Chapter 3

Genetic Evaluations

In chapter 2, we have already seen that the **breeding value** is a really important concept. The Definition 2.1 of the term breeding value has some important consequences.

- The breeding value is based on the average of a large number of offspring. This is necessary, because offspring inherit a random sample of a parents alleles. But not all offspring receive the same sample of alleles. Taking the average of a large number of offspring reduces the effect of sampling and thereby lets the breeding value converge to a stable value.
- The breeding value is defined as a deviation from the population mean. This population mean depends on allele frequencies which are specific for each population. Therefore breeding values can only be compared within one population.
- Because the breeding value is defined as a deviation, the expected value of the breeding values and the mean of all breeding values are 0 by definition.

3.1 Introduction

Because, in the more traditional setting[1] of livestock breeding, we do not have information about allele frequencies and about genotypic values, we have to predict breeding values. For this prediction we can use different sources of information. Currently, we are assuming that this information is all based on records of phenotypic observations.

[1] That means, at this moment, we are ignoring all recent developments made such as genomic selection.

3.1.1 The Basic Model

Although, the phenotypic observation might originate from different sources, we can use one basic model for all of the breeding value predictions. We have already seen a different form of this model in equation (1.1) in section 1.3.1. The original model from equation (1.1) is modified and extended to the model shown below.

$$y_{ij} = \mu_i + g_i + e_{ij} \tag{3.1}$$

where

y_{ij}	j^{th} record of animal i
μ_i	identifiable fixed environmental effect
g_i	sum of all additive (u), dominance (d) and epistatic effects of the genotype of animal i
e_{ij}	random environmental effects of animal i

Livestock species are mostly diploid and hence from a given parent only one allele of a given locus is passed to a gamete which can later be found in the parents offspring. Any interaction effects caused by dominance or epistasis are not preserved from parent to offspring. Only the additive effect of a given allele is passed from parent to offspring. The additive genetic part (u_i) of g_i in equation (3.1) represents the average genetic effect that animal i receives from its parents. It is therefore called the **breeding value**. Because the additive genetic effect is a function of the alleles passed from the parents to the progeny, it is the only component that can be selected for and is therefore the main component of interest from a livestock breeding perspective. Due to the major interest in the genetic additive component, the terms in the basic model in (3.1) are re-arranged as follows.

$$y_{ij} = \mu_i + u_i + e^*_{ij} \tag{3.2}$$

where

y_{ij}	j^{th} record of animal i
μ_i	identifiable fixed environmental effect
u_i	sum of all additive (u) genetic effects of the genotype of animal i
e^*_{ij}	dominance, epistatic and random environmental effects of the j^{th} record of animal i

The same re-arrangement of terms in the basic model is illustrated by Figure 3.1

Equation (3.2) constitutes the linear model that forms the basis for most problems of breeding value prediction in livestock breeding. Usually it is assumed that the phenotypic observations y_{ij} follow a multivariate normal distribution.

