# Estimation of Breeding Values

Guillaume Ramstein & Peter Sørensen

2022-03-03

## Contents



## Learning objective:

This section introduces the basic concepts of breeding value estimation such as:

- basic principle behind estimating breeding values
- accuracy of estimated breeding values
- use of genetic relationships for estimating breeding values
- the connection between genetic parameters and estimated breeding values
- different methods, data sources and experimental designs for estimating breeding values

# 1 Introduction

Breeding value estimation is a fundamental component of breeding programmes, in which the breeding value of each individual is predicted to inform subsequent selection decisions. The breeding value for an individual is the total additive genetic value which is passed on to the offspring. Thus the breeding value is not a measure of how good an individual is in itself, but rather of the effect its genes will have in the population. Breeding values are used for:

- comparing individuals in the breeding population and selecting parents for the next generation
- predicting the consequences of selection decisions
- describing genetic differences over time (result of previous selection)

The true breeding value (TBV) for an individual cannot be observed. It is only possible to measure its phenotypic value, which is influenced both by genotype and environment. Therefore, we need a way to infer the breeding value from the phenotypic value and select individuals based on an estimated breeding value (EBV). This is the objective of breeding value estimation.

# 2 Basic principles for breeding value estimation

Breeding values are estimated using information on phenotypes and genetic relationships for individuals in a breeding population. As introduced previously the phenotype for a quantitative trait is the sum of both genetic and environmental factors. In general the amount of information provided by the phenotype about the breeding value is determined by the narrow sense heritability ($h^2$), which measures the proportion of additive genetic variance contained in the total phenotypic variance). Furthermore phenotypes collected from close relatives provide more information about the breeding value of an individual. In this section we will illustrate these principles using phenotypic data and genetic relationships used for for estimating breeding values. We will now try to derive a general approach for predicting breeding values for any situation. Even though the procedure is general we will use a simple example to describe it.

## 2.1 Genetic model

The breeding value is based on an assumption of a specific genetic model. In general the total genetic effect for an individual is the sum of both additive and non-additive effects that affect the trait:

$$y = \mu + a + d + i + \epsilon \tag{1}$$

where $\mu$ is the population mean, $a$ is the breeding value (i.e. additive effect), $d$ is the dominance effect, $i$ is the epistasis effect, and $e$ is the environmental deviation (or residual) not explained by the genetic effects in the model. However, only the additive genetic effects are passed on to the offspring and therefore contributes to the breeding value. In contrast non-additive genetic effects (dominance and epistasis) are degraded by recombination and are not inherited, even though they may be important for the individual's phenotype. Therefore we only consider the additive genetic model as the basis for breeding value estimation;

$$y = \mu + a + e$$

The true breeding value for an individual is the sum of all additive genetic effects that affect the quantitative trait:

$$a_i = \sum_{i=1}^{q} a_{ij}$$

where $a_i$ is the total additive genetic effect and $a_{ij}$ is the additive genetic effect for loci j in individual i. We therefore assume (based on the central limit theory) that the true breeding values, a, and the residual term, $e$, are normally distributed which means that the observed phenotype is also normally distributed

$$\begin{aligned} a &\sim N(0, \sigma_a^2) \\ e &\sim N(0, \sigma_e^2) \\ y &\sim N(\mu, \sigma_a^2 + \sigma_e^2) \end{aligned}$$

## 2.2 Expected breeding value conditional on observed phenotype

The breeding value cannot be observed but must be estimated from phenotypic data and genetic relationships between individuals from the breeding population. Estimation of an unknown parameter using statistical modelling expresses the estimated quantity as a mathematical function of the observed data. The question is how this function should look like and what properties the estimated breeding values should fulfill. For breeding purposes one objective for the estimated breeding values is that the response to selection is maximized. [Henderson, 1963] found that the improvement of an offspring generation compared to the parent generation can be maximized when parents are selected based on the conditional expected value ($E(a|y)$) of the true breeding value $a$ given the observed phenotypic values $y$. Under the assumption of multivariate normality for $a$ and $y$ (which are justified under the central limit theorem and the assumptions of many genetic and environmental factors), the expected value of the breeding value conditional on the observed phenotype $y$ can be written as:

$$E(a|y) = E(a) + Cov(a, y)[Var(y)]^{-1}(y - E(y)) \tag{2}$$

The breeding value is defined as deviation from the general mean which means that the expected value $E(a)$ of the true breeding value $a$ is $E(a) = 0$. Therefore the expected value of the breeding value is:

$$E(a|y) = Cov(a, y)[Var(y)]^{-1}(y - E(y)) \tag{3}$$

The expression for the estimate of the breeding value consists of two parts; The term, $y - E(y)$, shows that the observed phenotypic values are corrected for the fixed effects represented by $\mu$. The term, $b_{a|y} = Cov(a, y)[Var(y)]^{-1}$, often referred to as the regression coefficient is a weighting factor with which the corrected phenotypic values are multiplied.

