5. Best Linear Unbiased Prediction

Julius van der Werf

Lecture 1: Best linear unbiased prediction

Learning objectives

On completion of Lecture 1 you should be able to:

- Understand the principle of mixed models
- Understand how in BLUP, different effects can be taken into account
- Understand how BLUP uses additive genetic relationships

Know that BLUP uses optimal selection index weights for the different sources of information to estimate the breeding value of an animal

Key terms and concepts

Fixed and random effects, Mixed model equations, Numerator relationship matrix.

Introduction to lecture 1

Before going to the full mixed model we take one intermediate step and consider a random model first, with observations only affected by animals' breeding values. This will show how random effects are estimated in linear models.

In contrast to fixed effects, which are estimated as (differences between) corrected means, random effects are somewhat regressed towards the mean, the same principle as when we estimate breeding values and assign only a part of a phenotypic difference toward the breeding value.

The other aspect about random effects is that they can be correlated to each other, e.g. two breeding values are correlated if the animals have an additive genetic relationship. This correlation can be taken into account using a matrix with additive genetic relationships between animals. We will show that using such a matrix results in using information from relatives in the estimation procedure.

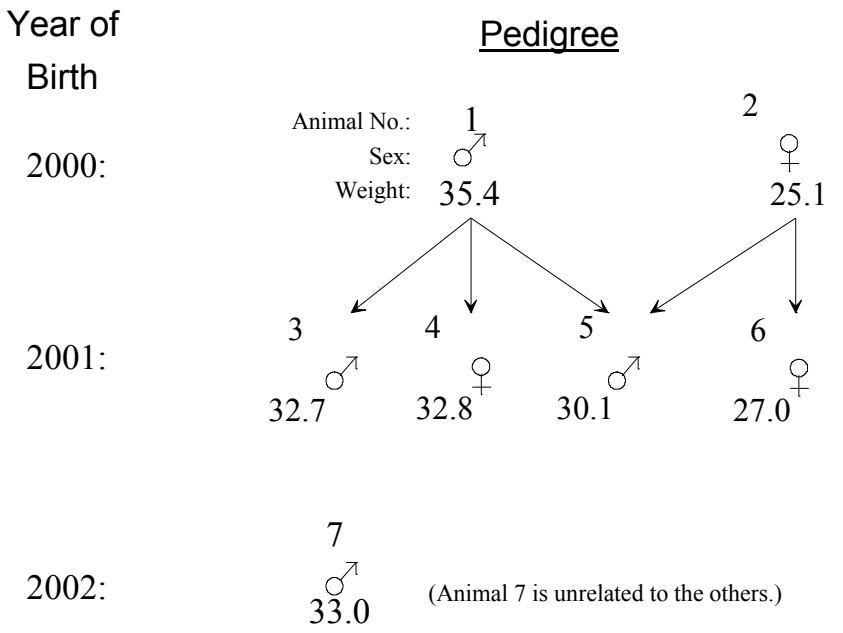
In the second part of this lecture, we present a mixed model, where random and fixed effects are jointly fitted and estimated. The solutions of the mixed model are BLUP EBV for individual animals, and an example will demonstrate that these solutions make sense based on what we have learned so far.

5.1 Example of a random model

We will fit a random model with observations only affected by animals' breeding values. A model with only random effects is strictly not possible as observations are always affected by at least one fixed effect (the mean), but we will consider observations as deviations from the mean (the mean is assumed known).

We will use the same example as in the previous topic, but additionally, we also consider a pedigree structure among the animals, as some of them are genetically related (Figure 5.1).

Figure 5.1 Random and fixed effects shown in a pedigree structure.
Source: van der Werf, (2006).



Using the same example data set as in the previous topic:

$$\begin{pmatrix} 35.4 \\ 25.1 \\ 32.7 \\ 32.8 \\ 30.1 \\ 27.0 \\ 33.0 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{pmatrix}$$

y = Z u + e

7 x 1 = 7 x 7 7 x 1 + 7 x 1

The Z matrix contains elements which relate the 7 observations (in the rows) to the 7 breeding values (in the columns). The Z matrix is a design matrix, like the X matrix in the fixed model. In this example, each animal in u has exactly one observation (no repeat measures or missing data), and Z is simply an identity matrix. If an animal would have no observations, we would have a column with zero's only (the animal would still be included in the model, because it may be genetically related to other animals, see later). Animals with more observations would have more "1" values in their column.

The need to shrink

Now, how do we find solutions for breeding values in this random model?
If we treated this as a fixed model and used

$\hat{u} = (Z'Z)^{-1}Z'y$ then we would have $\hat{u} = Y$

... as $(Z'Z)^{-1}Z'$ equals the identity matrix. **This is obviously wrong.** As the model does not fit the mean we should express data as deviations from the average:

$\hat{u}$ = $(Z'Z)^{-1}Z'(y-\bar{y})$ then we would have $\hat{u}$ = $y-\bar{y}$

Which is equivalent to $\hat{A}$ = P (P as a deviation) **which we know is still wrong!** P contains effects due to genes and to environment. We penalise for likely 'luck' in the environment by regressing by h^2, to give $\hat{A}$ = h^2P.

So: We cannot use $\hat{u}$ = $(Z'Z)^{-1}Z'(y-\bar{y})$ because it does not regress or *shrink* the observations to account for luck. Here is a solution:

The 'wrong' single-animal version is: $\hat{A} = \frac{V_A}{V_A} P$

(which we can see again is wrong as the correct EBV should be $\hat{A}$ = h^2P

Correct this to regress properly: $\hat{A} = \frac{V_A}{V_A + V_E} P$ = h^2P

Divide top and bottom by V_A: $\hat{A} = \frac{1}{1+\frac{V_E}{V_A}} P$

And in a linear model this is: $\hat{u}$ = $(Z'Z + \frac{V_E}{V_A} . I)^{-1} Z'(y - \bar{y})$

Or: $\hat{u}$ = $(Z'Z + \lambda.I)^{-1} Z'(y - \bar{y})$

where I is an identity matrix. Therefore, if we add the variance ratio $\lambda = \frac{V_E}{V_A}$ to the diagonals of Z'Z we achieve that each animal effect is estimated by regressing its deviation towards the mean. Note that λ is smaller for higher values of heritability, therefore the regression will be stronger for lower values of heritability.

To account for relationships a matrix with additive genetic relationships among animals is used. We add the inverse of this matrix, multiplied by λ, to Z'Z. An informal derivation is give below, and was first presented by Dr. C.R. Henderson in the early 1960's (Henderson, 1973). The random model equations are then

Or: $\hat{u}$ = $(Z'Z + \lambda A^{-1})^{-1} Z'(y - \bar{y})$

Where A^{-1} is the inverse numerator relationships matrix (also often called NRM). The term nominator refers to the definition of this relationship. The additive genetic relationship between animals (incl. with themselves) is between 0 and 1 with no inbreeding, and if inbreeding is accounted for, can be between 0 and 2.

Informal derivation of BLUP equations, and proof of equivalence between selection index and BLUP

$\hat{u}$ = *cov(u,y). var(y)$^{-1}$ y* *selection index*

var(u) = G
var(e) = R
var(y) = var(Zu+e) = ZGZ' + R
cov(u,y)= ZG

$\hat{u}$ = *G Z' (Z G Z' + R)$^{-1}$ y*

dimensions: *a.1 = a.a a.o (o.a a.a a.o o.o) o1*
(***o***bservations and ***a***nimals)

Note that $R = I.V_E$.

"Divide" by *G* and rearranging gives:
(note that this is not a formal derivation, there are some missing steps, but it gives the general idea)

$$\hat{u} = (Z'Z + G^{-1} I.V_E)^{-1} Z' y$$

Note that $G = A.V_A$, and $G^{-1} = A^{-1}. 1/V_A$

$$\hat{u} = (Z'Z + A^{-1} (\frac{V_E}{V_A}))^{-1} Z' y \qquad BLP$$

(BLP = Best Linear Prediction, and stands for BLUP without fixed effects)

The additive genetic relationships matrix

We want to estimate breeding values as $\hat{u}$ = $(Z'Z + \lambda A^{-1})^{-1} Z'(y - \bar{y})$, hence we need the inverse of the Numerator Relationships Matrix (A).

First we can calculate A, the Numerator Relationship Matrix for the example. Therefore we can calculate the coefficient of relationship for each pair of animals, then put the results in A. For example, the coefficient of relationship between parent and offspring (e.g. Animals 1 and 3) is 1/2. So $a_{1,3}$ equals 1/2, and so forth to give:

$$A = \begin{matrix} 1 & 0 & ½ & ½ & ½ & 0 & 0 \\ 0 & 1 & 0 & 0 & ½ & ½ & 0 \\ ½ & 0 & 1 & ¼ & ¼ & 0 & 0 \\ ½ & 0 & ¼ & 1 & ¼ & 0 & 0 \\ ½ & ½ & ¼ & ¼ & 1 & ¼ & 0 \\ 0 & ½ & 0 & 0 & ¼ & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{matrix}$$

Note that A is symmetrical: $a_{3,1}$ also equals 1/2. This way of building A is correct as long as no animal is inbred, but a simple correction is available to correct for inbreeding.