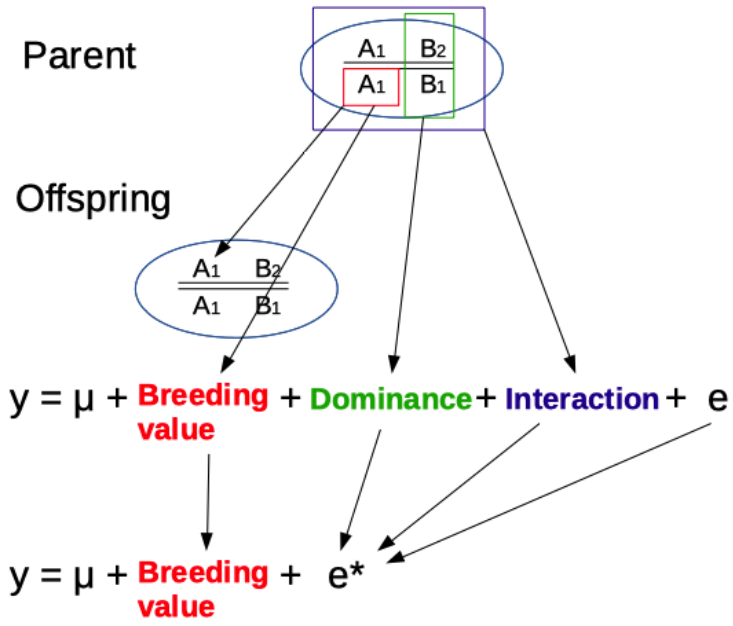

Figure 3.1: Re-arrangment of Terms Representing Genetic Effects

We have already seen in section 2.7 that the additive genetic effect (u_i) is thought to be the sum of a large number of unlinked loci that all contribute a very small amount to the total breeding value. Then by the central limit theorem it follows that u_i converges to a normal distribution. By the same reasoning that the environmental effect e^*_{ij} is composed of very many small contributions, also e^*_{ij} converges to a normal distribution. From distribution theory it is known that the sum of two normally distributed random variables (like u_i and e^*_{ij}) plus a fixed term (like μ) is again a random variable that follows a normal distribution. We can conclude that the assumption that all the random effects (y_{ij}, u_i and e^*_{ij}) in model (3.2) is consistent with distribution theory. Furthermore the central limit theorem implies that in principle the number of breeding values from single loci tends to infinity. That means the total breeding value u_i corresponds to a sum of infinitely many contributions. Based on the fact that in theory u_i is composed of an infinite number of infinitely small components, the model in (3.2) is called the **infinitesimal model**.

Concerning the variances, it is assumed that $var(y_{ij})$, $var(u_i)$ and $var(e_{ij})$ are all known. Covariances ($cov(u_i, e_{ij})$) between genetic and environmental effects and covariances ($cov(e^*_{ij}, e^*_{kl})$) between environmental effects of mates i and k are assumed to be zero, respectively.

Also μ_i which is used to represent the mean performance of animals in the same identifiable environment such as herd or management group or have the same sex or age, is assumed to be known.

3.1.2 Decomposition of Breeding Value

As already mentioned earlier, the breeding value u_i of an individual i represents the average additive genetic effect that animal i receives from its parents s and d. Hence u_i can be decomposed into

$$u_i = \frac{1}{2}u_s + \frac{1}{2}u_d + m_i \tag{3.3}$$

where u_s and u_d correspond to the breeding values of parents s and d, respectively and m_i is the deviation of u_i from the average breeding values of the parents and is called **Mendelian sampling**. The term m_i is necessary, because two fullsibs i and k both having parents s and d receive different random samples of the set of parental alleles. Hence the breeding values u_i and u_k of fullsibs i and k are not going to be the same. The difference between breeding values u_i and u_k is reflected in the different Mendelian sampling terms m_i and m_k for fullsibs i and k.

3.2 Basic Principle of Predicting Breeding Values

The prediction of breeding values mostly follows the same principles. From the point of view of statistics, estimations or predictions are always a function of the observed data. When looking at the model in (3.2), we can probably guess that the observed phenotypic records (y_{ij}) must be corrected somehow for the identifiable environmental effects represented by μ_i. The second influence that we want to consider when predicting breeding values is how "closely related" the observed record y_{ij} is to the breeding value. For traits where the influence of the genetic component is not very strong, it is probably a good idea to down-weigh the information from y_{ij}.

The two principles just described can be generalized as follows. Breeding values are predicted according to the following two steps.

1. Observations are corrected for the mean performance values of animals under the same environmental conditions. The conditions are described by the effects captured in μ_i.
2. The corrected observations are weighted by a factor that reflects the amount of information that is available for the prediction of an animals breeding value.

In what follows, we have a look at how breeding values are predicted from different sources of information.

3.3 Animal's Own Performance

3.3.1 Single Record

When one phenotypic observation per animal is the only information we have available, the predictor $\widehat{u}_i$ of the breeding value u_i of animal i can be derived according to the following line of argument. Let us assume for a moment that we know the true breeding value u_i for a population of animals. In addition to that each animal i has one observation y_i available. Then we plot the values of u_i against the values of y_i for the complete population.