To be able to estimate the breeding value we need to determine the values for the terms, $E(a)$, $E(y)$, $Var(y)$, and $Cov(a, y)$ in the expression above. It is possible to derive simple formula's for these terms based on:

- adjusted phenotypic observations for the quantitative trait of related individuals
- heritability of the quantitative trait
- knowledge of inheritance laws and genetic relationships (e.g. parents, grandparents, siblings) for individuals with phenotypic observations of the quantitative trait

We will distinguish between true and estimated breeding value using the following notation:

$a$ = additive genetic value = breeding value
$\hat{a} = E(a|y)$ = estimated additive genetic value = estimated breeding value

## 2.3 Accuracy of breeding value estimates ($r_{a,\hat{a}}$)

Estimated breeding values ($\hat{a}$) are estimates of the true breeding values ($a$), which cannot be observed directly. It is important to determine how well we have estimated the breeding value in relation to the true breeding value. This can be done using accuracy or reliability.

Accuracy is the correlation between the estimated and the true breeding value:

$$r_{a,\hat{a}} = \frac{\mathrm{Cov}(a,\hat{a})}{\sqrt{\mathrm{Var}(a)\,\mathrm{Var}(\hat{a})}} \tag{4}$$

A high correlation means that the estimated breeding value is very accurate. Reliability is the squared correlation, $r^2_{a,\hat{a}}$, between the estimated breeding value and the true breeding value.

To be able to compute the accuray or reliability of the estimated breeding value ($E(a|y)$) we need to determine the values for the terms, $Cov(a,\hat{a})$, $Var(\hat{a})$, and $Var(a)$ in the expression above. It can be shown that the variance of the estimated breeding value is the same as the covariance between the true and estimated breeding value (i.e. $Var(\hat{a}) = Cov(a,\hat{a})$). Therefore the reliability can be expressed as:

$$r^2_{a,\hat{a}} = \frac{\mathrm{Var}(\hat{a})}{\mathrm{Var}(a)} \tag{5}$$

Therefore reliability ($r^2_{a,\hat{a}}$) can be interpreted as the part of the genetic variation that we have explained by the estimated breeding values whereas the remainder (1-$r^2_{a,\hat{a}}$) is the uncertainty.

The variance of the estimated breeding value ($\hat{a}$) can be computed as:

$$\sigma^2_{\hat{a}} = r^2_{a,\hat{a}}\sigma^2_a$$

and from this expression it is clear that when the reliability increases, the variation in $\hat{a}$ increases (breeding value estimates are more variable over values of $y$), and the estimation becomes more precise (the residual variability of a over values of $y$, the part left to uncertainty, has decreased). If the reliability is 0, then we know nothing and the variance of $\hat{a}$ is 0. If the reliability is 1, then we know everything and the variance of $\hat{a}$ is $\sigma^2_{\hat{a}} = \sigma^2_a$ and the error variance (uncertainty) is 0. Reliability of the breeding value ($r^2_{a,\hat{a}}$) is important because it determines how well we can predict an individual's genetic value. It can be used to control the risk of a breeding plan: for example, low $r^2_{a,\hat{a}}$ leads to greater "risk" for both lower and higher true breeding value and we might consider more phenotypic records, in order to make better-informed selection decisions. Lastly reliability is one of the crucial factors that determines the genetic progress (e.g., breeders equation which will be introduced later in the course).