For large data sets, inverting A can be a real problem.

This problem does not exist for inverting $(Z'Z + \lambda.A^{-1})$, as some computing tricks have been eq \o(u,\s\up6(^)) iteratively, without making the inversion.

Fortunately, Henderson has shown how to 'build' A^{-1} directly. This saves a lot of computing time for doing the inversion (see Table 5.1 and example for more detail).

$$A^{-1} = \begin{pmatrix} 13/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1/2 & 11/6 & 0 & 0 & -1 & -2/3 & 0 \\ -2/3 & 0 & 4/3 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 0 & 4/3 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & -2/3 & 0 & 0 & 0 & 4/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

The last matrix is the final A^{-1}. Inverting it (on a computer) yields A as originally given!

Table 5.1 Summary of basic rules for creating the inverse of the relationships matrix. These rules assume no inbreeding, but they can be modified to account for it.
Source: van der Werf, (2006).

	For each animal which is to have an estimate of u, add to A^{-1}		
	Both parents known	One parent known	Neither parent known
Animal's Own diagonal	2	4/3	1
Parent x animal	-1	-2/3	
Parents' diagonals	1/2	1/3	
Parent x parent	1/2		

Example of building A^{-1}

Animals 1 and 2 have no known parents so our first 2 steps in building A are:

Step 1: add a "1" to the (1,1) element of A^{-1}
Step 2: add a "1" to the (2,2) element of A^{-1}

Animals 3 and 4 have animal 1 as a sire, but dams are unknown:

Step 3a: add a "4/3" to the (3,3) element of A^{-1}
Step 3b: add a "1/3" to the (1,1) element of A^{-1}
Step 3c: add a "-2/3" to the (1,3) and (3,1) elements of A^{-1}
Step 4a: add a "4/3" to the (4,4) element of A^{-1}
Step 4b: add a "1/3" to the (1,1) element of A^{-1}
Step 4c: add a "-2/3" to the (1,4) and (4,1) elements of A^{-1}

$$\begin{pmatrix} 5/3 & 0 & -2/3 & -2/3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 4/3 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 0 & 4/3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \quad \begin{pmatrix} 13/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1/2 & 11/6 & 0 & 0 & -1 & -2/3 & 0 \\ -2/3 & 0 & 4/3 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 0 & 4/3 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & -2/3 & 0 & 0 & 0 & 4/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

A^{-1} after step 4 A^{-1} *after step 7*

Animal 5 has both parents known: 1 and 2

Step 5a: add a "2" to the (5,5) element of A^{-1}
Step 5b: add a "1/2" to the (1,1) element of A^{-1}
add a "1/2" to the (2,2) element of A^{-1}
add a "1/2" to the (2,1) and (1,2) elements of A^{-1}
Step 5c: add a "-1" to the (1,5) and (5,1) elements of A^{-1}
add a "-1" to the (2,5) and (5,2) elements of A^{-1}

Animal 6 has one dam

Step 6a: add a "4/3" to the (6,6) element of A^{-1}
Step 6b: add a "1/3" to the (2,2) element of A^{-1}
Step 6c: add a "-2/3" to the (2,6) and (6,2) elements of A^{-1}

Animal 7 is unrelated

Step 7 add a "1" to position (7,1)

Solutions to the random model

What we want is $\hat{u} = (Z'Z + \lambda A^{-1})^{-1} Z'(y - \bar{y})$

Note that as Z = I in our example (since all animals have 1 record only), Z'Z = I and Z'Y = Y.

Also assuming $h^2 = 0.5$ then $\lambda = \frac{V_E}{V_A} = \frac{(1-h^2)V_P}{h^2 V_P} = \frac{(1-h^2)}{h^2} = 1$

$\hat{u}$ = (Z'Z + λA^{-1})$^{-1}$ Z'(y - $\bar{y}$)

and since in our case Z is the identity matrix I

$\hat{u}$ = (I + λA^{-1})$^{-1}$ (y - $\bar{y}$)

Note: $\bar{y}$ = 30.87

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \left(\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} + \begin{pmatrix} 13/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0 \\ 1/2 & 11/6 & 0 & 0 & -1 & -2/3 & 0 \\ -2/3 & 0 & 4/3 & 0 & 0 & 0 & 0 \\ -2/3 & 0 & 0 & 4/3 & 0 & 0 & 0 \\ -1 & -1 & 0 & 0 & 2 & 0 & 0 \\ 0 & -2/3 & 0 & 0 & 0 & 4/3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \right)^{-1} \begin{pmatrix} 4.53 \\ -5.77 \\ 1.83 \\ 1.93 \\ -.77 \\ -3.87 \\ 2.13 \end{pmatrix}$$

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \begin{pmatrix} .410 & 0.030 & .117 & .117 & .127 & -.008 & 0 \\ -.030 & .435 & -.008 & -.008 & .135 & .124 & 0 \\ .117 & -.008 & .462 & .033 & .036 & -.002 & 0 \\ .117 & -.008 & .033 & .462 & .036 & -.002 & 0 \\ .127 & .135 & .036 & .036 & .421 & .039 & 0 \\ -.008 & .124 & -.002 & -.002 & .039 & .464 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & .5 \end{pmatrix} \begin{pmatrix} 4.53 \\ -5.77 \\ 1.83 \\ 1.93 \\ -.77 \\ -3.87 \\ 2.13 \end{pmatrix}$$

The solution is BLP, as we have ignored fixed effects and just taken deviations from a general mean. BLP is the same as the classical selection index, except that there is a custom set of index weights for each candidate animal whose breeding value is to be estimated.

The result is:

$$\begin{pmatrix} \hat{u}_1 \\ \hat{u}_2 \\ \hat{u}_3 \\ \hat{u}_4 \\ \hat{u}_5 \\ \hat{u}_6 \\ \hat{u}_7 \end{pmatrix} = \begin{pmatrix} 2.40 \\ -3.26 \\ 1.47 \\ 1.53 \\ -.54 \\ -2.59 \\ 10.6 \end{pmatrix}$$

Some comments to the solutions

Refer to the coefficient matrix on the last page, we could call these the index weights. The weights for $\hat{u}_i$ are in the i[th] row.

1. Look at animal No. 7 in the example data set, and ignore any effects of year. With no relatives to help him, his EBV is quite simply:

$EBV_7 = h^2(y - \bar{y}) = 0.5(33.0 - 30.87) = 1.064$.

There are no off-diagonals in the NRM for animal 7. You can see that the 'custom index weights' for animal 7 involve no use of information from other animals, as expected.

2. Note also that the weights for all other animals make no use of information from animal 7, again as expected.

3. The diagonals of the index weight matrix have high values - as h^2 is high (0.5) animals gain most from their own phenotypes. Note that there are diminishing returns (lower diagonal values) from own phenotype as more information from relatives is available.

4. Animal 1 leans on its three offspring.

5. For animal 1, there is a negative weight on animal 2's phenotype. This makes sense: E.g. If animal 2 is very good indeed, then any superiority in animal 5 is likely to be due to animal 2 rather than animal 1. So, for a given phenotype of animal 5, the better animal 2 is, the lower animal 1's breeding value is likely to be.

5.2 Example of a mixed model

The mixed model

Overview

Fixed Model:	y = Xb + e	$\hat{b} = (X'X)^{-1} X'y$
Random Model:	y = Zu + e	$\hat{u} = (Z'Z + \lambda A^{-1})^{-1} Z'y$
	- where A = Relationships Matrix and $\lambda = \frac{V_E}{V_A}$	
	We say that $\hat{u}$ is BLP of u. u is the vector of TBVs and $\hat{u}$ is the vector of EBVs	
Mixed Model:	y = Xb + Zu + e = fixed effects + breeding values + residual	
Mixed Model Equations:	$\begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$	
	We say that $\hat{u}$ is BLUP of u.	

In this section we combine the two models in the two previous sections: the fixed model and the random model. In mixed models, the breeding values (the random effects) are estimated using the principle of regression, using information from all possible relatives, and correcting them for one, or possibly more fixed effects. We will use again the same example as before, estimating only one fixed effect (the year effect).

The mixed model is a mixture of a fixed and a random model. Both fixed effects (eg. the mean effect and the year effects) and random effects (usually animals' breeding values) **are fitted in the same model and estimated simultaneously in the same analysis.**

Example of a mixed model - following the same data as in the previous sections (with only year effect in the fixed model), just 'add' the two procedures:

y = X b + Z u + e

$$\begin{pmatrix} 354 \\ 251 \\ 327 \\ 328 \\ 301 \\ 270 \\ 330 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & 0 & 1 \\ 1 & -1 & -1 \end{pmatrix} \begin{pmatrix} b_{mean} \\ b_{Y2000} \\ b_{Y2001} \end{pmatrix} + \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ u_6 \\ u_7 \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \\ e_7 \end{pmatrix}$$

7 x 1 = 7 x 3 3 x 1 + 7 x 7 7 x 1 + 7 x 1

What does this mean?

For example, observation 1: The model says that 35.4Kg is made up of 1 'dose' of bmean, 1 dose of b_{2000}, no dose of b_{2001}, one dose of u_1, no dose of any of u_2 ... u_7, plus whatever is left over undescribed by these effects (e1). The task is to get estimates of b_{mean}, b_{2000}, b_{2001}, and u_1 ... u_7.