The plot in Figure 3.2 suggests that we fit a regression of the breeding values onto the phenotypic records. The fitted regression is represented by the red line. Hence as soon as we can draw the regression line, we can predict breeding values based on the phenotypic observations. The predicted breeding value $\widehat{u}_i$ for a given y_i corresponds to the value on the red line corresponding to the value of y_i. The slope of the regression line corresponds to the regression coefficient b. From regression theory, the coefficient b is computed as

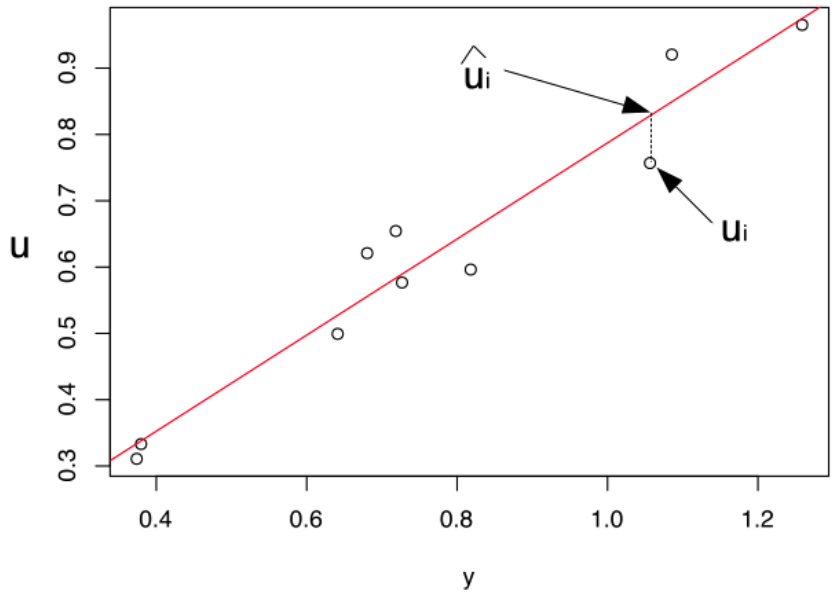

Figure 3.2: Regression of Breeding Values onto Phenotypic Observations

$$\begin{aligned} b &= \frac{cov(u,y)}{var(y)} \\ &= \frac{cov(u,\mu+u+e)}{var(y)} \\ &= \frac{cov(u,u)}{var(y)} \\ &= \frac{var(u)}{var(y)} \\ &= h^2 \end{aligned} \tag{3.4}$$

where h^2 corresponds to the ratio between the genetic additive and the phenotypic variance and is called **heritability**. We are using the regression coefficient to predict the breeding value for animal i based on a single record y_i.

$$\begin{aligned} \widehat{u_i} &= b * (y_i - \mu) \\ &= h^2 * (y_i - \mu) \end{aligned} \tag{3.5}$$

From that it follows that the predicted breeding value for an animal based on a single own performance record corresponds to the observation corrected for the general mean μ times the heritability. The correlation between the selection criterion, in our case the phenotypic record and the true breeding value is known as the accuracy of the prediction. It provides a means of evaluating the different selection criteria. The higher the correlation between selection criterion and breeding value, the better is the prediction. Sometimes the accuracy of evaluation is reported in terms of reliability (r^2) which corresponds to the squared correlation between selection criterion and true breeding value. With a single own performance record per animal, the correlation is

$$\begin{aligned} r_{u,y} &= \frac{cov(u,y)}{\sigma_u \ \sigma_y} \\ &= \frac{\sigma_u^2}{\sigma_u \ \sigma_y} \\ &= \frac{\sigma_u}{\sigma_y} \\ &= h \end{aligned} \tag{3.6}$$

An alternative way to assess the quality of the breeding value prediction is to compute the variance of the predicted breeding values.

$$\begin{aligned} var(\widehat{u_i}) &= var(by) = var(h^2 y) \\ &= h^4 var(y) \\ &= r^2_{u,y} h^2 \sigma^2_y \\ &= r^2_{u,y} \sigma^2_a \end{aligned} \tag{3.7}$$

Hence the variance of the predicted breeding values corresponds to the product of the reliability times the genetic additive variance. The expected response (R) to selection on the basis of one record per animal is

$$R = i * r^2_{u,y} * \sigma_y = i * h^2 * \sigma_y \tag{3.8}$$

where i, the selection intensity refers to the superiority of selected individuals above population mean expressed in phenotypic standard deviation.

