38
13	1	2	2	29
14	3	4	2	27
15	5	7	3	37
16	6	14	3	30
17	8	7	3	42
18	1	10	3	46
19	3	13	3	24
20	5	9	3	26

   (a) Construct $\mathbf{A}^{-1}$ and set up the MME.

   (b) Apply the following model,

$$y_{ij} = \mu + g_i + a_j + e_{ij},$$

   where group, animal additive genetic, and residual effects are random. Let

$$\sigma_e^2/\sigma_a^2 = 1.5, \quad \sigma_e^2/\sigma_g^2 = 5.0.$$

   (c) Compute SEP and reliabilities for all animals.

   (d) Estimate the average EBVs for each contemporary group.

   (e) Apply a reduced animal model to the same data. Compare solutions.

2. Generate phenotypes for animals 7 to 16 in the table below. Contemporary groups (g_i) are random effects with variance, $\sigma_g^2 = 120$. Age effects are fixed with differences $A_1 = 0$, $A_2 = 15$, $A_3 = 20$, and $A_4 = 22$. Let the overall mean (μ) be 300, and the

additive genetic variance be 2000, and the residual variance be 6500. The model equation is

$$y_{ijk} = \mu \; + \; g_i \; + \; A_j \; + \; a_k \; + \; e_{ijk},$$

where a_k are the animal additive genetic values, and e_{ijk} is a residual effect.

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

(a) Analyze the data you have created with the assumed model to obtain EBV on the animals.

(b) Correlate the EBVs with the true breeding values (which are known in a simulation exercise).

(c) Repeat the generation of new data sets, in the same manner using a different set of random numbers, a total of 10 times, and average the correlation of EBV with true breeding values over the 10 replicates. What is the variability of the correlation?

(d) Analyze the data with an appropriate Sire-Dam model and compare the sire and dam solutions with the solutions from the Animal model.

Reduced Animal Models

90 Introduction

In situations where many offspring can be generated from one mating (as in fish, poultry, or swine), or where only a few animals are retained for breeding, the genetic evaluation of all animals may not be necessary. Only animals that are candidates for becoming the parents of the next generation need to be evaluated. Pollak and Quaas (1980) came up with the reduced animal model or RAM to cover this situation. Consider an animal model with periods as a fixed factor and one observation per animal, as in the table below.

Animal Model Example Data.

Animal	Sire	Dam	Period	Observation
5	1	3	2	250
6	1	3	2	198
7	2	4	2	245
8	2	4	2	260
9	2	4	2	235
4	-	-	1	255
3	-	-	1	200
2	-	-	1	225

90.1 Usual Animal Model Analysis

Assume that the ratio of residual to additive genetic variances is 2. The MME for this data would be of order 11 (nine animals and two periods). The left hand sides and right hand sides of the MME are

$$\left(\begin{array}{ccccccccccc}
3 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\
0 & 5 & 0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & 1 \\
0 & 0 & 4 & 0 & 2 & 0 & -2 & -2 & 0 & 0 & 0 \\
1 & 0 & 0 & 6 & 0 & 3 & 0 & 0 & -2 & -2 & -2 \\
1 & 0 & 2 & 0 & 5 & 0 & -2 & -2 & 0 & 0 & 0 \\
1 & 0 & 0 & 3 & 0 & 6 & 0 & 0 & -2 & -2 & -2 \\
0 & 1 & -2 & 0 & -2 & 0 & 5 & 0 & 0 & 0 & 0 \\
0 & 1 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 & 0 \\
0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 5 & 0 & 0 \\
0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 5 & 0 \\
0 & 1 & 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 5
\end{array}\right), \quad \left(\begin{array}{r}
680 \\ 1188 \\ 0 \\ 225 \\ 200 \\ 255 \\ 250 \\ 198 \\ 245 \\ 260 \\ 235
\end{array}\right)$$

and the solutions to these equations are

$$\left(\begin{array}{c} \hat{b}_1 \\ \hat{b}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \\ \hat{a}_5 \\ \hat{a}_6 \\ \hat{a}_7 \\ \hat{a}_8 \\ \hat{a}_9 \end{array} \right) = \left(\begin{array}{r} 225.8641 \\ 236.3366 \\ -2.4078 \\ 1.3172 \\ -10.2265 \\ 11.3172 \\ -2.3210 \\ -12.7210 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{array} \right).$$

A property of these solutions is that

$$\mathbf{1}'\mathbf{A^{-1}}\hat{\mathbf{a}} = \mathbf{0},$$

which in this case means that the sum of solutions for animals 1 through 4 is zero.

90.2 Reduced AM

RAM results in fewer equations to be solved, but the solutions from RAM are exactly the same as from the full MME. In a typical animal model with $\mathbf{a}$ as the vector of additive genetic values of animals, there will be animals that have had progeny, and there will be other animals that have not yet had progeny (and some may never have progeny). Denote animals with progeny as $\mathbf{a_p}$, and those without progeny as $\mathbf{a_o}$, so that

$$\mathbf{a}' = \left(\begin{array}{cc} \mathbf{a_p}' & \mathbf{a_o}' \end{array} \right).$$

In terms of the example data,

$$\begin{aligned} \mathbf{a_p}' &= \left(\begin{array}{cccc} a_1 & a_2 & a_3 & a_4 \end{array} \right), \\ \mathbf{a_o}' &= \left(\begin{array}{ccccc} a_5 & a_6 & a_7 & a_8 & a_9 \end{array} \right). \end{aligned}$$

Genetically for any individual, i, the additive genetic value may be written as the average of the additive genetic values of the parents plus a Mendelian sampling effect, which is the animal's specific deviation from the parent average, i.e.

$$a_i = .5(a_s + a_d) + m_i.$$

Therefore,

$$\mathbf{a}_o = \mathbf{T}\mathbf{a}_p + \mathbf{m},$$

where $\mathbf{T}$ is a matrix that indicates the parents of each animal in $\mathbf{a}_o$, and $\mathbf{m}$ is the vector of Mendelian sampling effects. Then

$$\begin{aligned}\mathbf{a} &= \begin{pmatrix}\mathbf{a}_p \\ \mathbf{a}_o\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{I} \\ \mathbf{T}\end{pmatrix}\mathbf{a}_p + \begin{pmatrix}\mathbf{0} \\ \mathbf{m}\end{pmatrix},\end{aligned}$$

and

$$\begin{aligned}Var(\mathbf{a}) &= \mathbf{A}\sigma_a^2 \\ &= \begin{pmatrix}\mathbf{I} \\ \mathbf{T}\end{pmatrix}\mathbf{A}_{pp}\begin{pmatrix}\mathbf{I} & \mathbf{T}'\end{pmatrix}\sigma_a^2 + \begin{pmatrix}\mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}\end{pmatrix}\sigma_a^2\end{aligned}$$

where $\mathbf{D}$ is a diagonal matrix with diagonal elements equal to $(1 - .25d_i)$, and d_i is the number of identified parents, i.e. 0, 1, or 2, for the i^{th} animal, and

$$Var(\mathbf{a}_p) = \mathbf{A}_{pp}\sigma_a^2.$$

The animal model can now be written as

$$\begin{pmatrix}\mathbf{y}_p \\ \mathbf{y}_o\end{pmatrix} = \begin{pmatrix}\mathbf{X}_p \\ \mathbf{X}_o\end{pmatrix}\mathbf{b} + \begin{pmatrix}\mathbf{Z}_p & \mathbf{0} \\ \mathbf{0} & \mathbf{Z}_o\end{pmatrix}\begin{pmatrix}\mathbf{I} \\ \mathbf{T}\end{pmatrix}\mathbf{a}_p + \begin{pmatrix}\mathbf{e}_p \\ \mathbf{e}_o + \mathbf{Z}_o\mathbf{m}\end{pmatrix}.$$

Note that the residual vector has two different types of residuals and that the additive genetic values of animals without progeny have been replaced with $\mathbf{Ta}_p$. Because every individual has only one record, then $\mathbf{Z}_o = \mathbf{I}$, but $\mathbf{Z}_p$ may have fewer rows than there are elements of $\mathbf{a}_p$ because not all parents may have observations themselves. In the example data, animal 1 does not have an observation, therefore,

$$\mathbf{Z}_p = \begin{pmatrix}0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1\end{pmatrix}.$$

Consequently,

$$\begin{aligned}\mathbf{R} &= Var\begin{pmatrix}\mathbf{e}_p \\ \mathbf{e}_o + \mathbf{m}\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{I}\sigma_e^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_e^2 + \mathbf{D}\sigma_a^2\end{pmatrix} \\ &= \begin{pmatrix}\mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{R}_o\end{pmatrix}\sigma_e^2\end{aligned}$$

The mixed model equations for the reduced animal model are

$$\begin{pmatrix}\mathbf{X}_p'\mathbf{X}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{X}_p'\mathbf{Z}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{T} \\ \mathbf{Z}_p'\mathbf{X}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{X}_o & \mathbf{Z}_p'\mathbf{Z}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{T} + \mathbf{A}_{pp}^{-1}\alpha\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}} \\ \hat{\mathbf{a}}_p\end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}_p'\mathbf{y}_p + \mathbf{X}_o'\mathbf{R}_o^{-1}\mathbf{y}_o \\ \mathbf{Z}_p'\mathbf{y}_p + \mathbf{T}'\mathbf{R}_o^{-1}\mathbf{y}_o \end{pmatrix}.$$

Solutions for $\hat{\mathbf{a}}_o$ are derived from the following formulas.

$$\hat{\mathbf{a}}_o = \mathbf{T}\hat{\mathbf{a}}_p + \hat{\mathbf{m}},$$

where

$$\hat{\mathbf{m}} = (\mathbf{Z}_o'\mathbf{Z}_o + \mathbf{D}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\hat{\mathbf{b}} - \mathbf{T}\hat{\mathbf{a}}_p).$$

Using the example data,

$$\mathbf{T} = \begin{pmatrix} .5 & 0 & .5 & 0 \\ .5 & 0 & .5 & 0 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \\ 0 & .5 & 0 & .5 \end{pmatrix},$$

and

$$\mathbf{D} = \text{diag}\begin{pmatrix} .5 & .5 & .5 & .5 & .5 \end{pmatrix},$$

then the MME with $\alpha = 2$ are

$$\begin{pmatrix} 3 & 0 & 0 & 1 & 1 & 1 \\ 0 & 4 & .8 & 1.2 & .8 & 1.2 \\ 0 & .8 & 2.4 & 0 & .4 & 0 \\ 1 & 1.2 & 0 & 3.6 & 0 & .6 \\ 1 & .8 & .4 & 0 & 3.4 & 0 \\ 1 & 1.2 & 0 & .6 & 0 & 3.6 \end{pmatrix} \begin{pmatrix} \hat{b}_1 \\ \hat{b}_2 \\ \hat{a}_1 \\ \hat{a}_2 \\ \hat{a}_3 \\ \hat{a}_4 \end{pmatrix} = \begin{pmatrix} 680. \\ 950.4 \\ 179.2 \\ 521. \\ 379.2 \\ 551. \end{pmatrix}$$

The solutions are as before, i.e.

$$\begin{array}{lll} \hat{b}_1 = 225.8641, & \hat{a}_1 = \text{-}2.4078, & \hat{a}_3 = \text{-}10.2265, \\ \hat{b}_2 = 236.3366, & \hat{a}_2 = 1.3172, & \hat{a}_4 = 11.3172. \end{array}$$

91 Backsolving for Omitted Animals

To compute $\hat{\mathbf{a}}_o$, first calculate $\hat{\mathbf{m}}$ as:

$$(\mathbf{I} + \mathbf{D}^{-1}\alpha) \quad = \quad \begin{pmatrix} 5 & 0 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 0 & 5 \end{pmatrix}$$

$$\begin{aligned}
\mathbf{y}_o &= \left(\begin{array}{c} 250 \\ 198 \\ 245 \\ 260 \\ 235 \end{array}\right) \\
\mathbf{X}_o\hat{\mathbf{b}} &= \left(\begin{array}{cc} 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right) \left(\begin{array}{c} 225.8641 \\ 236.3366 \end{array}\right) \\
\mathbf{T}\hat{\mathbf{a}}_p &= \left(\begin{array}{r} -6.3172 \\ -6.3172 \\ 6.3172 \\ 6.3172 \\ 6.3172 \end{array}\right) \\
\hat{\mathbf{m}} &= (\mathbf{I} + \mathbf{D}^{-1}\alpha)^{-1}(\mathbf{y}_o - \mathbf{X}_o\hat{\mathbf{b}} - \mathbf{T}\hat{\mathbf{a}}_p) \\
&= \left(\begin{array}{r} 3.9961 \\ -6.4039 \\ .4692 \\ 3.4692 \\ -1.5308 \end{array}\right)
\end{aligned}$$

and

$$\mathbf{T}\hat{\mathbf{a}}_p + \hat{\mathbf{m}} = \left(\begin{array}{r} -2.3211 \\ -12.7211 \\ 6.7864 \\ 9.7864 \\ 4.7864 \end{array}\right).$$

The reduced animal model was originally described for models where animals had only one observation, but Henderson(1988) described many other possible models where this technique could be applied. Generally, with today's computers there is not much problem in applying regular animal models without the need to employ a reduced animal model.

92 EXERCISES

Below are data on animals with their pedigrees.

Animal	Sire	Dam	Year	Group	Observation
1	-	-	1920	1	24
2	-	-	1920	1	12
3	-	-	1920	1	33
4	-	-	1920	2	27
5	-	-	1920	2	8
6	-	-	1920	2	19
7	1	4	1922	3	16
8	1	4	1922	3	28
9	1	4	1922	3	30
10	1	5	1922	3	42
11	1	6	1922	3	37
12	1	6	1922	3	44
13	2	4	1922	4	11
14	2	4	1922	4	18
15	2	4	1922	4	23
16	2	5	1922	4	9
17	2	5	1922	4	2
18	2	6	1922	4	25
19	7	16	1924	5	14
20	7	16	1924	5	19
21	7	16	1924	5	17
22	10	13	1924	5	39
23	10	13	1924	5	43

Assume a heritability of 0.32 for this trait.

Analyze the data with both the usual animal model, and the reduced animal model.

$$y_{ijk} = Y_i + G_j + a_k + e_{ijk},$$

where Y_i is a year effect, G_j is a group effect, a_k is an animal effect, and e_{ijk} is a residual effect.

The solutions to both analyses should be identical.

In the RAM, backsolve for $\hat{a}_{23}$.

What about the prediction error variance for $\hat{a}_{23}$?

Estimation of Variances and Covariances

93 Variables and Distributions

Random variables are samples from a population with a given set of population parameters. Random variables can be **discrete**, having a limited number of distinct possible values, or **continuous**.

94 Continuous Random Variables

The *cumulative distribution function* of a random variable is

$$F(y) = Pr(Y \leq y),$$

for $-\infty < y < \infty$.

As y approaches $-\infty$, then $F(y)$ approaches 0. As y approaches ∞, then $F(y)$ approaches 1.

$F(y)$ is a nondecreasing function of y. If $a < b$, then $F(a) < F(b)$.

$p(y) = \frac{\partial F(y)}{\partial y} = F'(y)$, wherever the derivative exists.

$\int_{-\infty}^{\infty} p(y) \, \partial y = 1$.

$F(t) = \int_{-\infty}^{t} p(y) \, \partial y$.

$E(y) = \int_{-\infty}^{\infty} y \, p(y) \, \partial y$

$E(g(y)) = \int_{-\infty}^{\infty} g(y) \, p(y) \, \partial y$

$Var(y) = E(y^2) - [E(y)]^2$.

94.1 Normal Random Variables

Random variables in animal breeding problems are typically assumed to be samples from Normal distributions, where

$$p(y) = (2\pi)^{-.5}\sigma^{-1}\exp(-.5(y-\mu)^2\sigma^{-2})$$

for $-\infty < x < +\infty$, where σ^2 is the variance of y and μ is the expected value of y.

For a random vector variable, $\mathbf{y}$, the multivariate normal density function is

$$p(\mathbf{y}) = (2\pi)^{-.5n} \mid \mathbf{V} \mid^{-.5} \exp(-.5(\mathbf{y}-\boldsymbol{\mu})'\mathbf{V}^{-1}(\mathbf{y}-\boldsymbol{\mu}))$$

denoted as $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$ where $\mathbf{V}$ is the variance-covariance matrix of $\mathbf{y}$. The determinant of $\mathbf{V}$ must be positive, otherwise the density function is undefined.

94.2 Chi-Square Random Variables

If $\mathbf{y} \sim N(\mathbf{0}, \mathbf{I})$, then $\mathbf{y}'\mathbf{y} \sim \chi^2_n$, where χ^2_n is a *central chi-square* distribution with n degrees of freedom and n is the length of the random vector variable $\mathbf{y}$.

The mean is n. The variance is $2n$. If $s = \mathbf{y}'\mathbf{y} > 0$, then

$$p(s \mid n) = (s)^{(n/2)-1} \exp -0.5s/[2^{0.5n}\Gamma(0.5n)].$$

If $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{I})$, then $\mathbf{y}'\mathbf{y} \sim \chi^2_{n,\lambda}$ where λ is the noncentrality parameter which is equal to $.5\boldsymbol{\mu}'\boldsymbol{\mu}$. The mean of a noncentral chi-square distribution is $n + 2\lambda$ and the variance is $2n + 8\lambda$.

If $\mathbf{y} \sim N(\boldsymbol{\mu}, \mathbf{V})$, then $\mathbf{y}'\mathbf{Q}\mathbf{y}$ has a noncentral chi-square distribution only if $\mathbf{QV}$ is idempotent, i.e. $\mathbf{QVQV} = \mathbf{QV}$. The noncentrality parameter is $\lambda = .5\boldsymbol{\mu}'\mathbf{QVQ}\boldsymbol{\mu}$ and the mean and variance of the distribution are $tr(\mathbf{QV}) + 2\lambda$ and $2tr(\mathbf{QV}) + 8\lambda$, respectively.

If there are two quadratic forms of $\mathbf{y}$, say $\mathbf{y}'\mathbf{Q}\mathbf{y}$ and $\mathbf{y}'\mathbf{P}\mathbf{y}$, and both quadratic forms have chi-square distributions, then the two quadratic forms are independent if $\mathbf{QVP} = \mathbf{0}$.

94.3 The Wishart Distribution

The Wishart distribution is similar to a multivariate Chi-square distribution. An entire matrix is envisioned of which the diagonals have a Chi-square distribution, and the off-diagonals have a built-in correlation structure. The resulting matrix is positive definite. This distribution is needed when estimating covariance matrices, such as in multiple trait models, maternal genetic effect models, or random regression models.

94.4 The F-distribution

The F-distribution is used for hypothesis testing and is built upon two independent Chi-square random variables. Let $s \sim \chi^2_n$ and $v \sim \chi^2_m$ with s and v being independent, then

$$\frac{(s/n)}{(v/m)} \sim F_{n,m}.$$

The mean of the F-distribution is $m/(m-2)$. The variance is

$$\frac{2m^2(n+m-2)}{n(m-2)^2(m-4)}.$$

95 Expectations of Random Vectors

Let $\mathbf{y}_1$ be a random vector variable, then

$$E(\mathbf{y}_1) = \boldsymbol{\mu}_1 = \begin{pmatrix} \mu_{11} \\ \mu_{12} \\ \vdots \\ \mu_{1n} \end{pmatrix},$$

for a vector of length n. If c is a scalar constant, then

$$E(c\mathbf{y}_1) = c\boldsymbol{\mu}_1.$$

Similarly, if $\mathbf{C}$ is a matrix of constants, then

$$E(\mathbf{C}\mathbf{y}_1) = \mathbf{C}\boldsymbol{\mu}_1.$$

Let $\mathbf{y}_2$ be another random vector variable of the same length as $\mathbf{y}_1$, then

$$\begin{aligned} E(\mathbf{y}_1 + \mathbf{y}_2) &= E(\mathbf{y}_1) + E(\mathbf{y}_2) \\ &= \boldsymbol{\mu}_1 + \boldsymbol{\mu}_2. \end{aligned}$$

96 Variance-Covariance Matrices

Let $\mathbf{y}$ be a random vector variable of length n, then the *variance-covariance* matrix of $\mathbf{y}$ is:

$$\begin{aligned} Var(\mathbf{y}) &= E(\mathbf{y}\mathbf{y}') - E(\mathbf{y})E(\mathbf{y}') \\ &= \begin{pmatrix} \sigma^2_{y_1} & \sigma_{y_1y_2} & \cdots & \sigma_{y_1y_n} \\ \sigma_{y_1y_2} & \sigma^2_{y_2} & \cdots & \sigma_{y_2y_n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{y_1y_n} & \sigma_{y_2y_n} & \cdots & \sigma^2_{y_n} \end{pmatrix} \\ &= \mathbf{V} \end{aligned}$$

A variance-covariance (VCV) matrix is square, symmetric and should always be positive definite, i.e. all of the eigenvalues must be positive.

Another name for VCV matrix is a *dispersion* matrix or (co)variance matrix.

Let $\mathbf{C}$ be a matrix of constants conformable for multiplication with the vector $\mathbf{y}$, then

$$\begin{aligned} Var(\mathbf{Cy}) &= E(\mathbf{Cyy'C'}) - E(\mathbf{Cy})E(\mathbf{y'C'}) \\ &= \mathbf{C}E(\mathbf{yy'})\mathbf{C'} - \mathbf{C}E(\mathbf{y})E(\mathbf{y'})\mathbf{C'} \\ &= \mathbf{C}\left(E(\mathbf{yy'}) - E(\mathbf{y})E(\mathbf{y'})\right)\mathbf{C'} \\ &= \mathbf{C}Var(\mathbf{y})\mathbf{C'} = \mathbf{CVC'}. \end{aligned}$$

If there are two sets of functions of $\mathbf{y}$, say $\mathbf{C}_1\mathbf{y}$ and $\mathbf{C}_2\mathbf{y}$, then

$$Cov(\mathbf{C}_1\mathbf{y}, \mathbf{C}_2\mathbf{y}) = \mathbf{C}_1\mathbf{V}\mathbf{C}'_2.$$

If $\mathbf{y}$ and $\mathbf{z}$ represent two different random vectors, possibly of different orders, and if the (co)variance matrix between these two vectors is $\mathbf{W}$, then

$$Cov(\mathbf{C}_1\mathbf{y}, \mathbf{C}_2\mathbf{z}) = \mathbf{C}_1\mathbf{W}\mathbf{C}'_2.$$

97 Quadratic Forms

Variances are estimated using sums of squares of various normally distributed variables, and these are known as *quadratic forms*. The general quadratic form is

$$\mathbf{y'Qy},$$

where $\mathbf{y}$ is a random vector variable, and $\mathbf{Q}$ is a regulator matrix. Usually $\mathbf{Q}$ is a symmetric matrix, but not necessarily positive definite.

Examples of different $\mathbf{Q}$ matrices are as follows:

1. $\mathbf{Q} = \mathbf{I}$, then $\mathbf{y'Qy} = \mathbf{y'y}$ which is a total sum of squares of the elements in $\mathbf{y}$.
2. $\mathbf{Q} = \mathbf{J}(1/n)$, then $\mathbf{y'Qy} = \mathbf{y'Jy}(1/n)$ where n is the length of $\mathbf{y}$. Note that $\mathbf{J} = \mathbf{11'}$, so that $\mathbf{y'Jy} = (\mathbf{y'1})(\mathbf{1'y})$ and $(\mathbf{1'y})$ is the sum of the elements in $\mathbf{y}$.
3. $\mathbf{Q} = (\mathbf{I} - \mathbf{J}(1/n))/(n-1)$, then $\mathbf{y'Qy}$ gives the variance of the elements in $\mathbf{y}$, σ^2_y.

The expected value of a quadratic form is

$$E(\mathbf{y'Qy}) = E(tr(\mathbf{y'Qy})) = E(tr(\mathbf{Qyy'})) = tr(\mathbf{Q}E(\mathbf{yy'})),$$

and the covariance matrix is

$$Var(\mathbf{y}) = E(\mathbf{yy'}) - E(\mathbf{y})E(\mathbf{y'})$$

so that

$$E(\mathbf{yy}') = Var(\mathbf{y}) + E(\mathbf{y})E(\mathbf{y}'),$$

then

$$E(\mathbf{y}'\mathbf{Q}\mathbf{y}) = tr(\mathbf{Q}(Var(\mathbf{y}) + E(\mathbf{y})E(\mathbf{y}'))).$$

Let $Var(\mathbf{y}) = \mathbf{V}$ and $E(\mathbf{y}) = \boldsymbol{\mu}$, then

$$\begin{array}{rcl} E(\mathbf{y}'\mathbf{Q}\mathbf{y}) & = & tr(\mathbf{Q}(\mathbf{V} + \boldsymbol{\mu}\boldsymbol{\mu}\text{'})) \\ & = & tr(\mathbf{QV}) + tr(\mathbf{Q}\boldsymbol{\mu}\boldsymbol{\mu}\text{'}) \\ & = & tr(\mathbf{QV}) + \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu}. \end{array}$$

The expectation of a quadratic form was derived without knowing the distribution of $\mathbf{y}$. However, the variance of a quadratic form requires that $\mathbf{y}$ follows a multivariate normal distribution. Without showing the derivation, the variance of a quadratic form, assuming $\mathbf{y}$ has a multivariate normal distribution, is

$$Var(\mathbf{y}'\mathbf{Q}\mathbf{y}) = 2tr(\mathbf{QVQV}) + 4\boldsymbol{\mu}'\mathbf{QVQ}\boldsymbol{\mu}.$$

The quadratic form, $\mathbf{y}'\mathbf{Q}\mathbf{y}$, has a chi-square distribution if

$$tr(\mathbf{QVQV}) = tr(\mathbf{QV}), \text{ and } \boldsymbol{\mu}'\mathbf{QVQ}\boldsymbol{\mu} = \boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu},$$

or the single condition that $\mathbf{QV}$ is idempotent. Then if

$$m = tr(\mathbf{QV}) \text{ and } \lambda = .5\boldsymbol{\mu}'\mathbf{Q}\boldsymbol{\mu},$$

the expected value of $\mathbf{y}'\mathbf{Q}\mathbf{y}$ is $m + 2\lambda$ and the variance is $2m + 8\lambda$, which are the usual results for a noncentral chi-square variable.

The covariance between two quadratic forms, say $\mathbf{y}'\mathbf{Q}\mathbf{y}$ and $\mathbf{y}'\mathbf{P}\mathbf{y}$, is

$$Cov(\mathbf{y}'\mathbf{Q}\mathbf{y}, \mathbf{y}'\mathbf{P}\mathbf{y}) = 2tr(\mathbf{QVPV}) + 4\boldsymbol{\mu}'\mathbf{QVP}\boldsymbol{\mu}.$$

The covariance is zero if $\mathbf{QVP} = \mathbf{0}$, then the two quadratic forms are said to be independent.

98 Basic Model for Variance Components

The general linear model is described as

$$\begin{array}{rcl} \mathbf{y} & = & \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \\ \text{where } E(\mathbf{y}) & = & \mathbf{Xb}, \\ E(\mathbf{u}) & = & \mathbf{0}, \\ \text{and } E(\mathbf{e}) & = & \mathbf{0}. \end{array}$$

Often $\mathbf{u}$ is partitioned into s factors as

$$\mathbf{u}' = (\mathbf{u}_1' \;\; \mathbf{u}_2' \; \ldots \; \mathbf{u}_s' \;).$$

The (co)variance matrices are defined as

$$\mathbf{G} = Var(\mathbf{u}) = Var\begin{pmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_s \end{pmatrix} = \begin{pmatrix} \mathbf{G}_1\sigma_1^2 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \mathbf{G}_2\sigma_2^2 & \ldots & \mathbf{0} \\ \vdots & \vdots & & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{G}_s\sigma_s^2 \end{pmatrix}$$

and

$$\mathbf{R} = Var(\mathbf{e}) = \mathbf{I}\sigma_0^2.$$

Then

$$Var(\mathbf{y}) = \mathbf{V} = \mathbf{ZGZ}' + \mathbf{R},$$

and if $\mathbf{Z}$ is partitioned corresponding to $\mathbf{u}$, as

$$\begin{aligned} \mathbf{Z} &= [\ \mathbf{Z}_1 \;\; \mathbf{Z}_2 \; \ldots \; \mathbf{Z}_s\ ], \text{ then} \\ \mathbf{ZGZ}' &= \sum_{i=1}^{s} \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'\sigma_i^2. \\ \text{Let } \mathbf{V}_i &= \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i' \text{ and} \\ \mathbf{V}_0 &= \mathbf{I}, \text{ then} \\ \mathbf{V} &= \sum_{i=o}^{s} \mathbf{V}_i\sigma_i^2. \end{aligned}$$

98.1 Mixed Model Equations

Henderson's mixed model equations (MME) are written as

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2 & \ldots & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_s \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1 + \mathbf{G}_1^{-1}\sigma_1^{-2} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2 & \ldots & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_s \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2 + \mathbf{G}_2^{-1}\sigma_2^{-2} & \ldots & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_s \\ \vdots & \vdots & \vdots & & \vdots \\ \mathbf{Z}_s'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_s'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_s'\mathbf{R}^{-1}\mathbf{Z}_2 & \ldots & \mathbf{Z}_s'\mathbf{R}^{-1}\mathbf{Z}_s + \mathbf{G}_s^{-1}\sigma_s^{-2} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}}_1 \\ \hat{\mathbf{u}}_2 \\ \vdots \\ \hat{\mathbf{u}}_s \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y} \\ \vdots \\ \mathbf{Z}_s'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}$$

99 Unbiased Estimation of Variances

Assume that all $\mathbf{G}_i$ are equal to $\mathbf{I}$ for this example, so that $\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'$ simplifies to $\mathbf{Z}_i\mathbf{Z}_i'$. Let

$$\begin{aligned}
\mathbf{X} &= \left(\begin{array}{c} 1 \\ 1 \\ 1 \end{array}\right), \ \mathbf{Z}_1 = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{array}\right), \\
\mathbf{Z}_2 &= \left(\begin{array}{cc} 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{array}\right), \ \text{and } \mathbf{y} = \left(\begin{array}{c} 29 \\ 53 \\ 44 \end{array}\right),
\end{aligned}$$

Then

$$\mathbf{V}_1 = \mathbf{Z}_1\mathbf{Z}_1' = \left(\begin{array}{ccc} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{array}\right),$$

and

$$\mathbf{V}_2 = \mathbf{Z}_2\mathbf{Z}_2' = \left(\begin{array}{ccc} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{array}\right)$$

and $\mathbf{V}_0 = \mathbf{I}$.

99.1 Define the Necessary Quadratic Forms

At least three quadratic forms are needed in order to estimate the variances. Below are three arbitrary $\mathbf{Q}$-matrices that were chosen such that $\mathbf{Q}_k\mathbf{X} = \mathbf{0}$. Let

$$\begin{aligned}
\mathbf{Q}_1 &= \left(\begin{array}{rrr} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{array}\right), \\
\mathbf{Q}_2 &= \left(\begin{array}{rrr} 1 & 0 & -1 \\ 0 & 1 & -1 \\ -1 & -1 & 2 \end{array}\right), \\
\text{and } \mathbf{Q}_3 &= \left(\begin{array}{rrr} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{array}\right).
\end{aligned}$$

The numeric values of the quadratic forms are

$$\begin{aligned}
\mathbf{y}'\mathbf{Q}_1\mathbf{y} &= 657, \\
\mathbf{y}'\mathbf{Q}_2\mathbf{y} &= 306, \\
\text{and } \mathbf{y}'\mathbf{Q}_3\mathbf{y} &= 882.
\end{aligned}$$

For example,

$$\mathbf{y}'\mathbf{Q}_1\mathbf{y} = \begin{pmatrix} 29 & 53 & 44 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} 29 \\ 53 \\ 44 \end{pmatrix} = 657.$$

99.2 The Expectations of the Quadratic Forms

The expectations of the quadratic forms are

$$\begin{aligned} E(\mathbf{y}'\mathbf{Q}_1\mathbf{y}) &= tr\mathbf{Q}_1\mathbf{V}_0\sigma_0^2 \;+\; tr\mathbf{Q}_1\mathbf{V}_1\sigma_1^2 \;+\; tr\mathbf{Q}_1\mathbf{V}_2\sigma_2^2 \\ &= 4\sigma_0^2 \;+\; 2\sigma_1^2 \;+\; 2\sigma_2^2 \\ E(\mathbf{y}'\mathbf{Q}_2\mathbf{y}) &= 4\sigma_0^2 \;+\; 4\sigma_1^2 \;+\; 2\sigma_2^2, \\ E(\mathbf{y}'\mathbf{Q}_3\mathbf{y}) &= 6\sigma_0^2 \;+\; 4\sigma_1^2 \;+\; 4\sigma_2^2. \end{aligned}$$

99.3 Equate Expected Values to Numerical Values

Equate the numeric values of the quadratic forms to their corresponding expected values, which gives a system of equations to be solved, such as $\mathbf{F}\sigma = \mathbf{w}$. In this case, the equations would be

$$\begin{pmatrix} 4 & 2 & 2 \\ 4 & 4 & 2 \\ 6 & 4 & 4 \end{pmatrix} \begin{pmatrix} \sigma_0^2 \\ \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} = \begin{pmatrix} 657. \\ 306. \\ 882. \end{pmatrix},$$

which gives the solution as $\hat{\sigma} = \mathbf{F}^{-1}\mathbf{w}$, or

$$\begin{pmatrix} \hat{\sigma}_0^2 \\ \hat{\sigma}_1^2 \\ \hat{\sigma}_2^2 \end{pmatrix} = \begin{pmatrix} 216.0 \\ -175.5 \\ 72.0 \end{pmatrix}.$$

The resulting estimates are *unbiased.*

99.4 Variances of Quadratic Forms

The variance of a quadratic form is

$$Var(\mathbf{y}'\mathbf{Q}\mathbf{y}) = 2tr\mathbf{QVQV} + 4\mathbf{b}'\mathbf{X}'\mathbf{QVQXb}.$$

Only translation invariant quadratic forms are typically considered in variance component estimation, that means $\mathbf{b}'\mathbf{X}'\mathbf{QVQXb} = 0$. Thus, only $2tr\mathbf{QVQV}$ needs to be calculated.

Remember that $\mathbf{V}$ can be written as the sum of $s+1$ matrices, $\mathbf{V}_i\sigma_i^2$, then

$$\begin{aligned} tr\mathbf{QVQV} &= tr\mathbf{Q}\sum_{i=o}^{s}\mathbf{V}_i\sigma_i^2\ \mathbf{Q}\sum_{j=o}^{s}\mathbf{V}_j\sigma_j^2 \\ &= \sum_{i=o}^{s}\sum_{j=o}^{s} tr\ \mathbf{QV}_i\mathbf{QV}_j\ \sigma_i^2\ \sigma_j^2 \end{aligned}$$

For example, if $s=2$, then

$$\begin{aligned} tr\mathbf{QVQV} &= tr\mathbf{QV}_0\mathbf{QV}_0\sigma_0^4\ +\ 2tr\mathbf{QV}_0\mathbf{QV}_1\sigma_0^2\sigma_1^2 \\ &\quad +\ 2tr\mathbf{QV}_0\mathbf{QV}_2\sigma_0^2\sigma_2^2 +\ tr\mathbf{QV}_1\mathbf{QV}_1\sigma_1^4 \\ &\quad +\ 2tr\mathbf{QV}_1\mathbf{QV}_2\sigma_1^2\sigma_2^2\ +\ tr\mathbf{QV}_2\mathbf{QV}_2\sigma_2^4. \end{aligned}$$

The sampling variances depend on

1. The true magnitude of the individual components,
2. The matrices $\mathbf{Q}_k$, which depend on the method of estimation and the model, and
3. The structure and amount of the data through $\mathbf{X}$ and $\mathbf{Z}$.

For small examples, the calculation of sampling variances can be easily demonstrated. In this case,

$$Var(\mathbf{F}^{-1}\mathbf{w}) = \mathbf{F}^{-1}Var(\mathbf{w})\mathbf{F}^{-1'},$$

a function of the variance-covariance matrix of the quadratic forms.

