### 5.3.3 Animal Model

An extension of the sire model is called **animal model**. In an animal model not only the sires get predicted breeding values but all animals in the pedigree will be assigned a predicted breeding value. That is only possible, if we extend our dataset by the column of the dam of each animal for which we have observations. That leads to the following table.

Table 5.9: Pre-weaning Gain in kg for five beef animals

| Animal | Sire | Dam | Sex | WWG |
|---|---|---|---|---|
| 4 | 1 | NA | M | 4.5 |
| 5 | 3 | 2 | F | 2.9 |
| 6 | 1 | 2 | F | 3.9 |
| 7 | 4 | 5 | M | 3.5 |
| 8 | 3 | 6 | M | 5.0 |

In Table 5.9 the Dam of animal 4 is noted as `NA` which stands for `not available`. This means that the dam of animal 4 is not known.

Because in an animal model all animals in the pedigree will get a predicted breeding value, we write the vector of breeding values as $\mathbf{u}$ and the complete animal model can be written as follows

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e} \tag{5.17}$$

All other components have the same meaning as in the sire model. For the vector $\mathbf{u}$, we also have to define the expeted value ($E(\mathbf{u})$) and the variance-covariance matrix $var(\mathbf{u})$. Breeding values, by their very nature, are deviations from a population mean. This leads to the definition of the expected value of $\mathbf{u}$ to be $E(\mathbf{u}) = \mathbf{0}$. The variance-covariance matrix $var(\mathbf{u})$ of the vector of breeding values $\mathbf{u}$ is similar to the sire model given by the product of a matrix $\mathbf{A}$ and the variance component $\sigma_u^2$. The matrix $\mathbf{A}$ is called numerator relationship matrix. The diagonal elements $(\mathbf{A})_{ii}$ of the matrix $\mathbf{A}$ are computed as

$$(\mathbf{A})_{ii} = 1 + F_i \tag{5.18}$$

where $F_i$ is the inbreeding coefficient of animal $i$ which corresponds to half of the relationship coefficient of the parents $s$ and $d$ of animal $i$. As a formula this can be written as

$$F_i = \frac{1}{2} * (\mathbf{A})_{sd} \tag{5.19}$$

The offdiagonal elements of $\mathbf{A}$ are the proportionality constants which together with $\sigma_u^2$ form the covariance of the breeding values of two animals. If we look at two animals $i$ and $j$, the covariance $cov(u_i, u_j)$ can be written as

$$cov(u_i, u_j) = (\mathbf{A})_{ij} * \sigma_u^2 \tag{5.20}$$

The coefficients $(\mathbf{A})_{ij}$ can be determined by decomposing both breeding values $u_i$ and $u_j$ recursively into the breeding of their parents until some common ancestors are found in the pedigree. Based on this decomposition, the covariance $cov(u_i, u_j)$ and with that the coefficient $(\mathbf{A})_{ij}$ can be computed. If no common ancestors of $i$ and $j$ can be found in the pedigree the covariance $cov(u_i, u_j)$ is zero.

The example dataset shown in Table 5.9 cannot be analysed with the package `pedigreemm`. The problem is that `pedigreemm` does not allow to specify given variance components, but it wants to estimate the variance components from the dataset specified. In the small dataset with only one observation per animal, `pedigreemm` cannot estimate both variance components $\sigma_e^2$ and $\sigma_u^2$.

But as already shown with the sire model, it is possible to get estimates of the fixed effects and predicted breeding values for all animals using the solutions to the following mixed model equations.

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \lambda * A^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix} \tag{5.21}$$

where $\lambda = \sigma_e^2/\sigma_u^2$. For our example the matrix $X$ is the same as for the sire model. The matrix $Z$ is defined as

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

The inverse $\mathbf{A}^1$ of the numerator relationship matrix can be computed by the function `pedigreemm::getAinv()`

```
library(pedigreemm)
ped_ani <- pedigree(sire = c(rep(NA, n_nr_founder),1,3,1,4,3),
                    dam  = c(rep(NA, n_nr_founder),NA,2,2,5,6),
                    label = as.character(1:n_nr_animal))
mat_Ainv <- getAInv(ped = ped_ani)
```

$$\mathbf{A}^1 = \begin{bmatrix} 1.8333 & 0.5 & 0 & -0.6667 & 0 & -1 & 0 & 0 \\ 0.5 & 2 & 0.5 & 0 & -1 & -1 & 0 & 0 \\ 0 & 0.5 & 2 & 0 & -1 & 0.5 & 0 & -1 \\ -0.6667 & 0 & 0 & 1.8333 & 0.5 & 0 & -1 & 0 \\ 0 & -1 & -1 & 0.5 & 2.5 & 0 & -1 & 0 \\ -1 & -1 & 0.5 & 0 & 0 & 2.5 & 0 & -1 \\ 0 & 0 & 0 & -1 & -1 & 0 & 2 & 0 \\ 0 & 0 & -1 & 0 & 0 & -1 & 0 & 2 \end{bmatrix}$$

Assuming that variance components were estimated from a different dataset, the following values can be used, $\sigma_u^2 = 20$ and $\sigma_e^2 = 40$. With all this information, mixed model equations can be solved.