3.3.2 Repeated Records

When animals get older, it is likely that we can observe multiple measurements for the same trait. An example is milk yield in dairy cows where a cow might have repeated lactation records. The breeding value of an animal may be predicted based on the mean of the repeated records. With repeated records, an additional resemblance between the records of an animal due to permanent environmental factors occurs. The between-animal variance is partly genetic and partly caused by permanent environmental effects. The within-animal variance is attributed to differences between the successive measurements of the animal arising from temporary environmental variations, i.e., from environmental factors that change over time. The variance of observations ($var(y)$) can therefore be partitioned as

$$var(y) = var(g) + var(pe) + var(te) \tag{3.9}$$

where $var(g)$ is the genetic variance, $var(pe)$ the variance due to permanent environmental effects and $var(te)$ the variance due to temporary environmental effects.

The intra-class correlation t is defined as the ratio of the genetic plus the permanent environmental variance divided by the phenotypic variance.

$$t = \frac{var(g) + var(pe)}{var(y)} \tag{3.10}$$

The term t is also called **repeatability** and it measures the correlation between the records of an individual. From (3.10) it follows that

$$1 - t = \frac{var(te)}{var(y)} \tag{3.11}$$

With this model, it is assumed that the repeated records on the animal are the same trait. Therefore the genetic correlation between all pairs of records is one. We also assume that all records have equal variance and that the environmental correlations between all pairs of records are equal. Let $\tilde{y}$ represent the mean of n records on animal i which means

$$\tilde{y_i} = \frac{1}{n} \sum_{k=1}^{n} y_{ik} \tag{3.12}$$

In this case, we use the mean ($\tilde{y_i}$) to predict the breeding value ($\hat{u_i}$)

$$\hat{u_i} = b(\tilde{y_i} - \mu) \tag{3.13}$$

where

$$b = \frac{cov(u, \tilde{y})}{var(\tilde{y})} \tag{3.14}$$

The single elements are computed as

$$cov(u, \tilde{y}) = cov(u, g + pe + \frac{1}{n} \sum_{k=1}^{n} te_k) = \sigma_a^2 \tag{3.15}$$

and

$$var(\tilde{y}) = var(g) + var(pe) + \frac{1}{n} var(te) \tag{3.16}$$

Expressing (3.16) in terms of (3.10) and (3.11) leads to

$$\begin{aligned} var(\tilde{y}) &= t * \sigma_y^2 + \frac{1}{n}(1 - t) * \sigma_y^2 \\ &= \frac{1}{n} \left(n * t + (1 - t)\right) \sigma_y^2 \\ &= \frac{1 + (n - 1)t}{n} \sigma_y^2 \end{aligned} \tag{3.17}$$

Inserting this into (3.14) results in

$$\begin{aligned} b &= \frac{cov(u, \tilde{y})}{var(\tilde{y})} \\ &= \frac{n\sigma_u^2}{(1 + (n-1)t)\sigma_y^2} \\ &= \frac{nh^2}{1 + (n-1)t} \end{aligned} \tag{3.18}$$

When we predict the breeding value u_i of animal i using repeated records, the regression coefficient b depends on

1. the heritability (h^2)
2. the repeatability (t) and
3. the number (n) of repeated records per animal

The difference between repeated records of an animal is assumed to be due to temporary environmental differences between successive performances. However, if successive records are known to be affected by factors which influence performance, these must be corrected for. For instance, differences in age at calving in first and second lactations may influence milk yield in first and second lactation. Such age differences should be adjusted for before using the means of both lactations for breeding value prediction.