Using the small example of the previous section, the $Var(\mathbf{w})$ is a 3x3 matrix. The (1,1) element is the variance of $\mathbf{y}'\mathbf{Q}_1\mathbf{y}$ which is

$$\begin{aligned} Var(\mathbf{y}'\mathbf{Q}_1\mathbf{y}) &= 2tr\mathbf{Q}_1\mathbf{VQ}_1\mathbf{V} \\ &= 2tr\mathbf{Q}_1\mathbf{V}_0\mathbf{Q}_1\mathbf{V}_0\sigma_0^4\ +\ 4tr\mathbf{Q}_1\mathbf{V}_0\mathbf{Q}_1\mathbf{V}_1\sigma_0^2\sigma_1^2 \\ &\quad +4tr\mathbf{Q}_1\mathbf{V}_0\mathbf{Q}_1\mathbf{V}_2\sigma_0^2\sigma_2^2\ +\ 2tr\mathbf{Q}_1\mathbf{V}_1\mathbf{Q}_1\mathbf{V}_1\sigma_1^4 \\ &\quad +4tr\mathbf{Q}_1\mathbf{V}_1\mathbf{Q}_1\mathbf{V}_2\sigma_1^2\sigma_2^2\ +\ 2tr\mathbf{Q}_1\mathbf{V}_2\mathbf{Q}_1\mathbf{V}_2\sigma_2^4 \\ &= 20\sigma_0^4\ +\ 16\sigma_0^2\sigma_1^2\ +\ 16\sigma_0^2\sigma_2^2\ +\ 8\sigma_1^4\ +\ 0\sigma_1^2\sigma_2^2\ +\ 8\sigma_2^4 \end{aligned}$$

The (1,2) element is the covariance between the first and second quadratic forms,

$$Cov(\mathbf{y}'\mathbf{Q}_1\mathbf{y}, \mathbf{y}'\mathbf{Q}_2\mathbf{y}) = 2tr\mathbf{Q}_1\mathbf{VQ}_2\mathbf{V},$$

and similarly for the other terms. All of the results are summarized in the table below.

Forms	σ_0^4	$\sigma_0^2\sigma_1^2$	$\sigma_0^2\sigma_2^2$	σ_1^4	$\sigma_1^2\sigma_2^2$	σ_2^4
$Var(w_1)$	20	16	16	8	0	8
$Cov(w_1, w_2)$	14	24	8	16	0	8
$Cov(w_1, w_3)$	24	24	24	16	0	16
$Var(w_2)$	20	48	16	32	16	8
$Cov(w_2, w_3)$	24	48	24	32	16	16
$Var(w_3)$	36	48	48	32	16	32

To get numeric values for these variances, the true components need to be known. Assume that the true values are $\sigma_0^2 = 250$, $\sigma_1^2 = 10$, and $\sigma_2^2 = 80$, then the variance of w_1 is

$$\begin{aligned} Var(w_1) &= 20(250)^2 + 16(250)(10) + 16(250)(80) \\ &\quad +8(10)^2 + 0(10)(80) + 8(80)^2 \\ &= 1,662,000. \end{aligned}$$

The complete variance- covariance matrix of the quadratic forms is

$$Var \begin{pmatrix} w_1 \\ w_2 \\ w_3 \end{pmatrix} = \begin{pmatrix} 1,662,000 & 1,147,800 & 2,144,000 \\ 1,147,800 & 1,757,200 & 2,218,400 \\ 2,144,000 & 2,218,400 & 3,550,800 \end{pmatrix}.$$

The variance-covariance matrix of the estimated variances (assuming the above true values) would be

$$\begin{aligned} Var(\hat{\sigma}) &= \mathbf{F}^{-1} Var(\mathbf{w}) \mathbf{F}^{-1'} \\ &= \begin{pmatrix} 405,700 & -275,700 & -240,700 \\ -275,700 & 280,900 & 141,950 \\ -240,700 & 141,950 & 293,500 \end{pmatrix} = \mathbf{C}. \end{aligned}$$

99.5 Variance of A Ratio of Variance Estimates

Often estimates of ratios of functions of the variances are needed for animal breeding work, such as heritabilities, repeatabilities, and variance ratios. Let such a ratio be denoted as a/c where

$$a = \hat{\sigma_2^2} = (0 \;\; 0 \;\; 1)\hat{\sigma} = 72.$$

and

$$c = \hat{\sigma_0^2} + \hat{\sigma_1^2} + \hat{\sigma_2^2} = (1 \;\; 1 \;\; 1)\hat{\sigma} = 288.$$

(NOTE: the negative estimate for $\hat{\sigma}_1^2$ was set to zero before calculating c.

From Osborne and Patterson (1952) and Rao (1968) an approximation to the variance of a ratio is given by

$$Var(a/c) = (c^2 Var(a) + a^2 Var(c) - 2ac\ Cov(a, c))/c^4.$$

Now note that

$$\begin{aligned} Var(a) &= (0 \;\; 0 \;\; 1)\mathbf{C}(0 \;\; 0 \;\; 1)' = 293,500, \\ Var(c) &= (1 \;\; 1 \;\; 1)\mathbf{C}(1 \;\; 1 \;\; 1)' = 231,200, \\ Cov(a, c,) &= (0 \;\; 0 \;\; 1)\mathbf{C}(1 \;\; 1 \;\; 1)' = 194,750. \end{aligned}$$

Then

$$\begin{aligned} Var(a/c) &= [(288)^2(293,500) + (72)^2(231,200) \\ &\quad -2(72)(288)(194,750)]/(288)^4 \\ &= 2.53876 \end{aligned}$$

This result is very large, but could be expected from only 3 observations. Thus, $(a/c) = .25$ with a standard deviation of 1.5933.

Another approximation method assumes that the denominator has been estimated accurately, so that it is considered to be a constant, such as the estimate of σ_e^2. Then,

$$Var(a/c) \cong Var(a)/c^2.$$

For the example problem, this gives

$$Var(a/c) \cong 293,500/(288)^2 = 3.53853,$$

which is slightly larger than the previous approximation. The second approximation would not be suitable for a ratio of the residual variance to the variance of one of the other components. Suppose $a = \hat{\sigma}_0^2 = 216$, and $c = \hat{\sigma}_2^2 = 72$, then $(a/c) = 3.0$, and

$$\begin{aligned} Var(a/c) &= [(72)^2(405,700) + (216)^2(293,500) \\ &\quad -2(72)(216)(-240,700)]/(72)^4 \\ &= 866.3966, \end{aligned}$$

with the first method, and

$$Var(a/c) = 405,700/(72)^2 \;=\; 78.26,$$

with the second method. The first method is probably more realistic in this situation, but both are very large.

100 Useful Derivatives of Quantities

The following information is necessary for derivation of methods of variance component estimation based on the multivariate normal distribution.

1. The (co)variance matrix of $\mathbf{y}$ is

$$\begin{aligned}\mathbf{V} &= \sum_{i=1}^{s} \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'\sigma_i^2 + \mathbf{R}\sigma_0^2 \\ &= \mathbf{ZGZ}' + \mathbf{R}.\end{aligned}$$

   Usually, each $\mathbf{G}_i$ is assumed to be $\mathbf{I}$ for most random factors, but for animal models $\mathbf{G}_i$ might be equal to $\mathbf{A}$, the additive genetic relationship matrix. Thus, $\mathbf{G}_i$ does not always have to be diagonal, and will not be an identity in animal model analyses.

2. The inverse of $\mathbf{V}$ is

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}.$$

   To prove, show that $\mathbf{VV}^{-1} = \mathbf{I}$. Let $\mathbf{T} = \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}$, then

$$\begin{aligned}\mathbf{VV}^{-1} &= (\mathbf{ZGZ}' + \mathbf{R})[\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{ZT}^{-1}\mathbf{Z}'\mathbf{R}^{-1}] \\ &= \mathbf{ZGZ}'\mathbf{R}^{-1} - \mathbf{ZGZ}'\mathbf{R}^{-1}\mathbf{ZT}^{-1}\mathbf{Z}'\mathbf{R}^{-1} \\ &\quad + \mathbf{I} - \mathbf{ZT}^{-1}\mathbf{Z}'\mathbf{R}^{-1} \\ &= \mathbf{I} + [\mathbf{ZGT} - \mathbf{ZGZ}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}](\mathbf{T}^{-1}\mathbf{Z}'\mathbf{R}^{-1}) \\ &= \mathbf{I} + [\mathbf{ZG}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}) - \mathbf{ZGZ}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}](\mathbf{T}^{-1}\mathbf{Z}'\mathbf{R}^{-1}) \\ &= \mathbf{I} + [\mathbf{ZGZ}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{Z} - \mathbf{ZGZ}'\mathbf{R}^{-1}\mathbf{Z} - \mathbf{Z}](\mathbf{T}^{-1}\mathbf{Z}'\mathbf{R}^{-1}) \\ &= \mathbf{I} + [\mathbf{0}](\mathbf{T}^{-1}\mathbf{Z}'\mathbf{R}^{-1}) \\ &= \mathbf{I}.\end{aligned}$$

3. If k is a scalar constant and $\mathbf{A}$ is any square matrix of order m, then

$$\mid \mathbf{A}k \mid = \; k^m \mid \mathbf{A} \mid .$$

4. For general square matrices, say $\mathbf{M}$ and $\mathbf{U}$, of the same order then

$$\mid \mathbf{MU} \mid = \mid \mathbf{M} \mid \mid \mathbf{U} \mid .$$

5. For the general matrix below with $\mathbf{A}$ and $\mathbf{D}$ being square and non-singular (i.e. the inverse of each exists), then

$$\left| \begin{array}{cc} \mathbf{A} & -\mathbf{B} \\ \mathbf{Q} & \mathbf{D} \end{array} \right| = \mid \mathbf{A} \mid \mid \mathbf{D} + \mathbf{Q}\mathbf{A}^{-1}\mathbf{B} \mid = \mid \mathbf{D} \mid \mid \mathbf{A} + \mathbf{B}\mathbf{D}^{-1}\mathbf{Q} \mid .$$

Then if $\mathbf{A} = \mathbf{I}$ and $\mathbf{D} = \mathbf{I}$, then $\mid \mathbf{I} \mid = 1$, so that

$$\begin{aligned} \mid \mathbf{I} + \mathbf{QB} \mid &= \mid \mathbf{I} + \mathbf{BQ} \mid \\ &= \mid \mathbf{I} + \mathbf{B}'\mathbf{Q}' \mid \\ &= \mid \mathbf{I} + \mathbf{Q}'\mathbf{B}' \mid . \end{aligned}$$

6. Using the results in (4) and (5), then

$$\begin{aligned} \mid \mathbf{V} \mid &= \mid \mathbf{R} + \mathbf{ZGZ}' \mid \\ &= \mid \mathbf{R}(\mathbf{I} + \mathbf{R}^{-1}\mathbf{ZGZ}') \mid \\ &= \mid \mathbf{R} \mid \mid \mathbf{I} + \mathbf{R}^{-1}\mathbf{ZGZ}' \mid \\ &= \mid \mathbf{R} \mid \mid \mathbf{I} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{ZG} \mid \\ &= \mid \mathbf{R} \mid \mid (\mathbf{G}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z})\mathbf{G} \mid \\ &= \mid \mathbf{R} \mid \mid \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \mid \mid \mathbf{G} \mid . \end{aligned}$$

7. The mixed model coefficient matrix of Henderson can be denoted by

$$\mathbf{C} = \left(\begin{array}{cc} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{array} \right)$$

then the determinant of $\mathbf{C}$ can be derived as

$$\begin{aligned} \mid \mathbf{C} \mid &= \mid \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} \mid \\ &\quad \times \mid \mathbf{G}^{-1} + \mathbf{Z}'(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}^{-1})\mathbf{Z} \mid \\ &= \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid \\ &\quad \times \mid \mathbf{X}'(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1})\mathbf{X} \mid . \end{aligned}$$

Now let $\mathbf{S} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}^{-1}$ then

$$\begin{aligned} \mid \mathbf{C} \mid &= \mid \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} \mid \ \mid \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{SZ} \mid \\ &= \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid . \end{aligned}$$

8. A projection matrix, $\mathbf{P}$, is defined as

$$\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}.$$

Properties of $\mathbf{P}$:

$$\begin{aligned} \mathbf{PX} &= \mathbf{0}, \\ \mathbf{Py} &= \mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}), \quad \text{where} \\ \hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}. \end{aligned}$$

Therefore,

$$\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}).$$

9. Derivative of $\mathbf{V}^{-1}$ is

$$\frac{\partial \mathbf{V}^{-1}}{\partial \sigma_i^2} = -\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}$$

10. Derivative of $\ln \mid \mathbf{V} \mid$ is

$$\frac{\partial \ln \mid \mathbf{V} \mid}{\partial \sigma_i^2} = tr\left(\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\right)$$

11. Derivative of $\mathbf{P}$ is

$$\frac{\partial \mathbf{P}}{\partial \sigma_i^2} = -\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}.$$

12. Derivative of $\mathbf{V}$ is

$$\frac{\partial \mathbf{V}}{\partial \sigma_i^2} = \mathbf{Z}_i\mathbf{G}_i\mathbf{Z}_i'.$$

13. Derivative of $\ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid$ is

$$\frac{\partial \ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid}{\partial \sigma_i^2} = \mathrm{tr}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}\mathbf{X}.$$

101 Random Number Generators and R

R has some very good random number generators built into it. These functions are very useful for application of Gibbs sampling methods in Bayesian estimation. Generators for the uniform distribution, normal distribution, Chi-square distribution, and Wishart distribution are necessary to have. Below are some examples of the various functions in R. The package called "MCMCpack" should be obtained from the CRAN website.

```
require(MCMCpack)

# Uniform distribution generator
# num = number of variates to generate
# min = minimum number in range
# max = maximum number in range
x = runif(100,min=5,max=10)

# Normal distribution generator
# num = number of deviates to generate
# xmean = mean of the distribution you want
# xSD = standard deviation of deviates you want
w = rnorm(200,-12,16.3)

# Chi-square generator
# num = number of deviates to generate
# df = degrees of freedom
# ncp = non-centrality parameter, usually 0
w = rchisq(15,24,0)

# Inverted Wishart matrix generator
# df = degrees of freedom
# SS = matrix of sum of squares and crossproducts
U = riwish(df,SS)
# New covariance matrix is the inverse of U
V = ginv(U)
```

A Chi-square variate with m degrees of freedom is the sum of squares of m random normal deviates. The random number generator, however, makes use of a gamma distribution, which with the appropriate parameters is a Chi-square distribution.

The uniform distribution is the key distribution for all other distribution generators. R uses the Mersenne Twister (Matsumoto and Nishimura, 1997) with a cycle time of

$2^{19937}-1$. The Twister is based on a Mersenne prime number.

102 Positive Definite Matrices

A covariance matrix should be positive definite. To check a matrix, compute the eigenvalues and eigenvectors of the matrix. All eigenvalues should be positive. If they are not positive, then they can be modified, and a new covariance matrix constructed from the eigenvectors and the modified set of eigenvalues. The procedure is shown in the following R statements - should be improved.

```
# Compute eigenvalues and eigenvectors
GE = eigen(G)

nre = length(GE $values)
for(i in 1:nre)  {
qp = GE$ values[i]
if(qp < 0)qp = (qp*qp)/10000
GE$ values[i] = qp  }

# Re-form new matrix
Gh = GE$ vectors
GG = Gh %*% diag(GE$values) %*% t(Gh)
```

If the eigenvalues are all positive, then the new matrix, `GG`, will be the same as the input matrix, `G`.

103 EXERCISES

1. This is an example of Henderson's Method 1 of unbiased estimation of variance components. Let

$$\mathbf{y} = \mathbf{1}\mu + \mathbf{Z}_1\mathbf{u}_1 + \mathbf{Z}_2\mathbf{u}_2 + \mathbf{e},$$

with data as follows:

$$\begin{pmatrix} 15\\42\\20\\36\\50\\17\\34\\23\\28\\31\\45\\37 \end{pmatrix} = \begin{pmatrix} 1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1\\1 \end{pmatrix} \mu + \begin{pmatrix} 1&0&0\\0&1&0\\0&0&1\\1&0&0\\0&1&0\\0&0&1\\0&1&0\\0&1&0\\0&0&1\\0&1&0\\1&0&0\\1&0&0 \end{pmatrix} \begin{pmatrix} u_{11}\\u_{12}\\u_{13} \end{pmatrix} + \begin{pmatrix} 1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&0&1\\0&0&1&0\\0&1&0&0\\1&0&0&0\\0&0&1&0\\0&0&0&1\\0&0&0&1 \end{pmatrix} \begin{pmatrix} u_{21}\\u_{22}\\u_{23}\\u_{24} \end{pmatrix} + \mathbf{e}.$$

Also,

$$\mathbf{V} = \mathbf{Z}_1\mathbf{Z}_1'\sigma_1^2 + \mathbf{Z}_2\mathbf{Z}_2'\sigma_2^2 + \mathbf{I}\sigma_0^2.$$

Calculate the following:

(a) $\mathbf{M} = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}'$

(b) $\mathbf{A} = \mathbf{Z}_1(\mathbf{Z}_1'\mathbf{Z}_1)^{-1}\mathbf{Z}_1'$

(c) $\mathbf{B} = \mathbf{Z}_2(\mathbf{Z}_2'\mathbf{Z}_2)^{-1}\mathbf{Z}_2'$

(d) $\mathbf{Q}_0 = \mathbf{I} - \mathbf{M}$

(e) $\mathbf{Q}_1 = \mathbf{A} - \mathbf{M}$

(f) $\mathbf{Q}_2 = \mathbf{B} - \mathbf{M}$

(g) $\mathbf{y}'\mathbf{Q}_0\mathbf{y}$

(h) $\mathbf{y}'\mathbf{Q}_1\mathbf{y}$

(i) $\mathbf{y}'\mathbf{Q}_2\mathbf{y}$

(j) $E(\mathbf{y}'\mathbf{Q}_0\mathbf{y}) = tr(\mathbf{Q}_0\mathbf{V}_0)\sigma_0^2 + tr(\mathbf{Q}_0\mathbf{V}_1)\sigma_1^2 + tr(\mathbf{Q}_0\mathbf{V}_2)\sigma_2^2$

(k) $E(\mathbf{y}'\mathbf{Q}_1\mathbf{y})$

(l) $E(\mathbf{y}'\mathbf{Q}_2\mathbf{y})$

(m) Estimate the variances.

(n) Compute the variances of the estimated variances.

2. Check the following matrix for positive definiteness, and create a new modified matrix from it, that is positive definite (if it is not already positive definite).

$$\mathbf{R} = \begin{pmatrix} 1 & -2 & 3 & -4 & 5\\ -2 & 3 & -1 & 3 & 4\\ 3 & -1 & 7 & -3 & 5\\ -4 & 3 & -3 & 11 & -2\\ 5 & 4 & 5 & -2 & 15 \end{pmatrix}.$$

Likelihood Methods

104 Likelihood Functions

The multivariate normal distribution likelihood function is

$$L(\mathbf{y}) = (2\pi)^{-.5N} \mid \mathbf{V} \mid^{-.5} \exp(-.5(\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})).$$

The log of the likelihood, say L_1 is

$$L_1 = -0.5[N\ln(2\pi) + \ln \mid \mathbf{V} \mid + (\mathbf{y}-\mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{Xb})].$$

The term $N\ln(2\pi)$ is a constant that does not involve any of the unknown variances or effects in the model, and therefore, it is commonly omitted during maximization computations. Maximizing the log likelihood maximizes the original likelihood function.

Previously,

$$\mid \mathbf{V} \mid = \mid \mathbf{R} \mid \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid \mid \mathbf{G} \mid,$$

and therefore,

$$\ln \mid \mathbf{V} \mid = \ln \mid \mathbf{R} \mid + \ln \mid \mathbf{G} \mid + \ln \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid .$$

If $\mathbf{R} = \mathbf{I}\sigma_0^2$, then

$$\begin{aligned} \ln \mid \mathbf{R} \mid &= \ln \mid \mathbf{I}\sigma_0^2 \mid \\ &= \ln(\sigma_0^2)^N \mid \mathbf{I} \mid \\ &= N \ln \sigma_0^2 (1). \end{aligned}$$

Similarly, if $\mathbf{G} = \sum^{+} \mathbf{I}\sigma_i^2$, where $i = 1$ to s, then

$$\begin{aligned} \ln \mid \mathbf{G} \mid &= \sum_{i=1}^{s} \ln \mid \mathbf{I}\sigma_i^2 \mid \\ &= \sum_{i=1}^{s} q_i \ln \sigma_i^2. \end{aligned}$$

Except, that in animal models one of the $\mathbf{G}_i$ is equal to $\mathbf{A}\sigma_i^2$. In that case,

$$\ln \mid \mathbf{A}\sigma_i^2 \mid = \ln(\sigma_i^2)^{q_i} \mid \mathbf{A} \mid$$

which is

$$\ln \mid \mathbf{A}\sigma_i^2 \mid = q_i \ln \sigma_i^2 \mid \mathbf{A} \mid = \; q_i \ln \sigma_i^2 + \ln \mid \mathbf{A} \mid .$$

Recall that

$$\mathbf{C} \;=\; \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \end{pmatrix},$$

and

$$\mid \mathbf{C} \mid = \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid$$

so that

$$\ln \mid \mathbf{C} \mid = \ln \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid + \ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid .$$

105 Maximum Likelihood Method

Hartley and Rao (1967) described the maximum likelihood approach for the estimation of variance components. Let L_2 be equivalent to L_1 except for the constant involving π.

$$L_2 = -0.5[\ln \mid \mathbf{V} \mid + (\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb})].$$

The derivatives of L_2 with respect to $\mathbf{b}$ and to σ_i^2 for $i = 0, 1, \ \ldots \ s$ are

$$\frac{\partial L_2}{\partial \mathbf{b}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{Xb} - \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$$

and

$$\begin{aligned} \frac{\partial L_2}{\partial \sigma_i^2} &= -.5 \ tr[\mathbf{V}^{-1}(\partial \mathbf{V}/\partial \sigma_i^2)] \\ & \quad + .5(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}(\partial \mathbf{V}/\partial \sigma_i^2)\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) \\ &= -.5 \ tr[\mathbf{V}^{-1}\mathbf{V}_i] + .5(\mathbf{y} - \mathbf{Xb})'\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) \end{aligned}$$

Equating the derivatives to zero gives

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

and

$$tr[\mathbf{V}^{-1}\mathbf{V}_i] = (\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\mathbf{V}_i\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}).$$

Recall that

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

where $\mathbf{P}$ is the projection matrix, and that $\mathbf{V}_i = \mathbf{Z}_i\mathbf{Z}_i'$, then

$$tr[\mathbf{V}^{-1}\mathbf{Z}_i\mathbf{Z}_i'] = \mathbf{y}'\mathbf{PV}_i\mathbf{Py}.$$

In usual mixed model theory, the solution vector for a random factor may be written as

$$\hat{\mathbf{u}}_i = \mathbf{G}_i\mathbf{Z}_i'\mathbf{Py},$$

so that

$$\begin{aligned} \mathbf{y}'\mathbf{PV}_i\mathbf{Py} &= \mathbf{y}'\mathbf{PZ}_i\mathbf{G}_i\mathbf{G}_i^{-2}\mathbf{G}_i\mathbf{Z}_i'\mathbf{Py} \\ &= \hat{\mathbf{u}}_i'\mathbf{G}_i^{-2}\hat{\mathbf{u}}_i \\ &= \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^4. \end{aligned}$$

Also,

$$tr[\mathbf{V}^{-1}\mathbf{V}_i] \;=\; tr[(\mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1})\mathbf{Z}_i\mathbf{Z}_i'].$$

Let

$$\begin{aligned} \mathbf{T} &= (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}, \\ \mathbf{R} &= \mathbf{I}\sigma_0^2, \\ \text{and } \mathbf{G} &= \sum^{+}\mathbf{I}\sigma_i^2, \end{aligned}$$

then

$$tr[\mathbf{V}^{-1}\mathbf{V}_i] = tr(\mathbf{Z}_i'\mathbf{Z}_i)\sigma_0^{-2} \;-\; tr(\mathbf{Z}_i'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{Z}_i)\sigma_0^{-4}.$$

If $\mathbf{T}$ can be partitioned into submatrices for each random factor, then

$$\mathbf{T}\sigma_0^{-2}(\mathbf{Z}'\mathbf{Z} + \sum^{+}\mathbf{I}\alpha_i) \;=\; \mathbf{I},$$

and

$$\begin{aligned} \mathbf{T}\mathbf{Z}'\mathbf{Z}\sigma_0^{-2} &= \mathbf{I} - \mathbf{T}(\sum^{+}\mathbf{I}\sigma_i^{-2}), \\ \mathbf{T}\mathbf{Z}'\mathbf{Z}_i\sigma_0^{-2} &= \mathbf{I} - \mathbf{T}_{ii}\sigma_i^{-2}, \end{aligned}$$

which yields

$$tr(\mathbf{Z}_i'\mathbf{Z}\mathbf{T}\mathbf{Z}'\mathbf{Z}_i)\sigma_0^{-4} \;=\; tr(\mathbf{Z}_i'\mathbf{Z}_i)\sigma_0^{-2} - tr(\mathbf{I} - \mathbf{T}_{ii}\sigma_i^{-2})\sigma_i^{-2}.$$

Finally,

$$\begin{aligned} tr[\mathbf{V}^{-1}\mathbf{V}_i] &= tr(\mathbf{I} - \mathbf{T}_{ii}\sigma_i^{-2})\sigma_i^{-2} \\ &= tr\mathbf{I}\sigma_i^{-2} - tr\mathbf{T}_{ii}\sigma_i^{-4} \\ &= q_i\sigma_i^{-2} - tr\mathbf{T}_{ii}\sigma_i^{-4}. \end{aligned}$$

Combining results gives

$$\hat{\sigma}_i^2 = (\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i + tr\mathbf{T}_{ii}\hat{\sigma}_0^2)/q_i$$

for $i = 1, 2, \ \ldots \ , s$, and for $i = 0$ gives

$$\hat{\sigma}_0^2 = (\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'\mathbf{Z}'\mathbf{y})/\ N.$$

105.1 The EM Algorithm

EM stands for Expectation Maximization. The procedure alternates between calculating conditional expected values and maximizing simplified likelihoods. The actual data $\mathbf{y}$ are called the incomplete data in the EM algorithm, and the complete data are considered to be $\mathbf{y}$ and the unobservable random effects, $\mathbf{u}_i$. If the realized values of the unobservable random effects were known, then their variance would be the average of their squared values, i.e.,

$$\hat{\sigma}_i^2 = \mathbf{u}_i'\mathbf{u}_i/q_i.$$

However, in real life the realized values of the random effects are unknown.

The steps of the EM algorithm are as follows:

Step 0. Decide on starting values for the variances and set $m = 0$.

Step 1.(E-step) Calculate the conditional expectation of the sufficient statistics, conditional on the incomplete data.

$$\begin{aligned} E(\mathbf{u}_i'\mathbf{u}_i \mid \mathbf{y}) &= \sigma_i^{4(m)}\mathbf{y}'\mathbf{P}^{(m)}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}^{(m)}\mathbf{y} \\ &\quad + tr(\sigma_i^{2(m)}\mathbf{I} - \sigma_i^{4(m)}\mathbf{Z}_i'(\mathbf{V}^{(m)})^{-1}\mathbf{Z}_i) \\ &= \hat{t}_i^{(m)} \end{aligned}$$

Step 2.(M-step) Maximize the likelihood of the complete data,

$$\sigma_i^{2(m+1)} = \hat{t}_i^{(m)}/q_i, \quad i = 0, 1, 2, \ldots, s.$$

Step 3. If convergence is reached, set $\hat{\boldsymbol{\sigma}} = \boldsymbol{\sigma}^{(m+1)}$, otherwise increase m by one and return to Step 1.

This is equivalent to constructing and solving the mixed model equations with a given set of variances, $\boldsymbol{\sigma}^{(m)}$, and then

$$\begin{aligned} \sigma_0^{2(m+1)} &= (\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{u}}'\mathbf{Z}'\mathbf{y})/N, \\ \text{and } \sigma_i^{2(m+1)} &= (\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i + \sigma_0^{2(m+1)} tr\mathbf{T}_{ii})/q_i. \end{aligned}$$

106 Restricted Maximum Likelihood Method

Restricted (or Residual) maximum likelihood (REML), was first suggested by Thompson (1962), and described formally by Patterson and Thompson (1971). The procedure requires that $\mathbf{y}$ has a multivariate normal distribution. The method is translation invariant. The maximum likelihood approach automatically keeps the estimator within the allowable parameter space(i.e. zero to plus infinity), and therefore, REML is a biased procedure. REML was proposed as an improvement to ML in order to account for the degrees of freedom lost in estimating fixed effects.

The likelihood function used in REML is that for a set of error contrasts (i.e. residuals) assumed to have a multivariate normal distribution. The multivariate normal distribution likelihood function for the residual contrasts, $\mathbf{K}'\mathbf{y}$, where $\mathbf{K}'\mathbf{X} = 0$, and $\mathbf{K}'$ has rank equal to $N - r(\mathbf{X})$, is

$$L(\mathbf{K}'\mathbf{y}) = (2\pi)^{-.5(N-r(\mathbf{X}))} \mid \mathbf{K}'\mathbf{V}\mathbf{K} \mid^{-.5} \exp(-.5(\mathbf{K}'\mathbf{y})'(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}(\mathbf{K}'\mathbf{y})).$$

The natural log of the likelihood function is

$$L_3 = -.5(N - r(\mathbf{X}))\ln(2\pi) - .5\ln \mid \mathbf{K}'\mathbf{V}\mathbf{K} \mid -.5\mathbf{y}'\mathbf{K}(\mathbf{K}'\mathbf{V}\mathbf{K})^{-1}\mathbf{K}'\mathbf{y}.$$

Notice that $-.5(N - r(\mathbf{X}))\ln(2\pi)$ is a constant that does not depend on the unknown variance components or factors in the model, and therefore, can be ignored to give L_4. Searle (1979) showed that

$$\ln \mid \mathbf{K}'\mathbf{V}\mathbf{K} \mid \; = \; \ln \mid \mathbf{V} \mid + \ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid$$

and

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

for any $\mathbf{K}'$ such that $\mathbf{K}'\mathbf{X} = \mathbf{0}$. Hence, L_4 can be written as

$$L_4 = -.5\ln \mid \mathbf{V} \mid -.5\ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid -.5(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}}).$$

REML can be calculated a number of different ways.

1. **Derivative Free approach** is a search technique to find the parameters that maximize the log likelihood function. Two techniques will be described here.

2. **First Derivatives and EM** is where the first derivatives of the log likelihood are determined and set to zero in order to maximize the likelihood function. Solutions need to be obtained by iteration because the resulting equations are non linear.

3. **Second Derivatives** are generally more computationally demanding. Gradient methods are used to find the parameters that make the first derivatives equal to zero. Newton-Raphson (involves the observed information matrix) and Fishers Method of Scoring (involves the expected information matrix) have been used. Lately, the "average information" algorithm (averages the observed and expected information matrices) has been used to reduce the computational time.

106.1 Example Problem

All of the approaches attempt to maximize the log likelihood function of the error contrasts. To illustrate the methods, consider a single trait model with three factors (F, A, B), of which A and B are random factors. There were a total of 90 observations, and the total sum of squares was 356,000. The least squares equations for this small example are shown below.

$$\begin{pmatrix} 50 & 0 & 5 & 15 & 30 & 5 & 10 & 20 & 15 \\ 0 & 40 & 5 & 15 & 20 & 5 & 10 & 20 & 5 \\ 5 & 5 & 10 & 0 & 0 & 2 & 3 & 4 & 1 \\ 15 & 15 & 0 & 30 & 0 & 5 & 7 & 11 & 7 \\ 30 & 20 & 0 & 0 & 50 & 3 & 10 & 25 & 12 \\ 5 & 5 & 2 & 5 & 3 & 10 & 0 & 0 & 0 \\ 10 & 10 & 3 & 7 & 10 & 0 & 20 & 0 & 0 \\ 20 & 20 & 4 & 11 & 25 & 0 & 0 & 40 & 0 \\ 15 & 5 & 1 & 7 & 12 & 0 & 0 & 0 & 20 \end{pmatrix} \begin{pmatrix} F_1 \\ F_2 \\ A_1 \\ A_2 \\ A_3 \\ B_1 \\ B_2 \\ B_3 \\ B_4 \end{pmatrix} = \begin{pmatrix} 3200 \\ 2380 \\ 580 \\ 1860 \\ 3140 \\ 700 \\ 1320 \\ 2400 \\ 1160 \end{pmatrix}.$$

106.1.1 Derivative Free REML

Imagine an s dimensional array containing the values of the likelihood function for every possible set of values of the ratios of the components to the residual variance. The technique is to search this array and find the set of ratios for which the likelihood function is maximized. There is more than one way to conduct the search. Care must be taken to find the 'global' maximum rather than one of possibly many 'local' maxima. At the same time the number of likelihood evaluations to be computed must also be minimized.

Various alternative forms of L_4 can be derived. Note that

$$\ln \mid \mathbf{V} \mid = \ln \mid \mathbf{R} \mid + \ln \mid \mathbf{G} \mid + \ln \mid \mathbf{G}^{-1} + \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \mid$$

and that

$$\ln \mid \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} \mid = \ln \mid \mathbf{C} \mid - \ln \mid \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1} \mid$$

and that combining these results gives

$$L_4 = -.5 \ln \mid \mathbf{R} \mid -.5 \ln \mid \mathbf{G} \mid -.5 \ln \mid \mathbf{C} \mid -.5 \mathbf{y}'\mathbf{P}\mathbf{y}.$$

Now note that

$$\begin{aligned} \ln \mid \mathbf{R} \mid &= \ln \mid \mathbf{I}\sigma_0^2 \mid \\ &= N \ln \sigma_0^2, \\ \ln \mid \mathbf{G} \mid &= \sum_{i=1}^{s} q_i \ln \sigma_i^2, \end{aligned}$$

$$\begin{aligned}
\text{and} \quad \ln \mid \mathbf{C} \mid &= \ln \mid \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} \mid + \ln \mid \mathbf{Z}'\mathbf{S}\mathbf{Z} + \mathbf{G}^{-1} \mid \\
\text{where} \quad \ln \mid \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} \mid &= \ln \mid \mathbf{X}'\mathbf{X}\sigma_0^{-2} \mid \\
&= \ln(\sigma_0^{-2})^{r(\mathbf{X})} \mid \mathbf{X}'\mathbf{X} \mid \\
&= \ln \mid \mathbf{X}'\mathbf{X} \mid - r(\mathbf{X}) \ln \sigma_0^2, \\
\text{and} \quad \mathbf{Z}'\mathbf{S}\mathbf{Z} + \mathbf{G}^{-1} &= \sigma_0^{-2}\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{G}^{-1} \\
&= \sigma_0^{-2}(\mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{G}^{-1}\sigma_0^2).
\end{aligned}$$

Then

$$\ln \mid \mathbf{C} \mid = \ln \mid \mathbf{X}'\mathbf{X} \mid - r(\mathbf{X}) \ln \sigma_0^2 - q \ln \sigma_0^2 + \ln \mid \mathbf{Z}'\mathbf{M}\mathbf{Z} + \mathbf{G}^{-1}\sigma_0^2 \mid,$$

and finally, the log-likelihood function becomes

$$\begin{aligned}
L_4 &= -.5(N - r(\mathbf{X}) - q) \ln \sigma_0^2 - .5 \sum_{i=1}^{s} q_i \ln \sigma_i^2 \\
&\quad -.5 \ln \mid \mathbf{C}^\star \mid -.5\mathbf{y}'\mathbf{P}\mathbf{y},
\end{aligned}$$

where

$$\mathbf{C}^\star = \begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\sigma_0^2 \end{pmatrix}.$$

Note that

$$\begin{aligned}
q_i \ln \sigma_i^2 &= q_i \ln \sigma_0^2/\alpha_i \\
&= q_i(\ln \sigma_0^2 - \ln \alpha_i)
\end{aligned}$$

so that

$$L_4 = -.5[(N - r(\mathbf{X})) \ln \sigma_0^2 - \sum_{i=1}^{s} q_i \ln \alpha_i + \ln \mid \mathbf{C}^\star \mid + \mathbf{y}'\mathbf{P}\mathbf{y}].$$

The quantity $\mathbf{y}'\mathbf{P}\mathbf{y}$ is $\mathbf{y}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}})/\sigma_0^2$. The computations are achieved by constructing the following matrix,

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{G}^{-1}\sigma_0^2 & \mathbf{Z}'\mathbf{y} \\ \mathbf{y}'\mathbf{X} & \mathbf{y}'\mathbf{Z} & \mathbf{y}'\mathbf{y} \end{pmatrix} = \begin{pmatrix} \mathbf{C}^\star & \mathbf{W}'\mathbf{y} \\ \mathbf{y}'\mathbf{W} & \mathbf{y}'\mathbf{y} \end{pmatrix},$$

then by Gaussian elimination of one row at a time, the sum of the log of the non-zero pivots (using the same ordering for each evaluation of the likelihood) gives log $\mid \mathbf{C}^\star \mid$ and $\mathbf{y}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}})$. Gaussian elimination, using sparse matrix techniques, requires less computing time than inverting the coefficient matrix of the mixed model equations. The ordering of factors within the equations could be critical to the computational process and some experimentation may be necessary to determine the best ordering. The likelihood function can be evaluated without the calculation of solutions to the mixed model equations, without inverting the coefficient matrix of the mixed model equations, and without computing any of the σ_i^2. The formulations for more general models and multiple trait models are more complex, but follow the same ideas.