Henderson showed that the mixed model equations can be solved like this:

$$\begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z + \lambda A^{-1} \end{pmatrix}^{-1} \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

Notice that this is simply related to the equivalent estimates in fixed and random models:

Fixed Model: $\hat{b} = (X'X)^{-1}X'Y$

Random Model: $\hat{u} = (Z'Z + A^{-1}\boldsymbol{\lambda})^{-1}Z'Y$

The X'Z and Z'X blocks in the coefficient matrix (the matrix to be inverted) provide a connection between the fixed and random effects. If they were full of zeros, the results would be the same as if two separate models had been fitted (one fixed and one random, as in the previous sections). The values in these blocks let the analysis account for the fact that, for example, progeny which are very good because of being born in a good year do not overly increase their parents' EBV's. To solve for $\hat{b}$ and $\hat{u}$ we need X'Z and its transpose, Z'X. In the current example, Z is equal to the identity matrix (as all animals have 1 record only), and so X'Z = X' and Z'X = X.

$$\begin{pmatrix}\hat{b}\\ \hat{u}\end{pmatrix} = \begin{pmatrix}X'X & X'Z\\ Z'X & Z'Z+\lambda A^{-1}\end{pmatrix}^{-1} \begin{pmatrix}X'y\\ Z'y\end{pmatrix}$$

$$\begin{pmatrix}\hat{b}_{mean}\\ \hat{b}_{2000}\\ \hat{b}_{2001}\\ \\ \hat{u}_1\\ \hat{u}_2\\ \hat{u}_3\\ \hat{u}_4\\ \hat{u}_5\\ \hat{u}_6\\ \hat{u}_7\end{pmatrix} = \begin{bmatrix} 7 & 1 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1\\ 1 & 3 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & -1\\ 3 & 1 & 5 & 0 & 0 & 1 & 1 & 1 & 1 & -1\\ \\ 1 & 1 & 0 & 19/6 & 1/2 & -2/3 & -2/3 & -1 & 0 & 0\\ 1 & 1 & 0 & 1/2 & 17/6 & 0 & 0 & -1 & -2/3 & 0\\ 1 & 0 & 1 & -2/3 & 0 & 7/3 & 0 & 0 & 0 & 0\\ 1 & 0 & 1 & -2/3 & 0 & 0 & 7/3 & 0 & 0 & 0\\ 1 & 0 & 1 & -1 & -1 & 0 & 0 & 3 & 0 & 0\\ 1 & 0 & 1 & 0 & -2/3 & 0 & 0 & 0 & 7/3 & 0\\ 1 & -1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 2\end{bmatrix}^{-1} \begin{pmatrix}216.1\\ 27.5\\ 89.6\\ \\ 35.4\\ 25.1\\ 32.7\\ 32.8\\ 30.1\\ 27.0\\ 33.0\end{pmatrix} = \begin{pmatrix}31.194\\ -.915\\ -.890\\ \\ 2.826\\ -2.885\\ 1.834\\ 1.877\\ -.087\\ -2.240\\ 0\end{pmatrix}$$

Notice that:

1. The X'X and X'y blocks are as for the fixed effects analysis.
2. $Z'Z + A^{-1}$ is taken from the random model, with l = 1 as $h^2 = 0.5$, but presented already summed.
3. X'Z = X', Z'X = X and Z'Y = Y, as Z (and Z') is the identity matrix.
4. Because the mean is fitted, raw data can be used (in Z'y = y) rather than deviations from the overall mean, as used for the random model in the previous topic.
5. Once inverted, the coefficient matrix constitutes a set of custom index weights:

The elements of first three rows are multiplied by the elements of the (X'y plus Z'Y) vector to give the estimates of the three fixed effects.
The elements of the last seven rows are the index weights for estimating breeding values. Note that there are index weights in the "Z'X block" which are used to account for fixed effects when calculating EBV's.

Looking at the result vector, notice that:

1. b_{2002} = -(-.915) - (-.890) = 1.806 kg
2. Whereas the fixed model estimated 2001 to be .4 kg greater as an effect than 2000, this mixed model puts 2001 only about .025 kg ahead - a serious discrepancy of about .375 kg. This is because the average EBV of 2001 animals (+.346 kg) is about .375 kg ahead of the average EBV of 2000 animals (-.03 kg).

BLUP has determined that the difference observed in the means of 2000 and 2001 is largely due to genetic effects rather than environmental effects. This determination makes some sense when inspecting the pedigree diagram on the first page of the fixed model topic. Animal 1 looks to be quite superior to animal 2 (given the effects fitted in the BLUP model) and has a notably higher EBV. It leaves three offspring in 2001, to animal 2's two offspring - such that 2001 has a better representation of 'good genes' rather than 'bad genes'. The BLUP analysis takes this all into account, with the resulting discrepancy in year effects between the fixed and mixed models.

3 Animal 7 has a zero EBV. This is because it cannot be fairly compared to any other animal, as they are all born in different years. If animal 7 had had a relative born in a different year, it would have got a non-zero EBV, based on that relative's EBV.

Summary (Summary Slides are available on CD)

Best Linear Unbiased Prediction (BLUP) is the name of a method that is used worldwide to give estimated breeding values (EBV's) for commercially important traits.

BLUP uses all available information to estimate an animal's EBV, i.e. information from all genetically related animals, and possibly from correlated traits (if multi-trait BLUP is used). Furthermore, BLUP corrects for fixed effects such as flock, year or season of production etc., it accounts for unequal use of the best sires in different flocks, for selection and non-random mating.

BLUP relies on correct knowledge of genetic parameters (heritability, correlations), and on a good data structure.

The principle of BLUP is based on a combination of two techniques:

1) Selection index, where the phenotypic information about an animal is used to estimate a breeding value by regression.
2) Linear models as used in statistical analysis to estimate effects such as flock, year, season, age etc. in order to correct data when estimating breeding values.

Lecture 2: Properties of BLUP

Learning objectives

On completion of Lecture 2 you should be able to:

- Understand how BLUP accounts for selection of parents and non-random mating
- Understand how BLUP can estimate genetic trends
- Calculate accuracy from mixed model equations
- Understand accuracy from BLUP-EBV, in particular in relation to the contemporary group size and the 'effective information' about each animal.
- Understand the effects of BLUP selection on genetic progress and inbreeding

Key terms and concepts

BLUP accounting for selection and non-random mating, BLUP using optimal 'selection index' weight, Estimation of genetic trends, Accuracy and Prediction Error Variance of BLUP EBV, Effective information and contemporary group size, Optimising generation interval by selecting across age classes, BLUP selection leading to more inbreeding.

Introduction to lecture 2

The mixed model equations might seem a bit esoteric at first sight, but over the years researchers have discovered many interesting and useful properties, not only from a statistical point of view, but also from a genetic and animal breeding point of view. In addition, the mixed model equations are relatively easy to set up, and were a big improvement over previous methods, where complicated correction methods were applied, often one flock at a time. The main properties of BLUP are that

- Animals can be compared across herd/flocks unbiasedly and corrected appropriately for all identified fixed effects
- It uses all possible information about an animal's EBV, weighing all information sources optimally, and therefore is the most accurate estimate
- It corrects for selection of parents, assortative mating, and is able to distinguish between genetic and phenotypic trend.
- It automatically accounts for small contemporary group size

In this lecture we will look in more detail at some of these properties, which are important to understand when using BLUP-EBV in practice.

5.3 BLUP accounts for selection

We can understand more of the BLUP procedure for estimating breeding values if we write out individual equations, using the example in the previous lecture:

Looking at the equation for animal 6, who has one known parent (a dam):

$$\hat{\mu} + \hat{b}_{2001} - \tfrac{2}{3}\hat{u}_2 + \tfrac{7}{3}\hat{u}_6 = 27.0$$

$$\rightarrow \hat{u}_6 = \tfrac{3}{7}(27.0 - \hat{\mu} - \hat{b}_{2001}) + \tfrac{2}{7}\hat{u}_2$$

$$\rightarrow \hat{u}_6 = \tfrac{3}{7}(27.0 - \hat{\mu} - \hat{b}_{2001} - \tfrac{1}{2}\hat{u}_2) + \tfrac{1}{2}\hat{u}_2$$

Therefore, the breeding value of animal 6 is estimated from two parts:

1. a deviation of her phenotypic record from the expected mean and
2. the dam's (animal 2) breeding value.

The expected mean is the sum of all fixed effects plus the family mean. Since animal 6 has only one parent known, we take the deviation from one parent only (half sib family mean).