## 2.4 Prediction error variance (PEV) of estimated breeding values

Because every prediction is associated with an error, the same is true for the estimated breeding values $\hat{a}$. The variability of the error for the predicted breeding values are quantified by the prediction error variance (PEV). This is computed as:

$$Var(a-\hat{a}) = Var(a)(1-r^2_{a,\hat{a}}) \tag{6}$$

The standard error of prediction (SEP) can be a useful quantity. SEP corresponds just to the square root of PEV. Hence

$$\begin{aligned} SEP(\hat{a}) &= \sqrt{Var(a-\hat{a})} \\ &= \sqrt{Var(a)(1-r^2_{a,\hat{a}}))} \\ &= \sigma_a\sqrt{(1-r^2_{a,\hat{a}}))} \end{aligned}$$

# 3 Phenotypic data and genetic relationships are used to estimate breeding values

We will illustrate the basic principles of breeding value estimation using some simple examples where the trait has been measured on the individuals themselves or close relatives.

## 3.1 Estimation of breeding value and accuracy based on own phenotype:

An estimate of the breeding value (a) based on own phenotype (y) can be calculated as:

$$\begin{aligned} E(a|y) &= E(a) + Cov(a,y)[Var(y)]^{-1}(y - E(y)) \\ E(a|y) &= 0 + \sigma_a^2[\sigma_a^2 + \sigma_e^2]^{-1}(y - \mu) \\ E(a|y) &= h^2(y - \mu) \end{aligned}$$

Thus the estimated breeding value using own phenotypic record can be computed based on an estimate of the trait heritability ($h^2$) and the observed phenotype deviation ($y - \mu$). Use of records on the candidate itself is called performance testing. For performance testing to be efficient, the heritability should be at least moderately high (this can be derived from this equation: $E(a|y) = h^2(y - \mu)$).

The expression for expected value terms ($E(a)$ and $E(y)$) in the equation above are based on rules for expected value of a sum of (normally distributed) random variables:

$$\begin{aligned} \mathrm{E}(a) &= 0 \\ \mathrm{E}(e) &= 0 \\ \mathrm{E}(y) &= \mathrm{E}(\mu + a + e) \\ &= \mathrm{E}(\mu) + \mathrm{E}(a) + \mathrm{E}(e) \\ &= \mu + 0 + 0 \\ &= \mu \end{aligned}$$

The expression for (co)variance terms ($Var(y)$, and $Cov(a,y)$ ) in the equation above are based on rules for the variance of a sum of (normally distributed) random variables:

$$\begin{aligned} \mathrm{Var}(y) &= Var(a) + Var(e) + 2Cov(a,e) \\ \mathrm{Var}(a) &= \sigma_a^2 \\ \mathrm{Var}(e) &= \sigma_e^2 \\ \mathrm{Cov}(a,e) &= 0 \\ \mathrm{Var}(y) &= \sigma_a^2 + \sigma_e^2 \\ Cov(a,y) &= Cov(a, a+e) \\ &= Cov(a,a) + Cov(a,e) \\ &= \sigma_a^2 + 0 \\ &= \sigma_a^2 \end{aligned}$$

The accuracy for the breeding based on own phenotype (y) can be calculated as:

$$\begin{aligned} r_{a,\hat{a}} &= \frac{\mathrm{Cov}(a,\hat{a})}{\sqrt{\mathrm{Var}(a)}\sqrt{\mathrm{Var}(\hat{a})}} \\ r_{a,\hat{a}} &= \frac{(h^2)^2\sigma_y^2}{\sqrt{h^2\sigma_y^2}\sqrt{(h^2)^2\sigma_y^2}} \\ r_{a,\hat{a}} &= \sqrt{h^2} \end{aligned}$$

The variance of the estimated breeding value, $Var(\hat{a})$, can be expressed as:

$$Var(\hat{a}) = Var(h^2(y-\mu))$$
$$Var(\hat{a}) = (h^2)^2 Var(y-\mu) = (h^2)^2 Var(y) = (h^2)^2\sigma_y^2$$

The variance of the true breeding value, $Var(a)$, can be expressed by the heritability and phenotypic variance:

$$\sigma_a^2 = (\sigma_a^2)/(\sigma_y^2)\sigma_y^2 = h^2\sigma_y^2$$

Estimation of breeding value based on own phenotype is only possible when the trait in question can be measured (directly or indirectly) on the breeding individual, i.e., the candidate to be evaluated for selection. Sometimes this is not possible, e.g., traits that are sex-limited (milk production, female fertility, etc.) cannot be measured in male individuals. Traits like carcass composition and meat quality cannot be measured on live animals, unless an indirect method can be used (e.g. ultra-sonic measurement of carcass composition). In this situation it might be possible to use phenotypic information on relatives.