Searching the array of likelihood values for various values of α_i can be done in several different ways. One method is to fix the values of all but one of the s α_i, and then evaluate L_2 for four or more different values of the α_i that were not fixed. Then one can use a quadratic regression analysis to determine the value of that one ratio which maximizes L_2 given that the other ratios are fixed. This is repeated for each of the s ratios, and the process is repeated until a maximum likelihood is obtained. The calculations are demonstrated in the example that follows.

Begin by fixing the value of $\alpha_B = 10$ and letting the value of α_A take on the values of (5, 10, 20, 30, 40). Using L_4 to evaluate the likelihood, then the results were as follows:

α_A	L_4
5	-251.4442
10	-251.1504
20	-250.9822
30	-250.9274
40	-250.9019

For example, the likelihood value for $\alpha_A = 40$, would be

$$L_4 = -\frac{1}{2}[(N - r(\mathbf{X}))\ln\sigma_0^2 - q_A\ln\alpha_A - q_B\ln\alpha_B + \ln\mid\mathbf{C}^\star\mid + \mathbf{y}'(\mathbf{y} - \mathbf{X}\hat{\mathbf{b}} - \mathbf{Z}\hat{\mathbf{u}})/\sigma_0^2]$$

where

$$\begin{aligned}
\ln\mid\mathbf{C}^\star\mid &= 32.052454, \\
\mathbf{y}'\mathbf{P}\mathbf{y} &= 8483.176/\sigma_0^2 \ = \ 88, \\
q_A\ln\alpha_A &= 11.0666385, \\
q_B\ln\alpha_B &= 9.2103404, \\
\sigma_0^2 &= 96.399728, \\
\ln\sigma_0^2 &= 4.5685034, \\
(N - r(\mathbf{X})) &= 88,
\end{aligned}$$

then

$$\begin{aligned}
L_4 &= -0.5[88(4.5685) - 11.0666 - 9.2103 + 32.0525 + (8483.176/96.3997)] \\
&= -250.9019.
\end{aligned}$$

To find the value of α_A that maximizes L_4 for $\alpha_B = 10$, let

$$\mathbf{Q} = \begin{pmatrix} 1 & 5 & 25 \\ 1 & 10 & 100 \\ 1 & 20 & 400 \\ 1 & 30 & 900 \\ 1 & 40 & 1600 \end{pmatrix} \text{ and } \mathbf{Y} = \begin{pmatrix} -251.4442 \\ -251.1504 \\ -250.9822 \\ -250.9274 \\ -250.9019 \end{pmatrix}$$

then

$$\hat{\beta} = (\mathbf{Q}'\mathbf{Q})^{-1}\mathbf{Q}'\mathbf{Y} = \begin{pmatrix} -251.6016 \\ .0448877 \\ -.000698 \end{pmatrix}.$$

From this a prediction equation for L_4 can be written as

$$L_4 = -251.6016 + .04489\alpha_A - .000698\alpha_A^2.$$

This equation can be differentiated with respect to α_A and then equated to zero to find the value of the ratio that maximizes the prediction equation. This gives

$$\alpha_A = .04489/(2(.000698)) = 32.1546.$$

Now keep $\alpha_A = 32.1546$ and try a number of values of α_B from 2 to 10, which give the following results.

α_B	L_4
2	-250.2722
3	-250.1954
4	-250.2379
5	-250.3295
6	-250.4419
7	-250.5624
8	-250.6843
9	-250.8042
10	-250.9204

Applying the quadratic regression to these points gives

$$\alpha_B = 1.2625.$$

The next step would be to fix $\alpha_B = 1.2625$ and to try new values for α_A, such as 25 to 40 by units of 1. The range of values becomes finer and finer. To insure that one has found the global maximum, the entire process could be started with vastly different starting values for the ratios, such as $\alpha_B = 50$ and let values for α_A be 40, 50, 60, and

70. The more components there are to estimate, the more evaluations of the likelihood that are going to be needed, and the more probable that convergence might be to a local maximum rather than to the global maximum.

Please refer to the literature for specification of the log likelihood function for particular models and situations.

106.1.2 The Simplex Method

The Simplex Method (Nelder and Mead, 1965) is a procedure for finding the minimum of a function (i.e. the minimum of $-2L_4$ or the maximum of L_4) with respect to the unknown variances and covariances. The best way to describe the method is using the example data from the previous sections. Begin by constructing a set of 'points' for which L_4 is to be evaluated. A 'point' is a vector of values for the unknowns (α_A, α_B), for example,

$$\theta_1 = \left(\begin{array}{cc} 12.1 & 3.8 \end{array} \right),$$

then form two more points by changing one unknown at a time. Let the three points be as shown in the following table.

No.	α_A	α_B
1	12.1	3.8
2	13.1	3.8
3	12.1	4.3

Now calculate L_4 for each point and arrange from largest to lowest value.

No.	α_A	α_B	L_4
2	13.1	3.8	-250.3047
1	12.1	3.8	-250.3197
3	12.1	4.3	-250.3662

The idea now is to find another point to replace the last one(lowest L_4). This is done by a process called *reflection*. Compute the mean of all points excluding the one with the lowest L_4.

$$\theta_m = \left(\begin{array}{cc} 12.6 & 3.8 \end{array} \right),$$

then the reflection step is

$$\theta_4 = \theta_m + r * (\theta_m - \theta_{last}),$$

where r is recommended by Nelder and Mead (1965) to be 1, giving

$$\theta_4 = \left(\begin{array}{cc} 13.1 & 3.3 \end{array} \right).$$

The corresponding L_4 for this point was -250.2722. Compared to those in the table it has the largest value, and therefore, is a better point than the other three.

No.	α_A	α_B	L_4
4	13.1	3.3	-250.2722
2	13.1	3.8	-250.3047
1	12.1	3.8	-250.3197

Given this success, the Simplex method calls for an *expansion* step, i.e. to make a bigger change. Thus,

$$\theta_5 = \theta_m + E * (\theta_4 - \theta_m),$$

where E is suggested to be equal to 2. Hence

$$\theta_5 = \begin{pmatrix} 13.6 & 2.8 \end{pmatrix}.$$

Then $L_4 = -250.2546$, and the expanded point is better yet. Now drop θ_1 from the table and put θ_5 at the top.

No.	α_A	α_B	L_4
5	13.6	2.8	-250.2546
4	13.1	3.3	-250.2722
2	13.1	3.8	-250.3047

This completes one iteration. Begin the next iteration by computing the mean of all points excluding the point with the lowest L_4.

$$\theta_m = \begin{pmatrix} 13.35 & 3.05 \end{pmatrix}.$$

Another reflection step gives

$$\begin{aligned} \theta_6 &= \theta_m + r * (\theta_m - \theta_{last}), \\ &= \begin{pmatrix} 13.6 & 2.3 \end{pmatrix}. \end{aligned}$$

However, this gives $L_4 = -250.2761$, which is between θ_2 and θ_4, and can push out θ_2 from the table.

No.	α_A	α_B	L_4
5	13.6	2.8	-250.2546
4	13.1	3.3	-250.2722
6	13.6	2.3	-250.2761

Instead of an expansion step, a *contraction* step is needed because θ_6 did not give a greater L_4 than the first two. Thus,

$$\theta_7 = \theta_m + c * (\theta_6 - \theta_m),$$

where $c = 0.5$ is recommended. Hence,

$$\theta_7 = \begin{pmatrix} 13.475 & 3.05 \end{pmatrix}.$$

Then $L_4 = -250.2586$ is better than that given by θ_4, but not by θ_5, thus the new table becomes as follows:

No.	α_A	α_B	L_4
5	13.6	2.8	-250.2546
7	13.475	3.05	-250.2586
4	13.1	3.3	-250.2722

The following steps were taken in the next iteration.

1. The mean of the top two L_4 is

$$\theta_m = \begin{pmatrix} 13.5375 & 2.925 \end{pmatrix}.$$

2. A reflection step gives

$$\begin{aligned} \theta_8 &= \theta_m + r * (\theta_m - \theta_{last}), \\ &= \begin{pmatrix} 13.975 & 2.55 \end{pmatrix}, \end{aligned}$$

which gave $L_4 = -250.2563$, which is better than θ_7.

3. Add θ_8 to the table and drop θ_4.

No.	α_A	α_B	L_4
5	13.6	2.8	-250.2546
8	13.975	2.55	-250.2563
7	13.475	3.05	-250.2586

4. Because L_4 for θ_8 was not larger than L_4 for θ_5 or smaller than L_4 for θ_7, then no expansion or contraction step is necessary. Begin the next iteration.

The Simplex method continues in this manner until all point entries in the table are equal. The constants recommended by Nelder and Mead (1965) for reflection, expansion, and contraction could be adjusted for a particular data set. This method may converge to a local maximum, and so different starting values are needed to see if it converges to the same point. The Simplex method does not work well with a large number of parameters to be estimated.

106.2 First Derivatives and EM Algorithm

To derive formulas for estimating the variance components take the derivatives of L_4 with respect to the unknown components.

$$\begin{aligned}\frac{\partial L_4}{\partial \sigma_i^2} &= -.5tr\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2} - .5tr(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}\mathbf{X} \\ & +.5(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})'\mathbf{V}^{-1}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}})\end{aligned}$$

Combine the two terms involving the traces and note that

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

then

$$\frac{\partial L_4}{\partial \sigma_i^2} = -.5tr(\mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1})\frac{\partial \mathbf{V}}{\partial \sigma_i^2} + .5\mathbf{y}'\mathbf{P}\frac{\partial \mathbf{V}}{\partial \sigma_i^2}\mathbf{P}\mathbf{y}$$

$$= -.5tr\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i' + .5\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}$$

for $i = 1, \ldots, s$ or

$$= -.5tr\mathbf{P} + .5\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y}$$

for $i = 0$ for the residual component. Using $\mathbf{P}$ and the fact that

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{Z}(\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}\mathbf{Z}'\mathbf{R}^{-1}$$

then

$$tr\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i' = q_i/\sigma_i^2 - tr\mathbf{C}_{ii}\sigma_0^2/\sigma_i^4$$

and

$$tr\mathbf{P} = (N - r(\mathbf{X}))\sigma_0^2 - \sum_{i=1}^{s}\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^2.$$

The other terms, $\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}$ and $\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y}$, were simplified by Henderson (1973) to show that they could be calculated from the Mixed Model Equations. Note that Henderson (1973) showed

$$\begin{aligned}\mathbf{P}\mathbf{y} &= \mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\hat{\mathbf{b}}), \\ \hat{\mathbf{b}} &= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}, \\ \hat{\mathbf{u}}_i &= \mathbf{G}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y},\end{aligned}$$

then

$$\begin{aligned}\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} &= \mathbf{y}'\mathbf{P}\mathbf{Z}_i[\mathbf{G}_i\mathbf{G}_i^{-1}\mathbf{G}_i^{-1}\mathbf{G}_i]\mathbf{Z}_i'\mathbf{P}\mathbf{y} \\ &= (\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{G}_i)\mathbf{G}_i^{-2}(\mathbf{G}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}) \\ &= \hat{\mathbf{u}}_i'\mathbf{G}_i^{-2}\hat{\mathbf{u}}_i\end{aligned}$$

which when $\mathbf{G}_i = \mathbf{I}\sigma_i^2$ gives

$$\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^4.$$

Similarly for the residual component, Henderson showed that

$$\mathbf{y}'\mathbf{PPy} = [\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \sum_{i=1}^{s}(\hat{\mathbf{u}}_i'\mathbf{Z}_i'\mathbf{y} + \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i\alpha_i)]/\sigma_0^2,$$

where $\alpha_i = \sigma_0^2/\sigma_i^2$.

Equate the derivatives to zero incorporating the above simplifications and obtain

$$\begin{aligned}\hat{\sigma}_i^2 &= (\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i + tr\mathbf{C}_{ii}\sigma_0^2)/q_i,\\ \hat{\sigma}_0^2 &= \mathbf{y}'\mathbf{Py}/(N - r(\mathbf{X})).\end{aligned}$$

As with ML, solutions using the EM algorithm must be computed iteratively. Convergence is usually very slow, if it occurs, and the process may also diverge.

Notice the differences between REML and ML. The denominator for $\hat{\sigma}_0^2$ is $N - r(\mathbf{X})$ rather than N, and in $\hat{\sigma}_i^2$ is $tr\mathbf{C}_{ii}$ rather than $tr\mathbf{T}_{ii}$. The quadratic forms, however, are identical in REML and ML. Accounting for the degrees of freedom to estimate $\mathbf{b}$ has resulted in the REML algorithm.

A major computing problem with the EM algorithm is the calculation of $tr\mathbf{C}_{ii}$, which is the corresponding inverse elements of the mixed model equations for the i^{th} random factor. With most applications in animal breeding, the order of the mixed model equations are too large to be inverted, and solutions to the equations are obtained by Gauss-Seidel iterations. However, there have been several attempts to approximate $tr\mathbf{C}_{ii}$, but these have not been totally suitable.

To demonstrate the EM algorithm let $\alpha_A = 10$ and $\alpha_B = 5$ be the starting values of the ratios for factors A and B, respectively. There were $N = 90$ total observations, and $r(\mathbf{X}) = 2$. The solution vector is

$$\left(\begin{array}{c} F_1 \\ F_2 \\ A_1 \\ A_2 \\ A_3 \\ B_1 \\ B_2 \\ B_3 \\ B_4 \end{array}\right) = \left(\begin{array}{r} 64.6313 \\ 59.4225 \\ -2.1363 \\ .4955 \\ 1.6368 \\ 5.1064 \\ 2.6402 \\ -2.6433 \\ -5.1034 \end{array}\right).$$

Then

$$\mathbf{y}'(\mathbf{X}\hat{\mathbf{b}} + \mathbf{Z}\hat{\mathbf{u}}) = 347,871.2661$$

and from the inverse of the coefficient matrix,

$$tr\mathbf{C}_{AA} = .16493, \text{ and } tr\mathbf{C}_{BB} = .3309886$$

which give rise to the following estimates,

$$\begin{aligned}
\hat{\sigma}_0^2 &= (356,000 - 347,871.2661)/88 \\
&= 92.371976, \\
\hat{\sigma}_A^2 &= (7.4925463 + .16493(92.371976))/3 \\
&= 7.575855, \\
\hat{\sigma}_B^2 &= (66.0771576 + .3309886(92.371976))/4 \\
&= 24.16280774.
\end{aligned}$$

New ratios are formed as

$$\alpha_A = 92.371976/7.575855 = 12.192944,$$

and

$$\alpha_B = 92.371976/24.16280774 = 3.822899$$

and these are used to form the mixed model equations again, new solutions and traces are calculated, and so on, until the estimated ratios and the prior values of the ratios are equal. The estimates converge to

$$\begin{aligned}
\hat{\sigma}_0^2 &= 91.8639, \\
\hat{\sigma}_A^2 &= 2.5692, \\
\hat{\sigma}_B^2 &= 30.5190.
\end{aligned}$$

or

$$\alpha_A = 35.7558, \quad \text{and} \quad \alpha_B = 3.0101.$$

106.3 Second Derivatives, Average Information

Second derivatives of the log likelihood lead to the expectations of the quadratic forms. One technique, MIVQUE (Minimum Variance Quadratic Unbiased Estimation) equates the quadratic forms to their expectations. The estimates are unbiased and if all variances remain positive, then convergence will be to the REML estimates. However, due to a shortage of data or an inappropriate model, the estimates derived in this manner can be negative. Computing the expectations of the quadratic forms requires the inverse of the mixed model equations coefficient matrix, and then products and crossproducts of various parts of the inverse.

A gradient method using first and second derivatives can be used (Hofer, 1998). The gradient, $\mathbf{d}$ (the vector of first derivatives of the log likelihood), is used to determine the direction towards the parameters that give the maximum of the log likelihood, such that

$$\theta^{(t+1)} = \theta^{(t)} + \mathbf{M}^{(t)}\mathbf{d}^{(t)},$$

where $\mathbf{d}^{(t)}$ are the first derivatives evaluated at $\theta = \theta^{(t)}$, and $\mathbf{M}^{(t)}$ in the Newton-Raphson(NR) algorithm is the observed information matrix, and in the Fisher Method of Scoring(FS) it is the expected information matrix.

The first derivatives are as follows (from earlier in these notes):

$$\frac{\partial L_4}{\partial \sigma_i^2} = -.5tr\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i' + .5\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} = 0$$

for $i = 1, \ldots, s$ or

$$\frac{\partial L_4}{\partial \sigma_0^2} = -.5tr\mathbf{P} + .5\mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} = 0$$

for the residual component. Then from earlier results,

$$\begin{aligned} tr\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i' &= q_i/\sigma_i^2 - tr\mathbf{C}_{ii}\sigma_0^2/\sigma_i^4, \\ \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y} &= \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^4 \end{aligned}$$

which combined give

$$0.5(\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^4 - q_i/\sigma_i^2 + tr\mathbf{C}_{ii}\sigma_0^2/\sigma_i^4) = 0,$$

for $i = 1, \ldots, s$, and

$$\begin{aligned} tr\mathbf{P} &= (N - r(\mathbf{X}))\sigma_0^2 - \sum_{i=1}^{s}\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^2 \\ \mathbf{y}'\mathbf{P}\mathbf{P}\mathbf{y} &= [\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \sum_{i=1}^{s}(\hat{\mathbf{u}}_i'\mathbf{Z}_i'\mathbf{y} + \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i\alpha_i)]/\sigma_0^2 \end{aligned}$$

which combined give

$$0.5([\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \sum_{i=1}^{s}(\hat{\mathbf{u}}_i'\mathbf{Z}_i'\mathbf{y} + \hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i\alpha_i)]/\sigma_0^2 - (N - r(\mathbf{X}))\sigma_0^2 + \sum_{i=1}^{s}\hat{\mathbf{u}}_i'\hat{\mathbf{u}}_i/\sigma_i^2) = 0,$$

which simplifies to

$$0.5([\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \sum_{i=1}^{s}\hat{\mathbf{u}}_i'\mathbf{Z}_i'\mathbf{y}]/\sigma_0^2 - (N - r(\mathbf{X}))\sigma_0^2) = 0.$$

The second derivatives give a matrix of quantities. The elements of the *observed information* matrix (Gilmour et al. 1995) are

$$\begin{aligned}
-\frac{\partial^2 L_4}{\partial\sigma_i^2\partial\sigma_0^2} &= 0.5\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}/\sigma_0^4, \\
-\frac{\partial^2 L_4}{\partial\sigma_i^2\partial\sigma_j^2} &= 0.5tr(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_j') - 0.5tr(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j') \\
& +\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{y}/\sigma_0^2 - 0.5\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_j'\mathbf{P}\mathbf{y}/\sigma_0^2, \\
& \text{and} \\
-\frac{\partial^2 L_4}{\partial\sigma_0^2\partial\sigma_0^2} &= \mathbf{y}'\mathbf{P}\mathbf{y}/\sigma_0^6 - 0.5(N - r(\mathbf{X}))/\sigma_0^4.
\end{aligned}$$

The elements of the *expected information* matrix (Gilmour et al. 1995) are

$$\begin{aligned}
E[-\frac{\partial^2 L_4}{\partial\sigma_i^2\partial\sigma_0^2}] &= 0.5tr(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i')/\sigma_0^2, \\
E[-\frac{\partial^2 L_4}{\partial\sigma_i^2\partial\sigma_j^2}] &= 0.5tr(\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'), \\
& \text{and} \\
E[-\frac{\partial^2 L_4}{\partial\sigma_0^2\partial\sigma_0^2}] &= 0.5(N - r(\mathbf{X}))/\sigma_0^4.
\end{aligned}$$

As the name *Average Information* implies, average the *observed* and *expected* information matrices to give the following matrix of elements.

$$\begin{aligned}
I[\sigma_i^2, \sigma_0^2] &= 0.5\mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{y}/\sigma_0^4, \\
I[\sigma_i^2, \sigma_j^2] &= \mathbf{y}'\mathbf{P}\mathbf{Z}_i\mathbf{Z}_i'\mathbf{P}\mathbf{Z}_j\mathbf{Z}_j'\mathbf{P}\mathbf{y}/\sigma_0^2, \\
& \text{and} \\
I[\sigma_0^2, \sigma_0^2] &= 0.5\mathbf{y}'\mathbf{P}\mathbf{y}/\sigma_0^6.
\end{aligned}$$

The first derivatives form the vector, $\mathbf{d}^{(t)}$, and

$$\mathbf{M}^{(t)} = I[\sigma, \sigma]^{-1}.$$

The rest of this method is computational detail to simplify the requirements for inverse elements and solutions to MME. The calculations can not be illustrated very easily for the example data because the **y**-vector is not available.

106.4 Animal Models

The model commonly applied to estimation of variance components in livestock genetics since 1989 has been an animal model. The animal model assumes a large, random mating population, an infinite number of loci each with a small and equal effect on the trait, only additive genetic effects, and all relationships among animals are known and tracible to an unselected base population (somewhere in the past). Animals may have more than one record each. The equation of the model is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{Zp} + \mathbf{e},$$

where $\mathbf{a}$ is the vector of animal additive genetic effects (one per animal), and $\mathbf{p}$ is a vector of permanent environmental (p.e.) effects associated with each animal.

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{Xb}, \\ Var\begin{pmatrix} \mathbf{a} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} &= \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix}. \end{aligned}$$

The matrix $\mathbf{A}$ is called the numerator relationship matrix. Wright defined relationships among animals as correlations, but $\mathbf{A}$ is essentially relationships defined as covariances (the numerators of the correlation coefficients). Also, these only represent the additive genetic relationships between animals.

The MME for this model are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k_a & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}.$$

Note that k_a is the ratio of residual to additive genetic variances, and k_p is the ratio of residual to permanent environmental variances. Also, in MME the inverse of $\mathbf{A}$ is required.

The EM-REML procedure gives

$$\begin{aligned} \hat{\sigma}_e^2 &= (\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{a}}'\mathbf{Z}'\mathbf{y} - \hat{\mathbf{p}}'\mathbf{Z}'\mathbf{y})/(N - r(\mathbf{X})), \\ \hat{\sigma}_a^2 &= (\hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} + tr\mathbf{A}^{-1}\mathbf{C}_{aa}\hat{\sigma}_e^2)/n, \\ \hat{\sigma}_p^2 &= (\hat{\mathbf{p}}'\hat{\mathbf{p}} + tr\mathbf{C}_{pp}\hat{\sigma}_e^2)/n, \end{aligned}$$

where n is the total number of animals, N is the total number of records, and $\mathbf{C}_{aa}$ are the inverse elements of the MME for the animal additive genetic effects, and $\mathbf{C}_{pp}$ are the inverse elements of the MME for the animal permanent environmental effects. An example of this model will be given in later notes.

106.4.1 Quadratic Forms in an Animal Model

A necessary quadratic form in an animal model is $\hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}}$, and this can be computed very easily. Note that the inverse of $\mathbf{A}$ may be written as

$$\mathbf{A}^{-1} = \mathbf{T}^{-1}\mathbf{D}^{-2}\mathbf{T}'^{-1},$$

where $\mathbf{T}^{-1}$ is an upper triangular matrix, and diagonal matrix $\mathbf{D}^{-2}$ has elements equal to 1, 2, or 4/3 in noninbred situations, and values greater than 2 in inbred situations. In Henderson (1975), this inverse was shown to be composed of just three numbers, i.e. 0, 1's on the diagonals, and -.5 corresponding to the parents of an animal. For example,

$$\mathbf{T}'^{-1} = \left(\begin{array}{rrrr} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -.5 & -.5 & 1 & 0 \\ -.5 & 0 & -.5 & 1 \end{array} \right).$$

Then

$$\begin{aligned} \mathbf{T}'^{-1}\hat{\mathbf{a}} &= \hat{\mathbf{m}} \\ &= (\hat{a}_i - 0.5(\hat{a}_s + \hat{a}_d)), \end{aligned}$$

for the i^{th} animal, and $\hat{a}_s$ and $\hat{a}_d$ are the sire and dam estimated breeding values, respectively. Consequently,

$$\begin{aligned} \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}} &= \hat{\mathbf{a}}'\mathbf{T}^{-1}\mathbf{B}^{-2}\mathbf{T}'^{-1}\hat{\mathbf{a}} \\ &= \hat{\mathbf{m}}'\mathbf{B}^{-2}\hat{\mathbf{m}} \\ &= \sum_{i=1}^{q} \hat{m}_i^2 b^{ii}, \end{aligned}$$

where b^{ii} are the diagonal elements of $\mathbf{B}^{-2}$, and q is the number of animals.

107 EXERCISES

Below are pedigrees and data on 20 animals.

Animal	Sire	Dam	Group	Record
1	-	-		
2	-	-		
3	-	-		
4	-	-		
5	1	2	1	28
6	1	2	1	40
7	3	4	1	30
8	1	2	1	35
9	1	4	2	17
10	3	2	2	41
11	1	4	2	23
12	3	2	2	38
13	1	2	2	37
14	3	4	2	27
15	5	7	3	24
16	6	14	3	31
17	8	7	3	42
18	1	10	3	47
19	3	13	3	26
20	5	9	3	33

1. Construct $\mathbf{A}^{-1}$ and set up the MME.

2. Apply EM-REML to the model,

   $$y_{ij} = \mu + g_i + a_j + e_{ij},$$

   where group, animal additive genetic, and residual effects are random. Let

   $$\sigma_e^2/\sigma_a^2 = 1.5, \quad \sigma_e^2/\sigma_g^2 = 5.0,$$

   to start the iterations. Do five iterations of EM-REML.

3. Apply EM-REML again to the model with

   $$\sigma_e^2/\sigma_a^2 = 3.0, \quad \sigma_e^2/\sigma_g^2 = 2.5,$$

   and do five iterations of REML from these starting values.

4. Do the results of the previous two question tend to give similar answers? Comment on the results.

5. Use the Derivative Free method and compute the log likelihoods for various sets of parameters.

Bayesian Methods

Every variable in a linear model is a random variable derived from a distribution function. A fixed factor becomes a random variable with possibly a uniform distribution going from a lower limit to an upper limit. A component of variance is a random variable having a Gamma or Chi-square distribution with *df* degrees of freedom. In addition, the researcher may have information from previous experiments that strongly indicate the value that a variance component may have, and the Bayes approach allows the *apriori* information to be included in the analysis.

The Bayesian process is to

1. Specify distributions for each random variable of the model.
2. Combine the distributions into the joint posterior distribution.
3. Find the conditional marginal distributions from the joint posterior distribution.
4. Employ Markov Chain Monte Carlo (MCMC) methods to maximize the joint posterior distribution. Gibbs Sampling is a tool in MCMC methods for deriving estimates of parameters from the joint posterior distribution.

By determining conditional marginal distributions for each random variable of the model, then generating random samples from these distributions eventually converge to random samples from the joint posterior distribution. Computationally, any program that calculates solutions to Henderson's mixed model equations can be modified to implement Gibbs Sampling.

108 The Joint Posterior Distribution

Begin with a simple single trait animal model. That is,

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Za} + \mathbf{e}.$$

Let θ be the vector of random variables and $\mathbf{y}$ is the data vector, then

$$\begin{aligned} p(\theta, \mathbf{y}) &= p(\theta)\ p(\mathbf{y} \mid \theta) \\ &= p(\mathbf{y})\ p(\theta \mid \mathbf{y}) \end{aligned}$$

Re-arranging gives

$$p(\theta \mid \mathbf{y}) = \frac{p(\theta) p(\mathbf{y} \mid \theta)}{p(\mathbf{y})}$$

$$\begin{aligned} &= (\text{prior for } \theta)\frac{p(\mathbf{y} \mid \theta)}{p(\mathbf{y})} \\ &= \text{posterior probability function of } \theta \end{aligned}$$

In terms of the simple animal model, θ includes $\mathbf{b}$, $\mathbf{a}$, σ_a^2, and σ_e^2.

108.1 Conditional Distribution of Data Vector

The conditional distribution of $\mathbf{y}$ given θ is

$$\mathbf{y} \mid \mathbf{b}, \mathbf{a}, \sigma_a^2, \sigma_e^2 \sim N(\mathbf{Xb} + \mathbf{Za}, \mathbf{I}\sigma_e^2),$$

and

$$p(\mathbf{y} \mid \mathbf{b}, \mathbf{a}, \sigma_a^2, \sigma_e^2) \propto (\sigma_e^2)^{(-N/2)} \exp\left[-(\mathbf{y} - \mathbf{Xb} - \mathbf{Za})'(\mathbf{y} - \mathbf{Xb} - \mathbf{Za})/2\sigma_e^2\right].$$

108.2 Prior Distributions of Random Variables

108.2.1 Fixed Effects Vector

There is little prior knowledge about the values in $\mathbf{b}$ might have. This is represented by assuming

$$p(\mathbf{b}) \propto \text{constant}.$$

108.2.2 Random Effects and Variances

For $\mathbf{a}$, the vector of additive genetic values, quantitative genetics theory suggests that they follow a normal distribution, i.e.

$$\mathbf{a} \mid \mathbf{A}, \sigma_a^2 \sim N(\mathbf{0}, \mathbf{A}\sigma_a^2)$$

and

$$p(\mathbf{a}) \propto (\sigma_a^2)^{(-q/2)} \exp\left[-\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}/2\sigma_a^2\right],$$

where q is the length of $\mathbf{a}$.

A natural estimator of σ_a^2 is $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}/q$, call it S_a^2, where

$$S_a^2 \sim \chi_q^2 \sigma_a^2/q.$$

Multiply both sides by q and divide by χ_q^2 to give

$$\sigma_a^2 \sim qS_a^2/\chi_q^2$$

which is a scaled, inverted Chi-square distribution, written as

$$p(\sigma_a^2 \mid v_a, S_a^2) \propto (\sigma_a^2)^{-(\frac{v_a}{2}+1)} \exp\left(-\frac{v_a}{2}\frac{S_a^2}{\sigma_a^2}\right),$$

where v_a and S_a^2 are hyperparameters with S_a^2 being a prior guess about the value of σ_a^2 and v_a being the degrees of belief in that prior value. Usually q is much larger than v_a and therefore, the data provide nearly all of the information about σ_a^2.

108.2.3 Residual Effects

Similarly, for the residual variance,

$$p(\sigma_e^2 \mid v_e, S_e^2) \quad \propto \quad (\sigma_e^2)^{-(\frac{v_e}{2}+1)} \exp\left(-\frac{v_e}{2}\frac{S_e^2}{\sigma_e^2}\right).$$

108.2.4 Combining Prior Distributions

The joint posterior distribution is

$$p(\mathbf{b}, \mathbf{a}, \sigma_a^2, \sigma_e^2 \mid \mathbf{y}) \propto p(\mathbf{b})p(\mathbf{a} \mid \sigma_a^2)p(\sigma_a^2)p(\sigma_e^2)p(\mathbf{y} \mid \mathbf{b}, \mathbf{a}, \sigma_a^2, \sigma_e^2)$$

which can be written as

$$\propto (\sigma_e^2)^{-(\frac{N+v_e}{2}+1)} \exp\left[-\frac{1}{2\sigma_e^2}((\mathbf{y}-\mathbf{Xb}-\mathbf{Za})'(\mathbf{y}-\mathbf{Xb}-\mathbf{Za})+v_eS_e^2)\right]$$

$$(\sigma_a^2)^{-(\frac{q+v_a}{2}+1)} \exp\left[-\frac{1}{2\sigma_a^2}(\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}+v_aS_a^2)\right].$$

109 Fully Conditional Posterior Distributions

In order to implement Gibbs sampling, all of the fully conditional posterior distributions (one for each component of θ) need to be derived from the joint posterior distribution. The conditional posterior distribution is derived from the joint posterior distribution by picking out the parts that involve the unknown parameter in question.

109.1 Fixed and Random Effects of the Model

Let

$$\begin{aligned}
\mathbf{W} &= (\mathbf{X} \;\; \mathbf{Z}), \\
\beta' &= (\mathbf{b}' \;\; \mathbf{a}'), \\
\Sigma &= \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{-1}k \end{pmatrix}, \\
\mathbf{C} &= \text{Henderson}'\text{s Mixed Model Equations} \\
&= \mathbf{W}'\mathbf{W} + \Sigma \\
\mathbf{C}\hat{\beta} &= \mathbf{W}'\mathbf{y}
\end{aligned}$$

A new notation is introduced, let

$$\beta' = (\beta_i \;\; \beta'_{-i}),$$

where β_i is a scalar representing just one element of the vector β, and β_{-i} is a vector representing all of the other elements except β_i. Similarly, $\mathbf{C}$ and $\mathbf{W}$ can be partitioned in the same manner as

$$\begin{aligned}
\mathbf{W}' &= (\mathbf{W}_i \;\; \mathbf{W}_{-i})' \\
\mathbf{C} &= \begin{pmatrix} C_{i,i} & \mathbf{C}_{i,-i} \\ \mathbf{C}_{-i,i} & \mathbf{C}_{-i,-i} \end{pmatrix}.
\end{aligned}$$

In general terms, the conditional posterior distribution of β is a normal distribution,

$$\beta_i \mid \beta_{-i}, \sigma^2_a, \sigma^2_e, \mathbf{y} \;\; \sim \;\; N(\hat{\beta}_i, C^{-1}_{i,i}\sigma^2_e)$$

where

$$C_{i,i}\hat{\beta}_i = (\mathbf{W}'_i\mathbf{y} - \mathbf{C}_{i,-i}\beta_{-i}).$$

Then

$$b_i \mid \mathbf{b}_{-i}, \mathbf{a}, \sigma^2_a, \sigma^2_e, \mathbf{y} \;\sim\; N(\hat{b}_i, C^{-1}_{i,i}\sigma^2_e),$$

for

$$C_{i,i} = \mathbf{x}'_i\mathbf{x}_i.$$

Also,

$$a_i \mid \mathbf{b}, \mathbf{a}_{-i}, \sigma^2_a, \sigma^2_e, \mathbf{y} \;\sim\; N(\hat{a}_i, C^{-1}_{i,i}\sigma^2_e),$$

where $C_{i,i} = (\mathbf{z}'_i\mathbf{z}_i + A^{i,i}k)$, for $k = \sigma^2_e/\sigma^2_a$.

109.2 Variances

The conditional posterior distributions for the variances are inverted Chi-square distributions,

$$\sigma_a^2 \mid \mathbf{b}, \mathbf{a}, \sigma_e^2, \mathbf{y} \sim \tilde{v}_a \tilde{S}_a^2 \chi_{\tilde{v}_a}^{-2}$$

for $\tilde{v}_a = q + v_a$, and $\tilde{S}_a^2 = (\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + v_a S_a^2)/\tilde{v}_a$, and

$$\sigma_e^2 \mid \mathbf{b}, \mathbf{a}, \sigma_a^2, \mathbf{y} \sim \tilde{v}_e \tilde{S}_e^2 \chi_{\tilde{v}_e}^{-2}$$

for $\tilde{v}_e = N + v_e$, and $\tilde{S}_e^2 = (\mathbf{e}'\mathbf{e} + v_e S_e^2)/\tilde{v}_e$, and $\mathbf{e} = \mathbf{y} - \mathbf{X}\mathbf{b} - \mathbf{Z}\mathbf{a}$.

110 Computational Scheme

Gibbs sampling is much like Gauss-Seidel iteration. When a new solution is calculated in the Mixed Model Equations for a level of a fixed or random factor, a random amount is added to the solution based upon its conditional posterior distribution variance before proceeding to the next level of that factor or the next factor. After all equations have been processed, new values of the variances are calculated and a new variance ratio is determined prior to beginning the next round. The following MME for five animals will be used to illustrate the Gibbs sampling scheme:

$$\begin{pmatrix} 5 & 1 & 1 & 1 & 1 & 1 \\ 1 & 29 & 7 & -7 & -14 & 0 \\ 1 & 7 & 30 & -14 & 8 & -16 \\ 1 & -7 & -14 & 36 & -14 & 0 \\ 1 & -14 & 8 & -14 & 37 & -16 \\ 1 & 0 & -16 & 0 & -16 & 33 \end{pmatrix} \begin{pmatrix} \mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \end{pmatrix} = \begin{pmatrix} 238.2 \\ 38.5 \\ 48.9 \\ 64.3 \\ 50.5 \\ 36.0 \end{pmatrix},$$

where $k = \sigma_e^2/\sigma_a^2 = 14$, and

$$\mathbf{A}^{-1} = \frac{1}{14} \begin{pmatrix} 28 & 7 & -7 & -14 & 0 \\ 7 & 29 & -14 & 8 & -16 \\ -7 & -14 & 35 & -14 & 0 \\ -14 & 8 & -14 & 36 & -16 \\ 0 & -16 & 0 & -16 & 32 \end{pmatrix}.$$

The starting values for $\beta = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$, and for $v_a = v_e = 10$, and $S_e^2 = 93\frac{1}{3}$ and $S_a^2 = 6\frac{2}{3}$, so that $k = 14$. Let RND represent a random normal deviate from a random normal deviate generator, and let $CHI(idf)$ represent a random Chi-square variate from a random Chi-Square variate generator with idf degrees of freedom. Every time that RND and $CHI(idf)$ appear, a different random number is generated for that expression.