The accuracies of the predicted breeding value using repeated records is

$$\begin{aligned} r_{u,\tilde{y}} &= \frac{cov(u, \tilde{y})}{\sigma_u \sigma_y} \\ &= \frac{\sigma_u^2}{\sigma_u \sqrt{(1 + (n-1)t)/n\sigma_y^2}} \\ &= h\sqrt{n/(1 + (n-1)t)} \\ &= \sqrt{nh^2/(1 + (n-1)t)} \\ &= \sqrt{b} \end{aligned} \tag{3.19}$$

The expected response to selection using repeated records will be covered in an exercise.

3.4 Progeny Records

For traits that are recorded only on female animals, the prediction of breeding values for male animals (sires) is usually based on the mean of their female progeny. This is typical in dairy cattle, where bulls are evaluated on the basis

of their daughters. Let $\bar{y_i}$ be the mean of single records of n daughters of sire i with the assumption that the daughters are only related through the sire (paternal half-sibs), the predicted breeding value of sire i can then be computed as

$$\hat{u_i} = b * (\bar{y_i} - \mu) \tag{3.20}$$

where

$$b = \frac{cov(u, \bar{y})}{var(\bar{y})} \tag{3.21}$$

In the current case of using progeny records to predict a breeding value, we have

$$\begin{aligned} cov(u, \bar{y}) &= cov(u, \frac{1}{2}u_s + \frac{1}{2}u_d + \frac{1}{n}\sum_{k=1}^{n} m_k + \frac{1}{n}\sum_{k=1}^{n} e_k) \\ &= cov(u, \frac{1}{2}u_s) \\ &= \frac{1}{2}cov(u, u_s) \\ &= \frac{1}{2}\sigma_u^2 \end{aligned} \tag{3.22}$$

where u_s and u_d denote the breeding values of sire s and dam d, respectively and m_k and e_k stand for the mendelian sampling and the environmental effect of daughter k. Using the same principles as in section 3.3.2, we get

$$var(\bar{y}) = (t + (1 - t)/n)\sigma_y^2 \tag{3.23}$$

assuming there is no environmental covariance between half-sib records and the intra-class correlation t is $\frac{1/4\sigma_u^2}{\sigma_y^2}$. Then we can compute the regression coefficient as

$$
\begin{aligned}
b &= \frac{1/2\sigma_u^2}{(t+(1-t)/n)\sigma_y^2} \\
&= \frac{1/2h^2\sigma_y^2}{(\frac{1}{4}h^2+(1-\frac{1}{4}h^2)/n)\sigma_y^2} \\
&= \frac{2nh^2}{nh^2+(4-h^2)} \\
&= \frac{2n}{n+(4-h^2)/h^2} \\
&= \frac{2n}{n+k} \qquad (3.24)
\end{aligned}
$$

with $k = \frac{4-h^2}{h^2}$.

The term k is constant for any assumed heritability (h^2). The regression coefficient (b) depends on the heritability and number of progeny and converges towards a limit of 2 as the number of daughters increases.

The accuracy of the estimated breeding value is

$$
\begin{aligned}
r_{u,\bar{y}} &= \frac{cov(u,\bar{y})}{\sqrt{var(u)var(\bar{y})}} \\
&= \frac{1/2h^2\sigma_y^2}{\sqrt{h^2\sigma_y^2(\frac{1}{4}h^2+(1-\frac{1}{4}h^2)/n)\sigma_y^2}} \\
&= \frac{1/2h}{\sqrt{\frac{1}{4}h^2+(1-\frac{1}{4}h^2)/n}} \\
&= \sqrt{\frac{nh^2}{nh^2+(4-h^2)}} \\
&= \sqrt{\frac{n}{n+k}} \qquad (3.25)
\end{aligned}
$$

The term for $r_{u,\bar{y}}$ in (3.25) approaches 1 as the number of progeny increases. The reliability ($r_{u,\bar{y}}$) of the predicted breeding value is $n/(n+k)$ and corresponds to $1/2 * b$ computed in (3.24).