## 3.2 Estimation of breeding value and accuracy based on phenotypes of close relatives:

In practice we often use phenotypic records from close relatives, such as progenies, half-sibs, full-sibs, parents and grandparents. Phenotypes collected on close relatives (as compared to distant relatives) provide more information about the breeding value of an individual (as close relatives share more DNA in common). In the following we will provide a general formula for estimating breeding values and their accuracies using phenotypic information on different types of relatives.

**General formula for estimating breeding values using different sources of information:**

$$\hat{a} = b_{a|y}(y-\mu) \tag{7}$$

where the regression coefficient quantifies the weight (or importances) of the phenotypic information:

$$b_{a|y} = \frac{a' n h^2}{(1+(n-1)r} \tag{8}$$

where $a'$ is the genetic relationship between the breeding individual and individuals with phenotypes, $n$ is the number of phenotypic records, $h^2$ is the trait heritability, and $r$ is correlation between individuals with observations ($r = a^{''}h2 + c2$, where $a^{''}$ = genetic relationship between individuals with records and target, $c2$ = common environmental component).

Thus the importance given to a specific source of information depends on the additive genetic relationship ($a'$) with the breeding candidate, the heritability of the trait ($h^2$), and the amount of information ($n$), i.e. the number of progenies or sibs, etc.

**General formula for reliability of estimated breeding value using different sources of information:**

$$r_{a,\hat{a}}^2 = \frac{(a')^2 n h^2}{1+(n-1)r} \tag{9}$$

Thus reliability depends on the same factors as the estimated breeding value except for the phenotypic value. Although the reliability depends on the number of records it does not depend on the numerical value of phenotypes. From this formula is it clear that higher reliability (and accuracy) can be achived when:

- genetic relationship to individuals with information (a') is high
- there are many records (n)
- heritability ($h^2$) is high
- correlation between individuals with observations ($r = a''h^2 + c^2$) is low

## 3.3 Genetic relationship used for estimating breeding value

Related individuals share genes and thus resemble each other (have correlated phenotypes, to an extent that depends on additive genetic relationships). Consider a simple parent-offspring example. The offspring get half of the genes from each parent and therefore the breeding value for the offspring is the average of the parents' breeding values plus the Mendelian deviation (the part of the breeding value that is due to random segregation of the genes from each parent):

$$a_{\text{offspring}} = \frac{1}{2}a_{\text{father}} + \frac{1}{2}a_{\text{mother}} + a_{\text{mendelian}}$$

(a = additive genetic value = breeding value)

The term $a_{mendelian}$ is necessary, because two fullsibs $i$ and $j$ both having parents $father$ and $mother$ receive different random samples of the set of parental alleles. Hence the breeding values $a_i$ and $a_j$ of fullsibs $i$ and $j$ are not going to be the same. The Mendelian deviation reflects that random contribution of (Mendelian) segregation to breeding values of individuals.

In this equation the $\frac{1}{2}$ refers to the additive genetic relationship which in this example indicates that the offspring receives half of its genes from its parent. In general the weight given to a specific source of information depends on the additive genetic relationship with the candidate. Examples of different types of additive genetic relationships ($A_{ij}$) between the various sources (j) and the individual itself, i.e. the candidate to be evaluated (i), can be seen in the table below.

| Relative | $A_{ij}$ |
|---|---|
| Self | 1.0 |
| Unrelated | 0 |
| Mother | 0.5 |
| Father | 0.5 |
| Grandparent | 0.25 |
| Half-sib | 0.25 |
| Full-sib | 0.5 |
| Cousin | 0.0625 |
| Progeny | 0.5 |
| Twin(MZ/DZ) | 1/0.5 |

## 3.4 Estimation of breeding values using phenotypic information from multiple sources

Several factors influence which sources of information to use when estimating breeding values for a trait: what information is available, the heritability of the trait, and how and on what individuals the trait can be measured. Therefore in practice it is common to combine information from several sources. As already mentioned, all information available is usually utilized when an animal's breeding value is predicted. The weight given to a specific source of information depends on the additive genetic relationship with the candidate, the heritability and the amount of information, i.e. the number of progenies or sibs, etc.

- Using phenotypic records on progenies is generally the most accurate source of information for genetic evaluation (high genetic relationship $a'$ and high $n$). The average phenotypic value of a progeny group gives the best indication of the additive genetic value (i.e. the breeding value) of the candidate. The reliability (and accuracy) of breeding value estimates increases with the size of the progeny group. Progeny testing is useful also when the heritability is low. For example, it can be much more accurate than evaluation on own phenotype when the heritability is low (say, 0.1) and the progeny is large (~100-150). The disadvantage is that it takes resources (time and money) before results on progenies

are available. In animal breeding, progeny testing animals is often used for male animals as they usually get many more progenies than females, especially when artificial insemination is practiced.