For the fixed effects we get

Table 5.10: Solutions for fixed Effect of Sex

| Sex | Solution |
|---|---|
| F | 3.404430 |
| M | 4.358502 |

For the random animal breeding values, we get

Table 5.11: Solutions for random breeding values of all animals

| Animal | Solution |
|---|---|
| 1 | 0.0984446 |
| 2 | -0.0187701 |
| 3 | -0.0410842 |
| 4 | -0.0086631 |
| 5 | -0.1857321 |
| 6 | 0.1768721 |
| 7 | -0.2494586 |
| 8 | 0.1826147 |

Comparing the order of the breeding values of sires 1, 3 and 4, it can be seen that they are not the same for the sire model and the animal model. Although, it has to be noted that the differences are small, but the fact that in the animal model all available information are considered for the prediction of the breeding values, can make a difference when it comes to the ranking of animals as potential parents according to their predicted breeding values.

## 5.4 Genomic BLUP

Prediction of genomic breeding values can be done with two different modelling approaches.

1. Marker effect models (MEM): Linear mixed effects models with marker effects as random effects
2. Breeding-value based models (BVM): Genomic breeding values as random effects

### 5.4.1 Marker Effect Models

In MEM random effects of markers are directly included in the model. For an idealized data set we can write

$$y = 1_n\mu + Wq + e \tag{5.22}$$

where

| | |
|---|---|
| $y$ | vector of length $n$ with observations |
| $\mu$ | general mean denoting fixed effects |
| $1_n$ | vector of length $n$ of all ones |
| $q$ | vector of length $m$ of random SNP effects |
| $W$ | design matrix relating SNP-genotypes to observations |
| $e$ | vector of length $n$ of random error terms |

The vector $q$ contains a separate random effect for each SNP. Because the SNP effects are random, the expected value $E[q]$ and the variance $var(q)$ must be specified. In general, the random effects are defined as deviations and hence their expected value is 0. This means $E[q] = 0$. The variance explained by each SNP corresponds to $\sigma_q^2$ and is assumed to be constant. The variance $var(e)$ of the random error terms is taken to be $var(e) = I * \sigma_e^2$ where $I$ is the identity matrix and $\sigma_e^2$ is the error variance.

The random marker effects can be predicted using the following mixed model equations.

$$\begin{bmatrix} 1_n^T 1_n & 1_n^T W \\ W^T 1_n & W^T W + \lambda_q * I \end{bmatrix} \begin{bmatrix} \hat{\mu} \\ \hat{q} \end{bmatrix} = \begin{bmatrix} 1_n^T y \\ W^T y \end{bmatrix} \tag{5.23}$$

with $\lambda_q = \sigma_e^2/\sigma_q^2$.

The genomic breeding value for a given animal $i$ with given genotypes at all SNP-marker positions is computed by summing over the appropriate predicted marker effects solutions $\hat{q}$ determined by the genotypes of animal $i$.

### 5.4.2 Breeding Value Models

In a breeding value model a linear combination of all SNP effects are combined into a random genomic breeding value. This approach is meant when animal breeders are talking about Genomic BLUP (GBLUP). The mixed linear effects model in GBLUP corresponds to

$$y = Xb + Zg + e \tag{5.24}$$

where

| | |
|---|---|
| $y$ | vector of length $n$ with observations |
| $b$ | vector of length $r$ with fixed effects |
| $X$ | incidence matrix linking elements in $b$ to observations |
| $g$ | vector of length $t$ with random genomic breeding values |
| $Z$ | incidence matrix linking elements in $g$ to observations |
| $e$ | vector of length $n$ of random error terms |

The vector $g$ contains the genetic effects of all animals that are genotyped which means that they have genomic information based on SNP genotypes available. The expected values of all random effects is assumed to be 0. The variance $var(g)$ of the random genomic breeding values is given by $var(g) = G * \sigma_g^2$. This expression looks very similar to the variance of the breeding values in the traditional BLUP animal model. The matrix $G$ is called **genomic relationship matrix** (GRM). The variance $var(e)$ of the random error terms is given by $var(e) = I * \sigma_e^2$.