To begin, let $\sigma_e^2 = S_e^2$ and $\sigma_a^2 = S_a^2$. Below are descriptions of calculations in the first two rounds.

110.1 Round 1

110.1.1 Fixed and Random Effects of Model

- **Overall mean**

$$\begin{aligned}\hat{\mu} &= (238.2 - a_1 - a_2 - a_3 - a_4 - a_5)/5 \\ &= 47.64 \\ \mu &= \hat{\mu} + RND * (\sigma_e^2/5)^{.5} \\ &= 47.64 + (-1.21) * (4.32) \\ &= 42.41\end{aligned}$$

- **Animal 1**

$$\begin{aligned}\hat{a}_1 &= (38.5 - \mu - 7a_2 + 7a_3 + 14a_4)/29 \\ &= -.1349 \\ a_1 &= \hat{a}_1 + RND * (\sigma_e^2/29)^{.5} \\ &= -.1349 + (1.138)(1.794) \\ &= 1.9067\end{aligned}$$

- **Animal 2**

$$\begin{aligned}\hat{a}_2 &= (48.9 - \mu - 7a_1 + 14a_3 - 8a_4 + 16a_5)/30 \\ &= -6.8591/30 = -.2286 \\ a_2 &= \hat{a}_2 + RND * (\sigma_e^2/30)^{.5} \\ &= -.2286 + (.0047)(1.7638) \\ &= -.2203\end{aligned}$$

- **Animal 3**

$$\begin{aligned}\hat{a}_3 &= (64.3 - \mu + 7a_1 + 14a_2 + 14a_4)/36 \\ &= .8931 \\ a_3 &= \hat{a}_3 + RND * (\sigma_e^2/36)^{.5} \\ &= .8931 + (-1.1061)(1.6102) \\ &= -.8879\end{aligned}$$

- **Animal 4**

$$\begin{aligned}
\hat{a}_4 &= (50.5 - \mu + 14a_1 - 8a_2 + 14a_3 + 16a_5)/37 \\
&= .6518 \\
a_4 &= \hat{a}_4 + RND * (\sigma_e^2/37)^{.5} \\
&= .6518 + (-1.2293)(1.5882) \\
&= -1.3006
\end{aligned}$$

- **Animal 5**

$$\begin{aligned}
\hat{a}_5 &= (36.0 - \mu + 16a_2 + 16a_4)/33 \\
&= -.9316 \\
a_5 &= \hat{a}_5 + RND * (\sigma_e^2/33)^{.5} \\
&= -.9316 + (-.6472)(1.6817) \\
&= -2.0200
\end{aligned}$$

110.1.2 Residual Variance

Calculate the residuals and their sum of squares in order to obtain a new residual variance.

$$\begin{aligned}
e_1 &= 38.5 - 42.41 - 1.9067 = -5.8167 \\
e_2 &= 48.9 - 42.41 + .2203 = 6.7103 \\
e_3 &= 64.3 - 42.41 + .8879 = 22.7779 \\
e_4 &= 50.5 - 42.41 + 1.3006 = 9.3906 \\
e_5 &= 36.0 - 42.41 + 2.0200 = -4.3900 \\
\mathbf{e}'\mathbf{e} &= 705.1503
\end{aligned}$$

A new sample value of the residual variance is

$$\begin{aligned}
\sigma_e^2 &= (\mathbf{e}'\mathbf{e} + v_e S_e^2)/CHI(15) \\
&= (705.1503 + (10)(93.3333))/17.1321 \\
&= 95.6382.
\end{aligned}$$

110.1.3 Additive Genetic Variance

The additive genetic variance requires calculation of $\mathbf{a}'\mathbf{A}^{-1}\mathbf{a}$ using the a-values obtained above, which gives

$$\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} = 19.85586.$$

Then

$$\begin{aligned}
\sigma_a^2 &= (\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + v_a S_a^2)/CHI(15) \\
&= (19.85586 + (10)(6.66667))/10.7341 \\
&= 8.0605.
\end{aligned}$$

A new sample value of the variance ratio becomes

$$k = 95.6382/8.0605 = 11.8650.$$

110.2 Round 2

Round 2 begins by re-forming the MME using the new variance ratio. The equations change to

$$\left(\begin{array}{rrrrrr}
5 & 1 & 1 & 1 & 1 & 1 \\
1 & 24.73 & 5.93 & -5.93 & -11.86 & 0 \\
1 & 5.93 & 25.58 & -11.86 & 6.78 & -13.56 \\
1 & -5.93 & -11.86 & 30.66 & -11.86 & 0 \\
1 & -11.86 & 6.78 & -11.86 & 31.51 & -13.56 \\
1 & 0 & -13.56 & 0 & -13.56 & 28.12
\end{array}\right)
\left(\begin{array}{c}
\mu \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5
\end{array}\right)
=
\left(\begin{array}{r}
238.2 \\ 38.5 \\ 48.9 \\ 64.3 \\ 50.5 \\ 36.0
\end{array}\right).$$

110.2.1 Fixed and Random Effects of Model

The process is repeated using the last values of μ and $\mathbf{a}$ and σ_e^2.

$$\begin{aligned}
\hat{\mu} &= (238.2 - a_1 - a_2 - a_3 - a_4 - a_5)/5 \\
&= 48.14 \\
\mu &= \hat{\mu} + RND * (\sigma_e^2/5)^{.5} \\
&= 48.14 + (.7465) * (4.3735) \\
&= 51.41
\end{aligned}$$

$$\begin{aligned}
\hat{a}_1 &= (38.5 - \mu - 5.93a_2 + 5.93a_3 + 11.86_4)/24.73 \\
&= -1.3059 \\
a_1 &= \hat{a}_1 + RND * (\sigma_e^2/24.73)^{.5} \\
&= -1.3059 + (-.0478)(1.9665) \\
&= -1.3999
\end{aligned}$$

$$\hat{a}_2 = (48.9 - \mu - 5.93a_1 + 11.86a_3 - 6.78a_4 + 13.56a_5)/25.58$$

$$
\begin{aligned}
&= -.9113 \\
a_2 &= \hat{a}_2 + RND * (\sigma_e^2/25.58)^{.5} \\
&= -.9113 + (.8386)(1.9336) \\
&= .7102
\end{aligned}
$$

$$
\begin{aligned}
\hat{a}_3 &= -2.41355/30.66 \\
&= -.0787 \\
a_3 &= \hat{a}_3 + RND * (\sigma_e^2/30.66)^{.5} \\
&= -.0787 + (-1.8414)(1.7662) \\
&= -3.3309
\end{aligned}
$$

$$
\begin{aligned}
\hat{a}_4 &= -89.2236/31.51 = -2.8316 \\
a_4 &= -2.8316 + (-1.2549)(1.7422) \\
&= -5.0179
\end{aligned}
$$

$$
\begin{aligned}
\hat{a}_5 &= -73.8224/28.12 = -2.6253 \\
a_5 &= -2.6253 + (.8184)(1.8442) \\
&= -1.1160
\end{aligned}
$$

110.2.2 Residual Variance

The residuals and their sum of squares are

$$
\begin{aligned}
e_1 &= 38.5 - 51.41 + 1.3999 = -11.5101 \\
e_2 &= 48.9 - 51.41 - .7102 = -3.2202 \\
e_3 &= 64.3 - 51.41 + 3.3309 = 16.2209 \\
e_4 &= 50.5 - 51.41 + 5.0179 = 4.1079 \\
e_5 &= 36.0 - 51.41 + 1.1160 = -14.2940 \\
\mathbf{e}'\mathbf{e} &= 627.1630
\end{aligned}
$$

The new sample value of the residual variance is

$$
\begin{aligned}
\sigma_e^2 &= (\mathbf{e}'\mathbf{e} + v_e S_e^2)/CHI(15) \\
&= (627.1630 + (10)(93.3333))/20.4957 \\
&= 76.1377.
\end{aligned}
$$

110.2.3 Additive Genetic Variance

The new sample value of the additive genetic variance is

$$\begin{aligned}\sigma_a^2 &= (\mathbf{a}'\mathbf{A}^{-1}\mathbf{a} + v_a S_a^2)/CHI(15) \\ &= (36.8306 + (10)(6.66667))/16.6012 \\ &= 6.2343.\end{aligned}$$

The new variance ratio becomes

$$k = 76.1377/6.2343 = 12.2127.$$

Continue taking samples for thousands of rounds.

110.3 Burn-In Periods

The samples do not immediately represent samples from the joint posterior distribution. Generally, this takes anywhere from 100 to 10,000 samples depending on the model and amount of data. This period is known as the *burn-in period*. Samples from the burn-in period are discarded. The length of the burn-in period (i.e. number of samples) is usually judged by visually inspecting a plot of sample values across rounds.

A less subjective approach to determine convergence to the joint posterior distribution is to run two chains at the same time, both beginning with the same random number seed. However, the starting values (in variances) for each chain are usually greatly different, e.g. one set is greatly above the expected outcome and the other set is greatly below the expected outcome. When the two chains essentially become one chain, i.e. the squared difference between variance estimates is less than a specified value (like 10^{-5}), then convergence to the joint posterior distribution has occurred. All previous samples are considered to be part of the burn-in period and are discarded.

110.4 Post Burn-In Analysis

After burn-in, each round of Gibbs sampling is dependent on the results of the previous round. Depending on the total number of observations and parameters, one round may be positively correlated with the next twenty to three hundred rounds. The user can determine the effective number of samples by calculating lag correlations, i.e. the correlation of estimates between rounds, between every other round, between every third round, etc. Determine the number of rounds between two samples such that the correlation is zero. Divide the total number of samples (after burn-in) by the interval that gives a zero correlation, and that gives the effective number of samples. Suppose a total of 12,000 samples (after removing the burn-in rounds) and an interval of 240 rounds gives a zero correlation between samples, then the effective number of samples is 12,000 divided by 240 or 50 samples. There is no minimum number of independent samples that are required, just the need to know how many there actually were.

An overall estimate of a parameter can be obtained by averaging all of the 12,000 samples (after the burn-in). However, to derive a confidence interval or to plot the distribution of the samples or to calculate the standard deviation of the sample values, the variance of the independent samples should be used.

The final estimates are therefore, an average of the sample estimates. Some research has shown that the mode of the estimates might be a better estimate, which indicates that the distribution of sample estimates is skewed. One could report both the mean and mode of the samples, however, the mode should be based on the independent samples only.

110.5 Influence of the Priors

In the small example, $v_a = v_e = 10$ whereas N was only 5. Thus, the prior values of the variances received more weight than information coming from the data. This is probably appropriate for this small example, but if N were 5,000,000, then the influence of the priors would be next to nothing. The amount of influence of the priors is not directly determined by the ratio of v_i to N. In the small example, even though $v_a/(N + v_a) = \frac{2}{3}$, the influence of S_a^2 could be greater than $\frac{2}{3}$. When N is large there may be no need for v_a or v_e at all, or at least very small values would suffice.

110.6 Long Chain or Many Chains?

Early papers on MCMC (Monte Carlo Markov Chain) methods recommended running many chains of samples and then averaging the final values from each chain. This was to insure independence of the samples. Another philosophy recommends one single long chain. For animal breeding applications this could mean 100,000 samples or more. If a month is needed to run 50,000 samples, then maybe three chains of 50,000 would be preferable, all running simultaneously on a network of computers. If only an hour is needed for 50,000 samples, then 1,000,000 samples would not be difficult to run.

Two chains that utilize the same sequence of random numbers, but which use different starting variances, are recommended for determining the burn-in period, after which enough samples need to be run to generate a sufficient number of independent samples for obtaining standard deviations of the samples. A sufficient number of independent samples may be 100 or more depending on the amount of time needed to generate samples.

110.7 Heritabilities and Correlations

The Gibbs sampling process gives samples of variances, and usually each sample is saved (or every m^{th} sample). Thus, for each saved sample of variances, the heritability (or genetic correlation) could be calculated or the ratio of residual to additive variance, or any other quantity that may be of interest to the user. Then those values could be averaged and the variance of the samples calculated to give a standard error of the overall heritability estimate. This gives the user a good idea of the variability in these ratios.

110.8 Estimated Breeding Values

Although not common, Gibbs sampling can be used to get Estimated Breeding Values (EBV) of animals and their standard errors of prediction (across samples). The standard errors of prediction could then be converted to a reliability of the EBV rather than deriving an approximation for reliability. Only 100 to 500 additional rounds of Gibbs sampling are needed for this purpose.

111 EXERCISES

Below are data on progeny of 4 sires. Sires are assumed to be unrelated. Each number is an observation on one progeny.

Data For Assignment.

Sire	Contemporary Groups		
	1	2	3
1	13, 9	3, 9	12, 18
2	3	8, 13	6
3	-	18, 10	15
4	6, 8	-	9

Assume the model equation

$$y_{ijk} = c_i \; + \; s_j \; + e_{ijk},$$

where c_i is a fixed contemporary group effect, s_j is a random sire effect, and e_{ijk} is a random residual effect.

1. Set up the MME. Assume that $\sigma^2_e/\sigma^2_s = 10$.
2. Apply REML EM to the model (Just one iteration) to estimate the sire and residual variances. Calculate an estimate of heritability.
3. Perform many rounds of Gibbs sampling on the MME solutions and on the variances. The MME solutions have normal conditional posterior distributions, and the variances have inverted Chi-square distributions. Assume degrees of belief equal to the number of observations, and prior values equal to the estimates from the previous question.
4. Plot the sample values of the variances.
5. Calculate an estimate of heritability for each Gibbs sample, and compute the mean and variance of the sample heritability values.

Repeated Records Animal Model

112 Introduction

Animals are observed more than once for some traits, such as

- Fleece weight of sheep in different years.
- Calf records of a beef cow over time.
- Test day records within a lactation for a dairy cow.
- Litter size of sows over time.
- Antler size of deer in different seasons.
- Racing results of horses from several races.

Usually the trait is considered to be perfectly correlated over the ages of the animal. Besides an animal's additive genetic value for a trait, there is a common permanent environmental (PE) effect which is a non-genetic effect common to all observations on the same animal.

113 The Model

The model is written as

$$\mathbf{y} = \mathbf{Xb} + \begin{pmatrix} \mathbf{0} & \mathbf{Z} \end{pmatrix} \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{a}_r \end{pmatrix} + \mathbf{Zp} + \mathbf{e},$$

where

$$\begin{array}{rcl} \mathbf{b} & = & \text{vector of fixed effects,} \\ \begin{pmatrix} \mathbf{a}_0 \\ \mathbf{a}_r \end{pmatrix} & = & \begin{pmatrix} \text{animals without records} \\ \text{animals with records} \end{pmatrix}, \\ \mathbf{p} & = & \text{vector of PE effects of length equal to } \mathbf{a}_r \text{ , and} \\ \mathbf{e} & = & \text{vector of residual effects.} \end{array}$$

The matrices **X** and **Z** are design matrices that associate observations to particular levels of fixed effects and to additive genetic and PE effects, respectively. In a repeated records model, $\mathbf{Z}$ is not equal to an identity matrix. Also,

$$\begin{aligned}
\mathbf{a} \mid \mathbf{A}, \sigma_a^2 &\sim N(\mathbf{0}, \mathbf{A}\sigma_a^2) \\
\mathbf{p} \mid \mathbf{I}, \sigma_p^2 &\sim N(\mathbf{0}, \mathbf{I}\sigma_p^2) \\
\mathbf{e} &\sim N(\mathbf{0}, \mathbf{I}\sigma_e^2) \\
\mathbf{G} &= \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_p^2 \end{pmatrix}.
\end{aligned}$$

Repeatability is a measure of the average similarity of multiple records on animals across the population (part genetic and part environmental), and is defined as a ratio of variances as

$$r = \frac{\sigma_a^2 + \sigma_p^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2},$$

which is always going to be greater than or equal to heritability, because

$$h^2 = \frac{\sigma_a^2}{\sigma_a^2 + \sigma_p^2 + \sigma_e^2}.$$

114 Simulation of Records

Simulating multiple records on animals may help to understand this type of model. Let

$$\begin{aligned}
\sigma_a^2 &= 36 \\
\sigma_p^2 &= 16 \text{ and} \\
\sigma_e^2 &= 48
\end{aligned}$$

Thus,

$$h^2 = \frac{36}{36 + 16 + 48} = .36,$$

and

$$r = \frac{36 + 16}{36 + 16 + 48} = .52.$$

114.1 Data Structure

Animal	Sire	Dam	Year 1	Year 2	Year 3
7	1	2	$\surd$	$\surd$	$\surd$
8	3	4	$\surd$	$\surd$	
9	5	6	$\surd$		$\surd$
10	1	4		$\surd$	$\surd$
11	3	6			$\surd$
12	1	2		$\surd$	

None of the animals are inbred, so that the inverse of the additive genetic relationship matrix is

$$\mathbf{A}^{-1} = \frac{1}{2} \left(\begin{array}{rrrrrrrrrrrr} 5 & 2 & 0 & 1 & 0 & 0 & -2 & 0 & 0 & -2 & 0 & -2 \\ 2 & 4 & 0 & 0 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & -2 \\ 0 & 0 & 4 & 1 & 0 & 1 & 0 & -2 & 0 & 0 & -2 & 0 \\ 1 & 0 & 1 & 4 & 0 & 0 & 0 & -2 & 0 & -2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 1 & 0 & 0 & -2 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 4 & 0 & 0 & -2 & 0 & -2 & 0 \\ -2 & -2 & 0 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -2 & -2 & 0 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -2 & -2 & 0 & 0 & 4 & 0 & 0 & 0 \\ -2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 0 \\ 0 & 0 & -2 & 0 & 0 & -2 & 0 & 0 & 0 & 0 & 4 & 0 \\ -2 & -2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 4 \end{array} \right).$$

114.2 Additive Genetic Values of Animals

The first six animals are assumed to be base generation animals, and should be generated first. Let RND represent a random normal deviate, and b_i is $0.5 - 0.25(F_s + F_d)$.

Animal	Parent Ave.	RND	$(36 * b_i)^{.5}$	TBV
1	0.0	-2.5038	6.0	-15.0228
2	0.0	-.3490	6.0	-2.0940
3	0.0	-.2265	6.0	-1.3590
4	0.0	-.3938	6.0	-2.3628
5	0.0	1.4786	6.0	8.8716
6	0.0	2.3750	6.0	14.2500
7	-8.5584	-.8166	4.2426	-12.0229
8	-1.8609	1.0993	4.2426	2.8030
9	11.5608	1.5388	4.2426	18.0893
10	-8.6928	.0936	4.2426	-8.2957
11	6.4455	1.3805	4.2426	12.3024
12	-8.5584	-1.2754	4.2426	-13.9694

114.3 Permanent Environmental Effects

Each animal has a PE effect that is common to each of its own records, but is not transmitted to progeny. Genetic relationships have no bearing on PE effects. Generate a RND and multiply by $\sigma_p = 4$. These are shown in the table below.

Animal	TBV	PE
1	-15.02	2.97
2	-2.09	-9.04
3	-1.36	4.44
4	-2.36	-4.16
5	8.87	-5.68
6	14.25	6.85
7	-12.02	1.38
8	2.80	7.02
9	18.09	5.94
10	-8.30	-5.03
11	12.30	-1.06
12	-13.97	-2.69

114.4 Records

Records are generated according to the model equation,

$$y_{ijk} \;=\; t_i + a_j + p_j + e_{ijk},$$

where t_i is a year effect. Let $t_1 = 53$, $t_2 = 59$, and $t_3 = 65$. Note that $\sigma_p = 4$, and $\sigma_e = 6.9282$. Residual values are generated for each observation as $RND * \sigma_e$. Add together the pieces and round to the nearest whole number.

			Year 1	Year 2	Year 3
Animal	TBV	PE	y_{1jk}	y_{2jk}	y_{3jk}
1	-15.02	2.97			
2	-2.09	-9.04			
3	-1.36	4.44			
4	-2.36	-4.16			
5	8.87	-5.68			
6	14.25	6.85			
7	-12.02	1.38	39	51	62
8	2.80	7.02	48	72	
9	18.09	5.94	71		96
10	-8.30	-5.03		56	47
11	12.30	-1.06			86
12	-13.97	-2.69		46	

An interesting point to observe from the simulation is that the PE effects are present even for animals with only one record. Also, the same PE value is present in all records of one animal.

Another assumption is that the records have a genetic correlation of one, which is true in the way that the above records were simulated because the same TBV was used for each record. In real life, genes affecting a trait might change as the animal ages, and therefore, the genetic correlation between successive records could be less than unity.

115 Mixed Model Equations

Let

$$\mathbf{W} = \left[\begin{array}{ccc} \mathbf{X} & \left(\begin{array}{cc} \mathbf{0} & \mathbf{Z} \end{array} \right) & \mathbf{Z} \end{array} \right],$$

then

$$\mathbf{W}'\mathbf{W} = \left(\begin{array}{cccc} \mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \end{array} \right), \quad \mathbf{W}'\mathbf{y} = \left(\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{0} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{array} \right),$$

and

$$\Sigma = \left(\begin{array}{cccc} \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{00}k_a & \mathbf{A}^{0r}k_a & \mathbf{0} \\ \mathbf{0} & \mathbf{A}^{r0}k_a & \mathbf{A}^{rr}k_a & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}k_p \end{array} \right),$$

where $\mathbf{A}^{ij}$ are corresponding elements of the inverse of the additive genetic relationship matrix (given earlier) partitioned according to animals without and with records. In this example, each submatrix is of order 6. Also,

$$k_a = \sigma_e^2/\sigma_a^2 = 1.33333, \quad \text{and} k_p = \sigma_e^2/\sigma_p^2 = 3.$$

MME are therefore,

$$\begin{array}{rcl} (\mathbf{W}'\mathbf{W} + \Sigma)\beta & = & \mathbf{W}'\mathbf{y} \\ (\mathbf{W}'\mathbf{W} + \Sigma)\beta & = & \left(\begin{array}{cccc} \mathbf{X}'\mathbf{X} & \mathbf{0} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{0} & \mathbf{A}^{00}k_a & \mathbf{A}^{0r}k_a & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{A}^{r0}k_a & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{rr}k_a & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}k_p \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_0 \\ \hat{\mathbf{a}}_r \\ \hat{\mathbf{p}} \end{array} \right). \end{array}$$

Let a generalized inverse of the coefficient matrix be represented as

$$(\mathbf{W}'\mathbf{W} + \Sigma)^{-} = \left(\begin{array}{ccc} - & - & - \\ - & \mathbf{C}_{aa} & - \\ - & - & \mathbf{C}_{pp} \end{array} \right),$$

where $\mathbf{C}_{aa}$ is of order 12 in this case, and $\mathbf{C}_{pp}$ is of order 6.

The full HMME are too large to present here as a whole, so parts of the matrix are given as follows.

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{pmatrix}, \quad \mathbf{X}'\mathbf{y} = \begin{pmatrix} 158 \\ 225 \\ 291 \end{pmatrix},$$

$$\mathbf{X}'\mathbf{Z} = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 1 & 0 \end{pmatrix},$$

$$\mathbf{Z}'\mathbf{Z} = diag(3 \;\; 2 \;\; 2 \;\; 2 \;\; 1 \;\; 1),$$

and

$$\mathbf{Z}'\mathbf{y} = \begin{pmatrix} 152 \\ 120 \\ 167 \\ 103 \\ 86 \\ 46 \end{pmatrix}.$$

The solutions for animals are given in the table below. Solutions for year effects were

$$\begin{aligned} \hat{t}_1 &= 50.0858, \\ \hat{t}_2 &= 63.9612, \\ \hat{t}_3 &= 72.0582. \end{aligned}$$

Animal	TBV	PE	$\hat{\mathbf{a}}$	$\hat{\mathbf{p}}$
1	-15.02	2.97	-7.9356	
2	-2.09	-9.04	-4.4473	
3	-1.36	4.44	2.8573	
4	-2.36	-4.16	-2.6039	
5	8.87	-5.68	5.0783	
6	14.25	6.85	7.0512	
7	-12.02	1.38	-8.0551	-1.6566
8	2.80	7.02	1.0111	0.7861
9	18.09	5.94	11.1430	4.5140
10	-8.30	-5.03	-8.7580	-3.1007
11	12.30	-1.06	6.9271	1.7537
12	-13.97	-2.69	-8.7750	-2.2965

The correlation between $\hat{\mathbf{a}}$ and TBV was .9637, and between $\hat{\mathbf{p}}$ and true PE was .7215.

116 Reliability of EBVs

Owners of animals are very interested in Estimated Breeding Values (EBVs), and the main question is about its reliability or accuracy. The variance-covariance matrix of prediction errors is given by $\mathbf{C}_{aa}\sigma_e^2$. Reliability, R, of the i^{th} animal is defined as

$$R = (a_{ii}\sigma_a^2 - c_{ii}\sigma_e^2)/\sigma_a^2,$$

where a_{ii} is the diagonal of $\mathbf{A}$ for animal i, and c_{ii} is the diagonal of $\mathbf{C}_{aa}$ for animal i. Note that this is equivalent to

$$R = a_{ii} - c_{ii}k_a.$$

This procedure does not work when phantom groups are included in the formation of $\mathbf{A}^{-1}$ because then it is possible that

$$c_{ii}k_a > a_{ii}$$

for some situations. Below is a table of the reliabilities for the twelve animals in the example analysis.

Animal	$\hat{\mathbf{a}}$	c_{ii}	R
1	-7.9285	.6472	.1371
2	-4.4339	.6524	.1301
3	2.8289	.6476	.1365
4	-2.6365	.6398	.1469
5	5.0996	.6807	.0924
6	7.0703	.6653	.1129
7	-8.0155	.4740	.3680
8	.9544	.5012	.3317
9	11.1845	.4985	.3353
10	-8.7771	.5180	.3093
11	6.9204	.5656	.2459
12	-8.7807	.5592	.2544

Animals with records have a higher reliability than animals that have only progeny. Also, animal 7 had a higher reliability because it had three records while animals 11 and 12 had only one record. Reliability reflects the years in which the records were made and the number of contemporaries within a year, and specifically who the contemporaries actually were. Reliability also includes the fact that animals were related.

In the analysis of very large numbers of animals, the calculation of $\mathbf{C}_{aa}$ is virtually impossible. Thus, animal breeders have devised many ways of approximating the diagonals of $\mathbf{C}_{aa}$. The following method is due to Schaeffer and Jansen (1997).

Step 1 Account for contemporary group size and PE effects. Take animal 7 as an example. Animal 7's first record in year 1 was made with two other contemporaries, calculate

$$d_7 = 1 - \frac{1 * 1}{3} = \frac{2}{3}.$$

Animal 7's second and third records were made with three contemporaries each, so accumulate the following:

$$d_7 = d_7 + (1 - \frac{1}{4}) + (1 - \frac{1}{4})$$

or

$$d_7 = \frac{2}{3} + \frac{3}{4} + \frac{3}{4} = 2.1666667.$$

Now adjust for the fact that we must also estimate PE effects for this animal. The adjustment is

$$\begin{aligned} d_7 &= d_7 - \frac{d_7 * d_7}{d_7 + k_p} \\ &= 2.16667 - \frac{4.694444}{5.1666667} \\ &= 1.25806452. \end{aligned}$$

Finally, add $a^{ii}k_a$, diagonal element from $\mathbf{A}^{-1}k_a$, to give

$$d_7 = 1.25806 + 2(1.33333) = 3.92473.$$

This is done for all animals with records. Animals without records have $d_i = k_a$. For animals 1 through 12 the results are

$$\begin{aligned} d_1 &= 1.33333 \\ d_2 &= 1.33333 \\ d_3 &= 1.33333 \\ d_4 &= 1.33333 \\ d_5 &= 1.33333 \\ d_6 &= 1.33333 \\ d_7 &= 3.92473 \\ d_8 &= 3.62893 \\ d_9 &= 3.62893 \\ d_{10} &= 3.66667 \\ d_{11} &= 3.26667 \\ d_{12} &= 3.26667 \end{aligned}$$

Step 2 Convert the above numbers into a number that would represent an equivalent number of progeny, n_i, by

$$n_i = (d_i - k_a)/.5k_a.$$

This gives

$$\begin{aligned}
n_1 &= 0.00000 \\
n_2 &= 0.00000 \\
n_3 &= 0.00000 \\
n_4 &= 0.00000 \\
n_5 &= 0.00000 \\
n_6 &= 0.00000 \\
n_7 &= 3.88710 \\
n_8 &= 3.44339 \\
n_9 &= 3.44339 \\
n_{10} &= 3.50000 \\
n_{11} &= 2.90000 \\
n_{12} &= 2.90000
\end{aligned}$$

Step 3 Add contributions to parents. Animals must be processed from youngest to oldest. Let $\alpha = (4 - h^2)/h^2 = 10.11111$. The contribution to a parent is

$$q_i = .25\alpha t_i/(1 - .25t_i),$$

where

$$t_i = n_i/(n_i + \alpha),$$

or

$$q_i = .25\alpha * n_i/(.75n_i + \alpha).$$

The value q_i is added to the n_s of the sire of i and to the n_d of the dam of i. The q_i values of animals 7 through 12 are given below.

Animal	ne_i	t_i	q_i
7	3.88710	.277686	.754293
8	3.44339	.254040	.685706
9	3.44339	.254040	.685706
10	3.50000	.257143	.694657
11	2.90000	.222886	.596653
12	2.90000	.222886	.596653

For animal 7, for example, $q_i = .754293$ is added to n_1 and n_2 because the parents of 7 are animals 1 and 2. Animal 1 receives contributions from animals 7, 10, and 12, or

$$ne_1 = 0.0 + .754293 + .694657 + .596653 = 2.045603$$

Similarly for all parents,

$$\begin{aligned}
ne_2 &= 0.0 + .754293 + .596653 = 1.350946 \\
ne_3 &= 0.0 + .685706 + .596653 = 1.282359 \\
ne_4 &= 0.0 + .685706 + .694657 = 1.380363 \\
ne_5 &= 0.0 + .685706 = .685706 \\
ne_6 &= 0.0 + .685706 + .596653 = 1.282359
\end{aligned}$$

In real datasets, animals could very likely have both records and progeny.

Step 4 Set up selection index equations with progeny on the animal, on its sire, and on its dam. Assume the sire and dam are unrelated. Use n_i as the number of progeny for animal i. Take animal 7 as an example with sire equal to animal 1 and dam equal to animal 2, and calculate

$$\begin{aligned}
m_1 &= (n_1 - q_7)/(n_1 - q_7 + \alpha) \\
&= (2.045603 - .754293)/(1.291310 + 10.111111) \\
&= 1.291310/11.402421 \\
&= .113249 \\
m_2 &= (n_2 - q_7)/(n_2 - q_7 + \alpha) \\
&= (1.350946 - .754293)/(.596653 + 10.111111) \\
&= .055722 \\
t_{1,2} &= (m_1 + m_2)/4 \\
&= (.113249 + .055722)/4 \\
&= .042243 \\
q_{1,2} &= \alpha t_{1,2}/(1 - t_{1,2}) \\
&= .445958
\end{aligned}$$

Now $q_{1,2}$ is the contribution (in progeny equivalents) to the number of effective progeny for animal 7. Thus,

$$\begin{aligned}
R &= (n_7 + q_{1,2})/(n_7 + q_{1,2} + \alpha) \\
&= (3.88710 + .445958)/(4.333058 + 10.111111) \\
&= .299987.
\end{aligned}$$

This approximation is much less than the value of .3680 derived from the diagonal of $\mathbf{C}_{aa}$. The small number of records in this example could be a cause for the disagreement.

For animal 1, the parents are unknown and so

$$\begin{aligned}
R &= n_1/(n_1 + \alpha) \\
&= 2.045603/(2.045603 + 10.111111) \\
&= .168269,
\end{aligned}$$

which is greater than the value of .1371 from $\mathbf{C}_{aa}$ and shows that approximations do not work in all cases. For animal 12,with sire equal to animal 1 and dam equal to animal 2, and calculate

$$\begin{aligned}
m_1 &= (n_1 - q_{12})/(n_1 - q_{12} + \alpha) \\
&= (2.045603 - .596653)/(1.448950 + 10.111111) \\
&= 1.448950/11.560061 \\
&= .125341 \\
m_2 &= (n_2 - q_{12})/(n_2 - q_{12} + \alpha) \\
&= (1.350946 - .596653)/(1.754293 + 10.111111) \\
&= .117862 \\
t_{1,2} &= (m_1 + m_2)/4 \\
&= (.125341 + .117862)/4 \\
&= .060801 \\
q_{1,2} &= \alpha t_{1,2}/(1 - t_{1,2}) \\
&= .654562
\end{aligned}$$

Now $q_{1,2}$ is the contribution (in progeny equivalents) to the number of effective progeny for animal 12. Thus,

$$\begin{aligned}
R &= (n_{12} + q_{1,2})/(n_{12} + q_{1,2} + \alpha) \\
&= (2.9 + .654562)/(3.554562 + 10.111111) \\
&= .2601,
\end{aligned}$$

which is only slightly higher than .2544 given by $\mathbf{C}_{aa}$.

Reliabilities are required to determine the 'official' status of an animal's EBV. The approximations that are used should be on the conservative side for safety reasons.

There may also be a method of determining approximate reliabilities by using Gibbs sampling, but not allowing the variances to change in each round. The starting values would be the solutions to the MME and the known variances. Then about 200 rounds of sampling should give a good estimate of the prediction error variance of the EBVs for each animal, which can then be used to arrive at reliability. Two hundred samples would be calculated for each animal, and the standard deviation of those samples would estimate the square root of the prediction error variances. This would be similar to computing 200 more iterations in the solution phase of the program.

117 Selection of Animals to Have Later Records

Often records on animals are taken over time, as the animal ages. Consequently the observed value of the first record determines if that animal makes a second record, and so on. Thus, selection could affect the mean and variance of later repeated records. This type of selection is handled adequately by HMME **provided** that all first records of animals and pedigrees are known for all animals. If these provisions are not met, then EBVs and estimates of fixed effects from a repeated records analysis could be biased by culling, with the magnitude determined by the severity of the culling.

118 Permanent Effects?

Permanent environmental effects may not really be permanent. Environmental effects are encountered all through the life of any living animal. There are certain effects that become part of that animal's performance ability for the entirety of its life. However, there could be many effects that become incorporated into the animal's performance ability as the animal gains experience and encounters new events. For example, a tennis player may develop a wicked backhand shot due to a coach that forced the player to practice it endlessly. A few years later, a different coach may work more on the player's serve. Thus, environmental effects accumulate over time. Both coaches now affect that player's performance ability for the rest of his(her) life. Eventually age, injury, stress, success become other environmental effects.

A model with one permanent environmental effect per animal represents an average of all such environmental effects over the performance lifetime of the animal. There is not an easy way to model accumulated environmental effects with time. Putting a time effect into the model is not sufficient, because the time effects may be different for each animal.

The accumulated effects are averaged into the PE effect, and the deviations around that (animal by time interaction effects) become part of the residual effect. Either a random regression model or a multiple trait model are recommended for repeated record situations because of the above problems.

119 EXERCISES

Below are scores (out of 200) of dogs in three different agility competitions. Assume a repeated records animal model where the score has a heritability of 0.25, and a repeatability of 0.45. Each trial was held in a different city with a different judge and course setter.

Scores (out of 200) for three agility competitions for dogs.

Dog	Sire	Dam	Age	Trial 1	Trial 2	Trial 3
1	-	-	7	135		150
2	-	-	6	110		
3	1	2	4	127	134	130
4	1	2	4	108	140	
5	1	-	4	95		104
6	1	-	5	138	161	
7	3	-	2	154		166
8	3	4	2		155	140
9	5	-	1		128	
10	5	6	1			117

1. Write a complete repeated records model for these data.
2. Construct HMME and solve. Which dogs are tops?
3. Provide prediction error variances for the EBV.
4. Show how to estimate the variances for one round of EM-REML, or one round of Gibbs sampling.
5. Predict a record of animal 9 in Trial 3, if it had been entered. Also give a prediction error variance for that record.