- Phenotypic records on the candidate's sibs, half-sibs and full-sibs, are often used in addition to other information, or to give supplementary information, for example on traits that cannot be measured on the candidate itself. The accuracy of sib testing depends on the number of sibs that have records. Common-environment effects (e.g. full- sibs raised in the same herd) may bias the estimation of breeding values, unless we are able to adjust for them (e.g., by additional fixed effects in linear models).
- Parental information at different generations (parents, grandparents, etc.) is generally available even before the candidate is born, and can provide information very early. However, the genes from each locus of the parents are transmitted at random, so information based on pedigree alone is not very accurate, but can be valuable as additional information. Moreover, the additive genetic relationship, and thus the proportion of common genes between the candidate and the pedigree, is halved for every generation backwards (at least 0.5 for a parent, 0.25 for a grandparent, etc.). Finally, there is redundancy in the information provided by different generations of parents. Fore example, if there are accurate estimate of the parents breeding values then there is little to gain in using information on grandparents (actually, if the parents true breeding values are known, there would be no additional gain of information from grandparents).

As already mentioned, all information available is usually utilized when an individual's breeding value is predicted. The weight given to a specific source of information depends on the additive genetic relationship with the candidate, the heritability and the amount of information, i.e. the number of relatives (progenies, sibs, parents, etc.). In the following sections we will show how breeding values can be calculated when different types of phenotypic information (from different types of relatives) are available.

# 4 BLUP a general approach for estimation of breeding values

Breeding value estimation in animal and plant breeding programmes are nowadays based on the BLUP (Best Linear Unbiased Prediction) method. The estimation of breeding values based on multiple sources of informations must correct for the redundancy between them (e.g., the redundant information provided by parents and grandparents). Moreover, they need to be adjusted for average effects in the populations, "fixed effects". So far we have referred to that fixed effect as the population mean and we have assumed this adjustment $\mu$ to be known. Indeed, we defined the true breeding value $a$ and the non-identifiable environmental effects $e$ as deviations from a common mean, the average effect of all fixed genetic and environmental factors captured by the population mean $\mu$. But this is only true in a single idealized population where all selection candidates are kept in the same environment in which they deliver their performances at the same time. In practice the phenotypic records often need to be adjusted for systematic (fixed) environmental effects, such as age, parity, litter size, days open, sex, herd, year, season, management, etc. Several of those effects fluctuate very little over time, so accurate estimates of their effect may be obtained from previous ("historical") sets of data. Effects of factors like herd, year, season, and management fluctuate more and are therefore best estimated directly from the data to be used in the genetic evaluations.

Compared to the idealized cases described in the previous section, a practical breeding scenario poses two problems: accounting for heteregeneous sources of genetic information (different types of relatives); and adjusting for fixed effects in the breeding population(s) (fixed environmental or genetic effects). The BLUP solution to these problems was presented by Charles R. Henderson in several publications (e.g. Henderson1973a) and Henderson1975). The key idea behind the solution is to estimate the identifiable environmental factors as fixed effects and to predict the breeding values as random effects simultaneously in a linear mixed model. Here, mixed refers to the presence of two types of effects: fixed effects (identifiable effects from environmental or genetic factors) and random effects (non-identifiiable effects from segregating genetic factors and fluctuating environmental conditions). The methodology developed by Henderson is called **BLUP** and the properties of this methodology are directly incorporated into its name:

- **B** stands for **best** which means that the correlation between the true ($a$) and the predicted breeding value ($\hat{a}$) is maximal or the prediction error variance ($Var(a - \hat{a})$) is minimal.
- **L** stands for **linear** which means the predicted breeding values are linear functions of the observations ($y$)
- **U** stands for **unbiased** which means that the expected values of the predicted breeding values are equal to the true breeding values
- **P** stands for **prediction**

BLUP approaches are widely used in genetic evaluations, for both traditional predictions of breeding values and also for predicting genomic breeding values. The popularity of BLUP is not only due to the theoretical foundations behind BLUP, but also the efficient algorithms developed by Henderson for computing predicted breeding values, even in very large breeding populations. The theoretical foundations, and the development of efficient algorithms, together with the availability of large computational resources at a very low price, have made BLUP the de-facto standard for breeding value estimation.