Some more detail (reference only):

The weighting factor is the regression of the within half sib family deviation on the within half sib family breeding value. The variance of the within half sib breeding value is 0.75 times the genetic variance. The weight is therefore equal to the following regression coefficient:

$$\frac{\text{cov}(u_{wFSf}, (y - y_{fam\_mean}))}{\text{var}(y - y_{fam\_mean})} = \frac{¾ V_A}{V_E + ¾ V_A} = \frac{¾ h^2}{1 - h^2 + ¾ h^2} = 3/7$$

We also look in more detail at animal 5, who has 2 parents known:

$$\hat{\mu} + \hat{b}_{2001} - \hat{u}_1 - \hat{u}_2 + 3\hat{u}_5 = 30.1$$

$$\rightarrow \hat{u}_5 = ⅓(30.1 - \hat{\mu} - \hat{b}_{2001}) + ⅓(\hat{u}_1 + \hat{u}_2)$$

$$\rightarrow \hat{u}_5 = ⅓(30.1 - \hat{\mu} - \hat{b}_{2001} - ½(\hat{u}_1 + \hat{u}_2)) + ½(\hat{u}_1 + \hat{u}_2)$$

The breeding value of animal 5 is also estimated from 2 parts:

1. a deviation of his phenotypic record from the expected mean and
2. from the mean of his parents' (animals 1 and 2) breeding values.

The expected mean is again the sum of all fixed effects plus the family mean. Since animal 5 has both parents known, we take the deviation from the full sib family mean

More detail:

The weighting factor is the regression of the within full sib family deviation on the within full sib family breeding value. The variance of the within full sib breeding value is 0.5 times the genetic variance. The weight is therefore equal to the following regression coefficient:

$$\frac{\text{cov}(u_{wFSf}, (y - y_{fam\_mean}))}{\text{var}(y - y_{fam\_mean})} = \frac{½ V_A}{V_E + ½ V_A} = \frac{½ h^2}{1 - h^2 + ½ h^2} = 1/3$$

Writing out the equation for **animal 2** who is a parent with progeny:

$$\hat{\mu} + \hat{b}_{2000} + ½\hat{u}_1 + 17/6\,\hat{u}_2 - \hat{u}_5 - ⅔\hat{u}_6 = 25.1$$

$$\rightarrow \hat{u}_2 = 6/17(25.1 - \hat{\mu} - \hat{b}_{2000}) - 3/17\,\hat{u}_1 + 6/17\,\hat{u}_5 + 4/17\,\hat{u}_6$$

$$\rightarrow \hat{u}_2 = 6/17(25.1 - \hat{\mu} - \hat{b}_{2000}) - 6/17(\hat{u}_5 - \tfrac{1}{2}\hat{u}_1) + 4/17\,\hat{u}_6$$

Hence we see that the breeding value for animal 2 is estimated from

1. her own record as deviation from the fixed effects (we have no family mean since she has no parents known),

2. from the estimated breeding values of her progeny.

Notice that the breeding values are corrected for the other parent (i.e. there's a correction for the mate), if the mate is known. In this case, the EBV of animal 5 is corrected for the contribution of his sire.

More detail (reference only):

The weights are not very easy to recognise, but they are the same as selection index weights. We can check this by simplifying the example, and ignore animal 5 as a progeny.

If animal 2 had only one progeny (animal 6), then her BLUP equation would look like

$$\hat{u}_2 = \tfrac{6}{14}(25.1 - \hat{\mu} - \hat{b}_{1990}) + \tfrac{4}{14}\hat{u}_6$$

but we saw earlier how the breeding value of animal 6 is estimated, therefore:

$$\hat{u}_2 = \tfrac{6}{14}(25.1 - \hat{\mu} - \hat{b}_{2000}) + \tfrac{4}{14}[\tfrac{3}{7}(27.0 - \hat{\mu} - \hat{b}_{2001} - \tfrac{1}{2}\hat{u}_2) + \tfrac{1}{2}\hat{u}_2]$$

$$\rightarrow \hat{u}_2 = \tfrac{6}{14}(25.1 - \hat{\mu} - \hat{b}_{2000}) + \tfrac{6}{49}(27.0 - \hat{\mu} - \hat{b}_{2001}) + \tfrac{4}{49}\hat{u}_2$$

$$\rightarrow \hat{u}_2 = \tfrac{7}{15}(25.1 - \hat{\mu} - \hat{b}_{2000}) + \tfrac{2}{15}(27.0 - \hat{\mu} - \hat{b}_{2001})$$

and the two weights can be found exactly by selection index to estimate the breeding value of an animal based on her own and her progeny's phenotypic record using a heritability of 0.5.

The previous lecture has shown that BLUP combines information from different sources, using weights that are derived from regression. BLUP uses the same weights as selection index.

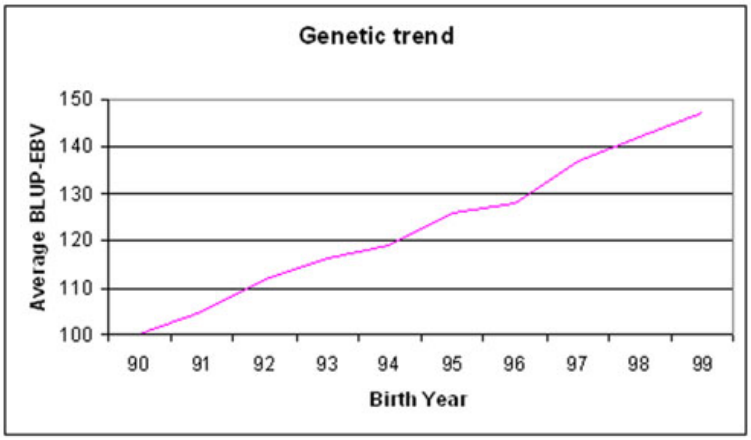
BLUP accounts also for possible genetic differences. It takes deviations from expected genetic means, and the EBV is regressed towards the expected genetic mean. This expected genetic mean is in most cases the family mean. This is an illustration of a very important property of BLUP: it corrects for selection. Since the better parents usually have more offspring, we expect that the average breeding value goes up in later generations (in the case of selection).

We also saw that in BLUP there is a correction for mates, i.e. BLUP corrects for assortative mating.

Genetic trend

Selection of parents over time leads to genetic improvement, i.e. there is an increase of average breeding value over time. This increase is referred to as 'genetic trend', and basically measures the success of breeding programs. The average of each generation is estimated by BLUP as the mean of the parents' EBV, and since BLUP is also able to work out the difference between all animals and those selected as parents, BLUP can properly estimate the genetic change over time. The genetic trend is estimated from the average EBV's over time, i.e. the EBV's are plotted against the birth year of the animals (Figure 5.2).

Figure 5.2 Hypothetical example of a genetic trend plotted as average EBV against birth-year. The average of 1990 is set to 100. Source: van der Werf, (2006).



The following example illustrates how BLUP distinguishes between genetic trend and environmental trend. Consider 3 sires performing in year 1 and having offspring in year 2. The average performance in years 1 and 2 are 30.0 and 31.3. If relationships within years were not considered, an animal's EBV would be estimated as deviation from the year mean (times h^2), and the average EBV per year would be equal to zero. The h^2 value used is 0.25.

year	year estimate	sire 1	sire 2	sire 3	
1	30.00	35.0	30.0	25.0	phenotype
		1.25	**0**	**-1.25**	**EBV**
		offspr. 1	offspr. 2	offspr. 3	
2	31.33	34.5	30.5	29.0	phenotype
		.79	**-.21**	**-.58**	**EBV**

Now consider the same example, but using the relationships between animals in a BLUP procedure. As the offspring have on average better parents, their average EBV is more than zero. The mean EBV in year 2 is now .45. The phenotypic trend from year 1 to year 2 is 1.3, and we now estimate that .89 of this is environmental and .45 of it is genetic trend.

year	year estimate	sire 1	sire 2	sire 3	
1	30.00	35.0	30.0	25.0	phenotype
		1.43	**-.18**	**-1.25**	**EBV**
		offspr. 1	offspr. 2	offspr. 3	
2	30.89	34.5	30.5	29.0	phenotype
		1.29	**.49**	**-.45**	**EBV**

5.4 Accuracy of BLUP EBV's

The accuracy of estimates from linear models is obtained from the inverse of the coefficient matrix. The coefficient matrix contains the amount of information about each effect (basically, the number of records for each sub class) and estimation error is inversely related to the amount of information.

The prediction error of estimates can be obtained from the mixed model equations:
For fixed effects the variance of prediction error is

$$Var(b) = C^{11}\sigma^2 e$$

where C^{11} is the part of the inverted coefficient matrix of the mixed model equations that refers to fixed effects. The standard error for a particular fixed effect estimate is then equal to the square root of this value.
In other words, if we have n observations in a subclass, by approximation, C^{11} ~ 1/n and the standard error of the estimate of the class effect ~ SQRT(1/n) times σ_E

For random effects the variance of prediction error is

$$Var(\hat{u} - u) = C^{22}\sigma^2 e$$

where C^{22} is the part of the inverted coefficient matrix of the mixed model equations that refers to random effects. The standard error of prediction (SEP) for a particular random effect (e.g. breeding value) is then again equal to the square root of this value.