Maternal Genetic Models

120 Introduction

In mammalian species of livestock, such as beef cattle, sheep or swine, the female provides an environment for its offspring to survive and grow in terms of protection and nourishment. Females vary in their ability to provide a suitable environment for their offspring, and this variability has a genetic basis. Offspring directly inherit an ability to grow (or survive) from both parents, and environmentally do better or poorer depending on their dam's genetic maternal ability. Maternal ability is a genetic trait expressed by the dam in the offsprings' performance, and is transmitted, like all genetic traits, from both parents. Maternal ability is only expressed by females when they have offspring (i.e. much like milk yield in dairy cows).

A model to account for maternal ability is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{m} + \mathbf{Z}_3\mathbf{p} + \mathbf{e},$$

where $\mathbf{y}$ is the growth trait of a young animal, $\mathbf{b}$ is a vector of fixed factors influencing growth, such as contemporary group, sex of the offspring, or age of dam, $\mathbf{a}$ is a vector of random additive genetic effects (i.e. direct genetic effects) of the animals, $\mathbf{m}$ is a vector of random maternal genetic (dam) effects, and $\mathbf{p}$, in this model, is a vector of maternal permanent environmental effects (because dams may have more than one offspring in the data).

The expectations of the random vectors, $\mathbf{a}$, $\mathbf{m}$, $\mathbf{p}$, and $\mathbf{e}$ are all null vectors in a model without selection, and the variance-covariance structure is

$$Var\begin{pmatrix} \mathbf{a} \\ \mathbf{m} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma_a^2 & \mathbf{A}\sigma_{am} & \mathbf{0} & \mathbf{0} \\ \mathbf{A}\sigma_{am} & \mathbf{A}\sigma_m^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix},$$

where σ_a^2 is the additive genetic variance, σ_m^2 is the maternal genetic variance, σ_{am} is the additive genetic by maternal genetic covariance, and σ_p^2 is the maternal permanent environmental variance. Also,

$$\left(\begin{array}{c|c} \mathbf{a} & \\ & \mathbf{A}, \mathbf{G} \\ \mathbf{m} & \end{array} \right) \sim N\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \quad \mathbf{G} \otimes \mathbf{A} \right),$$

where

$$\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix},$$

and

$$\mathbf{p} \mid \mathbf{I}, \sigma_p^2 \sim N(\mathbf{0}, \mathbf{I}\sigma_p^2),$$

and

$$\mathbf{e} \sim N(\mathbf{0}, \mathbf{I}\sigma_e^2).$$

In this model, a female animal, i, could have its own growth record for estimating $\hat{a}_i$. The same female could later have offspring for estimating $\hat{m}_i$ and $\hat{p}_i$, and the offspring would also contribute towards $\hat{a}_i$. The maternal effects model can be more complicated if, for example, embryo transfer is practiced. Recipient dams would have maternal effects, but would not have direct genetic effects on that calf, see Schaeffer and Kennedy (1989).

To better understand this model, simulation of records is again useful. Let

$$\mathbf{G} = \begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix} = \begin{pmatrix} 49 & -7 \\ -7 & 26 \end{pmatrix}.$$

Any positive definite matrix can be partitioned into the product of a matrix times its transpose (i.e. Cholesky decomposition), or

$$\begin{aligned} \mathbf{G} &= \mathbf{LL}' \\ \mathbf{L} &= \begin{pmatrix} 7 & 0 \\ -1 & 5 \end{pmatrix}. \end{aligned}$$

Let $\sigma_p^2 = 9$ and $\sigma_e^2 = 81$. This model differs from previous models in that both the additive genetic and maternal genetic effects need to be generated simultaneously because these effects are genetically correlated. The same is true for multiple trait models. Consider three animals, A, B, and C, where C is an offspring of sire A and dam B.

120.1 Genetic Values

1. For A, generate a vector of two random normal deviates which will be pre-multiplied by $\mathbf{L}$. Animals A and B are base population animals that are unrelated to each other. Let the vector of random normal deviates be $\mathbf{w}' = (2.533 \quad -.299)$, then for A

$$\begin{aligned} \begin{pmatrix} a_A \\ m_A \end{pmatrix} &= \mathbf{Lw} \\ &= \begin{pmatrix} 7 & 0 \\ -1 & 5 \end{pmatrix} \begin{pmatrix} 2.533 \\ -.299 \end{pmatrix} \\ &= \begin{pmatrix} 17.731 \\ -4.028 \end{pmatrix}. \end{aligned}$$

2. Similarly for animal B,

$$\begin{aligned}\begin{pmatrix} a_B \\ m_B \end{pmatrix} &= \begin{pmatrix} 7 & 0 \\ -1 & 5 \end{pmatrix}\begin{pmatrix} -1.141 \\ .235 \end{pmatrix} \\ &= \begin{pmatrix} -7.987 \\ 2.316 \end{pmatrix}.\end{aligned}$$

3. Creating a progeny's true breeding value is similar to the scalar version. Take the average of the parents' true breeding values and add a random Mendelian sampling term.

$$\begin{aligned}\begin{pmatrix} a_C \\ m_C \end{pmatrix} &= \frac{1}{2}\begin{pmatrix} a_A + a_B \\ m_A + m_B \end{pmatrix} + (b_{ii})^{.5}\mathbf{Lw} \\ &= \frac{1}{2}\begin{pmatrix} 17.731 - 7.987 \\ -4.028 + 2.316 \end{pmatrix} + (\frac{1}{2})^{.5}\mathbf{L}\begin{pmatrix} .275 \\ .402 \end{pmatrix} \\ &= \begin{pmatrix} 6.233 \\ .371 \end{pmatrix}.\end{aligned}$$

All animals have both direct and maternal genetic breeding values.

120.2 Maternal Permanent Environmental Values

For all dams, a maternal permanent environmental effect should be generated. In this case only for animal B, multiply a random normal deviate by $\sigma_p = 3$, suppose it is -4.491.

120.3 Phenotypic Record

An observation for animal C is created by following the model equation,

$$\begin{aligned} y &= \text{Fixed Effects} + a_C + m_B + p_B + \sigma_e * RND \\ &= 140 + 6.233 + 2.316 + (-4.491) + (9)(1.074) \\ &= 153.724. \end{aligned}$$

The Fixed Effects contribution of 140 was arbitrarily chosen for this example. The main point is that the observation on animal C consists of the direct genetic effect of animal C plus the maternal genetic effect of the dam (B) plus the maternal permanent environmental effect of the dam (B) plus a residual.

121 HMME

To illustrate the calculations, assume the data as given in the table below.

Animal	Sire	Dam	CG	Weight
5	1	3	1	156
6	2	3	1	124
7	1	4	1	135
8	2	4	2	163
9	1	3	2	149
10	2	4	2	138

CG stands for contemporary group, the only fixed effect in this example. Assume that the appropriate variance parameters are those which were used in the simulation in the previous section. Based on the matrix formulation of the model, the MME are

$$\left(\begin{array}{cccc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_1 & \mathbf{X}'\mathbf{Z}_2 & \mathbf{X}'\mathbf{Z}_3 \\ \mathbf{Z}_1'\mathbf{X} & \mathbf{Z}_1'\mathbf{Z}_1 + \mathbf{A}^{-1}k_{11} & \mathbf{Z}_1'\mathbf{Z}_2 + \mathbf{A}^{-1}k_{12} & \mathbf{Z}_1'\mathbf{Z}_3 \\ \mathbf{Z}_2'\mathbf{X} & \mathbf{Z}_2'\mathbf{Z}_1 + \mathbf{A}^{-1}k_{12} & \mathbf{Z}_2'\mathbf{Z}_2 + \mathbf{A}^{-1}k_{22} & \mathbf{Z}_2'\mathbf{Z}_3 \\ \mathbf{Z}_3'\mathbf{X} & \mathbf{Z}_3'\mathbf{Z}_1 & \mathbf{Z}_3'\mathbf{Z}_2 & \mathbf{Z}_3'\mathbf{Z}_3 + \mathbf{I}k_{33} \end{array}\right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{p}} \end{array}\right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_1'\mathbf{y} \\ \mathbf{Z}_2'\mathbf{y} \\ \mathbf{Z}_3'\mathbf{y} \end{array}\right),$$

where

$$\begin{aligned} \left(\begin{array}{cc} k_{11} & k_{12} \\ k_{12} & k_{22} \end{array}\right) &= \left(\begin{array}{cc} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{array}\right)^{-1} \sigma_e^2, \\ &= \left(\begin{array}{cc} 49 & -7 \\ -7 & 26 \end{array}\right)^{-1} (81), \\ &= \left(\begin{array}{cc} 1.7192 & .4628 \\ .4628 & 3.2400 \end{array}\right). \end{aligned}$$

Note that these numbers are not equal to

$$\left(\begin{array}{cc} 81/49 & 81/(-7) \\ 81/(-7) & 81/26 \end{array}\right).$$

Finally, $k_{33} = \sigma_e^2/\sigma_p^2 = 81/9 = 9$.

The matrices are

$$\mathbf{X} = \left(\begin{array}{cc} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{array}\right), \quad \mathbf{X}'\mathbf{y} = \left(\begin{array}{c} 415 \\ 450 \end{array}\right),$$

$$\mathbf{Z}_1 = \left(\begin{array}{cccccccccc} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right),$$

$$\mathbf{Z}_2 = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

$$\mathbf{Z}_3 = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Z}_3'\mathbf{y} = \begin{pmatrix} 429 \\ 436 \end{pmatrix}.$$

The other two right hand side matrices can be easily obtained from $\mathbf{y}$ and $\mathbf{Z}_3'\mathbf{y}$. Thus, the order of the MME will be 24. The inverse of the relationship matrix is

$$\mathbf{A}^{-1} = \frac{1}{2} \begin{pmatrix} 5 & 0 & 2 & 1 & -2 & 0 & -2 & 0 & -2 & 0 \\ 0 & 5 & 1 & 2 & 0 & -2 & 0 & -2 & 0 & -2 \\ 2 & 1 & 5 & 0 & -2 & -2 & 0 & 0 & -2 & 0 \\ 1 & 2 & 0 & 5 & 0 & 0 & -2 & -2 & 0 & -2 \\ -2 & 0 & -2 & 0 & 4 & 0 & 0 & 0 & 0 & 0 \\ 0 & -2 & -2 & 0 & 0 & 4 & 0 & 0 & 0 & 0 \\ -2 & 0 & 0 & -2 & 0 & 0 & 4 & 0 & 0 & 0 \\ 0 & -2 & 0 & -2 & 0 & 0 & 0 & 4 & 0 & 0 \\ -2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 4 & 0 \\ 0 & -2 & 0 & -2 & 0 & 0 & 0 & 0 & 0 & 4 \end{pmatrix}.$$

The solutions to the MME are

$$\hat{\mathbf{b}} = \begin{pmatrix} 137.8469 \\ 150.4864 \end{pmatrix}, \quad \hat{\mathbf{p}} = \begin{pmatrix} .0658 \\ -.0658 \end{pmatrix},$$

$$\hat{\mathbf{a}} = \begin{pmatrix} 2.3295 \\ -2.3295 \\ .1280 \\ -.1280 \\ 5.1055 \\ -4.1143 \\ .2375 \\ 2.0161 \\ .5447 \\ -3.7896 \end{pmatrix}, \quad \text{and} \quad \hat{\mathbf{m}} = \begin{pmatrix} -.3328 \\ .3328 \\ .1646 \\ -.1646 \\ -.6379 \\ .6792 \\ -.1254 \\ -.3795 \\ .0136 \\ .4499 \end{pmatrix}.$$

No correlations with true values were calculated for this small example.

The influence of the correlation between direct and maternal effects can be a matter of concern. If the correlation between direct and maternal true breeding values is negative (-0.1), for example, and if an animal has a high direct EBV based on its own growth record, then the maternal EBV could be very negative due to the correlation alone. Thus, if few of the animals with growth records have progeny too, then the relationship between direct and maternal EBVs will be strongly negative (like -0.8)(reflecting the assumed negative correlation amongst true breeding values). However, if the data are complete and animals have both their own records and those of several progeny, then the correlation between direct and maternal EBVs should more closely follow the assumed genetic correlation. This situation can also affect the correct estimation of this genetic correlation. Estimates of this correlation in beef cattle has ranged from -0.5 to +0.5, and this mostly reflects the differences in quality (completeness) of data used.

In experimental station herds with several generations and fairly complete data, the estimates have tended to be zero or slightly positive between direct and maternal effects. On the other hand, in field data with almost no ties between growth of calves with performance of offspring as a dam, the estimates of the correlation have tended to be negative. To determine if your data are complete, create a file that has an animal's own record plus the average growth record of its progeny, to do a sort of dam-offspring phenotypic correlation. If you have 3 million records, but only 100 dam-offspring pairs, then the reliability of the estimated correlation between direct and maternal effects will be low.

122 Cytoplasmic Effects

Cytoplasmic effects are created by mitochondrial DNA that is passed through the ooctye to each offspring of a female. This DNA does not undergo meiosis, but is transmitted directly and wholly to each oocyte. The size of this maternal effect is not known, but several attempts were made to estimate the effect. Brian Kennedy was noted for debunking the incorrect models that were used to estimate cytoplasmic effects (J. Dairy Sci. 1986, 69:3100-3105). Kennedy suggested the use of an animal model (with additive genetic relationships), which included a female line of origin effect. That is, each animal would need to be traced back to a female in the base population with unknown parents. All such females would represent a different line of cytoplasmic effects. The variance of the line effects would need to be estimated. Most studies using this model showed very small levels of cytoplasmic effects in dairy cattle.

123 Embryo Transfer

Embryo transfer is used widely in dairy and beef cattle breeding. A genetically superior female is identified, and the owner desires to have many offspring from this individual. The cow is superovulated and the multiple embryos are harvested and inseminated, then at a certain stage of development, the embryos are implanted into recipient cows (of a different breed possibly). The recipients serve as a surrogate mother. When the calf is born, the maternal environment that it lives with is that of the recipient cow, and not that of its biological female parent. This can be handled in the maternal genetics effects model. The biological parent is included in the inverse of the relationship matrix, but the design matrix ($\mathbf{Z}_2$) indicates the surrogate dam of that calf. The maternal genetic effects in the calf are inherited from the biological parents, but the maternal environment that influences its own growth is from the surrogate dam.

124 Data Structure

Data structure is very important for estimating the covariance between direct and maternal genetic effects. Mahyar Heydarpour (2006) recently completed a study of this problem with multiple trait models, and showed that estimates could be biased by poor data structure. At least two features must be present to enable proper estimation of the covariance matrices.

1. Females must appear in the data as a calf with their own early growth records, and they must appear later as the dam of other calves where they express their maternal genetic influence. There needs to be a high percentage of such ties (through a common ID number) in the data structure in order for the covariance between direct and maternal genetic effects to be estimable.

2. Sires must have daughters that appear as dams of calves.

Often the identification of a calf (registration number) is unknown in dairy and therefore, the link between the female calf growth record with that animal's progeny in later years is lost. There is a good link for a dam with all of her progeny, but not with her own growth data when she was a calf. Without this link, the covariance between direct and maternal genetic effects is often very highly negative. This can be shown to be a mathematical consequence of the data structure, if a negative correlation is used as a prior value.

If the data structure is poor, then use of a zero covariance may be better than trying to estimate the covariance

125 EXERCISES

Analyze the following data sets using an appropriate maternal effects model. Also, perform one iteration or one round of Gibbs sampling or EM REML.

125.1 Data Set 1

Weaning weights (pounds) of Hereford, female, beef calves from three contemporary groups (CG) and at different ages at weaning.

Calf	Sire	Dam	CG	Age	Weight(lb)
5	17	2	1	205	500
6	18	1	1	216	580
7	18	3	1	190	533
8	17	4	1	196	535
9	18	1	2	210	507
10	17	2	2	221	555
11	19	3	2	175	461
12	19	4	2	184	467
13	18	5	3	212	548
14	17	7	3	214	605
15	20	3	3	202	480
16	20	4	3	236	576

Assume that

$$\mathbf{G} = \begin{pmatrix} 2122 & 338 \\ 338 & 1211 \end{pmatrix},$$

$\sigma_p^2 = 476$, and $\sigma_e^2 = 5962$.

What are the direct and maternal heritabilities, and direct-maternal correlation?

125.2 Data Set 2

This is an example of some animals being produced from embryo transfer. An animal from ET has a recipient female as the "dam" providing the maternal environment. Assume recipients are unrelated to any other animals genetically.

Length of rat pups after one week (cm).

Pup	Sire	Dam	Sex	ET	Length
5	1	3	No	M	7.3
6	1	3	Yes	F	6.5
7	2	4	No	F	7.5
8	2	4	No	M	8.7
9	2	4	Yes	M	9.4

Assume that

$$\mathbf{G} = \begin{pmatrix} 0.05 & -0.01 \\ -0.01 & 0.02 \end{pmatrix},$$

$\sigma_p^2 = 0.014$, and $\sigma_e^2 = 0.12$.

What are the direct and maternal heritabilities, and direct-maternal correlation?

Random Regression Models

126 Introduction

All biological creatures grow and perform over their lifetime. Traits that are measured at various times during that life are known as *longitudinal* data. Examples are body weights, body lengths, milk production, feed intake, fat deposition, and egg production. On a biological basis there could be different genes that turn on or turn off as an animal ages causing changes in physiology and performance. Also, an animal's age can be recorded in years, months, weeks, days, hours, minutes, or seconds, so that, in effect, there could be a continuum or continuous range of points in time when an animal could be observed for a trait. These traits have also been called *infinitely dimensional* traits.

Take body weight on gilts from a 60-day growth test as an example.

Animal	Days on Test					
	10	20	30	40	50	60
1	42	53	60	72	83	94
2	30	50	58	68	76	85
3	38	44	51	60	70	77
SD	1.6	3.7	3.9	5.0	5.3	5.6

The differences among the three animals increase with days on test as the gilts become heavier. As the mean weight increases, so also the standard deviation of weights increases. The weights over time could be modeled as a mean plus covariates of days on test and days on test squared. Depending on the species and trait, perhaps a cubic or spline function would fit the data better. The point is that the means can be fit by a linear model with a certain number of parameters.

127 Multiple Trait Approach

The data presented in the previous table have typically been analyzed such that the weights at each day on test are different traits. If t is the day on test, i.e. 10, 20, 30, 40, 50, or 60, then a model for any one of the weights could be

$$\mathbf{y}_t = \mathbf{X}\mathbf{b}_t + \mathbf{a}_t + \mathbf{e}_t,$$

which is just a simple, single record, animal model. Analyses are usually done so that the genetic and residual variances and covariances are estimated among the six weights.

Suppose that an estimate of the genetic variances and covariances was

$$\mathbf{G} = \begin{pmatrix} 2.5 & 4.9 & 4.6 & 4.6 & 4.3 & 4.0 \\ 4.9 & 13.5 & 12.1 & 12.3 & 11.9 & 10.7 \\ 4.6 & 12.1 & 15.2 & 14.5 & 14.6 & 12.5 \\ 4.6 & 12.3 & 14.5 & 20.0 & 19.0 & 16.9 \\ 4.3 & 11.9 & 14.6 & 19.0 & 25.0 & 20.3 \\ 4.0 & 10.7 & 12.5 & 16.9 & 20.3 & 30.0 \end{pmatrix}.$$

Let the residual covariance matrix be

$$\mathbf{R} = \begin{pmatrix} 3.8 & 7.4 & 6.9 & 6.8 & 6.4 & 6.0 \\ 7.4 & 20.3 & 18.2 & 18.4 & 17.9 & 16.1 \\ 6.9 & 18.2 & 22.8 & 21.8 & 21.9 & 18.8 \\ 6.8 & 18.4 & 21.8 & 30.0 & 28.5 & 25.4 \\ 6.4 & 17.9 & 21.9 & 28.5 & 37.5 & 30.5 \\ 6.0 & 16.1 & 18.8 & 25.4 & 30.5 & 45.0 \end{pmatrix}.$$

Assuming a model with only an intercept, and that the three animals are unrelated, then

$$(\mathbf{X} \;\; \mathbf{Z}) = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix} \otimes \mathbf{I}_6,$$

where the identity is of order 6 and $\otimes$ is the direct product operator. The observations would be ordered by days on test within animals, i.e.,

$$\mathbf{y}' = \begin{pmatrix} 42 & 53 & 60 & 72 & 83 & 94 & \cdots & 60 & 70 & 77 \end{pmatrix}.$$

The resulting MME would be of order 24 by 24, and the solutions would be as follows.

Days on Test	Mean	Animal 1	Animal 2	Animal 3
10	36.67	2.10	-2.61	0.51
20	49.00	1.57	0.45	-2.02
30	56.33	1.48	0.64	-2.12
40	66.67	2.21	0.39	-2.60
50	76.33	2.72	-0.24	-2.48
60	85.33	3.48	-0.16	-3.32

Animal 1 clearly grew faster than the other two animals and its superiority grew larger with time. Animals 2 and 3 switched rankings after the first 10 days, and Animal 3 was the slower growing animal. The estimates for the mean give an average growth curve for the 3 animals.

A multiple trait approach may be appropriate here because every animal was weighed on exactly the same number of days on test throughout the trial. However, suppose the animals were of different ages at the start of test, and suppose that instead of days on test, the ages for each weight were given. Assume at start of test that Animal 1 was 18 days old, Animal 2 was 22, and Animal 3 was 25. The multiple trait model could include a factor (classification or covariable) to account for different starting ages. The differences observed at any point in time could be due to the ages of the animals rather than just on the number of days on test. The analysis shown above would have an implied assumption that all animals began the test at the same age.

128 Covariance Functions

Let the example data be as shown below, allowing for the different ages at each test. Note that the ages range from 28 days to 85 days, and that none of the animals were ever weighed at exactly the same age.

Animal 1		Animal 2		Animal 3	
Age	Wt	Age	Wt	Age	Wt
28	42	32	30	35	38
38	53	42	50	45	44
48	60	52	58	55	51
58	72	62	68	65	60
68	83	72	76	75	70
78	94	82	85	85	77

Kirkpatrick et al.(1991) proposed the use of covariance functions for longitudinal data of this kind. A covariance function (CF) is a way to model the variances and covariances of a longitudinal trait. Orthogonal polynomials are used in this model and the user must decide the order of fit that is best. Legendre polynomials (1797) are the easiest to apply.

To calculate Legendre polynomials, first define

$$\begin{aligned} P_0(x) &= 1, \text{ and} \\ P_1(x) &= x, \end{aligned}$$

then, in general, the $n+1$ polynomial is described by the following recursive equation:

$$P_{n+1}(x) = \frac{1}{n+1}\left((2n+1)xP_n(x) - nP_{n-1}(x)\right).$$

These quantities are "normalized" using

$$\phi_n(x) = \left(\frac{2n+1}{2}\right)^{.5} P_n(x).$$

This gives the following series,

$$\begin{aligned}
\phi_0(x) &= \left(\frac{1}{2}\right)^{.5} P_0(x) = .7071 \\
\phi_1(x) &= \left(\frac{3}{2}\right)^{.5} P_1(x) \\
&= 1.2247x \\
P_2(x) &= \frac{1}{2}(3xP_1(x) - 1P_0(x)) \\
\phi_2(x) &= \left(\frac{5}{2}\right)^{.5} (\frac{3}{2}x^2 - \frac{1}{2}) \\
&= -.7906 + 2.3717x^2,
\end{aligned}$$

and so on. The first six can be put into a matrix, Λ, as

$$\Lambda' = \begin{pmatrix} .7071 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1.2247 & 0 & 0 & 0 & 0 \\ -.7906 & 0 & 2.3717 & 0 & 0 & 0 \\ 0 & -2.8062 & 0 & 4.6771 & 0 & 0 \\ .7955 & 0 & -7.9550 & 0 & 9.2808 & 0 \\ 0 & 4.3973 & 0 & -20.5206 & 0 & 18.4685 \end{pmatrix}.$$

Now define another matrix, $\mathbf{M}$, as a matrix containing the polynomials of standardized time values. Legendre polynomials are defined within the range of values from -1 to +1. Thus, ages or time periods have to be standardized (converted) to the interval between -1 to +1. The formula is

$$q_\ell = -1 + 2\left(\frac{t_\ell - t_{min}}{t_{max} - t_{min}}\right).$$

Let the minimum starting age for pigs on test be 15 days and the maximum starting age be 28 days, then the maximum age at end of test was 88 days. Thus, $t_{min} = 25 = (15+10)$ and $t_{max} = 88 = (28+60)$.

The matrix $\mathbf{G}$ was based on weights taken on pigs that were all 21 days of age at start of test. The table below shows the ages and standardized time values for the six weigh dates.

Days on Test	Age	Standardized Value
10	31	-1.000
20	41	-.600
30	51	-.200
40	61	.200
50	71	.600
60	81	1.000

Therefore,

$$\mathbf{M} = \begin{pmatrix} 1 & -1 & 1 & -1 & 1 & -1 \\ 1 & -.600 & .360 & -.216 & .130 & -.078 \\ 1 & -.200 & .040 & -.008 & .002 & -.000 \\ 1 & .200 & .040 & .008 & .002 & .000 \\ 1 & .600 & .360 & .216 & .130 & .078 \\ 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}.$$

This gives

$$\begin{aligned} \Phi &= \mathbf{M}\Lambda, \\ &= \begin{pmatrix} .7071 & -1.2247 & 1.5811 & -1.8708 & 2.1213 & -2.3452 \\ .7071 & -.7348 & .0632 & .6735 & -.8655 & .3580 \\ .7071 & -.2449 & -.6957 & .5238 & .4921 & -.7212 \\ .7071 & .2449 & -.6957 & -.5238 & .4921 & .7212 \\ .7071 & .7348 & .0632 & -.6735 & -.8655 & -.3580 \\ .7071 & 1.2247 & 1.5811 & 1.8708 & 2.1213 & 2.3452 \end{pmatrix}. \end{aligned}$$

which can be used to specify the elements of **G** as

$$\begin{aligned} \mathbf{G} &= \Phi\mathbf{H}\Phi' \\ &= \mathbf{M}(\Lambda\mathbf{H}\Lambda')\mathbf{M}' \\ &= \mathbf{MTM}'. \end{aligned}$$

Note that Φ, **M**, and Λ are matrices defined by the Legendre polynomial functions and by the standardized time values and do not depend on the data or values in the matrix **G**. Therefore, it is possible to estimate either **H** or **T**,

$$\begin{aligned} \mathbf{H} &= \Phi^{-1}\mathbf{G}\Phi^{-T}, \\ &= \begin{pmatrix} 27.69 & 5.29 & -1.95 & 0.05 & -1.17 & 0.52 \\ 5.29 & 4.99 & 0.42 & -0.25 & -0.30 & -0.75 \\ -1.95 & 0.42 & 1.51 & 0.20 & -0.33 & -0.07 \\ 0.05 & -0.25 & 0.20 & 1.19 & 0.06 & -0.71 \\ -1.17 & -0.30 & -0.33 & 0.06 & 0.58 & 0.15 \\ 0.52 & -0.75 & -0.07 & -0.71 & 0.15 & 1.12 \end{pmatrix}, \end{aligned}$$

and

$$\begin{aligned} \mathbf{T} &= \mathbf{M}^{-1}\mathbf{G}\mathbf{M}^{-T} \\ &= \begin{pmatrix} 16.44 & 6.48 & -5.93 & -11.49 & -0.93 & 10.02 \\ 6.48 & 49.87 & -2.05 & -155.34 & 1.44 & 111.23 \\ -5.93 & -2.05 & 57.71 & 28.62 & -50.06 & -25.73 \\ -11.49 & -155.34 & 28.62 & 635.49 & -26.91 & -486.90 \\ -0.93 & 1.44 & -50.06 & -26.91 & 49.80 & 26.49 \\ 10.02 & 111.23 & -25.73 & -486.90 & 26.49 & 382.79 \end{pmatrix}. \end{aligned}$$

Why orthogonal polynomials? Convert **T** and **H** to correlation matrices.

$$\mathbf{T}_{cor} = \left(\begin{array}{rrrrrr} 1.00 & .23 & -.19 & -.11 & -.03 & .13 \\ .23 & 1.00 & -.04 & -.87 & .03 & .81 \\ -.19 & -.04 & 1.00 & .15 & -.93 & -.17 \\ -.11 & -.87 & .15 & 1.00 & -.15 & -.99 \\ -.03 & .03 & -.93 & -.15 & 1.00 & .19 \\ .13 & .81 & -.17 & -.99 & .19 & 1.00 \end{array} \right),$$

and

$$\mathbf{H}_{cor} = \left(\begin{array}{rrrrrr} 1.00 & .45 & -.30 & .01 & -.29 & .09 \\ .45 & 1.00 & .15 & -.10 & -.17 & -.32 \\ -.30 & .15 & 1.00 & .15 & -.36 & -.05 \\ .01 & -.10 & .15 & 1.00 & .07 & -.62 \\ -.29 & -.17 & -.36 & .07 & 1.00 & .19 \\ .09 & -.32 & -.05 & -.62 & .19 & 1.00 \end{array} \right).$$

The largest absolute correlation in **T** was .99, while the largest absolute correlation in **H** was only .62. Orthogonal polynomials tend to reduce the correlations between estimated regression coefficients. This is advantageous when trying to estimate **H** by REML or Bayesian methods, because the estimates would converge faster to the maximum or appropriate posterior distribution than trying to estimate **T**. The matrix **T** actually had four correlations greater than 0.80 in absolute value, while **H** had none. There are other kinds of orthogonal polynomials, but Legendre polynomials are probably the easiest to calculate and utilize.

H can be used to calculate the covariance between any two days on test between 10 and 60 days. To compute the covariance between days 25 and 55, calculate the Legendre polynomial covariates as in calculating a row of Φ. The standardized time values for days 25 and 55 are -0.4 and 0.8, respectively. The Legendre polynomials (stored in **L** are

$$\mathbf{L} = \left(\begin{array}{rrrrrr} .7071 & -.4899 & -.4111 & .8232 & -.2397 & -.6347 \\ .7071 & .9798 & .7273 & .1497 & -.4943 & -.9370 \end{array} \right).$$

Then the variances and covariance for those two ages are

$$\mathbf{LHL}' = \left(\begin{array}{rr} 14.4226 & 13.7370 \\ 13.7370 & 28.9395 \end{array} \right).$$

Thus, the genetic correlation between days 25 and 55 is 0.67. The same calculations could be repeated for the residual variance-covariance matrix. Let

$$\begin{aligned} \mathbf{S} &= \Phi^{-1}\mathbf{R}\Phi^{-T}, \\ &= \left(\begin{array}{rrrrrr} 41.57 & 7.94 & -2.91 & 0.11 & -1.76 & 0.76 \\ 7.94 & 7.45 & 0.62 & -0.41 & -0.44 & -1.07 \\ -2.91 & 0.62 & 2.29 & 0.31 & -0.52 & -0.12 \\ 0.11 & -0.41 & 0.31 & 1.76 & 0.08 & -1.04 \\ -1.76 & -0.44 & -0.52 & 0.08 & 0.88 & 0.24 \\ 0.76 & -1.07 & -0.12 & -1.04 & 0.24 & 1.64 \end{array} \right), \end{aligned}$$

then the residual variances and covariances for days 25 and 55 would be

$$\mathbf{LSL}' = \begin{pmatrix} 21.6645 & 20.6166 \\ 20.6166 & 43.3442 \end{pmatrix}.$$

128.1 Reduced Orders of Fit

Although the order of **G** in the previous example was six and polynomials of standardized ages to the fifth power were used to derive the covariance functions, perhaps only squared or cubed powers are needed to adequately describe the elements of **G**. That is, find Φ^* such that it is rectangular and $\mathbf{H}^*$ has a smaller order, $m < k$, but still

$$\mathbf{G} = \Phi^* \mathbf{H}^* \Phi^{*'}.$$

To determine $\mathbf{H}^*$, first pre-multiply **G** by $\Phi^{*'}$ and post-multiply that by Φ^* as

$$\begin{aligned} \Phi^{*'} \mathbf{G} \Phi^* &= \Phi^{*'} (\Phi^* \mathbf{H}^* \Phi^{*'}) \Phi^* \\ &= (\Phi^{*'} \Phi^*) \mathbf{H}^* (\Phi^{*'} \Phi^*). \end{aligned}$$

Now pre- and post- multiply by the inverse of $(\Phi^{*'}\Phi^*) = \mathbf{P}$ to determine $\mathbf{H}^*$,

$$\mathbf{H}^* = \mathbf{P}^{-1} \Phi^{*'} \mathbf{G} \Phi^* \mathbf{P}^{-1}.$$

To illustrate, let $m = 3$, then

$$\Phi^* = \begin{pmatrix} .7071 & -1.2247 & 1.5811 \\ .7071 & -.7348 & .0632 \\ .7071 & -.2449 & -.6957 \\ .7071 & .2449 & -.6957 \\ .7071 & .7348 & .0632 \\ .7071 & 1.2247 & 1.5811 \end{pmatrix},$$

and

$$\begin{aligned} \Phi^{*'} \Phi^* &= \begin{pmatrix} 3.0000 & 0.0000 & 1.3415 \\ 0.0000 & 4.1997 & 0.0000 \\ 1.3415 & 0.0000 & 5.9758 \end{pmatrix}, \\ (\Phi^{*'} \Phi^*)^{-1} &= \begin{pmatrix} .3705 & .0000 & -.0832 \\ .0000 & .2381 & .0000 \\ -.0832 & .0000 & .1860 \end{pmatrix}. \end{aligned}$$

Also,

$$\Phi^{*'} \mathbf{G} \Phi^* = \begin{pmatrix} 220.2958 & 78.0080 & 61.4449 \\ 78.0080 & 67.5670 & 44.9707 \\ 61.4449 & 44.9707 & 50.5819 \end{pmatrix}.$$

The matrix $\mathbf{H}^*$ is then

$$\left(\begin{array}{rrr} 26.8082 & 5.9919 & -2.9122 \\ 5.9919 & 3.8309 & .4468 \\ -2.9122 & .4468 & 1.3730 \end{array}\right).$$

What order of reduced fit is sufficient to explain the variances and covariances in $\mathbf{G}$? Kirkpatrick et al.(1990) suggested looking at the eigenvalues of the matrix $\mathbf{H}$ from a full rank fit. Below are the values. The sum of all the eigenvalues was , and also shown is the percentage of that total.

H	
Eigenvalue	Percentage
29.0357	.7831
4.2922	.1158
1.8161	.0490
1.3558	.0366
.5445	.0147
.0355	.0010

The majority of change in elements in $\mathbf{G}$ is explained by a constant, and by a linear increment. Both suggest that a quadratic function of the polynomials is probably sufficient. Is there a way to statistically test the reduced orders of fit to determine which is sufficient? A goodness of fit statistic is $\hat{\mathbf{e}}'\hat{\mathbf{e}}$ where

$$\hat{\mathbf{e}} = \mathbf{g} - \hat{\mathbf{g}}$$

and $\mathbf{g}$ is a vector of the half-stored elements of the matrix $\mathbf{G}$, i.e.,

$$\mathbf{g}' = \left(\begin{array}{ccccccc} g_{11} & g_{12} & \cdots & g_{16} & g_{22} & \cdots & g_{66} \end{array}\right).$$

A half-stored matrix of order k has $k(k+1)/2$ elements. For $k = 6$ there are 21 values. Likewise, $\hat{\mathbf{g}}$ is a vector of half stored elements of the matrix $\Phi^*\mathbf{H}^*\Phi^{*'}$. Although this matrix also has 21 values, because $\mathbf{M}$ has only $m < k$ columns, the number of independent values is $m(m+1)/2$. For $m = 3$ this number is 6.

The test statistic, $\hat{\mathbf{e}}'\hat{\mathbf{e}}$, has a Chi-square distribution with $k(k+1)/2 - m(m+1)/2$ degrees of freedom. In the example with $m = 3$,

$$\Phi^*\mathbf{H}^*\Phi^{*'} = \left(\begin{array}{rrrrrr} 3.9622 & 4.7467 & 5.2006 & 5.3239 & 5.1165 & 4.5786 \\ 4.7467 & 8.9493 & 11.4058 & 12.1162 & 11.0804 & 8.2986 \\ 5.2006 & 11.4058 & 15.2402 & 16.7038 & 15.7966 & 12.5186 \\ 5.3239 & 12.1162 & 16.7038 & 19.0868 & 19.2650 & 17.2386 \\ 5.1165 & 11.0804 & 15.7966 & 19.2650 & 21.4857 & 22.4586 \\ 4.5786 & 8.2986 & 12.5186 & 17.2386 & 22.4586 & 28.1786 \end{array}\right),$$

and the residuals (differences from the original $\mathbf{G}$) are

$$\begin{pmatrix} -1.4622 & .1533 & -.6006 & -.7239 & -.8165 & -.5786 \\ .1533 & 4.5507 & .6942 & .1838 & .8196 & 2.4014 \\ -.6006 & .6942 & -.0402 & -2.2038 & -1.1966 & -.0186 \\ -.7239 & .1838 & -2.2038 & .9132 & -.2650 & -.3386 \\ -.8165 & .8196 & -1.1966 & -.2650 & 3.5143 & -2.1586 \\ -.5786 & 2.4014 & -.0186 & -.3386 & -2.1586 & 1.8214 \end{pmatrix},$$

so that the goodness of fit statistic is

$$\hat{\mathbf{e}}'\hat{\mathbf{e}} = 59.3476,$$

with 21-6=15 degrees of freedom.