## 4.1 Linear Mixed Effects Models

A simple linear model contains fixed effects such as *herd* or *sex* of an animal or location of a plant and tries to explain the observations as linear functions of such effects. Because the effects considered in a model cannot account for all factors influencing a given set of observations, every model must have a random residual component. If a linear model contains besides any additional random effects (e.g., random genetic effects at the level of families, or locations), then this model is called a **linear mixed effect model**.

### 4.1.1 Numeric Example

We want to use a concrete numeric example of a small population to explain how breeding values are predicted using the BLUP methodology. The phenotypic observations consist of measurements of the trait **weaning weight** in beef cattle. Table 2 gives an overview of the dataset.

Table 1: (#tab:Table 2)Example Data Set for Weaning Weight in Beef Cattle

| Animal | Sire | Dam | Herd | Weaning Weight |
|---|---|---|---|---|
| 12 | 1 | 4 | 1 | 2.61 |
| 13 | 1 | 4 | 1 | 2.31 |
| 14 | 1 | 5 | 1 | 2.44 |
| 15 | 1 | 5 | 1 | 2.41 |
| 16 | 1 | 6 | 2 | 2.51 |
| 17 | 1 | 6 | 2 | 2.55 |
| 18 | 1 | 7 | 2 | 2.14 |
| 19 | 1 | 7 | 2 | 2.61 |
| 20 | 2 | 8 | 1 | 2.34 |
| 21 | 2 | 8 | 1 | 1.99 |
| 22 | 2 | 9 | 1 | 3.10 |
| 23 | 2 | 9 | 1 | 2.81 |
| 24 | 2 | 10 | 2 | 2.14 |
| 25 | 2 | 10 | 2 | 2.41 |
| 26 | 3 | 11 | 2 | 2.54 |
| 27 | 3 | 11 | 2 | 3.16 |

We assume the phenotypic variance ($\sigma_p^2$) to be 0.1014 and the heritability ($h^2$) to be 0.25.

#### 4.1.1.1 Fixed Versus Random Effects

Unfortunately, there is no unique and generally accepted definition of which effects should be fixed and which should be random. There are generally accepted guidelines of how to classify effects as fixed or random. Certain factors such as herd, sex, breed or feeding regimes can be classified unambiguously as fixed effects. On the other hand, breeding values are always treated as random effects. Breeding values have an expected value (of 0) and have a certain variance (the additive genetic variance, to be estimated). They must therefore be modeled as random effects and these properties must be integrated into the linear mixed model. Furthermore, each individual has a different realization of a breeding value. Exceptions are mono-clonal twins and clones (e.g., vegetatively reproduced plants like potato or tree species).

#### 4.1.1.2 Model Specification

In a linear mixed effects model a single observation $y_{ijk}$ is decomposed according to equation (10)

$$y_{ijk} = b_i + a_j + e_{ijk} \tag{10}$$

where $b_i$ is the fixed effect of class i (specific herd, population, location, etc.)stands for the $i-^{th}$ level of a fixed effect, $a_j$ is the $j-th$ realization of the random effect $a$ and $e_{ijk}$ is the residual effect of the $k-^{th}$ observation}. To include all observations of a data set, it is helpful to represent the model in (10) by a matrix-vector notation. This is shown in equation (11)

$$y = Xb + Za + e \tag{11}$$

where

| | |
|---|---|
| $y$ | vector of length $n$ of all observations |
| $b$ | vector of length $p$ of all fixed effects |
| $X$ | $n \times p$ design matrix linking observations to fixed effects |
| $a$ | vector of length $n_a$ of random effects |
| $Z$ | $n \times n_a$ design matrix linking observations to random effects |
| $e$ | vector of length $n$ of random residual effects. |

Furthermore, we assume the following expected values and for the variances. As already mentioned the random effects are defined as deviations and hence their expected value is set to zero.

$$E(a) = 0$$
$$E(e) = 0$$

From this it follows that $E(y) = Xb$. The variance-covariance matrices for the random effects are set to

$$Var(a) = G$$
$$Var(e) = R$$

Under the assumption that $Cov(a, e^T) = 0$, we can compute $Var(y) = Z * Var(a) * Z^T + Var(e) = ZGZ^T + R = V$.