A simple recipe for calculating the standard errors (= the Prediction Error) of BLUP-EBV is

1. Find or approximate the element relating to the animal's EBV in the inverted coefficient matrix C_{ii}.

2. Multiply this by V_E. This is known as the Prediction Error Variance,

3. The square root is the Standard Error (of Prediction) of the EBV.

4. From "Properties of EBVs", the accuracy is:

$$r_{IA} = \sqrt{1 - \frac{PEV}{V_A}} = \sqrt{1 - \lambda C^{ii}}$$

where λ is equal to $\frac{V_E}{V_A} = \frac{(1-h^2)V_P}{h^2 V_P} = \frac{(1-h^2)}{h^2}$

Example: in the random model (Topic 5 Lecture 1) the diagonal elements for animals 1 and 7 were 0.41 and 0.5 and $\lambda = 1$. Hence, the accuracy of EBV is:

for animal 1: $\sqrt{1 - \lambda C^{ii}}$ is $\sqrt{1 - 0.41}$ = 0.77

for animal 7: $\sqrt{1 - \lambda C^{ii}}$ is $\sqrt{1 - 0.5}$ = 0.71

Hence, the value of animal 7 is equal to h, as we would expect if only information from own performance is used. The accuracy of animal 1 is higher as it contained additional information from 3 progeny.

The squared value of accuracy of EBV (r) gives the reliability. Reliabilities are sometimes published (e.g. with sire summaries) instead of accuracies. The reliability is lower than the accuracy. The reliability reflects the squared correlation between true and estimated BV, or better, it reflects the proportion of variation in the TBVs which is explained by the data. In this sense, reliability is comparable with R^2 as used in statistical models.

Let the diagonal for animal I be	C^{ii}
The Prediction Error Variance of the EBV:	$PEV = C^{ii} \cdot V_E$
The Reliability of the EBV	$r^2_{IA} = 1 - C^{ii}\lambda$
The Accuracy of the EBV	$r_{IA} = \sqrt{(1 - C^{ii}\lambda)}$

Notice that the accuracy of breeding values is independent of the data! So the accuracy of an extremely good animal is not lower than that of an average animal, provided they both have the same information.

Example: Accuracy of the EBV of animal 3 in the random model example (assuming a known mean)

$C_{ii} = 0.462$.

$V_E = (1-h^2)V_P$, and with $h^2 = 0.5$ and $V_P = 50$ (for yearling weight in sheep, then $V_E = 25$, giving a result for PEV = 0.462 x 4.50 = 11.55 Kg^2.

Standard error of prediction: $SE_{EBV} = \sqrt{11.55} = 3.4$ Kg

Accuracy r_{IA} = = $\sqrt{1-1*0.462}$ = 0.733

Note that this seems only marginally more accurate than EBV from own phenotype alone (where r = h = $\sqrt{h^2}$ = $\sqrt{0.5}$ = 0.707). This is expected, as with high heritability there is little to be gained in accuracy from relatives.

Size of contemporary groups

Note that in a mixed model, the accuracy will be lower then above, as part of an animal's information is getting 'lost' in estimating the comparative mean (see next section).

As you can imagine, the information about an EBV is more accurate if an animal is compared with a larger number of contemporaries. We'll see in the next section that the information content is measured by 'effective number', and one observation in a group of 100 is worth effectively 0.99 whereas it is worth only 0.5 in a group of 2 and 0 in a group of 1, i.e. if there is no other contemporary in the group.

5.5 Obtaining accurate BLUP EBVs

The accuracy of an EBV depends on 'the amount of information used'. Generally, this depends on the number of records (observations) known about an animal and the number of records on relatives. The value of information depends on

1) the heritability, more accurate EBVs with higher heritability

2) the type of additive genetic relationship between the animal and its relatives that we have records on.

Since in BLUP we use information on ancestors, it is useful to know how much ancestral information actually contributes to increased accuracy of BLUP EBV.

Table 5.2 Accuracy of BLUP EBV for varying number of ancestral generations being present in the data. Source: van der Werf, (2006).

	N° of Generations Used in Evaluation			
h^2	1	2	3	4
0.1	0.14	0.16	0.16	0.17
0.3	0.38	0.40	0.40	0.41
0.5	0.57	0.58	0.59	0.59

The conclusion is that using ancestral information in BLUP has a small effect of the EBV accuracy. However, the main argument for including records of ancestors (and their contemporaries) in BLUP evaluation is to account for selection of parents and genetic trend.

3) the genetic and phenotypic correlation with other traits that we have records on. This will be discussed in Topic 6

4) common environmental effects between different records.

These effects of 1) to 4) can all be assessed using selection index theory. However, in practice, a very important criterion for the value of observations is the number of contemporaries that an animal could be compared with. It is not the number of records that is important, but

5) the effective number of records.

If an observation is compared with only one other record in a flock, we have a lot less certainty about the animals' performance compared with having 99 records in the flock to compare it with.

The effective information of a record is equal to

$$N_e = 1 - \frac{1}{N}$$

where N is the number of records in the contemporary group (Table 5.2).

Table 5.2 Effective information depending on contemporary group size. Source: van der Werf, (2006).

Contemporary Group Size	1 Record is Effectively
1	0
2	0.5
4	0.75
20	0.95

According to similar rules, we can work out the effective progeny of a sire. If a sire has n_i progeny in a certain flock of N animals, the effective number of progeny is:

$$\frac{n_i(N - n_i)}{N} = \frac{nr.prog\_sire * nr.comparable\_progeny}{total\_progeny}$$

Note that if 2 sires have all their progeny in one flock, they will have equal accuracy (even if one ram would have 10 times more progeny than the other ram!)

Table 5.3 Effective progeny for a Sire, depending on actual number and contemporary group size. Source: van der Werf, (2006).

Contemporary Group Size	Nr. Progeny Sire A	Effective Nr Progeny Sire A
1	1	0
10	10	0
10	9	0.9
10	1	0.9
10	5	2.5

Not only the data structure, but also the model used for genetic evaluation can therefore have a large impact on the accuracy of the BLUP EBV's. When too many effects are fitted, or contemporary group sizes are too small, the accuracy of EBV decreases. Therefore, setting up a proper model for genetic evaluation requires a statistical modeling approach where biasedness and accuracy have to be balanced; both biasedness and 'accuracy' increase if less effects are fitted.

The following example shows the accuracies of the EBV's for the 7 animals in the example for 4 different models. Note that the accuracy of animal 7 decreases to zero if it has no contemporaries in its group.

Table 5.4 Accuracies for the 7 animals in the example (Topic 5 Lecture 1) for 4 different models: 1 = random model (assuming mean known), 2 = mixed model fitting only the mean, 3 = mixed model fitting the year effect, and 4 = mixed model fitting year and sex.
Source: van der Werf, (2006).

	Model			
Animal	**1**	**2**	**3**	**4**
1	0.7683	0.6082	0.5404	0.3809
2	0.7516	0.6264	0.5561	0.5181
3	0.7335	0.609	0.5647	0.4164
4	0.7335	0.609	0.5647	0.5554
5	0.7612	0.5664	0.504	0.4243
6	0.7321	0.6175	0.5788	0.5124
7	0.7071	0.6304	0	0

5.6 Selection on BLUP breeding values

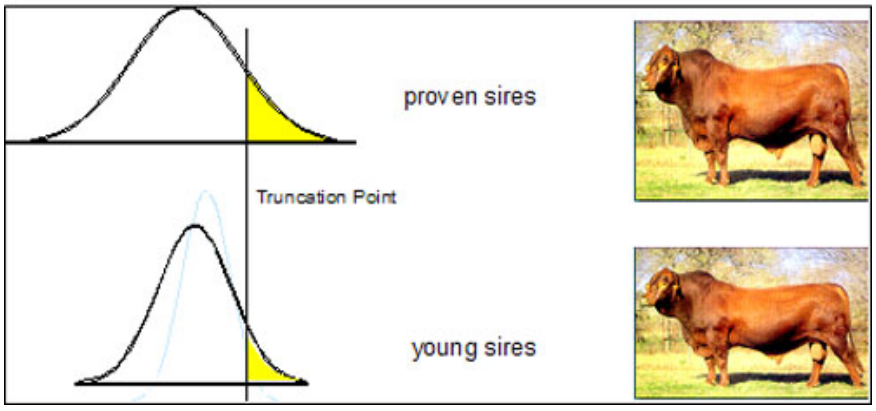
Maximum accuracy

Selection on BLUP breeding values should give the highest selection response, because using all possible information maximises the accuracy of EBV. Note that this is only the highest response on the short term, i.e. with respect to the next generation. In the longer term, response also depends on inbreeding, and an optimal strategy in the long term needs to balance (optimise) selection for merit while maintaining sufficient genetic diversity.

Optimising generation interval

When optimising a breeding program, a dilemma often arises as to whether young animals should be selected or older animals. Selecting young animals is good for achieving a short generation interval. However, younger animals have usually less accurate EBV. Older animals have generally more accurate EBV but selecting them would lead to longer generation intervals. Another (but essentially similar) argument against selecting older animals is that they are expected to have lower EBV. If there is a genetic gain per year, animals born x years apart are expected to differ x times the annual genetic gain.

Figure 5.3 Truncation points for older proven sires and younger unproven sires.
Source: van der Werf, (2006).



It may appear not easy to optimise selection over different age classes. However, the solution is remarkably simple. Because BLUP accounts for genetic trend, an important practical consequence is that animals can be selected across age classes based on their EBV by truncation selection. For example, all rams with an index value above 140 should be selected, independent of their age. It has been proven that BLUP selection optimises the use of rams across age classes. Hence, selection on BLUP EBV, irrespective of age, automatically optimises the generation interval, and the use of old versus young rams, as demonstrated in Figure 5.3.