Is a fit of order 3 poorer than a fit of order 5? An F-statistic is possible by taking the difference in the goodness of fit statistics, divided by an estimate of the residual variance. The residual variance is estimated from a fit of order $k-1$ or in this case of order 5. The goodness of fit statistic for order 5 was 7.2139 with 21-15=6 degrees of freedom. Hence the residual variance is

$$\sigma^2 = 7.2139/6 \ = \ 1.2023.$$

The F-statistic to test if a fit of order 3 is different from a fit of order 5 is

$$\begin{aligned} F &= \frac{(\hat{\mathbf{e}}'\hat{\mathbf{e}}_{m=3} - \hat{\mathbf{e}}'\hat{\mathbf{e}}_{m=5})/(15-6)}{\sigma^2} \\ &= \frac{(59.3476 - 7.2139)/9}{1.2023} \\ &= 5.7926/1.2023 \ = \ 4.8180, \end{aligned}$$

with (9,6) degrees of freedom. The table F-value at the ($P = .05$) level is 4.10. Thus, the difference is significant, and a fit of order 5 is better than a fit of order 3.

129 Basic Structure of RRM

Random regression models have a basic structure that is similar in most applications. A simplified RRM for a single trait can be written as

$$y_{ijkn:t} = F_i + g(t)_j + r(a, x, m1)_k + r(pe, x, m2)_k + e_{ijkn:t},$$

where

$y_{ijkn:t}$ is the n^{th} observation on the k^{th} animal at time t belonging to the i^{th} fixed factor and the j^{th} group;

F_i is a fixed effect that is independent of the time scale for the observations, such as a cage effect, a location effect or a herd-test date effect;

$g(t)_j$ is a function or functions that account for the phenotypic trajectory of the average observations across all animals belonging to the j^{th} group;

$r(a, x, m1)_k = \sum_{\ell=0}^{m_1} a_{k\ell} x_{ijk:\ell}$ is the notation adopted for a random regression function. In this case, a denotes the additive genetic effects of the k^{th} animal, x is the vector of time covariates, and $m1$ is the order of the regression function. So that $x_{ijk:\ell}$ are the covariables related to time t, and $a_{k\ell}$ are the animal additive genetic regression coefficients to be estimated;

$r(pe, x, m2)_k = \sum_{\ell=0}^{m_2} p_{k\ell} x_{ijk:\ell}$ is a similar random regression function for the permanent environmental (pe) effects of the k^{th} animal; and

$e_{ijkn:t}$ is a random residual effect with mean null and with possibly different variances for each t or functions of t.

The function, $g(t)_j$, can be either linear or nonlinear in t. Such a function is necessary in a RRM to account for the phenotypic relationship between y and the time covariables (or other types of covariables that could be used in a RRM). In a test day model, $g(t)_j$ accounts for different lactation curve shapes for groups of animals defined by years of birth, parity number, and age and season of calving within parities, for example. With growth data, $g(t)_j$ accounts for the growth curve of males or females of breed X or breed Y from young or old dams.

If the shape of the phenotypic relationship is not known or is nonlinear, then $g(t)_j$ could be a set of classification variables. Classification variables take up more degrees of freedom and require a large number of observations per level, but they do not force the user to explicitly define the shape of the trajectory. A mathematical function, on the other hand, does not use many degrees of freedom and gives a smooth trajectory over time regardless of the number of observations. The choice of classification variables or mathematical function is up to the researcher. If data are very numerous, and the mathematical function fits the data well, then either approach will generally lead to the same results. The phenotypic relationships, $g(t)_j$, are important to a RRM analysis and deserve care and effort in their correct specification.

The random regressions are intended to model the deviations around the phenotypic trajectories. The pattern of variation may be very different in shape or appearance from the phenotypic relationships, and may be more simple than $g(t)_j$. Orthogonal polynomials of standardized units of time have been recommended as covariables (Kirkpatrick et al., 1990). Orthogonal polynomials have computational advantages. The primary general advantage is the reduced correlations among the estimated coefficients. A standardized unit of time, w, ranges from -1 to +1, and is derived as

$$w = \frac{2 * (t - t_{min})}{(t_{max} - t_{min})} - 1,$$

where t_{min} is the earliest date (or the youngest age) and t_{max} is the latest date (or oldest age) represented in the data. The order of the orthogonal polynomials would be m_1 and m_2, i.e. the highest power of polynomial. Note that m_1 and m_2 do not need to be equal, but often (for simplicity of computing) they are chosen to be the same. Meyer(2000) and Pool et al. (2000), for example, compared many RRM models with different orders of orthogonal polynomials for the genetic and *pe* effects. Several types of orthogonal polynomials are available, but Legendre polynomials have been utilized (Kirkpatrick et al., 1990). The first 6 Legendre polynomial functions of standardized units of time are given in Table 1. Thus, if $w = -0.2$, then the covariables that would go into the model (for order equal to 5) are shown in the last column of Table 1. Covariables based upon orthogonal polynomials are small numbers that reduce problems with rounding errors, and they provide relatively small correlations between the estimated regression coefficients.

The residual variance should not be assumed to be constant from t_{min} to t_{max}. The residual effect is also known as a temporary environmental effect. Changes in residual variance might be predictable depending on the trajectory of the phenotypic data. For example, if RRM were being applied to growth data, weights may increase linearly with age, and the variance of weights may increase quadratically with age. Thus, the residual variance would be expected to increase in a similar manner as the phenotypic variance. Residual variances can be fit with a function of t, or assumed to have an autoregressive structure, or can be grouped into intervals having equal variance within the intervals. Research in this area is needed.

In matrix notation the RRM is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{a} + \mathbf{Z}_2\mathbf{p} + \mathbf{e},$$

where $\mathbf{b}$ contains F_i and $g(t)_j$ effects, $\mathbf{a}$ contains $m_1 + 1$ additive genetic regression coefficients for each animal, $\mathbf{p}$ contains $m_2 + 1$ permanent environmental regression coefficients for each animal with data, and $\mathbf{e}$ contains the temporary environmental effects. Also,

$$Var\begin{pmatrix}\mathbf{a}\\ \mathbf{p}\\ \mathbf{e}\end{pmatrix}=\begin{pmatrix}\mathbf{A}\otimes\mathbf{G} & \mathbf{0} & \mathbf{0}\\ \mathbf{0} & \mathbf{I}\otimes\mathbf{P} & \mathbf{0}\\ \mathbf{0} & \mathbf{0} & \mathbf{R}\end{pmatrix},$$

where $\mathbf{G}$ is the variance-covariance matrix of the additive genetic random regression coefficients of order m_1+1; $\mathbf{P}$ is the variance-covariance matrix of the permanent environmental random regression coefficients of order m_2+1; and $\mathbf{R}$ is a diagonal matrix of temporary environmental variances which could vary depending on t, or $\mathbf{R}$ could be block diagonal with an autocorrelation structure for each animal's records. The mixed model equations (MME) are represented as

$$\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_1+\mathbf{A}^{-1}\otimes\mathbf{G}^{-1} & \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{Z}_2\\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_1 & \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{Z}_2+\mathbf{I}\otimes\mathbf{P}^{-1}\end{pmatrix}\begin{pmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{a}}\\ \hat{\mathbf{p}}\end{pmatrix}=\begin{pmatrix}\mathbf{X}'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_1'\mathbf{R}^{-1}\mathbf{y}\\ \mathbf{Z}_2'\mathbf{R}^{-1}\mathbf{y}\end{pmatrix}.$$

Assumptions about the distributions of $\mathbf{y}$ and other random variables are not necessary to derive best linear unbiased predictors (BLUP)(Goldberger, 1962; Henderson, 1984) or the MME, but when $\mathbf{y}$ is normally distributed then BLUP is also BLP if the model is correct and variances and covariances are known. In order to estimate the elements of $\mathbf{G}$, $\mathbf{P}$, and $\mathbf{R}$ via Bayesian methods or restricted maximum likelihood, then normality of the random variables must be assumed (See for example Jamrozik and Schaeffer, 1997). This paper will concentrate on the applications of RRM and not on the estimation of (co)variance parameters, nor on the computational details of estimating (co)variances or solving mixed model equations. Some of the applications in this paper have applied RRM to discrete data, and therefore, a BLUP analysis would not be optimum. However, the presentation of these ideas may stimulate others to find better solutions.

130 Example Data Analysis By RRM

Below are the data structure and pedigrees of four dairy cows. Given is the age at which they were observed for a trait during four visits to one herd.

				Age, Obs. at Visit		
Cow	Sire	Dam	Visit 1	Visit 2	Visit 3	Visit 4
1	7	5	22,224	34,236	47,239	
2	7	6	30,244	42,247	55,241	66,244
3	8	5	28,224	40,242		
4	8	1		20,220	33,234	44,228

The model equation might be

$$\begin{aligned} y_{jik:t} &= V_j + b_0 + b_1(A) + b_2(A)^2 \\ &\quad +(a_{i0}z_0 + a_{i1}z_1 + a_{i2}z_2) \\ &\quad +(p_{i0}z_0 + p_{i1}z_1 + p_{i2}z_2) + e_{jik:t} \end{aligned}$$

where

V_j is a random contemporary group effect which is assumed to follow a normal distribution with mean 0 and variance, $\sigma_c^2 = 4$.

b_0, b_1, and b_2 are fixed regression coefficients on (A) = age and age squared which describes the general relationship between age and the observations,

a_{i0}, a_{i1}, and a_{i2} are random regression coefficients for animal i additive genetic effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix, $\mathbf{G}$,

p_{i0}, p_{i1}, and p_{i2} are random regression coefficients for animal i permanent environmental effects, assumed to follow a multivariate normal distribution with mean vector null and variance-covariance matrix, $\mathbf{P}$,

z_0, z_1, and z_2 are the Legendre polynomials based on standardized ages and derived as indicated earlier. The minimum age was set at 18 and the maximum age was set at 68 for calculating the Legendre polynomials.

and e_{jik} is a temporary residual error term assumed to follow a normal distribution with mean 0 and variance, $\sigma_e^2 = 9$. In this example, the residual variance is assumed to be constant across ages.

The model in matrix notation is

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wv} + \mathbf{Za} + \mathbf{Zp} + \mathbf{e},$$

where

$$\mathbf{X} = \begin{pmatrix} 1 & 22 & 484 \\ 1 & 30 & 900 \\ 1 & 28 & 784 \\ 1 & 34 & 1156 \\ 1 & 42 & 1764 \\ 1 & 40 & 1600 \\ 1 & 20 & 400 \\ 1 & 47 & 2209 \\ 1 & 55 & 3025 \\ 1 & 33 & 1089 \\ 1 & 66 & 4356 \\ 1 & 44 & 1936 \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} 224 \\ 244 \\ 224 \\ 236 \\ 247 \\ 242 \\ 220 \\ 239 \\ 241 \\ 234 \\ 244 \\ 228 \end{pmatrix}, \quad \mathbf{W} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and

$$\mathbf{Z} = \left(\begin{array}{rrrrrrrrrrrr} .7071 & -1.0288 & .8829 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & -.6369 & -.1493 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.7348 & .0632 & 0 & 0 & 0 \\ .7071 & -.4409 & -.4832 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & -.0490 & -.7868 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.1470 & -.7564 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -1.1268 & 1.2168 \\ .7071 & .1960 & -.7299 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & .7071 & .5879 & -.2441 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & -.4899 & -.4111 \\ 0 & 0 & 0 & .7071 & 1.1268 & 1.2168 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & .7071 & .0490 & -.7868 \end{array} \right).$$

In order to reduce rounding errors the covariates of age for the fixed regressions can be forced to have a mean of approximately zero by subtracting 38 from all ages and 1642 from all ages squared. Then

$$\mathbf{X} = \left(\begin{array}{rrr} 1 & -16 & -1158 \\ 1 & -8 & -742 \\ 1 & -10 & -858 \\ 1 & -4 & -486 \\ 1 & 4 & 122 \\ 1 & 2 & -42 \\ 1 & -18 & -1242 \\ 1 & 9 & 567 \\ 1 & 17 & 1383 \\ 1 & -5 & -553 \\ 1 & 28 & 2714 \\ 1 & 6 & 294 \end{array} \right).$$

The mixed model equations that need to be constructed to provide estimated breeding values are as follows;

$$\left(\begin{array}{lllll} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{W} & \mathbf{X}'\mathbf{Z} & \mathbf{0} & \mathbf{X}'\mathbf{Z} \\ \mathbf{W}'\mathbf{X} & \mathbf{W}'\mathbf{W} + \mathbf{I}\frac{9}{4} & \mathbf{W}'\mathbf{Z} & \mathbf{0} & \mathbf{W}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{nn} \otimes \mathbf{G}^{-1} & \mathbf{A}^{nb} \otimes \mathbf{G}^{-1} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}^{bn} \otimes \mathbf{G}^{-1} & \mathbf{A}^{bb} \otimes \mathbf{G}^{-1} & \mathbf{0} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{W} & \mathbf{Z}'\mathbf{Z} & \mathbf{0} & \mathbf{Z}'\mathbf{Z} + \mathbf{I} \otimes \mathbf{P}^{-1} \end{array} \right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{c}} \\ \hat{\mathbf{a}}_n \\ \hat{\mathbf{a}}_b \\ \hat{\mathbf{p}} \end{array} \right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{W}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{0} \\ \mathbf{Z}'\mathbf{y} \end{array} \right).$$

The entire MME can not be presented, but parts of the MME are given below.

$$\mathbf{W}'\mathbf{W} = \left(\begin{array}{cccc} 3 & 0 & 0 & 0 \\ 0 & 4 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 2 \end{array} \right),$$

$$\mathbf{W}'\mathbf{X} = \left(\begin{array}{rrr} 3 & -34 & -2758 \\ 4 & -16 & -1648 \\ 3 & 21 & 1397 \\ 2 & 34 & 3008 \end{array} \right),$$

$$\mathbf{X}'\mathbf{X} = \begin{pmatrix} 12 & 5 & -1 \\ 5 & 1995 & 166,883 \\ -1 & 166,883 & 14,415,319 \end{pmatrix},$$

$\mathbf{Z}'\mathbf{Z}$ is composed of the following four blocks of order 3, for the four animals with records;

$$\text{Animal 1} \quad \begin{pmatrix} 1.5 & -.9006 & -.2335 \\ -.9006 & 1.2912 & -.8383 \\ -.2335 & -.8383 & 1.5457 \end{pmatrix},$$

$$\text{Animal 2} \quad \begin{pmatrix} 2 & .7275 & .0259 \\ .7275 & 2.0233 & 1.3612 \\ .0259 & 1.3612 & 2.1815 \end{pmatrix},$$

$$\text{Animal 3} \quad \begin{pmatrix} 1 & -.6235 & -.4902 \\ -.6235 & .5615 & .0648 \\ -.4902 & .0648 & .5761 \end{pmatrix},$$

$$\text{Animal 4} \quad \begin{pmatrix} 1.5 & -1.1085 & .0134 \\ -1.1085 & 1.5121 & -1.2082 \\ .0134 & -1.2082 & 2.2687 \end{pmatrix}.$$

and $\mathbf{Z}'\mathbf{X}$ is

$$\mathbf{Z}'\mathbf{X} = \begin{pmatrix} 2.1213 & -7.7781 & -761.5467 \\ -1.2737 & 19.9884 & 1516.7598 \\ -.3302 & -18.7627 & -1201.416 \\ 2.8284 & 28.9911 & 2458.5867 \\ 1.0288 & 46.4439 & 4337.8027 \\ .0366 & 27.9679 & 2979.5959 \\ 1.4142 & -5.6568 & -636.3900 \\ -.8818 & 7.0540 & 636.6324 \\ -.6932 & -2.1448 & -22.4568 \\ 2.1213 & -12.0207 & -1061.3570 \\ -1.5677 & 23.0259 & 1684.8063 \\ .0189 & -24.5677 & -1515.2470 \end{pmatrix}.$$

The right hand sides of the MME are

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 2823 \\ 2070 \\ 68,064 \end{pmatrix},$$

$$\mathbf{W}'\mathbf{y} = \begin{pmatrix} 692 \\ 945 \\ 714 \\ 472 \end{pmatrix},$$

and

$$\mathbf{Z}'\mathbf{y} = \begin{pmatrix} 494.2629 \\ -287.6596 \\ -90.7117 \\ 690.1296 \\ 249.1165 \\ 7.3023 \\ 329.5086 \\ -200.1692 \\ -168.8920 \\ 482.2422 \\ -351.3606 \\ -7.8918 \end{pmatrix}.$$

The variance-covariance matrices of the additive and permanent environmental effects need to be known for BLUP. Normally, these are not well known and must be estimated simultaneously with the other effects of the model. Let

$$\mathbf{G} = \begin{pmatrix} 94.0000 & -3.8500 & .03098 \\ -3.8500 & 1.5000 & -.0144 \\ .03098 & -.0144 & .0014 \end{pmatrix},$$

and

$$\mathbf{P} = \begin{pmatrix} 63.0000 & -2.1263 & .0447 \\ -2.1263 & .5058 & -.00486 \\ .0447 & -.00486 & .0005 \end{pmatrix}.$$

The solutions to MME are

$$\hat{\mathbf{b}}' = \begin{pmatrix} 234.9797 & 1.4670 & -.01399 \end{pmatrix},$$

$$\hat{\mathbf{c}}' = \begin{pmatrix} -.8630 \\ 1.2885 \\ .1443 \\ -.5698 \end{pmatrix}.$$

Let the solutions for the animal additive genetic random regression coefficients be presented in tabular form as follows.

Animal	a_0	a_1	a_2
1	-2.021529	.175532	-.002696
2	5.751601	-2.139115	.025848
3	-2.474456	2.554412	-.029269
4	-5.376687	-.370873	.002174
5	-1.886714	1.464975	-.016963
6	3.333268	-1.065525	.013047
7	1.503398	-1.081654	.012555
8	-2.948511	.681643	-.008633

Similarly, the solutions for the animal permanent environmental random regression coefficients can be given in tabular form.

Animal	p_0	p_1	p_2
1	-.296786	.246946	-.002521
2	3.968256	-.730659	.009430
3	-.834765	.925329	-.008164
4	-4.505439	-.441805	.001257

The problem is to rank the animals for selection purposes. If animals are ranked on the basis of a_0, then animal 2 would be the highest (if that was desirable). If ranked on the basis of a_1, then animal 3 would be the highest, and if ranked on the basis of a_2, then animal 2 would be the highest. To properly rank the animals, an EBV at different ages could be calculated, and then these could be combined with appropriate economic weights. Calculate EBVs for 24, 36, and 48 mo of age, and use economic weights of 2, 1, and .5, respectively, for the three EBVs. A Total Economic Value can be calculated as

$$\text{TEV} = 2 * \text{EBV}(24) + 1 * \text{EBV}(36) + .5 * \text{EBV}(48).$$

The Legendre polynomials for ages 24, 36, and 48 mo are given in the rows of the following matrix $\mathbf{L}$,

$$\mathbf{L} = \begin{pmatrix} .7071 & -.8328 & .3061 \\ .7071 & -.3429 & -.6046 \\ .7071 & .2449 & -.6957 \end{pmatrix}.$$

The results are shown in the following table.

Animal	EBV(24)	EBV(36)	EBV(48)	TEV
1	-1.58	-1.49	-1.38	-5.33
2	5.86	4.78	3.53	18.26
3	-3.89	-2.61	-1.10	-10.93
4	-3.49	-3.68	-3.89	-12.61
5	-2.56	-1.83	-.96	-7.43
6	3.25	2.71	2.09	10.25
7	1.97	1.43	.79	5.76
8	-2.66	-2.31	-1.91	-8.58

The animal with the highest TEV was animal 2. All animals ranked rather similarly at each age on their EBVs. Rankings of animals could change with age. Thus, the pattern of growth could be changed one that is desirable.

Estimation of the residual variance is

$$\hat{\sigma}_e^2 = (\mathbf{y}'\mathbf{y} - \hat{\mathbf{b}}'\mathbf{X}'\mathbf{y} - \hat{\mathbf{c}}'\mathbf{W}'\mathbf{y} - \hat{\mathbf{a}}_n'\mathbf{M}'\mathbf{y} - \hat{\mathbf{p}}'\mathbf{M}'\mathbf{y})/(N - r(\mathbf{X})),$$

where

$$\begin{aligned}
\mathbf{y}'\mathbf{y} &= 665,035, \\
\hat{\beta}'\mathbf{W}'\mathbf{y} &= 664902.89, \\
N - r(\mathbf{X}) &= 12 - 3 = 9, \\
\hat{\sigma}_e^2 &= 14.6788.
\end{aligned}$$

131 EXERCISES

1. A biologist studied chipmunks in southern Ontario. He planted small TV cameras inside four nests of chipmunks. The territories of the occupants of the four nests did not overlap. With the cameras he had students monitor the nests for a day at several times during the summer and fall. Students counted the number of nuts that were collected and stored in the nest on a given day during a set 8 hour period. Below are the observations on the four nests for various days of the year.

Number of nuts collected in 8 hours.

Day of Year	Nest	Number of Nuts	Day of Year	Nest	Nuts
123	1	25	161	1	11
124	2	37	155	2	14
127	3	16	153	3	9
129	4	42	151	4	12
192	1	13	225	1	38
194	2	15	227	2	29
198	3	10	233	3	37
200	4	16	246	4	44

(a) Write a complete random regression model for these data, including the assumptions and limitations.

(b) Compute a simple 4 by 4 phenotypic covariance matrix for the four nest sites. Then use covariance functions and tests to determine the appropriate order for a random regression model, and to obtain initial covariance matrices for the model.

(c) Apply the random regression model to the data, and plot the differences between the nests.

Multiple Traits

132 Introduction

Animals are commonly observed for more than one trait because many traits affect overall profitability. A multiple trait (MT) model is one in which two or more traits are analyzed simultaneously in order to take advantage of genetic and environmental correlations between traits.

Low Heritability Traits: MT models are useful for traits where the difference between genetic and residual correlations are large (e.g. greater than 0.5 difference) or where one trait has a much higher heritability than the other trait. EBVs for traits with low heritability tend to gain more in accuracy than EBVs for traits with high heritability, although all traits benefit to some degree from the simultaneous analysis.

Culling: Another use of MT models is for traits that occur at different times in the life of the animal, such that culling of animals results in fewer observations on animals for traits that occur later in life compared to those at the start. Consequently, animals which have observations later in life tend to have been selected based on their performance for earlier traits. Thus, analysis of later life traits by themselves could suffer from the effects of culling bias, and the resulting EBV could lead to errors in selecting future parents. An MT analysis that includes all observations on an animal upon which culling decisions have been based, has been shown to account, to some degree, for the selection that has taken place.

MT models do not offer great increases in accuracy for cases where heritabilities of traits are similar in magnitude, where both genetic and residual correlations are relatively the same magnitude and sign, or where every animal is measured for all traits. However, if culling bias may exist, then an MT analysis should be performed even if the parameters are similar. An MT analysis relies on the accuracy of the genetic and residual correlations that are assumed.

If the parameter estimates are greatly different from the underlying, unknown true values, then an MT analysis could do as much harm as it might do good.

Computer programs are more complicated, require more memory and disk storage for MT analyses. Verification of results might be more complicated. These have to be balanced against the benefits of an MT analysis. If culling bias is the main concern, then an MT model must be used regardless of the costs or no analysis should be done at all, except for the traits not affected by culling bias.

133 Models

Consider two traits with a single observation per trait on animals. A model should be specified separately for each trait. Let the model equation for trait 1 be

$$y_{1ij} = B_{1i} + a_{1j} + e_{1ij},$$

where B_{1i} is a fixed effect with p_B levels, a_{1j} is a random, animal additive genetic effect for trait 1, and e_{1ij} is a random residual environmental effect for trait 1.

The model equation for trait 2 might be

$$y_{2ij} = C_{2i} + a_{2j} + e_{2ij},$$

where C_{2i} is a fixed effect (different from B_{1i} for trait 1) with p_C levels, a_{2j} is a random, animal additive genetic effect for trait 2, and e_{2ij} is a random residual environmental effect for trait 2.

For example, y_{1ij} could be birthweight, so that B_{1i} could identify animals born in the same season. Trait 2 could be yearling weights and C_{2i} could identify contemporary groups of animals of the same sex, same herd, and same rearing unit within herd.

Because the two traits will be analyzed simultaneously, the variances and covariances need to be specified for the traits together. For example, the additive genetic variance-covariance (VCV) matrix could be written as

$$\mathbf{G} = \begin{pmatrix} g_{11} & g_{12} \\ g_{12} & g_{22} \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 15 \end{pmatrix},$$

and the residual environmental VCV matrix as

$$\mathbf{R} = \begin{pmatrix} r_{11} & r_{12} \\ r_{12} & r_{22} \end{pmatrix} = \begin{pmatrix} 10 & 5 \\ 5 & 100 \end{pmatrix}.$$

The genetic and residual correlations are, respectively,

$$\begin{aligned} \rho_g &= 2/(15)^{.5} = .516, \\ \rho_r &= 5/(1000)^{.5} = .158 \end{aligned}$$

with

$$h_1^2 = \frac{1}{11} = .0909,$$

and

$$h_2^2 = \frac{15}{115} = .1304.$$

For all data, then

$$Var \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{A}g_{11} & \mathbf{A}g_{12} \\ \mathbf{A}g_{12} & \mathbf{A}g_{22} \end{pmatrix}.$$

The structure of the residual VCV matrix over all observations can be written several ways depending on whether allowance is made for missing observations on either trait for some animals. If all animals were observed for both traits, then

$$Var\begin{pmatrix} \mathbf{e}_1 \\ \mathbf{e}_2 \end{pmatrix} = \begin{pmatrix} \mathbf{I}r_{11} & \mathbf{I}r_{12} \\ \mathbf{I}r_{12} & \mathbf{I}r_{22} \end{pmatrix}.$$

134 Simulation of Trait Records

When simulating data for a multiple trait problem, observations for all animals for all traits should be generated. Then one can go through the simulated data and delete observations to simulate a missing data situation, or selectively delete observations to imitate culling decisions. Another simplification is to assume that the model for each trait is the same, and then for a factor that does not belong with a given trait just make the true values of levels for that factor and trait equal to zero. In matrix form, the model equation for one animal would be

$$\begin{aligned}\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} &= \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right)\begin{pmatrix} B_{11} \\ B_{12} \\ B_{21} \\ B_{22} \end{pmatrix} + \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right)\begin{pmatrix} C_{11} \\ C_{12} \\ C_{13} \\ C_{21} \\ C_{22} \\ C_{23} \end{pmatrix} \\ & + \begin{pmatrix} \text{Parent Ave. Trait 1} \\ \text{Parent Ave. Trait 2} \end{pmatrix} + (b_{ii})^{.5}\mathbf{L}_G\begin{pmatrix} \text{RND(1)} \\ \text{RND(2)} \end{pmatrix} \\ & + \mathbf{L}_R\begin{pmatrix} \text{RND(3)} \\ \text{RND(4)} \end{pmatrix},\end{aligned}$$

where $\mathbf{L}_G$ and $\mathbf{L}_R$ are lower triangular matrices such that

$$\begin{aligned}\mathbf{G} &= \mathbf{L}_G\mathbf{L}_G' \\ &= \begin{pmatrix} 1 & 0 \\ 2 & (11)^{.5} \end{pmatrix}\mathbf{L}_G' \\ \text{and}\ \mathbf{R} &= \mathbf{L}_R\mathbf{L}_R' \\ &= \begin{pmatrix} (10)^{.5} & 0 \\ (2.5)^{.5} & (97.5)^{.5} \end{pmatrix}\mathbf{L}_R'.\end{aligned}$$

Let $B_{11} = 6.7$ and $B_{12} = 6.3$ for trait 1, and because factor B is not in the model for trait 2, then $B_{21} = 0$ and $B_{22} = 0$. Similarly, $C_{21} = 25$, $C_{22} = 40$, and $C_{23} = 55$ for trait 2, and because factor C is not in the model for trait 1, then $C_{11} = 0$, $C_{12} = 0$, and

$C_{13} = 0$. Suppose the animal is a base animal, then $b_{ii} = 1$ and the parent averages for traits 1 and 2 are assumed to be zero, then the observations would be

$$\begin{aligned}
\begin{pmatrix} y_{1ij} \\ y_{2ij} \end{pmatrix} &= \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{array}\right) \begin{pmatrix} 6.7 \\ 6.3 \\ 0 \\ 0 \end{pmatrix} + \left(\begin{array}{ccc|ccc} 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right) \begin{pmatrix} 0 \\ 0 \\ 0 \\ 25 \\ 40 \\ 55 \end{pmatrix} \\
&+ \begin{pmatrix} \text{Parent Ave. Trait 1=0} \\ \text{Parent Ave. Trait 2=0} \end{pmatrix} + (1)^{.5} \begin{pmatrix} 1 & 0 \\ 2 & (11)^{.5} \end{pmatrix} \begin{pmatrix} .6942 \\ -1.3027 \end{pmatrix} \\
&+ \begin{pmatrix} (10)^{.5} & 0 \\ (2.5)^{.5} & (97.5)^{.5} \end{pmatrix} \begin{pmatrix} -.5324 \\ -.9468 \end{pmatrix}, \\
&= \begin{pmatrix} 6.7 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ 40 \end{pmatrix} + \begin{pmatrix} .6942 \\ -2.9322 \end{pmatrix} + \begin{pmatrix} -1.6836 \\ -10.1907 \end{pmatrix} \\
&= \begin{pmatrix} 5.7106 \\ 26.8771 \end{pmatrix}.
\end{aligned}$$

The following data (rounded off) were simulated according to the preceeding scheme and parameters.

Animal	Sire	Dam	B-level	C-level	Trait 1	Trait 2
1	0	0	1	1	2.3	39
2	0	0	1	2	2.6	39
3	0	0	1	3	9.8	53
4	0	0	1	1	4.7	4
5	0	0	1	2	5.5	63
6	1	3	2	3	2.5	64
7	1	4	2	2	8.4	35
8	1	5	2	3	8.2	41
9	2	3	2	1	9.0	27
10	2	4	2	1	7.8	32
11	2	5	2	2	2.8	46
12	6	10	2	3	7.4	67

To simulate selection, assume that all animals had trait 1 observed, but any animal with a trait 1 value below 3.0, then trait 2 observation was removed. Four trait 2 observations were deleted, giving the results in the table below.

Animal	Sire	Dam	B-level	C-level	Trait 1	Trait 2
1	0	0	1	1	2.3	
2	0	0	1	2	2.6	
3	0	0	1	3	9.8	53
4	0	0	1	1	4.7	4
5	0	0	1	2	5.5	63
6	1	3	2	3	2.5	
7	1	4	2	2	8.4	35
8	1	5	2	3	8.2	41
9	2	3	2	1	9.0	27
10	2	4	2	1	7.8	32
11	2	5	2	2	2.8	
12	6	10	2	3	7.4	67

135 HMME

Organize the data by traits within animals. With two traits there are three possible residual matrices per animal, i.e.,

$$\begin{aligned}\mathbf{R}_{12} &= \begin{pmatrix} 10 & 5 \\ 5 & 100 \end{pmatrix}, \\ \mathbf{R}_{1} &= \begin{pmatrix} 10 & 0 \\ 0 & 0 \end{pmatrix}, \\ \mathbf{R}_{2} &= \begin{pmatrix} 0 & 0 \\ 0 & 100 \end{pmatrix},\end{aligned}$$

depending on whether both traits, trait 1, or trait 2, respectively, were observed. In the example data, only $\mathbf{R}_{12}$ and $\mathbf{R}_1$ are needed. To simplify notation, let

$$\begin{aligned}\mathbf{E}_{12} &= \mathbf{R}_{12}^{-1} \\ &= \frac{1}{975}\begin{pmatrix} 100 & -5 \\ -5 & 10 \end{pmatrix} \\ &= \begin{pmatrix} .102564 & -.005128 \\ -.005128 & .010256 \end{pmatrix},\end{aligned}$$

and

$$\begin{aligned}\mathbf{E}_{1} &= \mathbf{R}_{1}^{-} \\ &= \begin{pmatrix} .1 & 0 \\ 0 & 0 \end{pmatrix}, \\ \mathbf{E}_{2} &= \mathbf{R}_{2}^{-} \\ &= \begin{pmatrix} 0 & 0 \\ 0 & .01 \end{pmatrix}.\end{aligned}$$

Again, to simplify construction of the MME, pretend that both traits have the same model equation, so that

$$y_{tijk} = B_{ti} + C_{tj} + a_{tk} + e_{tijk}.$$

There are 2 levels of factor B, three levels of factor C, and 12 animals. For a single trait model this would give MME of order 17. Construct a table of order 17. The elements of this table will be matrices of order 2(in general, order t). Start with animal 1, then

$$\mathbf{y}_1 = \begin{pmatrix} 2.3 \\ - \end{pmatrix}, \text{ and } \mathbf{E}_1\mathbf{y}_1 = \begin{pmatrix} .23 \\ 0 \end{pmatrix}.$$

Now add $\mathbf{E}_1$ to the boxes in the MME table as follows:

	B_1	B_2	C_1	C_2	C_3	a_1	$\cdots$	RHS
B_1	$\mathbf{E}_1$		$\mathbf{E}_1$			$\mathbf{E}_1$		$\mathbf{E}_1\mathbf{y}_1$
B_2								
C_1	$\mathbf{E}_1$		$\mathbf{E}_1$			$\mathbf{E}_1$		$\mathbf{E}_1\mathbf{y}_1$
C_2								
C_3								
a_1	$\mathbf{E}_1$		$\mathbf{E}_1$			$\mathbf{E}_1$		$\mathbf{E}_1\mathbf{y}_1$

Similarly for animal 2,

$$\mathbf{y}_2 = \begin{pmatrix} 2.6 \\ - \end{pmatrix}, \text{ and } \mathbf{E}_1\mathbf{y}_2 = \begin{pmatrix} .26 \\ 0 \end{pmatrix}.$$

Accumulating into the MME table gives

	B_1	B_2	C_1	C_2	C_3	$\cdots$	a_2	RHS
B_1	$2\mathbf{E}_1$		$\mathbf{E}_1$	$\mathbf{E}_1$			$\mathbf{E}_1$	$\mathbf{E}_1(\mathbf{y}_1 + \mathbf{y}_2)$
B_2								
C_1	$\mathbf{E}_1$		$\mathbf{E}_1$					$\mathbf{E}_1\mathbf{y}_1$
C_2	$\mathbf{E}_1$			$\mathbf{E}_1$			$\mathbf{E}_1$	$\mathbf{E}_1\mathbf{y}_2$
C_3								
a_2	$\mathbf{E}_1$			$\mathbf{E}_1$			$\mathbf{E}_1$	$\mathbf{E}_1\mathbf{y}_2$

For animal 3,

$$\mathbf{y}_3 = \begin{pmatrix} 9.8 \\ 53 \end{pmatrix}, \text{ and } \mathbf{E}_{12}\mathbf{y}_3 = \begin{pmatrix} .7333 \\ .4933 \end{pmatrix},$$

and the MME table becomes

	B_1	B_2	C_1	C_2	C_3	$\cdots$	a_3	RHS
B_1	$2\mathbf{E}_1 + \mathbf{E}_{12}$		$\mathbf{E}_1$	$\mathbf{E}_1$	$\mathbf{E}_{12}$		$\mathbf{E}_{12}$	$\mathbf{E}_1(\mathbf{y}_1 + \mathbf{y}_2) + \mathbf{E}_{12}\mathbf{y}_3$
B_2								
C_1	$\mathbf{E}_1$		$\mathbf{E}_1$					$\mathbf{E}_1\mathbf{y}_1$
C_2	$\mathbf{E}_1$			$\mathbf{E}_1$				$\mathbf{E}_1\mathbf{y}_2$
C_3	$\mathbf{E}_{12}$				$\mathbf{E}_{12}$		$\mathbf{E}_{12}$	$\mathbf{E}_{12}\mathbf{y}_3$
a_3	$\mathbf{E}_{12}$				$\mathbf{E}_{12}$		$\mathbf{E}_{12}$	$\mathbf{E}_{12}\mathbf{y}_3$

The remaining animals are processed in the same manner. The resulting equations are of order 34 by 34. To these $\mathbf{A}^{-1} \otimes \mathbf{G}^{-1}$ must be added to the animal by animal submatrix in order to form the full HMME. However, solutions for the B-factor for trait 2 are not needed because the B-factor does not affect trait 2, and solutions for the C-factor for trait 1 are not needed because the C-factor does not affect trait 1. Therefore, remove rows (and columns) 2, 4, 5, 7, and 9, or if an iterative solution is being computed, then require that the solutions for B_{21}, B_{22}, C_{11}, C_{12}, and C_{13} are always equal to zero. The solutions to the HMME, for this example, were

$$\begin{aligned}
B_{11} &= 5.0209 \\
B_{12} &= 6.5592 \\
C_{21} &= 20.0882 \\
C_{22} &= 49.0575 \\
C_{23} &= 51.9553
\end{aligned}$$

The animal additive genetic solutions are shown in the table below.