Mostly the older animals for which SNP information is available may have observations ($y$) in the dataset. The younger animals may have SNP information but in most cases no information is available for them. The goal of GBLUP is to predict genomic breeding values for these animals. Depending on the number of genotyped animals which is in most cases smaller compared to the number of SNP loci, the BVM model has the following advantages over the MEM model

1. The length of the vector $g$ is $t$ which corresponds to the number of genotyped animals which in most cases is smaller than the length of the vector $q$ which is $m$ corresponding to the number of SNPs.
2. Accuracies of genomic breeding values can be computed analogously to the traditional BLUP animal model. This is analogy of accuracies does not exist in MEM.
3. BVM can be combined with pedigree-based animal model analysis which is then referred to as **single step** approach.

More recently with the number of genotyped animals growing very fast, these advantages are no longer as important as they used to be.

Genomic breeding values from a BVM can be predicted by solving the following mixed model equations.

$$\begin{bmatrix} X^TX & X^TZ \\ Z^TX & Z^TZ + \lambda_g * G^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \end{bmatrix} \tag{5.25}$$

with $\lambda_g = \sigma_e^2/\sigma_g^2$.

## 5.5 Genomic Relationship Matrix

The variance-covariance matrix between the genetic effects $g$ in model (5.24) is proportional to the genomic relationship matrix $G$. Analogously to the traditional BLUP animal model where the variance-covariance matrix of the random breeding values is proportional to the numerator relationship matrix $A$.

### 5.5.1 Derivation of $G$

Because the traditional pedigree-based BLUP animal model is very well respected in animal breeding and the defined model (5.24) produces an analogy of the genomic evaluation model to the already known animal model the following properties of $g$ and the genomic relationship matrix $G$ are essential.

1. The genetic effects $g$ should correspond to a linear combination of the single SNP-effects $q$
2. The genetic effects $g$ should be defined as deviations from a common mean, leading to the expected value $E[g] = 0$.
3. The variance-covariance matrix of the vector $g$ corresponds to the product of $G$ times a common variance component $\sigma_g^2$.
4. The genomic relationship matrix $G$ should be similar to the numerator relationship matrix $A$. The diagonal elements should be close to 1 and off-diagonal elements of animals that are related should have higher values than elements between unrelated animals.

The matrix $G$ can be computed based on SNP genotypes. In what follows the material of [VanRaden, 2008] and [Gianola et al., 2009] is used to derive the genomic relationship matrix.

### 5.5.2 Linear Combination of SNP Effects

Based on the SNP marker information the marker effects in the vector $q$ can be estimated. Hence, we assume that the vector $q$ is known. The property that $g$ should be a linear combination of the effects in $q$ means that there exists a matrix $U$ for which we can write

$$g = U \cdot q \tag{5.26}$$

The matrix $U$ is determined based on the desired properties described above.

### 5.5.3 Deviation

The genetic effects $g$ should be defined as deviation from a common basis. Due to this definition the expected value of the genetic effect is determined by $E[g] = 0$. This requirement has the following consequences for the matrix $U$.

Let us have a look at the random variable $w$ which takes the SNP-genotype codes in the matrix $W$ in the MEM model given in (5.22). Let us further assume that the SNP loci are in Hardy-Weinberg equilibrium. Then $w$ can take the following values

$$w = \begin{cases} -1 & \text{with probability} \quad (1-p)^2 \\ 0 & \text{with probability} \quad 2p(1-p) \\ 1 & \text{with probability} \quad p^2 \end{cases} \tag{5.27}$$

The expected value of $w$ corresponds to

$$E[w] = (-1)*(1-p)^2+0*2p(1-p)+1*p^2 = -1+2p-p^2+p^2 = 2p-1 \tag{5.28}$$

The matrix $U$ is computed as the difference between the matrix $W$ and the matrix $P$ where the matrix $P$ corresponds to column vectors which have elements corresponding to $2p_j - 1$ where $p_j$ corresponds to the allele frequency of the positive allele at SNP locus $j$. The following table gives an overview of the elements of matrix $U$ for the different genotypes at SNP locus $j$.

| Genotype | Genotypic Value | Coding in Matrix $U$ |
|---|---|---|
| $(G_2G_2)_j$ | $-2p_jq_j$ | $-1-2(p_j-0.5) = -2p_j$ |
| $(G_1G_2)_j$ | $(1-2p_j)q_j$ | $-2(p_j-0.5) = 1-2p_j$ |
| $(G_1G_1)_j$ | $(2-2p_j)q_j$ | $1-2(p_j-0.5) = 2-2p_j$ |