In model (11) the vectors $b$ and $a$ are unknown. They contain the fixed and random effects which must be estimated (or predicted). The solution of the model (11) for the unknowns $b$ and $a$ leads to estimates $\hat{b}$ for the fixed effects $b$ and for predicted random effects $\hat{a}$.

## 4.2 Estimating effects in the linear mixed model

The solutions to the effects $b$ and $a$ in the mixed model presented above are derived based on statistical techniques based on matrix algebra and multivariate normal theory. The details of these derivation are not important. Therefore, we are presenting here directly the results. For $\hat{a}$, the best linear unbiased prediction (BLUP) is:

$$\hat{a} = GZ^T V^{-1}(y - X\hat{b}) \tag{12}$$

We call $\hat{a}$ the best linear unbiased prediction of $a$ or shorter $\hat{a} = BLUP(a)$. For $\hat{b}$, we insert the generalized least squares estimator (GLS) which corresponds to

$$\hat{b} = (X^T V^{-1} X)^{-1} X^T V^{-1} y \tag{13}$$

The matrix $(X^T V^{-1} X)^{-1}$ denotes the inverse of the matrix $(X^T V^{-1} X)$. Analogously to $\hat{a}$, $\hat{b}$ is called the best linear unbiased estimator of the fixed effects $b$. In short, we can state $\hat{b} = BLUE(b)$.

## 4.3 Mixed Model Equations

The solutions shown in (12) for $\hat{a}$ and in (13) for $\hat{b}$ are not suitable for practical purposes. Both solutions contain the inverse $V^{-1}$ of matrix $V$. The matrix $V$ corresponds to the variance-covariance matrix of all observations $y$. The inverse matrix $V^{-1}$ is not easy to compute. Furthermore, procedures to invert general matrices are computationally expensive and are prone to rounding errors. In one of his many papers, Henderson has shown that the results for $\hat{a}$ and $\hat{b}$ are the same when solving the following system of equations simultaneously:

$$\begin{bmatrix} X^T R^{-1} X & X^T R^{-1} Z \\ Z^T R^{-1} X & Z^T R^{-1} Z + G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{a} \end{bmatrix} = \begin{bmatrix} X^T R^{-1} y \\ Z^T R^{-1} y \end{bmatrix} \tag{14}$$

The above shown equations are called **mixed model equations** (MME). They no longer contain the inverse $V^{-1}$ and hence these MME are much simpler to solve. Instead, the MME contain the inverses $R^{-1}$ and $G^{-1}$, which are easier to invert: R is often a very simple matrix, and G is usually smaller than (or the same size as) V. As a consequence, whenever we have to estimate breeding values using BLUP we will usually use the mixed model equations shown in (14).

## 4.4 BLUP breeding values are useful for ranking and selection

BLUP estimates of breeding values (EVBs), especially from the linear mixed model including all relationships, are useful tools in selection. Selection on BLUP breeding values maximizes the probability for correct ranking of breeding individuals and selection on them maximizes genetic gain from one generation to another. There are many factors that contribute to this:

- The linear mixed model which makes full use of information from all relatives increases accuracy (precision).
- The breeding values are adjusted for systematic environmental effects in an optimal way. This means that animals can also be compared across herds, age classes etc, assuming the data is connected.
- The procedure is flexible, various practical situations can be handled.
- Non-random mating can be accounted for.
- Several traits can be analyzed simultanously

- Bias due to culling within generation (e.g., between the 1st and 2nd lactations in dairy cattle) and selection (over generations) is accounted for, assuming that also non-selected animals' data are included in the analysis.

It should, however, be noted that the genetic evaluation is based on phenotypic observations, and that regardless of how splendid the BLUP procedure may be, it cannot compensate for bad data. So a good recording is necessary for a reliable genetic evaluation and subsequent genetic gain. It should also not be forgotten that BLUP assumes that the genetic parameters used are the true ones. In practice that means that EBVs will only be accurate if the estimated genetic parameters are close enough to their true value.
It should be noted that there is a potential risk for increased inbreeding when selection is based on information from all relatives. The probability that several family members are selected jointly is increased, which may result in increased inbreeding. To avoid this, and to optimize longterm selection response, selection on BLUP breeding values might be combined with some restriction on average relationship of the selected individuals.
A useful side effect of genetic evaluation by BLUP estimates is that it gives estimates of the realized genetic trend. This trend can be observed by comparing BLUP breeding values of individuals from different years.