Younger animals have on average better EBV, but also generally less variation in EBV. The optimum proportion of younger animals depends on the difference in the variance of the EBV within age classes (i.e. on the accuracy) and on the genetic lag between age classes (i.e. on the genetic gain per year and the number of years).

Truncation selection across age classes will automatically optimise the proportion of young rams used. The proportion that should be optimally selected from each age class is automatically achieved if simply the best animals are selected based on their BLUP EBV. Selecting animals on BLUP EBV irrespective of their age automatically optimises the generation interval.

With a larger genetic trend or with increased accuracy of the young ram EBV, the proportion of young rams will increase (Figure 5.2.2).

BLUP and inbreeding

Selection on BLUP EBV maximises response to selection. However, response is only maximised with respect to the next generation. Long term selection response is not necessarily optimised with BLUP selection. The reason is that BLUP selection leads to more inbreeding than for example mass selection (selection based on own phenotype). More inbreeding leads in the longer term to loss of genetic variation, and possibly inbreeding depression, and therefore reduces and offsets genetic gain.

Why does BLUP selection lead to more inbreeding? This can be explained by the fact that BLUP uses information on all possible relatives to estimate an animals' EBV. Using information from family members implies that members of the same (good) family have more chance to be jointly selected. With lower heritability there is relatively more weight on the family information (compared with own performance), and the lower the heritability, the more BLUP selection resembles family selection and the more it leads to inbreeding. This is illustrated in Table 5.5 with a simulation study of a closed swine herd (Belonsky and Kennedy, 1988). Table 5.5 shows that:

- BLUP selection leads to significantly more selection response than selection on individual phenotype. The difference is larger for lower heritability.
- BLUP leads to more inbreeding than selection on individual phenotype. With BLUP selection the rate of inbreeding is considerably higher with lower heritability. With selection on individual phenotype there is (slightly) more inbreeding with higher heritability.
- Even after 10 years of selection, BLUP hadn't lost its superiority over individual selection. Loss of variance due to inbreeding was offset by increased used of relatives' information. However, effects of inbreeding depression were not simulated in this study.

To optimise selection in the longer term, it is useful to combine selection on BLUP EBV with some restriction on the average relatedness of the selected animals. For optimal selection rules see Wray and Goddard (1994) or Meuwissen (1997).

Table 5.5 Average genetic merit and average inbreeding of progeny after 5 and 10 years of selection on individual phenotype and on Best Linear Unbiased Prediction of breeding value (BLUP). Source: Belonsky and Kennedy, (1988).

		Average Genetic Merit		Average Inbreeding	
Heritability	Year	Phenotypic Selection	BLUP Selection	Phenotypic Selection	BLUP Selection
0.10	5	0.38	0.63	0.067	0.167
	10	0.78	1.41	0.174	0.383
0.30	5	1.10	1.41	0.078	0.141
	10	2.40	3.14	0.193	0.332
0.60	5	2.25	2.29	0.087	0.130
	10	5.16	5.31	0.205	0.293

Summary **(Summary slides are available on CD).**

Mixed model equations are relatively easy to set up, and their solutions appear to have a number of properties that are desirable from a practical point of view.

- BLUP accounts for selection of parents and non-random mating
- BLUP can estimate (and correct) for genetic trend, making EBV comparable across age classes.

The accuracy of BLUP EBV can be derived from the mixed model equations. The accuracy is higher with:

- higher heritability
- more information from relatives and correlated traits
- more “effective records” per animal.

Small contemporary groups decrease effective information per record, and therefore decrease accuracy.

Truncation selection based on BLUP EBV leads to:

- Increased accuracy in the short term
- Optimised generation intervals
- More inbreeding

For optimal selection, BLUP EBV (being the best estimate of an animal’s EBV) can be used in an optimal selection procedure where animals’ co-ancestry is also taken into account to maintain sufficient genetic diversity.

Lecture 3: Evaluation of Animals in Practice

Learning objectives

On completion of Lecture 3 you should be able to:

- Understand the level of complexity of commercial genetic evaluation programs
- Understand the requirements about data quality and the quality of the models used for genetic evaluation
- Know about the different extensions that are possible based on the animal model
- Understand across breed evaluation
- Be aware of computational methods used to solve large scale genetic evaluation systems

Key terms and concepts

Animal model, General formulation of mixed model , Sire model, Reduced animal model, Multiple trait model, Repeated records model, Maternal effects model, A model containing genetic groups, Iterative methods to solve mixed model equations.

Introduction to lecture 3

Application of mixed models has become a useful tool to evaluate animals in breeding programs of various breeding organisations. The methodology consists of a framework with justifiable statistical and genetic properties and it potentially delivers the most accurate and least biased prediction of breeding values.

The quality of evaluations depends on:

1. The data
    - Recording standards
    - Consistency of recording
    - Recording of pedigree
    - Recording of fixed effects
    - Number of different traits recorded simultaneously

2. The model
    - Accounting for all fixed effects
    - Accounting for maternal effects and repeatability effects
    - Accounting for base population heterogeneity
    - Accounting for heterogeneous variation
    - Correctness of genetic parameters

The basis for a good genetic evaluation system is good data. For a good recording system, rules need to be set, e.g. about age at recording of certain traits (e.g. ‘post weaning weight’), recording of the age (or date), management groups, the number of traits to be recorded, and whether full pedigree is required. It is difficult to impose rules on farmers, and a better strategy is often to give incentives for breeders for doing the right thing. For example, to become acknowledged as ‘gold standard’ a breeder would have to meet a set of criteria in relation to data and pedigree recording to qualify.

In Australia, the main system in the beef cattle industry is BREEDPLAN. LAMBPLAN has been set up for meat sheep, and for wool sheep there are several systems, such as Merino Genetic Services, Merino Benchmark and Select Sires. Currently (2005) an important development is the establishment of the Australian Sheep Genetic Database, with one evaluation system for all Australian sheep. In dairy cattle, genetic evaluation is carried out through the Australian Dairy Herd Improvement Scheme.

With international genetic evaluation, it is also important that different countries have comparable systems. The International Committee on Animal Recording (ICAR) has been established to coordinate such efforts. ICAR is currently mainly active in relation to the dairy industry, being the most internationalised industry, but also in sheep and beef cattle there are initiative for international genetic evaluation. The recording of 'type traits' has been especially hard to standardise across countries.

Appropriate methodology and an appropriate model are required for accurate and unbiased EBV (although it is not possible to fix up bad data with a good method). The method of choice is based on mixed model (BLUP) methodology but there are a large range of models possible. The BLUP methodology has the ability to account for selection of parents in a breeding population. Hence, it accounts for the fact that some animals are from better parents than other animals. Note that a requirement is that the pedigree, and data on selected parents as well as non-selected contemporaries are included in the analysis.

Models can be extended to account for more complicated effects, such as:

- different breeds (useful for an 'across-breed' evaluation or when there are animals imported from other countries)
- heterogeneity of genetic mean of base population
- maternal effects: important in all pre-weaning traits
- correlations between repeated records
- correlated traits: useful for higher accuracy or to account for selection on a second trait (e.g. first lactation vs. later lactation or weaning weight vs. yearling weight)
- interactions between environment and genotype: some sires may have a different effect in different environments
- heterogeneous variance: the differences in one flock may be much larger (on average) than the differences in another flock.

Some factors are more difficult to include in the model: e.g. preferential treatment of some animals (e.g. with supplements), or serious illness at the time of measurement. It is up the flock-recording scheme to design rules for when measurements can be considered as 'valid'. It is important here that the flock recording is unbiased and non-selective.

In this lecture, we will discuss in more detail the model extensions that are possible to account for certain effects that are common in practice.

5.7 Other models for BLUP evaluation

So far we have dealt with a single trait animal model BLUP. It is an 'animal' model because we fit a breeding value for each animal. It is good to realise that this is the simplest model; it is called a single trait animal model. It means that animals have only observations on one character (trait) and there are only fixed effects and additive genetic effects, and no other random effects such as maternal effects or interactions between sire and flock. It is important to understand the principles of the simplest model. Less simple models are based on the same principles, and therefore are not really that much more difficult to understand. Like in any other statistical model, building more complicated models largely requires more knowledge of the data, and determining the significance of the different effects that could possibly explain part of the variation observed.

More detail (reference only):

The mixed model equations in a general form look like:

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

where b is the vector with fixed effects with design matrix X (relating observations to fixed effects); u is the vector with random effects with design matrix Z (relating observations to random effects) and e are residual effects associated with error.

The variances of these effects (in vectors) are defined in variance-covariance matrices as follows:

var(u)= G
var(e)= R
var(y)= ZGZ' + R

In the single trait animal model with breeding value as the only random effect, we assume that the matrix R = $I.\sigma^2_e$ and G = $A\sigma^2_a$. The simple equations were therefore obtained by multiplying the equations with the factor σ^2_e.

$$\begin{pmatrix} X'X & X'Z \\ Z'X & Z'Z+\lambda A^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \end{pmatrix}$$

where $\lambda = \sigma^2_e / \sigma^2_a$.