Animal	Sire	Dam	Trait 1	Trait 2
1	0	0	-.3573	-1.6772
2	0	0	-.0730	1.0418
3	0	0	.4105	1.1707
4	0	0	-.0449	-1.4922
5	0	0	.0646	.9570
6	1	3	-.1033	-.1410
7	1	4	-.1975	-2.2983
8	1	5	-.1410	-.9633
9	2	3	.3079	1.6227
10	2	4	.1426	1.1273
11	2	5	-.1830	.6418
12	6	10	.1554	1.5089

The correlation between the animal additive genetic solutions for traits 1 and 2 was .74 which is greater than the .52 assumed in the original $\mathbf{G}$.

136 Bruce Tier's MT Trick

There is another way to construct the MME without the need of forming different inverses of $\mathbf{R}$ for missing traits. If a trait is missing, then that observation is assigned to its own contemporary group in the model for that trait. In the example data there were four missing observations. Animal 1 would be assigned to C_{24}, animal 2 to C_{25}, animal 6 to C_{26} and animal 11 to C_{27}, respectively. In this case only trait 2 observations were missing. If trait 1 observations were also missing, then animals would be assigned to separate levels of factor B. In this way, only one residual VCV matrix is needed, i.e. $\mathbf{R}_{12}$. Let $\mathbf{X}$ represent the design matrix for fixed effects (factors B and C) for either trait. Note the four extra columns for factor C for the animals with missing trait 2 observations.

$$\mathbf{Xb} = \left(\begin{array}{ccccccccc} 1 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array} \right) \left(\begin{array}{c} B_{11} \\ B_{12} \\ C_{21} \\ C_{22} \\ C_{23} \\ C_{24} \\ C_{25} \\ C_{26} \\ C_{27} \end{array} \right).$$

A missing observation is replaced with zero. The resulting solutions are identical to those given earlier, except that now there are also solutions for the single observation contemporary groups. That is, $C_{24} = 2.8590$, $C_{25} = .1322$, $C_{26} = 2.1189$, and $C_{27} = 1.1463$. These solutions do not mean anything, but must be calculated.

To prove that this trick will work, take $\mathbf{R}^{-1}$ and do a Gaussian elimination (i.e. absorption) of the row and column corresponding to the missing trait, say trait 2,

$$\left(\begin{array}{cc} e^{11} & e^{12} \\ e^{12} & e^{22} \end{array} \right) - \left(\begin{array}{c} e^{12} \\ e^{22} \end{array} \right) (e^{22})^{-1} \left(\begin{array}{cc} e^{12} & e^{22} \end{array} \right),$$

$$= \left(\begin{array}{cc} e^{11} & e^{12} \\ e^{12} & e^{22} \end{array} \right) - \left(\begin{array}{cc} e^{12}(e^{22})^{-1}e^{12} & e^{12} \\ e^{12} & e^{22} \end{array} \right),$$

$$= \left(\begin{array}{cc} e^{11} - e^{12}(e^{22})^{-1}e^{12} & 0 \\ 0 & 0 \end{array} \right).$$

Recall that for a matrix of order 2 that

$$e^{11} = e_{22}/ \mid \mathbf{R} \mid,$$

$$\begin{aligned} e^{12} &= -e_{12}/\mid \mathbf{R} \mid, \\ e^{22} &= e_{11}/\mid \mathbf{R} \mid, \\ \mid \mathbf{R} \mid &= (e_{11}e_{22} - e_{12}e_{12}) \end{aligned}$$

then

$$\begin{aligned} e^{11} - e^{12}(e^{22})^{-1}e^{12} &= (e_{22} - e_{12}(e_{11})^{-1}e_{12})/\mid \mathbf{R} \mid \\ &= e_{11}(e_{22} - e_{12}(e_{11})^{-1}e_{12}/e_{11}(e_{11}e_{22} - e_{12}e_{12}) \\ &= (e_{11})^{-1} \end{aligned}$$

which is exactly the weight applied to records on animals with only trait 1 observed. This proof can be extended to any number of traits recorded and any number missing, by partitioning $\mathbf{R}$ into

$$\begin{pmatrix} \mathbf{R}_{oo} & \mathbf{R}_{om} \\ \mathbf{R}_{mo} & \mathbf{R}_{mm} \end{pmatrix},$$

where the subscript o refers to traits that were observed and m refers to traits that were missing on an animal. Then it can be easily shown that

$$\mathbf{R}_{oo}^{-1} = \mathbf{R}^{oo} - \mathbf{R}^{om}(\mathbf{R}^{mm})^{-1}\mathbf{R}^{mo}.$$

This trick is not very practical, for example, when one trait has 1 million observations and trait 2 has only 100,000 observations, then there would be 900,000 extra single observation subclasses created for trait 2. However, if the percentages of missing observations are relatively small, or if many traits are being considered, then pretending all observations are present may make programming easier.

137 Estimation of Covariance Matrices

Derivative free REML is one option for estimating variances and covariances in a multitrait situation. The EM algorithm is not suitable due to the requirement for the traces of inverse elements that are needed. Even DF REML takes considerably more time as the number of parameters to be estimated increases.

Another option is the Bayesian approach, where operations are performed in t dimensions, for t being the number of traits. Thus, for a solution to the MME, the $t \times 1$ vector for any one fixed effect, for example, would be

$$\hat{\beta}_i = (\mathbf{X}_i'\mathbf{R}^{-1}\mathbf{X}_i)^{-1}\mathbf{X}_i'\mathbf{R}^{-1}(\mathbf{y}_i - \mathbf{W}_{-i}\beta_{-i}),$$

then

$$\beta_i = \hat{\beta}_i + \mathbf{L}\mathbf{v},$$

where

$$\mathbf{LL}' = (\mathbf{X}_i'\mathbf{R}^{-1}\mathbf{X}_i)^{-1},$$

and $\mathbf{v}$ is a $t \times 1$ vector of random normal deviates. Similar formulas can be derived for the random factors of the model. The conditional distributions for these factors are assumed to be normally distributed.

If $\mathbf{a}_i$ is the $q \times 1$ vector of animal solutions for trait i, then form

$$\mathbf{U} = \left(\begin{array}{cccc} \mathbf{a}_1 & \mathbf{a}_2 & \cdots & \mathbf{a}_t \end{array} \right),$$

followed by

$$\mathbf{S}_a = (\mathbf{U}'\mathbf{A}^{-1}\mathbf{U} + \nu_a\mathbf{G}_a),$$

which is then inverted and a Cholesky decomposition is applied to the inverse, i.e.

$$\mathbf{L}_a\mathbf{L}_a' = \mathbf{S}_a^{-1},$$

where $\mathbf{L}_a$ is supplied to a Wishart distribution random number generator to give a new sample matrix for the inverse of the additive genetic variances and covariances.

A difficult part of a multiple trait analysis, when missing traits are possible, is the calculation of the appropriate residual matrix of sums of squares and cross products. The residual effect for any one trait is

$$e_{ti} = y_{ti} - \mathbf{w}_i'\beta_t,$$

and for a particular animal, k,

$$\mathbf{res}_k = \left(\begin{array}{cc} \mathbf{e}_o' & \mathbf{e}_m' \end{array} \right),$$

where the subscripts o and m refer to observed and missing, respectively, and where $\mathbf{e}_m' = \mathbf{0}'$. In order to calculate the correct sum of squares and crossproducts of the residuals for REML or the Bayesian approach, a prediction of $\mathbf{e}_m$ is needed. This can be easily done by

$$\mathbf{res}_{k*}' = \left(\begin{array}{cc} \mathbf{R}_{oo} & \mathbf{R}_{om} \\ \mathbf{R}_{mo} & \mathbf{R}_{mm} \end{array} \right) \left(\begin{array}{cc} \mathbf{R}_{oo}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{array} \right) \left(\begin{array}{c} \mathbf{e}_o \\ \mathbf{0} \end{array} \right).$$

For the first animal in the example data, for example, the estimated residual for trait 1 was -2.363638, then

$$\begin{aligned} \mathbf{res}_{1*}' &= \left(\begin{array}{rr} 10 & 5 \\ 5 & 100 \end{array} \right) \left(\begin{array}{rr} \frac{1}{10} & 0 \\ 0 & 0 \end{array} \right) \left(\begin{array}{r} -2.363638 \\ 0 \end{array} \right), \\ &= \left(\begin{array}{rr} 1 & 0 \\ .5 & 0 \end{array} \right) \left(\begin{array}{r} -2.363638 \\ 0 \end{array} \right), \\ &= \left(\begin{array}{r} -2.363638 \\ -1.181819 \end{array} \right). \end{aligned}$$

If you use Bruce Tier's MME with the single observation contemporary groups for missing trait observations, then the residuals can be calculated directly by $(\mathbf{y}-\mathbf{W}\beta)$ using zero as the observation for the missing traits and using the solutions for the single observation contemporary groups. This gives the exact same residual estimates as the above methodology. Therefore, Tier's approach is handy for the Gibb's sampling algorithm.

Once the residuals are calculated for all animals with records, then calculate

$$\mathbf{res}'_{k*}\mathbf{res}_{k*},$$

and sum across N animals. Then let

$$\mathbf{S}_e = (\sum_{k=1}^{N}(\mathbf{res}'_{k*}\mathbf{res}_{k*}) + \nu_e\mathbf{R}_e),$$

which is then inverted and a Cholesky decomposition is applied to the inverse, i.e.

$$\mathbf{L}_e\mathbf{L}'_e = \mathbf{S}_e^{-1},$$

where $\mathbf{L}_e$ is supplied to a Wishart distribution random number generator to give a new sample matrix for the inverse of the residual variances and covariances.

138 EXERCISES

Below are data on three blood component traits (just called Trait 1, Trait 2, and Trait 3) taken on 6 mink adults. The lab technician slid on a banana peel, dropped some of the blood samples, and so some data were lost. The technician was fired for eating in the lab.

Animal	Sire	Dam	Age(yrs)	Trait 1	Trait 2	Trait 3
6	1	3	3	2.5	53	
7	1	4	2	8.4		175
8	1	5	3		41	197
9	2	3	1	9.0		
10	2	4	1		32	156
11	2	5	2			168

Let

$$\mathbf{R} = \begin{pmatrix} 11 & 5 & 29 \\ 5 & 97 & 63 \\ 29 & 63 & 944 \end{pmatrix},$$

and

$$\mathbf{G} = \begin{pmatrix} 0.94 & 1.62 & -5.87 \\ 1.62 & 14.33 & -11.79 \\ -5.87 & -11.79 & 803.41 \end{pmatrix}.$$

1. Write a model for the analysis.
2. Analyze the data with that model.
3. Rank the animals in some way.
4. Perform one iteration or one round of sampling to estimate the covariance matrices.

Non-Additive Animal Models

139 Non-Additive Genetic Effects

Non-additive genetic effects (or epistatic effects) are the interactions among loci in the genome. There are many possible degrees of interaction (involving different numbers of loci), but the effects and contributions of those interactions have been shown to diminish as the degree of complexity increases. Thus, mainly dominance, additive by additive, additive by dominance, and dominance by dominance interactions have been considered in animal studies.

To estimate the variances associated with each type of interaction, relatives are needed that have different additive and dominance relationships among themselves. Computation of dominance relationships is more difficult than additive relationships, but can be done as shown in earlier notes.

If non-additive genetic effects are included in an animal model, then an assumption of random mating is required. Otherwise non-zero covariances can arise between additive and dominance genetic effects, which complicates the model enormously.

140 Interactions at a Single Locus

Let the model for the genotypic values be given as

$$G_{ij} = \mu + a_i + a_j + d_{ij},$$

where

$$\begin{aligned}
\mu &= G_{..} = \sum_{i,j} f_{ij} G_{ij}, \\
a_i &= G_{i.} - G_{..}, \\
G_{i.} &= Pr(A_1)G_{11} + Pr(A_2)G_{12}, \\
a_j &= G_{.j} - G_{..} \\
G_{.j} &= Pr(A_1)G_{12} + Pr(A_2)G_{22}, \\
d_{ij} &= G_{ij} - a_i - a_j - \mu
\end{aligned}$$

Thus, there are just additive effects and dominance effects to be estimated at a single locus. A numerical example is given below.

Genotype	Frequency	Value
A_1A_1	$f_{11} = 0.04$	$G_{11} = 100$
A_1A_2	$f_{12} = 0.32$	$G_{12} = 70$
A_2A_2	$f_{22} = 0.64$	$G_{22} = 50$

Then

$$\begin{aligned}
\mu &= 0.04(100) + 0.32(70) + 0.64(50) \ = \ 58.4, \\
G_{1.} &= 0.2(100) + 0.8(70) \ = \ 76.0, \\
G_{.2} &= 0.2(70) + 0.8(50) \ = \ 54.0, \\
a_1 &= G_{1.} - \mu \ = \ 17.6, \\
a_2 &= G_{.2} - \mu \ = \ -4.4, \\
d_{11} &= G_{11} - a_1 - a_1 - \mu \ = \ 6.4, \\
d_{12} &= G_{12} - a_1 - a_2 - \mu \ = \ -1.6, \\
d_{22} &= G_{22} - a_2 - a_2 - \mu \ = \ 0.4.
\end{aligned}$$

Now a table of breeding values and dominance effects can be completed.

Genotype	Frequency	Total	Additive	Dominance
A_1A_1	0.04	$G_{11} = 100$	$a_1 + a_1 = 35.2$	$d_{11} = 6.4$
A_1A_2	0.32	$G_{12} = 70$	$a_1 + a_2 = 13.2$	$d_{12} = -1.6$
A_2A_2	0.64	$G_{22} = 50$	$a_2 + a_2 = -8.8$	$d_{22} = 0.4$

The additive genetic variance is

$$\sigma_a^2 = 0.04(35.2)^2 + 0.32(13.2)^2 + 0.64(-8.8)^2 \ = \ 154.88,$$

and the dominance genetic variance is

$$\sigma_d^2 = 0.04(6.4)^2 + 0.32(-1.6)^2 + 0.64(0.4)^2 \ = \ 2.56,$$

and the total genetic variance is

$$\begin{aligned}
\sigma_G^2 &= \sigma_a^2 + \sigma_d^2 \\
&= 157.44, \\
&= 0.04(100 - \mu)^2 + 0.32(70 - \mu)^2 + 0.64(50 - \mu)^2 \\
&= 0.04(41.6)^2 + 0.32(11.6)^2 + 0.64(-8.4)^2 \\
&= 157.44.
\end{aligned}$$

This result implies that there is a zero covariance between the additive and dominance deviations. This can be shown by calculating the covariance between additive and dominance deviations,

$$\begin{aligned} Cov(A,D) &= 0.04(35.2)(6.4) + 0.32(13.2)(-1.6) + 0.64(-8.8)(0.4) \\ &= 0 \end{aligned}$$

The covariance is zero under the assumption of a large, random mating population without selection.

141 Interactions for Two Unlinked Loci

Consider two loci each with two alleles, and assume that the two loci are on different chromosomes and therefore unlinked. Let $p_A = 0.4$ be the frequency of the A_1 allele at locus A, and let $p_B = 0.8$ be the frequency of the B_1 allele at locus B. Then the possible genotypes, their expected frequencies assuming joint equilibrium, and genotypic values would be as in the table below. Joint equilibrium means that each locus is in Hardy-Weinberg equilibrium and that the probabilities of the possible gametes are equal to the product of the allele frequencies as shown in the table below.

Possible Gametes	Expected Frequencies	
A_1 B_1	$p_A p_B$	$= 0.32$
A_1 B_2	$p_A q_B$	$= 0.08$
A_2 B_1	$q_A p_B$	$= 0.48$
A_2 B_2	$q_A q_B$	$= 0.12$

Multiplying these gametic frequencies together, to simulate random mating gives the frequencies in the table below. The genotypic values were arbitrarily assigned to illustrate the process of estimating the genetic effects.

Genotypes		Frequencies		Genotypic
A-Locus	B-locus		$f_{ijk\ell}$	Value,$G_{ijk\ell}$
i,j	k,ℓ			
11	11	$p_A^2 p_B^2$	=.1024	$G_{1111} = 108$
11	12	$p_A^2 2p_B q_B$	=.0512	$G_{1112} = 75$
11	22	$p_A^2 q_B^2$	=.0064	$G_{1122} = -80$
12	11	$2p_A q_A p_B^2$	=.3072	$G_{1211} = 95$
12	12	$4p_A q_A p_B q_B$	=.1536	$G_{1212} = 50$
12	22	$2p_A q_A q_B^2$	=.0192	$G_{1222} = -80$
22	11	$q_A^2 p_B^2$	=.2304	$G_{2211} = 48$
22	12	$q_A^2 2p_B q_B$	=.1152	$G_{2212} = 36$
22	22	$q_A^2 q_B^2$	=.0144	$G_{2222} = -100$

141.1 Estimation of Additive Effects

$$\begin{aligned}
\alpha_{A1} &= G_{1\cdots} - \mu_G = 16.6464, \\
\alpha_{A2} &= -11.0976, \\
\alpha_{B1} &= 10.4384, \\
\alpha_{B2} &= -41.7536.
\end{aligned}$$

The additive genetic effect for each genotype is

$$\begin{aligned}
a_{ijk\ell} &= \alpha_{Ai} + \alpha_{Aj} + \alpha_{Bk} + \alpha_{B\ell} \\
a_{1111} &= \alpha_{A1} + \alpha_{A1} + \alpha_{B1} + \alpha_{B1}, \\
&= 16.6464 + 16.6464 + 10.4384 + 10.4384, \\
&= 54.1696, \\
a_{1112} &= 1.9776, \\
a_{1122} &= -50.2144, \\
a_{1211} &= 26.4256, \\
a_{1212} &= -25.7664, \\
a_{1222} &= -77.9584, \\
a_{2211} &= -1.3184, \\
a_{2212} &= -53.5104, \\
a_{2222} &= -105.7024.
\end{aligned}$$

The additive genetic variance is then

$$\sigma_a^2 = \sum_{ijk\ell} f_{ijk\ell} a_{ijk\ell}^2 \;=\; 1241.1517.$$

141.2 Estimation of Dominance Effects

There are six conditional means to compute, one for each single locus genotype.

$$\begin{aligned}
G_{11..} &= Pr(B_1B_1)G_{1111} + Pr(B_1B_2)G_{1112} + Pr(B_2B_2)G_{1122}, \\
&= (.64)(108) + (.32)(75) + (.04)(-80), \\
&= 89.92, \\
G_{12..} &= 73.60, \\
G_{22..} &= 38.24, \\
G_{..11} &= 80.16, \\
G_{..12} &= 48.96, \\
G_{..22} &= -87.20.
\end{aligned}$$

The dominance genetic effects are given by

$$\delta_{Aij} = G_{ij..} - \mu_G - \alpha_{Ai} - \alpha_{Aj},$$

so that

$$\begin{aligned}
\delta_{A11} &= 89.92 - 63.4816 - 16.6464 - 16.6464, \\
&= -6.8544, \\
\delta_{A12} &= 4.5696, \\
\delta_{A22} &= -3.0464, \\
\delta_{B11} &= -4.1984, \\
\delta_{B12} &= 16.7936, \\
\delta_{B22} &= -67.1744.
\end{aligned}$$

The dominance deviations for each genotype are

$$\begin{aligned}
d_{ijk\ell} &= \delta_{Aij} + \delta_{Bk\ell}, \\
d_{1111} &= -11.0528, \\
d_{1112} &= 9.9392, \\
d_{1122} &= -74.0288, \\
d_{1211} &= 0.3712, \\
d_{1212} &= 21.3632, \\
d_{1222} &= -62.6048, \\
d_{2211} &= -7.2448, \\
d_{2212} &= 13.7472, \\
d_{2222} &= -70.2208.
\end{aligned}$$

The dominance genetic variance is therefore,

$$\sigma_d^2 = \sum_{ijk\ell} f_{ijk\ell} d_{ijk\ell}^2 \ = \ 302.90625.$$

141.3 Additive by Additive Effects

These are the interactions between alleles at different loci. There are four conditional means to calculate,

$$\begin{aligned}
G_{1.1.} &= Pr(A_1B_1)G_{1111} + Pr(A_1B_2)G_{1112} \\
&\quad +Pr(A_2B_1)G_{1211} + Pr(A_2B_2)G_{1212}, \\
&= .32(108) + .08(75) + .48(95) + .12(50), \\
&= 92.16, \\
G_{1..2} &= 84.192, \\
G_{.21.} &= 61.76, \\
G_{.2.2} &= 14.88.
\end{aligned}$$

The additive by additive genetic effect is

$$\begin{aligned}
\alpha\alpha_{A1B1} &= G_{1.1.} - \mu_G - \alpha_{A1} - \alpha_{B1} \ = \ 1.5936, \\
\alpha\alpha_{A1B2} &= -6.3744, \\
\alpha\alpha_{A2B1} &= -1.0624, \\
\alpha\alpha_{A2B2} &= 4.2496.
\end{aligned}$$

The additive by additive deviations for each genotype are

$$\begin{aligned}
aa_{ijk\ell} &= \alpha\alpha_{AiBk} + \alpha\alpha_{AiB\ell} + \alpha\alpha_{AjBk} + \alpha\alpha_{AjB\ell}, \\
aa_{1111} &= 6.3744, \\
aa_{1112} &= -9.5616, \\
aa_{1122} &= -25.4976, \\
aa_{1211} &= 1.0624, \\
aa_{1212} &= -1.5936, \\
aa_{1222} &= -4.2496, \\
aa_{2211} &= -4.2496, \\
aa_{2212} &= 6.3744, \\
aa_{2222} &= 16.9984.
\end{aligned}$$

The additive by additive genetic variance is

$$\sigma^2_{aa} = \sum_{ijk\ell} f_{ijk\ell} aa^2_{ijk\ell} \ = \ 27.08865.$$

141.4 Additive by Dominance Effects

This is the interaction between a single allele at one locus with the pair of alleles at a second locus. There are twelve possible conditional means for twelve possible different A by D interactions. Not all are shown,

$$\begin{aligned}
G_{1.11} &= Pr(A_1)G_{1111} + Pr(A_2)G_{1211}, \\
&= .4(108) + .6(95), \\
&= 100.2, \\
G_{.211} &= Pr(A_1)G_{1211} + Pr(A_2)G_{2211}, \\
&= .4(95) + .6(48), \\
&= 66.8.
\end{aligned}$$

The specific additive by dominance effects are

$$\alpha\delta_{A1B11} = G_{1.11} - \mu_G - \alpha_{A1} - 2\alpha_{B1} - \delta_{B11} \ = \ 0.2064.$$

Finally, the additive by dominance genetic values for each genotype are

$$\begin{aligned}
ad_{ijk\ell} &= \alpha\delta_{AiBk\ell} + \alpha\delta_{AjBk\ell} + \alpha\delta_{BkAij} + \alpha\delta_{B\ell Aij}, \\
ad_{1111} &= -3.8784, \\
ad_{1112} &= 4.7856, \\
ad_{1122} &= 23.7696, \\
ad_{1211} &= 2.9296, \\
ad_{1212} &= -4.5664, \\
ad_{1222} &= -10.3424, \\
ad_{2211} &= -2.1824, \\
ad_{2212} &= 3.9616, \\
ad_{2222} &= 3.2256.
\end{aligned}$$

The additive by dominance genetic variance is

$$\sigma^2_{ad} = \sum_{ijk\ell} f_{ijk\ell} ad^2_{ijk\ell} \ = \ 17.2772.$$

141.5 Dominance by Dominance Effects

Dominance by dominance genetic effects are the interaction between a pair of alleles at one locus with another pair of alleles at a second locus. These effects are calculated as the genotypic values minus all of the other effects for each genotype. That is,

$$dd_{ijk\ell} = G_{ijk\ell} - \mu_G - a_{ijk\ell} - d_{ijk\ell} - aa_{ijk\ell} - ad_{ijk\ell}.$$

The dominance by dominance genetic variance is the sum of the frequencies of each genotype times the dominance by dominance effects squared, $\sigma^2_{dd} = 8.5171$. The table of all genetic effects are given below.

Genotypes		$f_{ijk\ell}$	$G_{ijk\ell}$	$a_{ijk\ell}$	$d_{ijk\ell}$	$aa_{ijk\ell}$	$ad_{ijk\ell}$	$dd_{ijk\ell}$
A-Locus	B-Locus							
11	11	.1024	108	54.1696	-11.0528	6.3744	-3.8784	-1.0944
11	12	.0512	75	1.9776	9.9392	-9.5616	4.7856	4.3776
11	22	.0064	-80	-50.2144	-74.0288	-25.4976	23.7696	-17.5104
12	11	.3072	95	26.4256	0.3712	1.0624	2.9296	0.7296
12	12	.1536	50	-25.7664	21.3632	-1.5936	-4.5664	-2.9184
12	22	.0192	-80	-77.9584	-62.6048	-4.2496	-10.3424	11.6736
22	11	.2304	48	-1.3184	-7.2448	-4.2496	-2.1824	-0.4864
22	12	.1152	36	-53.5104	13.7472	6.3744	3.9616	1.9456
22	22	.0144	-100	-105.7024	-70.2208	16.9984	3.2256	-7.7824

A summary of the genetic variances is

Total Genetic	1596.9409,
Additive	1241.1517,
Dominance	302.9062,
Add by Add	27.0886,
Add by Dom	17.2772,
Dom by Dom	8.5171.

142 More than Two Loci

Interactions can occur between several loci. The maximum number of loci involved in simultaneous interactions is unknown, but the limit is the number of gene loci in the genome. Many geneticists believe that the higher order interactions are few in number, and that if they exist the magnitude of their effects is small. Someday the measurement of all of these interactions may be possible, but modelling them may be impossible, and practical utilization of that information may be close to impossible.

143 Linear Models for Non-Additive Genetic Effects

Consider a simple animal model with additive, dominance, and additive by dominance genetic effects, and repeated observations per animal, i.e.,

$$y_{ij} = \mu + a_i + d_i + (ad)_i + p_i + e_{ij},$$

where μ is the overall mean, a_i is the additive genetic effect of animal i, d_i is the dominance genetic effect of animal i, $(ad)_i$ is the additive by dominance genetic effect of animal i, p_i is the permanent environmental effect for an animal with records, and e_i is the residual effect. Also,

$$Var\begin{pmatrix} \mathbf{a} \\ \mathbf{d} \\ \mathbf{ad} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\sigma^2_{10} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}\sigma^2_{01} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{A}\odot\mathbf{D}\sigma^2_{11} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_p & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma^2_e \end{pmatrix}.$$

143.1 Simulation of Data

Data should be simulated to understand the model and methodology. The desired data structure is given in the following table for four animals.

Animal	Number of Records
1	3
2	2
3	1
4	4

Assume that

$$\begin{aligned} \sigma^2_{10} &= 324, \quad \sigma^2_{01} = 169, \\ \sigma^2_{11} &= 49, \quad \sigma^2_p = 144, \\ \sigma^2_e &= 400. \end{aligned}$$

The additive genetic relationship matrix for the four animals is

$$\begin{aligned} \mathbf{A} &= \mathbf{L}_{10}\mathbf{L}'_{10} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ .5 & .866 & 0 & 0 \\ .25 & 0 & .9682 & 0 \\ .75 & .433 & 0 & .7071 \end{pmatrix} \mathbf{L}'_{10} \end{aligned}$$

$$= \begin{pmatrix} 1 & .5 & .25 & .75 \\ .5 & 1 & .125 & .75 \\ .25 & .125 & 1 & .1875 \\ .75 & .75 & .1875 & 1.25 \end{pmatrix}.$$

The dominance genetic relationship matrix (derived from the genomic relationship matrix) is

$$\begin{aligned} \mathbf{D} &= \mathbf{L}_{01}\mathbf{L}'_{01} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ .25 & .9682 & 0 & 0 \\ .0625 & -.0161 & .9979 & 0 \\ .125 & .0968 & -.00626 & .9874 \end{pmatrix} \mathbf{L}'_{01} \\ &= \begin{pmatrix} 1 & .25 & .0625 & .125 \\ .25 & 1 & 0 & .125 \\ .0625 & 0 & 1 & 0 \\ .125 & .125 & 0 & 1 \end{pmatrix}. \end{aligned}$$

The additive by dominance genetic relationship matrix is the Hadamard product of **A** and **D**, which is the element by element product of matrices.

$$\begin{aligned} \mathbf{A} \odot \mathbf{D} &= \mathbf{L}_{11}\mathbf{L}'_{11} \\ &= \begin{pmatrix} 1 & 0 & 0 & 0 \\ .125 & .9922 & 0 & 0 \\ .015625 & -.00197 & .999876 & 0 \\ .09375 & .08268 & -.0013 & 1.1110 \end{pmatrix} \mathbf{L}'_{11} \\ &= \begin{pmatrix} 1 & .125 & .015625 & .09375 \\ .125 & 1 & 0 & .09375 \\ .015625 & 0 & 1 & 0 \\ .09375 & .09375 & 0 & 1.25 \end{pmatrix}. \end{aligned}$$

The Cholesky decomposition of each of these matrices is necessary to simulate the separate genetic effects. The simulated genetic effects for the four animals are (with $\mathbf{v}_a$, $\mathbf{v}_d$, and $\mathbf{v}_{ad}$ being vectors of random normal deviates)

$$\begin{aligned} \mathbf{a} &= (324)^{.5}\mathbf{L}_{10}\mathbf{v}_a, \\ &= \begin{pmatrix} 12.91 \\ 13.28 \\ -10.15 \\ 38.60 \end{pmatrix}, \\ \mathbf{d} &= (169)^{.5}\mathbf{L}_{01}\mathbf{v}_d, \\ &= \begin{pmatrix} 15.09 \\ 5.32 \\ -17.74 \\ 3.89 \end{pmatrix}, \end{aligned}$$

$$\begin{aligned}(\mathbf{ad}) &= (49)^{.5}\mathbf{L}_{11}\mathbf{v}_{ad}, \\ &= \begin{pmatrix} -12.22 \\ -1.32 \\ -4.30 \\ 5.76 \end{pmatrix}.\end{aligned}$$

In the additive genetic animal model, base population animals were first simulated and then progeny were simulated by averaging the additive genetic values of the parents and adding a random Mendelian sampling effect to obtain the additive genetic values. With non-additive genetic effects, such a simple process does not exist. The appropriate genetic relationship matrices are necessary and these need to be decomposed. The alternative is to determine the number of loci affecting the trait, and to generate genotypes for each animal after defining the loci with dominance genetic effects and those that have additive by dominance interactions. This might be the preferred method depending on the objectives of the study.

Let the permanent environmental effects for the four animals be

$$\mathbf{p} = \begin{pmatrix} 8.16 \\ -8.05 \\ -1.67 \\ 15.12 \end{pmatrix}.$$

The observations on the four animals, after adding a new residual effect for each record, and letting $\mu = 0$, are given in the table below.

Animal	**a**	**d**	**(ad)**	**p**	1	2	3	4
1	12.91	15.09	-12.22	8.16	36.21	45.69	49.41	
2	13.28	5.32	-1.32	-8.05	9.14	-14.10		
3	-10.15	-17.74	-4.30	-1.67	-20.74			
4	38.60	3.89	5.76	15.12	24.13	83.09	64.67	50.13

143.2 HMME

Using the simulated data, the MME that need to be constructed are as follows.

$$\begin{pmatrix} \mathbf{X'X} & \mathbf{X'Z} & \mathbf{X'Z} & \mathbf{X'Z} & \mathbf{X'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z}+\mathbf{A}^{-1}k_{10} & \mathbf{Z'Z} & \mathbf{Z'Z} & \mathbf{Z'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} & \mathbf{Z'Z}+\mathbf{D}^{-1}k_{01} & \mathbf{Z'Z} & \mathbf{Z'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} & \mathbf{Z'Z} & \mathbf{Z'Z}+(\mathbf{A}\odot\mathbf{D})^{-1}k_{11} & \mathbf{Z'Z} \\ \mathbf{Z'X} & \mathbf{Z'Z} & \mathbf{Z'Z} & \mathbf{Z'Z} & \mathbf{Z'Z}+\mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{d}} \\ \hat{\mathbf{ad}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X'y} \\ \mathbf{Z'y} \\ \mathbf{Z'y} \\ \mathbf{Z'y} \\ \mathbf{Z'y} \end{pmatrix},$$

where $k_{10} = 400/324$, $k_{01} = 400/169$, $k_{11} = 400/49$, and $k_p = 400/144$. Thus, the order is 17 for these four animals, with only 10 observations. Note that

$$\mathbf{X'y} = (327.63),$$

and

$$\mathbf{Z}'\mathbf{y} = \left(\begin{array}{r} 131.31 \\ -4.96 \\ -20.74 \\ 222.02 \end{array}\right).$$

The solutions are

$$\hat{\mathbf{a}} = \left(\begin{array}{r} 12.30 \\ 1.79 \\ -8.67 \\ 15.12 \end{array}\right), \quad \hat{\mathbf{d}} = \left(\begin{array}{r} 4.18 \\ -4.86 \\ -6.20 \\ 8.19 \end{array}\right), \quad \hat{\mathbf{ad}} = \left(\begin{array}{r} 1.49 \\ -1.68 \\ -1.87 \\ 3.00 \end{array}\right), \quad \hat{\mathbf{p}} = \left(\begin{array}{r} 4.56 \\ -6.18 \\ -5.57 \\ 7.18 \end{array}\right),$$

and $\hat{\mu} = 17.02$.

The total genetic merit of an animal can be estimated by adding together the solutions for the additive, dominance, and additive by dominance genetic values,

$$\hat{\mathbf{g}} = \left(\begin{array}{r} 17.97 \\ -4.75 \\ -16.73 \\ 26.32 \end{array}\right) = (\hat{\mathbf{a}} + \hat{\mathbf{d}} + \hat{\mathbf{ad}}).$$

On the practical side, the solutions for the individual dominance and additive by dominance solutions should be used in breeding programs, but how? Dominance effects arise due to particular sire-dam matings, and thus, dominance genetic values could be used to determine which matings were better. However, additive by dominance genetic solutions may be less useful. Perhaps the main point is that if non-additive genetic effects are significant, then they should be removed through the model to obtain more accurate estimates of the additive genetic effects, assuming that these have a much larger effect than the non-additive genetic effects.

144 Computing Simplification

Take the MME as shown earlier, i.e.

$$\left(\begin{array}{ccccc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k_{10} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{D}^{-1}k_{01} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + (\mathbf{A} \odot \mathbf{D})^{-1}k_{11} & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}k_p \end{array}\right) \left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \\ \hat{\mathbf{d}} \\ \hat{\mathbf{ad}} \\ \hat{\mathbf{p}} \end{array}\right) = \left(\begin{array}{c} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{array}\right),$$

Now subtract the equation for dominance genetic effects from the equation for additive genetic effects, and similarly for the additive by dominance and permanent environmental effects, giving

$$\begin{aligned}
\mathbf{A}^{-1}k_{10}\hat{\mathbf{a}} - \mathbf{D}^{-1}k_{01}\hat{\mathbf{d}} &= \mathbf{0} \\
\mathbf{A}^{-1}k_{10}\hat{\mathbf{a}} - (\mathbf{A}\odot\mathbf{D})^{-1}k_{11}\hat{\mathbf{ad}} &= \mathbf{0} \\
\mathbf{A}^{-1}k_{10}\hat{\mathbf{a}} - \mathbf{I}^{-1}k_{p}\hat{\mathbf{p}} &= \mathbf{0}
\end{aligned}$$

Re-arranging terms, then

$$\begin{aligned}
\hat{\mathbf{d}} &= \mathbf{DA}^{-1}(k_{10}/k_{01})\hat{\mathbf{a}} \\
\hat{\mathbf{ad}} &= (\mathbf{A}\odot\mathbf{D})\mathbf{A}^{-1}(k_{10}/k_{11})\hat{\mathbf{a}} \\
\hat{\mathbf{p}} &= \mathbf{A}^{-1}(k_{10}/k_{p})\hat{\mathbf{a}}
\end{aligned}$$

The only inverse that is needed is for **A**, and the equations to solve are only as large as the usual animal model MME. The steps in the procedure would be iterative.