Here we assume that for a locus $G_j$, the allele $(G_1)_j$ has a positive effect and occurs with frequency $p_j$. We can now verify that with this definition of $U$, the expected value for a genetic effect determined by the locus $j$ corresponds to

$$\begin{aligned} E[g]_j &= \left[(1-p_j)^2 * (-2p_j) + 2p_j(1-p_j)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= 0 \end{aligned} \tag{5.29}$$

### 5.5.4 Variance of Genetic Effects

As already postulated the variance-covariance matrix of the genetic effects should be proportional to the genomic relationship matrix $G$.

$$var(g) = G * \sigma_g^2 \tag{5.30}$$

Computing the same variance-covariance matrix based on equation (5.26)

$$var(g) = U \cdot var(q) \cdot U^T \tag{5.31}$$

The variance-covariance matrix of the SNP effects is $var(q) = I * \sigma_q^2$. Inserting this into (5.31) we get $var(g) = UU^T\sigma_q^2$.

In [Gianola et al., 2009] the variance component $\sigma_g^2$ was derived from $\sigma_q^2$ leading to

$$\sigma_g^2 = 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 \tag{5.32}$$

Now we combine all relationships for $var(g)$ leading to

$$var(g) = G * \sigma_g^2 = UU^T\sigma_q^2 \tag{5.33}$$

In (5.33), $\sigma_g^2$ is replaced by the result of (5.32).

$$G * 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 = UU^T\sigma_q^2 \tag{5.34}$$

Dividing both sides of (5.34) by $\sigma_q^2$ and solving for $G$ gives us a formula for the genomic relationship matrix $G$

$$G = \frac{UU^T}{2\sum_{j=1}^{m} p_j(1-p_j)} \tag{5.35}$$

## 5.6 How Does GBLUP Work

The genomic relationship matrix $G$ allows to predict genomic breeding values for animals with SNP-Genotypes without any observation in the dataset. This fact is the basis of the large benefit of genomic selection. As soon as a young animal is born, its SNP genotypes can be determined and a genomic breeding value can be predicted. This genomic breeding value is much more accurate then the traditional breeding value based only on ancestral information.

The BVM model given in (5.24) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (5.36). In this form the Inverse $G^{-1}$ of $G$ and the vector $\hat{g}$ of predicted genotypic breeding values are split into one part corresponding to the animals with observations and a second part for the animals without phenotypic information.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g}_1 \\ \hat{g}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{5.36}$$

The matrix $G^{(11)}$ denotes the part of $G^{-1}$ corresponding to the animals with phenotypic observations. Similarly, $G^{(22)}$ stands for the part of the animals without genotypic observations. The matrices $G^{(12)}$ and $G^{(21)}$ are the parts of $G^{-1}$ which link the two groups of animals. The same partitioning holds for the vector of predicted breeding values. The vector $\hat{g}_1$ contains the predicted breeding values for the animals with observations and the vector $\hat{g}_2$ contains the predicted breeding values of all animals without phenotypic observations.

Based on the last line of (5.36) the predicted breeding values $\hat{g}_2$ of all animals without phenotypic observations can be computed from the predicted breeding values $\hat{g}_1$ from the animals with observations.

$$\hat{g}_2 = -\left(G^{22}\right)^{-1} G^{21}\hat{g}_1 \tag{5.37}$$

Equation (5.37) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations.

## 5.7 Single Step Genomic BLUP

In real-world livestock breeding datasets not all animals are genotyped. But we want to have predicted breeding values for all animals in a population. Futhermore, the genomic information of the genotyped animals should also give more accurate predicted breeding values for related animals without genomic information.

The single step genomic BLUP model can be specified as

$$y = Xb + Zg + e \tag{5.38}$$

with $var(g) = H * \sigma_g^2$ and $var(e) = I * \sigma_e^2$. At this point it is important to note that the vector $g$ of genomic breeding values can be split into two parts

$$g = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$$

where $g_1$ is the vector of breeding values for non-genotyped animals and $g_2$ is the vector of genotyped animals.

$$\begin{bmatrix} X^T X & X^T Z \\ Z^T X & Z^T Z + \lambda * H^{-1} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{g} \end{bmatrix} = \begin{bmatrix} X^T y \\ Z^T y \end{bmatrix} \tag{5.39}$$

where here $\lambda = \sigma_e^2/\sigma_g^2$.

The above required inverse matrix $H^{-1}$ can be shown (e.g. in [Legarra et al., 2014]) to correspond to

$$H^{-1} = A^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{pmatrix}$$

where $A^{-1}$ is the inverse numerator relationship matrix and $A_2 2$ corresponds to the part of the numerator relationship matrix containing all genotyped animals.