Note that from now on we use σ^2_a for V_A and σ^2_e for V_E as this is more common in the mixed model literature

Other models are extensions (or sometimes simplifications) of the single-trait animal model:

Sire model

In this model, only effects of sires are fitted on records of their progeny (these are 0.5 times their breeding values), making for computational ease. We may have only 100 sires in a data set on 100,000 recorded animals, hence needing 0.1% of the number of equations of an animal model. The EBV may be slightly less reliable: because of lower accuracy (only in the case of few progeny per sire) and potential bias, because there is no correction for differences between dams. The model basically assumes that all progeny of a sire are from a different dam and all dams are expected to be equally good.

Reduced animal model (RAM)

In this model, breeding values are only fitted for animals that have progeny records. This makes for faster computing (only equations for animals that are parents), and the EBV's for all other animals are simply derived from those of their parents, plus their own corrected phenotypes. The results are the same as for a full animal model. Less computing time at the cost of some extra computer programming time is needed.

Multi trait model

This is an extension of the single trait case. Data on a number of traits are available in y, and EBV are calculated for each trait. The results are generally different from what would be obtained from a number of separate single-trait BLUP EBV, because each trait is used to help give information about all other traits, much as with a selection index. In a later topic, the multiple trait BLUP procedure will be demonstrated in more detail. The benefit from multiple trait models comes from

- more accuracy as information from correlated traits is used
- less bias as the analysis takes into account that for traits that are measured after sequential rounds of selection, only the better progeny are evaluated.

An example of potential selection bias. Compare a good ram and a bad ram, each having 40 progeny at weaning. From the good ram, no progeny are culled, whereas from the bad ram 50% are culled. Comparing the progeny of these rams at post-weaning will give a huge advantage to the bad ram, as his bad progeny have been removed. Multi-trait BLUP would correct for this bias.

Repeated records model

This is used where animals can have more than one record, such as multiple fleece weight records in sheep. The phenotypic correlation between recordings is equal to repeatability, and genetic correlation between recordings is assumed one (if the genetic correlation was less than one, then the multi-trait approach outline above is applicable!).

The approach is to create a permanent environmental effect for each animal, i.e. when the animal has a second record, not only his breeding value but also part the environmental effects are repeated. This can represent the effect of the environment during raising of the animal (a good 'development' guarantees a consistently good performance later on), or the occurrence of a disease that happened to a particular animal, with permanent effects.

More detail (reference only):

The mixed model with repeated records can look like:

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

where **y** is the vector of the observations, **b** is the vector of fixed effects, **u** is a vector of additive genetic effects, **p** is a vector of permanent environmental effects and **e** is a vector of residual effects. The matrix **X** is the incidence matrix for the fixed effects and **Z** is the incidence matrix relating observations to animals. Each animal has an additive genetic as well as a permanent environmental effect, hence both effects have the same design matrix.
The three random effects have the following distribution

$$\text{var}\begin{pmatrix} u \\ p \\ e \end{pmatrix} = \begin{pmatrix} A\sigma_a^2 & 0 & 0 \\ 0 & I\sigma_c^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix} \qquad G = \begin{pmatrix} A\sigma_a^2 & 0 \\ 0 & I\sigma_c^2 \end{pmatrix}$$

where σ^2_a is the direct additive genetic variance and σ^2_c is the variance due to permanent environmental effects variance and σ^2_e is the variance due to random environmental effects. The model assumes that those permanent environmental effects for different animals are uncorrelated, and within an animal there is no correlation between its additive and its permanent environmental effect. The total phenotypic variance is the sum of the three variance components.

The mixed model equations for a model with repeated records look like:

$$\begin{pmatrix} X'X & X'Z & X'Z \\ Z'X & Z'Z + \lambda A^{-1} & Z'Z \\ Z'X & Z'Z & Z'Z + \gamma I \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \\ \hat{p} \end{pmatrix} = \begin{pmatrix} X'y \\ Z'y \\ Z'y \end{pmatrix}$$

where $\lambda = \sigma^2_e / \sigma^2_a$ and $\gamma = \sigma^2_e / \sigma^2_c$

Maternal effects model

Some traits such as survival of piglets or early growth in beef cattle and meat sheep are influenced by maternal effects. The mother has an influence on the performance of her offspring over and above that of her direct additive genetic contribution, i.e. through maternal effects. These maternal effects are strictly environmental for the offspring, but can have both a genetic and environmental component. In selection of animals, and especially in dam lines, it is important to consider the maternal genetic effects. Beef cattle and lamb producers are interested in animals which have a high breeding value for growth (direct genetic effect) but also in cows and ewes with good mothering abilities (milk production). Including maternal effects in the model allows us to estimate maternal effects and to correct for possible biases in genetic evaluation of the growing animal. It is usually assumed that maternal effects are genetic, although part of it might also be a permanent environmental effect (e.g. a dam missing a teat).

More detail (reference only):

In the following model the direct genetic and maternal genetic effects are considered:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Z}_1\mathbf{u} + \mathbf{Z}_2\mathbf{m} + \mathbf{e}$$

where **y** is the vector of the observations, **b** is a vector of fixed effects, **u** is a vector of additive genetic effects, **m** is a vector of maternal genetic effects and **e** is a vector of residual effects. **X** is the incidence matrix for the fixed effects and $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are incidence matrices relating observations to random effects of animal (additive genetic) and dam (maternal genetic), respectively. The random effects have the following distribution:

$$\text{var}\begin{pmatrix} u \\ m \\ e \end{pmatrix} = \begin{pmatrix} A\sigma_a^2 & A\sigma_{am} & 0 \\ A\sigma_{am} & A\sigma_m^2 & 0 \\ 0 & 0 & I\sigma_e^2 \end{pmatrix} = \begin{pmatrix} G & 0 \\ 0 & R \end{pmatrix}$$

$$G = \begin{pmatrix} A\sigma_a^2 & A\sigma_{am} \\ A\sigma_{am} & A\sigma_m^2 \end{pmatrix} = G_0 \otimes A$$

where G_0 is a 2 by 2 matrix: $\begin{pmatrix} \sigma_a^2 & \sigma_{am} \\ \sigma_{am} & \sigma_m^2 \end{pmatrix}$ and $\otimes$ is a direct product (it 'blows up' a matrix).

Further σ^2_a is the direct genetic variance, σ^2_m the maternal genetic variance, σ_{am} the covariance between direct and maternal genetic effects and σ^2_e the error variance. The model shows that both random effects have a covariance structure depending on the genetic relationships. Related dams have related maternal genetic effects, and there is a correlation between a dam's direct additive genetic effect and her maternal genetic effect. The total phenotypic variance is equal to

$$\sigma_p^2 = \sigma_a^2 + \sigma_m^2 + \sigma_{am} + \sigma_e^2$$

The mixed model equations are:

$$\begin{pmatrix} X'X & X'Z_1 & X'Z_2 \\ Z_1'X & Z_1'Z_1 + \alpha_{11}A^{-1} & Z_1'Z_2 + \alpha_{12}A^{-1} \\ Z_2'X & Z_2'Z_1 + \alpha_{21}A^{-1} & Z_2'Z_2 + \alpha_{22}A^{-1} \end{pmatrix} \begin{pmatrix} \hat{b} \\ \hat{u} \\ \hat{m} \end{pmatrix} = \begin{pmatrix} X'y \\ Z_1'y \\ Z_2'y \end{pmatrix}$$

where $\begin{pmatrix} \alpha_{11} & \alpha_{12} \\ \alpha_{21} & \alpha_{22} \end{pmatrix} = G_0^{-1}.\sigma_e^2$. Hence the α-values are still ratios of variances, as in the simple mixed model, but they are now dependent on the covariance between the two random effects.

Since the full (inverse) relationships matrix is used with relationships between all animals (progeny as well as dams), estimates will be obtained for additive effects for progeny (with records) as well as for dams (with possibly no own record). Equally, estimates for maternal effects will be obtained for dams (with progeny) as well as for progeny (which may not have expressed their maternal ability yet).

Genetic groups model

This sort of model is important where the animals in the data set come from widely divergent sources. Examples are:

- different strains of Merino sheep
- different breeds (Hereford and Angus, Poll Dorset and Suffolk)
- different base populations (US Angus, Australian Angus)
- different selection groups (based on birth year or breeding status)

The mixed model assumes that the breeding values to be estimated come from a homogeneous population, and all have the same expected mean, that is for the animals with unknown parents (the expectation of animals with parents known is equal to the parental average EBV). Animals without parents are called 'base animals', and if they are not from a homogeneous population, genetic groups are needed to distinguish between different genetic levels of base animals.

Notice that the relationships matrix takes care of all genetic differences due to selection since the base population. For example, in analysing data of a selection experiment with a high and low line, but both stemming from the same base population, genetic groups are not needed as long as pedigree and data since the start of selection are included in the analysis. Genetic groups are therefore needed for those cases where we can't explain genetic differences between animals by pedigree and data. This is typically the case if animals arise from different breeds or populations.

Consider Finn sheep (F, average litter size about 3) mixed in with Merinos (M, litter size = 0.9). Litter size is a lowly heritable trait, and so any genetic evaluation ignoring breed will regress all EBV to close to the average - clearly wrong, as the breed effect on litter size is strong and reliable.