1. Adjust the observation vector for solutions to $\hat{\mathbf{d}}$, $\hat{\mathbf{ad}}$, and $\hat{\mathbf{p}}$ (initially these would be zero) as
$$\tilde{\mathbf{y}} = \mathbf{y} - \mathbf{Z}(\hat{\mathbf{d}} + \hat{\mathbf{ad}} + \hat{\mathbf{p}}).$$

2. Solve the following equations:
$$\left(\begin{array}{cc} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z} + \mathbf{A}^{-1}k_{10} \end{array}\right)\left(\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{a}} \end{array}\right) = \left(\begin{array}{c} \mathbf{X}'\tilde{\mathbf{y}} \\ \mathbf{Z}'\tilde{\mathbf{y}} \end{array}\right).$$

3. Obtain solutions for $\hat{\mathbf{d}}$, $\hat{\mathbf{ad}}$, and $\hat{\mathbf{p}}$ using
$$\begin{aligned}
\hat{\mathbf{d}} &= \mathbf{DA}^{-1}(k_{10}/k_{01})\hat{\mathbf{a}} \\
\hat{\mathbf{ad}} &= (\mathbf{A}\odot\mathbf{D})\mathbf{A}^{-1}(k_{10}/k_{11})\hat{\mathbf{a}} \\
\hat{\mathbf{p}} &= \mathbf{A}^{-1}(k_{10}/k_{p})\hat{\mathbf{a}}.
\end{aligned}$$

4. Go to step 1 and begin again until convergence is reached.

145 Estimation of Variances

Given the new computing algorithm, and using Gibbs sampling as a tool the variances can be estimated. Notice from the above formulas that

$$\begin{aligned}
\mathbf{w}_{01} &= (\mathbf{D}^{-1}\hat{\mathbf{d}}) = \mathbf{A}^{-1}(k_{10}/k_{01})\hat{\mathbf{a}} \\
\mathbf{w}_{11} &= ((\mathbf{A}\odot\mathbf{D})^{-1}\hat{\mathbf{ad}}) = \mathbf{A}^{-1}(k_{10}/k_{11})\hat{\mathbf{a}} \\
\mathbf{w}_{p} &= (\mathbf{I}\hat{\mathbf{p}}) = \mathbf{A}^{-1}(k_{10}/k_{p})\hat{\mathbf{a}}.
\end{aligned}$$

Again, the inverses of $\mathbf{D}$ and $(\mathbf{A} \odot \mathbf{D})$ are not needed. The necessary quadratic forms are then

$$\begin{aligned} \hat{\mathbf{d}}'\mathbf{w}_{01} &= \hat{\mathbf{d}}'\mathbf{D}^{-1}\hat{\mathbf{d}}, \\ \hat{\mathbf{ad}}'\mathbf{w}_{11} &= \hat{\mathbf{ad}}'(\mathbf{A} \odot \mathbf{D})^{-1}\hat{\mathbf{ad}}, \\ \hat{\mathbf{p}}'\mathbf{w}_{p} &= \hat{\mathbf{p}}'\hat{\mathbf{p}}, \end{aligned}$$

and $\hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}}$. Generate 4 random Chi-Square variates, C_i, with degrees of freedom equal to the number of animals in $\hat{\mathbf{a}}$, then

$$\begin{aligned} \sigma^2_{10} &= \hat{\mathbf{a}}'\mathbf{A}^{-1}\hat{\mathbf{a}}/C_1 \\ \sigma^2_{01} &= \hat{\mathbf{d}}'\mathbf{w}_{01}/C_2 \\ \sigma^2_{11} &= \hat{\mathbf{ad}}'\mathbf{w}_{11}/C_3 \\ \sigma^2_{p} &= \hat{\mathbf{p}}'\mathbf{w}_{p}/C_4. \end{aligned}$$

The residual variance would be estimated from

$$\sigma^2_e = \hat{\mathbf{e}}'\hat{\mathbf{e}}/C_5,$$

where C_5 is a random Chi-square variate with degrees of freedom equal to the total number of observations. This may not be a totally correct algorithm and some refinement may be necessary, but this should be the basic starting point.

146 EXERCISES

1. Below are data on five animals for two traits.

Animal	F_i	Trait 1	Trait 2
1	0	14.3	34.4
2	.25		38.9
3	.125	15.3	
4	.375	21.1	38.7
5	.344	12.1	45.9

Let the model for trait t be

$$y_{ti} = \mu_t + a_{ti} + d_{ti} + b_t F_i + e_{ti},$$

where μ_t is the overall mean, a_{ti} is the additive genetic merit of animal i for trait t, d_{ti} is the dominance genetic merit of animal i for trait t, F_i is the inbreeding coefficient of animal i, b_t is a regression of the trait on the inbreeding coefficient (inbreeding depression), and e_{ti} is a random residual effect.

$$Var\begin{pmatrix} \mathbf{a} \\ \mathbf{d} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{A}\bigotimes\mathbf{G}_{10} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}\bigotimes\mathbf{G}_{01} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\bigotimes\mathbf{R} \end{pmatrix},$$

where

$$\mathbf{G}_{10} = \begin{pmatrix} 16 & 3 \\ 3 & 33 \end{pmatrix}, \quad \mathbf{G}_{01} = \begin{pmatrix} 9 & 4 \\ 4 & 24 \end{pmatrix},$$

and

$$\mathbf{R} = \begin{pmatrix} 49 & 11 \\ 11 & 64 \end{pmatrix}.$$

The additive and dominance relationships among the five animals are as follows:

$$\mathbf{A} = \frac{1}{32}\begin{pmatrix} 32 & 24 & 20 & 28 & 24 \\ 24 & 40 & 24 & 32 & 28 \\ 20 & 24 & 36 & 22 & 29 \\ 28 & 32 & 22 & 44 & 33 \\ 24 & 28 & 29 & 33 & 43 \end{pmatrix},$$

and

$$\mathbf{D} = \frac{1}{1024}\begin{pmatrix} 1024 & 256 & 256 & 384 & 288 \\ 256 & 1088 & 160 & 480 & 376 \\ 256 & 160 & 1040 & 208 & 396 \\ 384 & 480 & 208 & 1168 & 484 \\ 288 & 376 & 396 & 484 & 1145 \end{pmatrix}.$$

(a) Set up the appropriate MME for this MT model and solve. Also show the variances of prediction error for both additive and dominance effects.

(b) Test is the regression coefficient is significantly different from zero using the general linear hypothesis, for each trait separately.

(c) Show the formulas for the EM REML algorithm for this model to estimate $\mathbf{G}_{10}$, $\mathbf{G}_{01}$, and $\mathbf{R}$.

(d) Perform the calculations of EM REML for one iteration using the results from part i.

(e) Do another multiple trait model without the dominance genetic effects in the model and compare EBVs for the additive genetic merit. Do the animals rank the same?

(f) Do another iteration of EM REML for this second model. Where does the dominance genetic variance go, (into the additive or residual variances)?

2. Assume a model with the following genetic factors,

$$y_{ijk} = Y_i + a_j + d_j + (aa)_j + p_j + e_{ijk}.$$

Let the underlying parameters be

$$\begin{array}{rcl} \sigma_a^2 & = & 64, \\ \sigma_{aa}^2 & = & 9, \\ \sigma_p^2 & = & 36 \end{array} \quad \text{and} \quad \begin{array}{rcl} \sigma_d^2 & = & 25, \\ \sigma_e^2 & = & 169, \end{array}$$

Below are the observations on animals.

Animal	Sire	Dam	Year 1	Year 2	Year 3
1	-	-	102		
2	-	-	66		
3	1	2	90	79	
4	1	2	43	58	
5	3	4		76	63
6	3	4		60	84
7	3	2			59
8	1	4			97

(a) Construct the gametic relationship matrix and from that obtain **A**, the additive genetic relationship matrix and **D**, the dominance genetic relationship matrix.

(b) Construct the matrix for additive by additive genetic effects, **A#A**.

(c) Analyze the data with the above model by setting up the complete MME. Estimate the total genetic effect as $g_j = a_j + d_j + (aa)_j$.

(d) Use the shortcut computing algorithm as given in class.

(e) Try to estimate the variance components of the model.

Effects of Selection

147 Background History

Animal models assume that matings amongst animals have been random. However, livestock animals are continually selected to produce the next generation of better performing individuals. Selection is considered to be any actions taken to change the probability that an animal reproduces. Once a mating occurs, the genotypes are determined by the laws of Mendelian sampling, a random process.

Gail Belonsky and Brian Kennedy (1988) did a simulation study that compared selection of males based on random selection, phenotypic selection, or selection on EBVs from a BLUP animal model in terms of the amount of inbreeding that would be created. The population simulated was 100 females mated to 5 males. Each mating resulted in 2 offspring. Twenty generations of matings were simulated. The results are in the table below.

Average Inbreeding Coefficients after 20 Generations

Selection	$h^2 = 0.1$	$h^2 = 0.3$	$h^2 = 0.6$
Random	.15	.15	.15
Phenotypic	.17	.21	.22
BLUP EBVs	.29	.30	.27

Selection on BLUP EBVs increased the level of inbreeding after 20 generations by almost 2 times compared to random selection of males. However, at a heritability of 0.6, the level of inbreeding was actually less. This is because the amount of weight given to the parent average in the EBV decreases as heritability goes up. This can be seen in the thesis of Longyang Wu (2000). Note that

$$\hat{a} = w_1(\text{Data}) + w_2(\text{ParentAve}) + w_3(\text{ProgenyAve}).$$

Let

$$\begin{aligned} D &= \text{Records} + 2k + 0.5kp \\ k &= \sigma^2_e/\sigma^2_a \\ p &= \text{number of progeny} \\ w_1 &= \text{Records}/D \\ w_2 &= 2k/D \\ w_3 &= 0.5kp/D \end{aligned}$$

The effect of heritability can be seen in the following table.

Weights on Different Components of BLUP EBVs

One record per animal, two progeny per animal.

h^2	k	D	Data, w_1	Parents, w_2	Progeny, w_3
.1	9	28	.036	.643	.321
.3	$2\frac{1}{3}$	8	.125	.583	.292
.6	$\frac{2}{3}$	3	.333	.444	.222

At low heritability, there is a high weight on the parent average, but as heritability increases that weight decreases, and the weight on the animal's own record increases.

If the number of progeny was increased to 10, then w_3 changes to .703 at heritability of .1. Thus, increasing the number of progeny is the best way to increase the accuracy of an EBV.

148 Selection Theory

Consider a bivariate normal distribution for variables, y and x, where

$$E\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \mu_y \\ \mu_x \end{pmatrix},$$

and

$$Var\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{yx} & \sigma_x^2 \end{pmatrix}.$$

The conditional variance of x given y is

$$Var(x \mid y) = \sigma_x^2 - \frac{(\sigma_{yx})^2}{\sigma_y^2}.$$

Assume truncation selection on y (pick the top fraction from a sorted list of y), then

$$\begin{aligned} y_s &= \text{the selected } y \\ E(y_s) &= \mu + i\sigma_y \\ i &= \text{the mean in a N}(0,1) \text{ distribution} \\ & \quad \text{of the top fraction} \\ Var(y_s) &= \sigma_{y_s}^2 \\ &= (1-k)\sigma_y^2 \\ k &= i(i-t) \end{aligned}$$

where t is the trunction point for that fraction on a N(0,1) distribution.

148.1 Variance of x_s

The variance of the selected x values, given that selection was on y is given by

$$Var(x_s) = Var(x \mid y_s) + Var(E(x \mid y_s)),$$

where

$$\begin{aligned}
E(x \mid y_s) &= \mu_x + \frac{\sigma_{yx}}{\sigma^2_y}(y_s - E(y_s)) \\
&= E(x \mid y)
\end{aligned}$$

$$\begin{aligned}
Var(x \mid y_s) &= \sigma^2_x - \frac{(\sigma_{yx})^2}{\sigma^2_y}\frac{\sigma^2_x}{\sigma^2_x} \\
&= (1 - r^2_{yx})\sigma^2_x,
\end{aligned}$$

$$\begin{aligned}
Var(E(x \mid y_s)) &= Var(\frac{\sigma_{yx}}{\sigma^2_y}y_s) \\
&= (\frac{\sigma_{yx}}{\sigma^2_y})^2\sigma^2_{y_s} \\
&= (\frac{\sigma_{yx}}{\sigma^2_y})^2(1-k)\sigma^2_y \\
&= (1-k)\frac{(\sigma_{yx})^2}{\sigma^2_y}\frac{\sigma^2_x}{\sigma^2_x} \\
&= (1-k)r^2_{yx}\sigma^2_x
\end{aligned}$$

Putting the pieces together, then

$$\begin{aligned}
Var(x_s) &= (1 - r^2_{yx})\sigma^2_x + (1-k)r^2_{yx}\sigma^2_x \\
&= [(1 - r^2_{yx}) + (1-k)r^2_{yx}]\sigma^2_x \\
&= (1 - kr^2_{yx})\sigma^2_x
\end{aligned}$$

148.2 Example Simulation

Let

$$Var\begin{pmatrix} y \\ x \end{pmatrix} = \begin{pmatrix} 100 & -20 \\ -20 & 40 \end{pmatrix},$$

which gives a correlation of

$$\begin{aligned}
r_{yx} &= -.31622777 \\
r^2_{yx} &= .1
\end{aligned}$$

One thousand pairs of y and x were generated, sorted from highest to lowest for y, and then the top 100 y were used to form y_s. The intensity of selection is the top 10%, so that $i = 1.755$, $t = 1.2821$, and $k = i(i - t) = .83$. Below is a table of the predicted variances and actual variances (average over 1000 replicates).

Parameter	Expected		Actual
$\sigma^2_{y_s}$	$(1\text{-}.83)\sigma^2_y =$	17.00	16.95
$\sigma_{y_s x_s}$	$(1\text{-}.83)\sigma_{yx} =$	-3.40	-3.26
$\sigma^2_{x_s}$	$(1\text{-}.83(.1))\sigma^2_x =$	36.68	36.59

149 Selection and Animal Models

A simple animal model, one record per animal, is

$$y_i = \mu + a_i + e_i.$$

Selection is on y and now let $x = a$ be the correlated variable. The expected changes in the additive variance would be as derived below.

$$\begin{aligned}
\sigma^2_{a_s} &= \sigma^2_a - k\frac{(\sigma_{ya})^2}{\sigma^2_y} \\
\sigma_{ya} &= \sigma^2_a = h^2\sigma^2_y \\
\sigma^2_{a_s} &= \sigma^2_a - k\frac{(\sigma^2_a)^2}{\sigma^2_y} \\
&= (1 - k\ h^2)\sigma^2_a
\end{aligned}$$

Now change x to be e_i, which is also correlated with y.

$$\begin{aligned}
\sigma^2_{e_s} &= \sigma^2_e - k\frac{(\sigma_{ye})^2}{\sigma^2_y} \\
\sigma_{ye} &= \sigma^2_e = (1 - h^2)\sigma^2_y \\
\sigma^2_{e_s} &= \sigma^2_e - k\frac{(\sigma^2_e)^2}{\sigma^2_y} \\
&= (1 - k\ (1 - h^2))\sigma^2_e
\end{aligned}$$

As an example to illustrate, let

$$\begin{aligned}\sigma^2_y &= 700,\\ \sigma^2_a &= 210,\\ \sigma^2_e &= 490,\end{aligned}$$

so that $h^2 = 0.3$. Assume 50% truncation selection where $i = 0.8$, $t = 0$, and $k = 0.64$. Then

$$\begin{aligned}\sigma^2_{y_s} &= (1 - .64)(700) = 252,\\ \sigma^2_{a_s} &= (1 - .64(.3))(210) = 169.68,\\ \sigma^2_{e_s} &= (1 - .64(1 - .3))(490) = 270.48.\end{aligned}$$

Note that

$$\sigma^2_{y_s} \neq \sigma^2_{a_s} + \sigma^2_{e_s},$$

because selection creates a covariance between selected animal and residual effects. The correct result is

$$\sigma^2_{y_s} = \sigma^2_{a_s} + \sigma^2_{e_s} + 2\sigma_{a_se_s}.$$

Thus,

$$\sigma_{a_se_s} = 0.5(252 \ - \ 169.68 \ - \ 270.48) \ = \ -94.08.$$

150 Mating of the Selected Animals

The animals in y_s are randomly mated to produce progeny generation 1. The variances in the progeny generation 1 are as follows:

$$\begin{aligned}\sigma^2_{e_1} &= \sigma^2_e = 490.\\ \sigma^2_{a_1} &= \frac{1}{2}(\sigma^2_{a_s}) + (\text{Mendelian sampling})\\ &= \frac{1}{2}(\sigma^2_{a_s}) \ + \ \frac{1}{2}(\sigma^2_a)\\ &= \frac{1}{2}(1 - k\ h^2)\sigma^2_a \ + \ \frac{1}{2}\sigma^2_a\\ &= (1 - \frac{1}{2}k\ h^2)\sigma^2_a = (1 - 0.5(.64)(.3))(210) = 189.84,\end{aligned}$$

$$
\begin{aligned}
&= \sigma^2_a - d_1 \\
d_1 &= 20.16,
\end{aligned}
$$

where d_1 is the amount of linkage disequilibrium generated by selection of the parents.

$$
\begin{aligned}
\sigma^2_{y_1} &= \sigma^2_{a_1} + \sigma^2_{e_1} \\
&= 189.84 + 490 \;=\; 679.84, \\
&= (1 - 0.5k\ h^4)\sigma^2_y.
\end{aligned}
$$

Now, allow the progeny generation to mate randomly again, without any selection on generation 1 animals. Generation 2 animals would then have the following variances:

$$
\begin{aligned}
\sigma^2_{e_2} &= \sigma^2_e = 490. \\
\sigma^2_{a_2} &= \frac{1}{2}(\sigma^2_{a_1}) + (\text{Mendelian sampling}) \\
&= \frac{1}{2}(\sigma^2_{a_1}) \;+\; \frac{1}{2}(\sigma^2_a) \\
&= (1 - \frac{1}{4}k\ h^2)\sigma^2_a = (1 - 0.25(.64)(.3))(210) = 199.92, \\
&= \sigma^2_a - 0.5\ d_1 \\
\sigma^2_{y_2} &= 689.92.
\end{aligned}
$$

Continued generations of random mating would re-generate the lost genetic variance due to the original selection. Each generation of random mating reduces the original disequilibrium by one half.

If the population was small in size, then Mendelian sampling variance would decrease due to rising inbreeding levels.

151 Another Cycle of Selection

Instead of randomly mating individuals from generation 1, suppose that another selection of 50% was imposed. The heritability of the trait in generation 1 is

$$h^2_1 = \sigma^2_{a_1}/\sigma^2_{y_1} = 189.84/679.84 \;=\; 0.27924.$$

Then the new variances in the selected animals would be

$$\begin{aligned}\sigma^2_{a_{1s}} &= (1-h^2_1\ k)\sigma^2_{a_1} \\ &= 0.821285(189.84) \\ &= 155.91,\end{aligned}$$

$$\begin{aligned}\sigma^2_{a_2} &= 0.5\sigma^2_{a_{1s}} + 0.5\sigma^2_a \\ &= 0.5(155.91) + 0.4(210) \\ &= 182.955, \\ &= \sigma^2_a - 0.5d_1 - 0.5h^2_1k\sigma^2_{a_1} \\ &= \sigma^2_a - d_2,\end{aligned}$$

$$\begin{aligned}d_2 &= 0.5d_1 + 0.5h^2_1k\sigma^2_{a_1} \\ &= 0.5(20.16) + 16.96 \ = \ 27.04.\end{aligned}$$

Continuing to make further cycles of selection, then the additive genetic variance at time t would be

$$\sigma^2_{a_t} = \sigma^2_a \ - \ d_t,$$

where

$$d_{t+1} \ = \ 0.5\ d_t + 0.5h^2_tk\sigma^2_{a_t},$$

or the disequilibrium at time $t+1$ is half the disequilibrium at time t plus some new disequilibrium. If we equate d_{t+1} to d_t and call it d, then

$$\begin{aligned}d &= 0.5\ d + 0.5h^2_tk\sigma^2_{a_t} \\ &= h^2_tk\sigma^2_{a_t},\end{aligned}$$

as t goes to infinity. Thus, the amount of new disequilibrium equals one half the old disequilibrium, so that the loss due to selection is counteracted equally by the gain in genetic variability due to Mendelian sampling. This is known as the **Bulmer effect, 1970**. Of course, this ignores small population size and the effects of inbreeding.

The Bulmer effect was demonstrated by a simulation experiment by Kennedy. The percentage selected was the top 20%, where $k = .7818$ with an initial phenotypic variance of 100 and an initial heritability of 0.5.

Generation	$\sigma^2_{y_t}$	h^2_t	d_t	$R = ih^2_t\sigma_{y_t}$
0	100.0	.500	0.0	0.00
1	90.2	.446	-9.8	7.00
2	88.1	.432	-11.9	5.93
3	87.6	.429	-12.4	5.68
4	87.5	.428	-12.5	5.63
inf	87.5	.428	-12.5	5.61

The total amount of genetic change in 4 generations would be

$$R = 7.00 \ + \ 5.93 \ + \ 5.68 \ + \ 5.63 \ = \ 24.24.$$

Normally, (before Bulmer's effect was shown) the total response in four generations would have been predicted to be

$$R = 4 \ \times \ 7.00 \ = \ 28.00,$$

but the actual response would have been 15.5% less.

152 Summary of Selection Effects

Non-random mating results in the following consequences.

1. Causes changes to gene frequencies at all loci.
2. Changes in gene frequencies cause a change in genetic variance. Recall from quantitative genetics that

$$\sigma^2_G = 2pq[a + d(q - p)]^2 + [2pqd]^2.$$

3. In finite populations, non-random mating causes a reduction in the effective population size, which subsequently causes an increase in levels of inbreeding.
4. Joint equilibrium becomes joint disequilibrium, and therefore, non-zero covariances between additive and dominance genetic effects are created.
5. If the pedigrees of animals are not complete nor traceable to the base generation, then non-random mating causes genetic evaluations by BLUP to be biased, and causes estimates of genetic variances, by any method, to be biased.
6. Because the genetic variance decreases due to joint disequilibrium and inbreeding, response to selection is generally lower than expected by selection index over several generations.

153 Effects of Selection on Genetic Evaluation

The effects of non-random mating on genetic evaluation are minimal **IF**

- Complete (no missing parent information) pedigrees are known back to a common base population which was mating randomly,

- Data on all candidates for selection are available, and
- Genetic parameters from the base population are known.

If the above conditions hold, then application of BLUP does not lead to bias in EBVs, but selection increases the variance of prediction error over populations that are randomly mating. However, in animal breeding, the practical situation is that complete pedigrees seldom exist. Thus, bias can creep into estimates of fixed effects and EBVs.

Recall that HMME for a simple animal model are

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} & \mathbf{0} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{A}^{nn}k_a & \mathbf{A}^{no}k_a \\ \mathbf{0} & \mathbf{A}^{on}k_a & \mathbf{A}^{oo}k_a \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{a}}_n \\ \hat{\mathbf{a}}_o \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{0} \end{pmatrix},$$

where $k_a = \sigma_a^{-2}$. A generalized inverse of the coefficient matrix can be represented as

$$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xn} & \mathbf{C}_{xo} \\ \mathbf{C}_{nx} & \mathbf{C}_{nn} & \mathbf{C}_{no} \\ \mathbf{C}_{ox} & \mathbf{C}_{on} & \mathbf{C}_{oo} \end{pmatrix}.$$

Then remember that

$$Var\begin{pmatrix} \hat{\mathbf{a}}_n - \mathbf{a}_n \\ \hat{\mathbf{a}}_o - \mathbf{a}_o \end{pmatrix} = \begin{pmatrix} \mathbf{C}_{nn} & \mathbf{C}_{no} \\ \mathbf{C}_{on} & \mathbf{C}_{oo} \end{pmatrix},$$

and that

$$\begin{aligned} Cov(\hat{\mathbf{b}}, \hat{\mathbf{a}}_n) &= \mathbf{0}, \\ Cov(\hat{\mathbf{b}}, \mathbf{a}_n) &= -\mathbf{C}_{xn}. \end{aligned}$$

These results indicate that HMME forces the covariance between estimates of the fixed effects and estimates of additive genetic effects to be null. However, there is a non-zero covariance between estimates of the fixed effects and the true additive genetic values of animals. Hence, any problem with the true additive genetic values will cause problems with estimates of fixed effects.

Consider the equation for $\hat{\mathbf{b}}$,

$$\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{y} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}\hat{\mathbf{a}}_n),$$

and the expectation of this vector is

$$E(\hat{\mathbf{b}}) = (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}(\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\mathbf{b} - \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z}E(\hat{\mathbf{a}}_n)).$$

The fixed effects solution vector contains a function of the expectation of the additive genetic solution vector. Normally, because the BLUP methodology requires

$$E(\hat{\mathbf{a}}_n) = E(\mathbf{a}_n) = \mathbf{0},$$

then the fixed effects solution vector is also unbiased.

Selection, however, can cause a change in expectations where

$$E(\mathbf{a}_n) \neq \mathbf{0},$$

and therefore, the expectation of the fixed effects solution vector contains a function of $E(\mathbf{a}_n)$ and is consequently biased. If $\hat{\mathbf{b}}$ is biased, then this will cause a bias in $\hat{\mathbf{a}}$.

153.1 Derivation of A Different Method

Re-state the model (in general terms) as

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

where

$$E\begin{pmatrix} \mathbf{u} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{u} \\ \mathbf{0} \end{pmatrix},$$

and therefore,

$$E(\mathbf{y}) = \mathbf{Xb} + \mathbf{Zu}.$$

To simplify, assume that $\mathbf{G} = Var(\mathbf{u})$ and $\mathbf{R} = Var(\mathbf{e})$ and that neither is drastically affected by non-random mating.

The prediction problem is the same as before. Predict a function of $\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}$ by a linear function of the observation vector, $\mathbf{L}'\mathbf{y}$, such that

$$E(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u}) = E(\mathbf{L}'\mathbf{y}),$$

and such that $Var(\mathbf{K}'\mathbf{b} + \mathbf{M}'\mathbf{u} - \mathbf{L}'\mathbf{y})$ is minimized. Form the variance of prediction errors and add a LaGrange multiplier to ensure the unbiasedness condition, then differentiate with respect to the unknown $\mathbf{L}$ and the matrix of LaGrange multipliers and equate to zero. The solution gives the following equations.

$$\begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{V}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{V}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{V}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{V}^{-1}\mathbf{y} \end{pmatrix}.$$

Because $\mathbf{V} = \mathbf{ZGZ}' + \mathbf{R}$, and

$$\mathbf{V}^{-1} = \mathbf{R}^{-1} - \mathbf{R}^{-1}\mathbf{ZTZ}'\mathbf{R}^{-1},$$

for $\mathbf{T} = (\mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1})^{-1}$, then the following equations give the exact same solutions as the previous equations.

$$\begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}'\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}'\mathbf{R}^{-1}\mathbf{Z} \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}'\mathbf{R}^{-1}\mathbf{y} \end{pmatrix}.$$

If a generalized inverse to the above coefficient matrix is represented as

$$\begin{pmatrix} \mathbf{C}_{xx} & \mathbf{C}_{xz} \\ \mathbf{C}_{zx} & \mathbf{C}_{zz} \end{pmatrix},$$

then some properties of these equations are

$$\begin{aligned} Cov(\hat{\mathbf{b}}, \mathbf{u}) &= \mathbf{0}, \\ E(\hat{\mathbf{b}}) &= (\mathbf{X}'\mathbf{R}^{-1}\mathbf{X})^{-}\mathbf{X}'\mathbf{R}^{-1}\mathbf{X}\mathbf{b}, \\ Cov(\hat{\mathbf{b}}, \hat{\mathbf{u}}) &= \mathbf{C}_{xz}. \end{aligned}$$

Firstly, these results suggest that if non-random mating has occurred and has changed the expectation of the random vector, then an appropriate set of equations is the generalized least squares equations. However, GLS equations give a lower correlation with true values and large mean squared errors (when matings are at random) compared to BLUP and HMME. Secondly, the estimates of the fixed effects have null covariances with the true random effects, and the covariances between estimates of the fixed effects and estimates of the random effects are non-zero, which is **opposite** to the results from BLUP. With the least squares solutions, application of the regressed least squares procedure could be subsequently used to give EBVs.

There is another problem with these equations. If $\mathbf{u} = \mathbf{a}$ as in an animal model, then $\mathbf{Z} = \mathbf{I}$, and the GLS equations do not have a solution unless $\hat{\mathbf{a}} = \mathbf{0}$. This is not very useful for genetic evaluation purposes.

153.2 An Alternative Model

The Mendelian sampling variance was assumed to be unaffected by non-random mating, but could be reduced by the accumulation of inbreeding. The animal model equation is

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{a} + \mathbf{Z}\mathbf{p} + \mathbf{e}.$$

The animal additive genetic effect can be written as

$$\mathbf{a} = \mathbf{T}_s\mathbf{s} + \mathbf{T}_d\mathbf{d} + \mathbf{m},$$

where $\mathbf{T}_s$ and $\mathbf{T}_d$ are matrices of ones and zeros, such that each row has an element that is 1 and all others are 0, and these indicate the sire and dam of the animal, respectively, and $\mathbf{m}$ is the Mendelian sampling effect. Due to non-random mating then,

$$E(\mathbf{a}) = \mathbf{T}_s\mathbf{s} + \mathbf{T}_d\mathbf{d},$$

which is not a null vector, in general. Let

$$\begin{aligned} \mathbf{Z}_s &= \mathbf{Z}\mathbf{T}_s, \\ \mathbf{Z}_d &= \mathbf{Z}\mathbf{T}_d, \end{aligned}$$

then the model becomes

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_s\mathbf{s} + \mathbf{Z}_d\mathbf{d} + \mathbf{Zm} + \mathbf{Zp} + \mathbf{e}.$$

Also,

$$\begin{aligned} E(\mathbf{y}) &= \mathbf{Xb} + \mathbf{Z}_s\mathbf{s} + \mathbf{Z}_d\mathbf{d}, \\ E(\mathbf{m}) &= \mathbf{0}, \\ E(\mathbf{p}) &= \mathbf{0}, \\ E(\mathbf{e}) &= \mathbf{0}, \end{aligned}$$

and

$$Var\begin{pmatrix} \mathbf{m} \\ \mathbf{p} \\ \mathbf{e} \end{pmatrix} = \begin{pmatrix} \mathbf{B}\sigma_a^2 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}\sigma_p^2 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}\sigma_e^2 \end{pmatrix},$$

where $\mathbf{B}$ is from

$$\mathbf{A} = \mathbf{TBT}'.$$

If all animals were non inbred then all of the diagonals of $\mathbf{B}$ would be equal to .5.

Note that the matrix $\mathbf{A}$ or its inverse are not necessary in this model, and that sires and dams (resulting from selection) are fixed effects in this model. The equations to solve are

$$\begin{pmatrix} \mathbf{X}'\mathbf{X} & \mathbf{X}'\mathbf{Z}_s & \mathbf{X}'\mathbf{Z}_d & \mathbf{X}'\mathbf{Z} & \mathbf{X}'\mathbf{Z} \\ \mathbf{Z}_s'\mathbf{X} & \mathbf{Z}_s'\mathbf{Z}_s & \mathbf{Z}_s'\mathbf{Z}_d & \mathbf{Z}_s'\mathbf{Z} & \mathbf{Z}_s'\mathbf{Z} \\ \mathbf{Z}_d'\mathbf{X} & \mathbf{Z}_d'\mathbf{Z}_s & \mathbf{Z}_d'\mathbf{Z}_d & \mathbf{Z}_d'\mathbf{Z} & \mathbf{Z}_d'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}_s & \mathbf{Z}'\mathbf{Z}_d & \mathbf{Z}'\mathbf{Z} + \mathbf{B}^{-1}k_a & \mathbf{Z}'\mathbf{Z} \\ \mathbf{Z}'\mathbf{X} & \mathbf{Z}'\mathbf{Z}_s & \mathbf{Z}'\mathbf{Z}_d & \mathbf{Z}'\mathbf{Z} & \mathbf{Z}'\mathbf{Z} + \mathbf{I}k_p \end{pmatrix} \begin{pmatrix} \hat{\mathbf{b}} \\ \hat{\mathbf{s}} \\ \hat{\mathbf{d}} \\ \hat{\mathbf{m}} \\ \hat{\mathbf{p}} \end{pmatrix} = \begin{pmatrix} \mathbf{X}'\mathbf{y} \\ \mathbf{Z}_s'\mathbf{y} \\ \mathbf{Z}_d'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \\ \mathbf{Z}'\mathbf{y} \end{pmatrix}.$$

Thus, for each animal with a record, both parents must be known as well as the inbreeding coefficients of all animals.

This model was applied to data on 12 animals with records, four sires and four dams. The solutions for the sires and dams are shown below, after forcing their sum to be zero.

$$\begin{aligned} \hat{\mu} &= 49.558333, \\ \hat{s}_1 &= -7.079167, \\ \hat{s}_3 &= 7.2875, \\ \hat{s}_5 &= -16.14583, \\ \hat{s}_7 &= 15.9375, \\ \hat{d}_2 &= 1.7375, \\ \hat{d}_4 &= 0.0875, \\ \hat{d}_6 &= 3.9375, \\ \hat{d}_8 &= -5.7625, \end{aligned}$$

The solutions for sires and dams represent estimated transmitting abilities and should be multiplied by 2 to give EBV. The estimates of the Mendelian sampling effects for animals 5 through 16 were

$$\hat{\mathbf{m}} = \left(\begin{array}{r} -1.254878 \\ -1.763415 \\ 1.254878 \\ 1.7414634 \\ -0.296341 \\ 1.4634146 \\ 0.2743902 \\ 0.7829268 \\ -0.486585 \\ 0.5085366 \\ 0.0219512 \\ -2.246341 \end{array} \right).$$

The general property of these solutions would be $\mathbf{1}'\mathbf{B}^{-1}\hat{\mathbf{m}} = 0$. In this example all of the diagonal elements of $\mathbf{B}^{-1}$ were equal to 2, but with inbred individuals this would not be the case. The $\hat{\mathbf{m}}$ sum to zero in this example.

EBV are created by summing sire and dam solutions with the Mendelian sampling estimates. The results for animals 5 through 16 were

$$EBV = \left(\begin{array}{r} -6.60 \\ 5.61 \\ 10.28 \\ -5.25 \\ -12.50 \\ 11.64 \\ -21.63 \\ 20.66 \\ -3.63 \\ 2.03 \\ -16.04 \\ 7.93 \end{array} \right).$$

Animals 5, 6, 7, and 8 have two sets of solutions. For example, animal 5 has $\hat{s}_5 = -16.14$ and $EBV = -6.60$. The former solution is based on the progeny of animal 5 only, and the EBV is based upon its own record plus information from its parents. If $\hat{s}_i$ is based upon many progeny, then it could be more accurate than the EBV based on just one record and parent information. Therefore, as in HMME a combination of progeny, own record, and parents should be possible, and maybe more accurate than either piece alone.

Let $w_1 = 0.5q_ik_a$, and $w_2 = (n_i + 2k_a)$ where q_i is the number of progeny for animal i and n_i is the number of records on animal i, then a combined EBV on a cow could be calculated as

$$cEBV = (w_1 \hat{d}_i + w_2 \hat{a}_i)/(w_1 + w_2).$$

Using animal 5 as an example, $q_5 = 3$ and $n_5 = 1$, then

$$\begin{aligned} w_1 &= 0.5(3)(64/36) = 2.6667, \\ w_2 &= (1 + 2(64/36)) = 4.5556. \end{aligned}$$

Then

$$cEBV = (2.6667(-16.14) + 4.5556(-6.60))/(2.6667 + 4.5556) \; = \; -10.12.$$

Non-random mating is taken into account because the sire and dam of each animal with a record is included in the model. The solutions for sires and dams from this model are valid estimates of transmitting abilities provided that the progeny are a random sample of their progeny. The Mendelian sampling estimates for animals provides a means of estimating the additive genetic variance. Inbreeding is still accounted for in the matrix **B**. This model also avoids the problem of forming phantom parent groups for animals with missing parent information. If an animal with a record has an unknown dam (or sire), then a phantom dam (sire) can be created which has this animal as its only progeny. If both parents are unknown, then both a phantom sire and phantom dam need to be assumed, with this animal as their only progeny. Further study of this model is warranted.