The solution is to fit animal source as a fixed effect. With ongoing breeding, individual animals can be a mixture of sources - but this is not a problem. Table 5.6 shows an example of entries in the X matrix for the F(inn) and M(erino) fixed effects.

Table 5.6 Example entries for the X matrix for the F(inn) and M(erino) fixed effects. Source: van der Werf, (2006).

Type of animal	F effect	M effect
Finn	1	0
Merino	0	1
F x M	1/2	1/2
M x (FxM)	1/4	3/4

Other examples of genetic groupings are:
"animals imported from Canada"
"animals born before 1980"

The EBV of an animal is now:

EBV = genetic group solution + random additive genetic effect (within group)

For example, if the fixed effect estimate of F is +0.7 compared to M, animals fully belonging to the Finn breed get 0.7 added to their random within breed breeding value, so that EBV of Finns and Merino's can be directly compared to each other.

Another example:

Consider the yearling weight of 2 bulls from different breeds:

Michael Angus	315	Mean Angus	300
Whiskey Hereford	315	Mean Hereford	320

The within breed EBV is regressed towards the breed mean, and the across breed EBV contains the breed difference:

	EBV_{within}	EBV_{across}
Michael (Angus)	+ 6	+ 6
Whiskey (Hereford)	- 2	+ 18

5.8 Computational issues in genetic evaluation

Related to the problem of finding the optimal model, is the question of whether or not it is feasible to compute all evaluations for the population in mind. Many breeding organisations have a well organised system of data collection. In that case, central evaluation of breeding values can be done for all animals in a region or country. Besides the animals of the present population, it is also desirable to include the parents and grandparents in the evaluation, since it enables us to account for selection and genetic relationships. The number of animals to be evaluated in such a data set can be several thousand but some dairy populations contain several million animals to be evaluated. A complete genetic evaluation, considering multiple traits and multiple effects (additive genetic, maternal, permanent environment, etc.) can be a gigantic task, even with modern high speed computers.

The above methods of estimating fixed and random effects involve inversion of matrices. In the earlier days (before 1965), people would lock themselves away for a fortnight with a calculating machine, just to invert a 40 x 40 matrix. When breeding values are fitted, as in the animal model BLP and BLUP we considered in the last two lectures, then for large data sets we may have to invert matrices with very large dimensions which is still not practical, even with today's computers.

Think of a genetic evaluation for half a million animals for 5 traits with maternal effects. We would have 10 equations per animal plus additional fixed effects, say at least 5 million equations. The coefficient matrix would have the 25.10^{12} elements. So many coefficients cannot be stored and such matrices cannot be inverted. There are very many numerical and programming tricks to get around such problems. For example, only non-zero coefficients are stored and matrix equations are solved by iteration. Animal models can be solved by 'iteration on data' where right hand sides are corrected each round for all possible effects, i.e. the flock total is corrected for the effects of all animals in the flock and the animal total is corrected for the effects of flock (and other fixed effects).

More detail (reference only):

Solving equations by iteration.
You are not asked to be able to remember details here. All you need to remember is that there is a wealth of computing strategies which make it possible to run BLUP evaluations for large populations.

As a simple example, consider the fixed model example given in Topic 4.
eq \o(b,\s\up6(^)) = X'y:

$$X'X \quad \hat{b} \quad = \quad X'y$$

$$\begin{pmatrix} 7 & 1 & 3 \\ 1 & 3 & 1 \\ 3 & 1 & 5 \end{pmatrix} \begin{pmatrix} 31.3 \\ \hat{b}_{2000} \\ -.65 \end{pmatrix} = \begin{pmatrix} 216.1 \\ 27.5 \\ 89.6 \end{pmatrix}$$

For illustration it is assumed that $\hat{\mu}$ and $\hat{b}_{2001}$ are known (31.3Kg and -.65Kg) but that $\hat{b}_{2000}$ is not.

The simple rules of matrix multiplication tells us where the 27.5 comes from:

$$1 \times 31.3 + 3 \times \hat{b}_{2000} + 1 \times (-.65) = 27.5$$

giving

$$\hat{b}_{2000} = \frac{27.5 - 1x31.3 - 1x(-.65)}{3} = \frac{-3.15}{3} = -1.05$$

as we saw earlier.

In practice, we do not have the luxury of knowing all the other estimates. So, all estimates are initially set to zero, and the above operation is carried out for each effect to be estimated, in turn. Thus when the second effect is being estimated, it has the benefit of a non-zero entry for the first estimate.

After one cycle of doing this, the estimates are, of course, not correct - but they will be more correct after a second cycle. The procedure ***iterates*** through as many cycles as needed to stop the estimates changing from the last cycle.

Thus the problem has been solved without inverting the coefficient matrix.

Here are the values of the estimates at each cycle or 'iteration' for the example:

round 1: $\hat{\mu} = \frac{216.1 - 0 - 0}{7} =$ -30.871

round 2: $\hat{\mu} = \frac{216.1 - 1.(-1.1238) - 3.(-.3781)}{7} =$ -31.194

Iteration Round	$\hat{\mu}$	$\hat{b}_{2000}$	$\hat{b}_{2001}$
0	0	0	0
1	30.87143	-1.12381	-.3780957
2	31.19401	-1.105306	-.5753463
3	31.27591	-1.066854	-.6321748
4	31.29477	-1.054198	-.6460219
5	31.29896	-1.050958	-.6491456
6	31.29977	-1.050209	-.6498209

Readings

The following readings are available on CD:

1. Simm, G. 2000, 'BLUP', in *Genetic improvement of cattle and sheep*, Farming Press, Miller Freeman, UK, pp. 165-182.

Activities

@ Available on WebCT

Multi-Choice Questions

@ Submit answers via WebCT

Useful Web Links

@ Available on WebCT

Assignment Questions

@ Choose ONE question from ONE of the topics as your assignment. Short answer questions appear on WebCT. Submit your answer via WebCt

Summary

The quality of genetic evaluation systems depends on the quality of the data as well as on the models used.

The simple animal model can easily be extended to models that account for maternal effects, repeated records or multiple traits.

Genetic groups are used to account for animals of different genetic mean, e.g. due to breed, region or year of birth.

Computational methods such as iteration are used to be able to solve large systems of equations.

References

Lecture 1

Henderson, C.R. 1973, 'Sire evaluation and genetic trends', in *Proceedings of the Animal Breeding and Genetics Symposium in Honor of J. L. Lush,* American Society for Animal Science, Blackburgh, Champaign, Illinois, pp. 10-41.

Simm, G. 2000, *Genetic improvement of cattle and sheep*, Farming Press, Miller Freeman, UK.

Lecture 2

Belonsky, G.M. and Kennedy, B.W. 1988, 'Selection on individual phenotype and best linear unbiased prediction of breeding value in a closed swine herd', *Journal of Animal Science,* vol. 66, p. 1124.

Meuwissen, T.H.E. 1997, 'Maximizing the response of selection with a predefined rate of inbreeding', *Journal of Animal Science*, vol. 75, p. 934.

Wray, N.R., and M.E. Goddard. 1994, 'Increasing long term response to selection', *Genetics Selection Evolution,* vol. 26, p. 431.

Simm, G. 2000, *Genetic improvement of cattle and sheep*. Farming Press, Miller Freeman, UK.

Glossary of terms

Accuracy	The correlation between true and estimated breeding value
Assortative mating[1]	Mating can be assortative with respect to a certain genotype (e.g. individuals with genotype AA tend to mate with other individuals of genotype AA) or phenotype (e.g. tall individuals mate with other tall individuals)
BLP	Best Linear Prediction – Prediction of EBVs ignoring any fixed effects
Contemporary group	Group of animals of about the same age and performing in about the same environment
Effective records	The number of records that an animal can be compared with. The more effective records to compare against the more certainty about the animals performance
Genetic groups model	A model used when the animals in the data set come from highly diverse backgrounds eg different breeds or one breed in different countries
Generation interval[1]	The weighted average age of parents when their offspring are born
Genetic trend[1]	The change in mean EBV in a population of animals over time
Identity matrix	Similar to an incidence matrix in a fixed model, the identity matrix shows what observations relate to which animal in a random model
Iteration	During statistical computation, the estimation of values progresses through numerous cycles or iterations with each iteration producing estimates slightly closer to the correct values. The iterations continue until the estimates remain the same as the previous iteration
Maternal effects model	A model that accounts for maternal effects on the records of an animal eg mothers milk production on weaning weight
Mixed model	A model that fits both fixed and random effects simultaneously
Multi trait model	Similar to a single trait model except a number of traits may be included
Pedigree[1]	A record of the ancestry of an animal
Reduced animal model	A model that is only used for animals that have progeny records
Regression[1]	A procedure that measures the direction and strength of an association between two characters
Reliability[1]	The squared accuracy of EBVs
Repeated records model	A model used for those traits that might be measured a number of times over the life of an animal eg fleece weight
Selection index[1]	An overall score of genetic merit which combines information on several measured traits, with an emphasis on strength of association with traits in the breeding objective and their relative economic value
Single trait model	A model that contains only one trait and has only fixed effects and additive genetic effects but has no random effects included
Sire model	A model that fits only the effects of sires on their progeny

[1] Glossary terms taken from Simm (2000).