4.1.3 Random Effects Models

Before, we introduce the mixed linear effects model, we first have a look at how the repeated measurements data can be modelled with a random effects model. For the demonstration of the random effects model, we use the dataset in Table 4.3, but we are ignoring the factor `Breed` for a moment. Then this dataset just looks like a repeated measurement of the body weight of some beef cattle animals (see Table 4.4).

Table 4.4: Repeated Measurements of Body Weight for Beef Cattle Animals

Animal	Body Weight
2	463.0000
2	468.8940
2	467.8753
5	496.0000
5	495.0033
5	493.6563
7	518.0000
7	509.3221
7	506.5958
10	541.0000
10	547.3609
10	533.9288

In a random effects model with repeated observations, the expected value $E(y_{ij})$ for body weight y_{ij} of animal i with the j^{th} observation can be written as

$$E(y_{ij}) = \mu + \alpha_i \tag{4.2}$$

Algebraically the expression for $E(y_{ij})$ given in (4.2) is not different from what we have seen for the fixed linear effects model in chapter 3. But the assumptions are different. In (4.2), α_i is the effect of animal i on the observed body weight. Because the animals in the dataset (Table 4.4) is a random sample of a large population of animals, the effect α_i is a so-called **random effect**. A random effect in a model is to be treated as a random variable for which, we have to specify its distributional properties such as expected value and variance. For our example of the repeated measurements data, we assume the following three properties for the α_i effects

1. they are indepentently and identically distributed (i.i.d.)
2. they all have expected value of 0, $E(\alpha_i) = 0 \quad \forall i$

3. they all have the same variance σ_{α}^2, $var(\alpha_i) = E\left[\alpha_i - E(\alpha_i)\right]^2 = E(\alpha_i^2) = \sigma_{\alpha}^2$ with $cov(\alpha_i, \alpha_k) = 0 \quad \forall i \neq k$

A further consequence of choosing α_i as a random effect is that, the expected value in (4.2) must be considered a second time and must be specified with more details. Assuming that α^* denotes the general random animal effect on the observed body weight. For a given animal i, the effect is then α_i which is a realized but unobservable value of the distribution of the α^* effects. Therefore in (4.2) the expected value of y_{ij} is conditional on the fact that the random variable α^* takes the value α_i. Hence (4.2) is a conditional mean

$$E(y_{ij}|\alpha^* = \alpha_i) = \mu + \alpha_i \tag{4.3}$$

For notational simplicity, the α^* is often ommitted. Taking expectation over α^* leads to

$$E_{\alpha^*}\left[E(y_{ij}|\alpha_i)\right] = E(y_{ij}) = \mu \tag{4.4}$$

The residuals are defined as

$$e_{ij} = y_{ij} - E(y_{ij}|\alpha_i) = y_{ij} - (\mu + \alpha_i) \tag{4.5}$$

With that definition, we can establish the model equation for an observation y_{ij} as

$$y_{ij} = \mu + \alpha_i + e_{ij} \tag{4.6}$$

The properties of the residuals are assumed analogously to the fixed effects model. In summary, the properties are listed as

- the expected value of the residuals are all 0, $E(e_{ij}) = 0$
- the variances of the residuals are all equal to σ_e^2, $var(e_{ij}) = E(e_{ij}^2) = \sigma_e^2$
- all residuals are independent, $cov(e_{ij}, e_{i'j'}) = 0 \quad \forall i, i'$ and $\forall j, j'$ except $i = i'$ and $j = j'$
- residuals are independen of α_i effects, $cov(e_{ij}, \alpha_k) = 0 \quad \forall i, j, k$

Together with (4.6), we can establish the total variance of all observations y_{ij} as

$$var(y_{ij}) = var(\mu + \alpha_i + e_{ij}) = \sigma_{\alpha}^2 + \sigma_e^2 = \sigma_y^2 \tag{4.7}$$

This shows that the variance (σ_y^2) can be decomposed into the two variance components σ_{α}^2 and σ_e^2. It is also noted that the intra-class covariance which corresponds to the covariance between body weights for the same animal can be written as

$$cov(y_{ij}, y_{ij'}) = cov(\mu + \alpha_i + e_{ij}, \mu + \alpha_i + e_{ij'}) = \sigma_{\alpha}^2 \quad \text{for } j \neq j' \tag{4.8}$$

4.1.3.1 Package lme4

In R, one of the packages that can handle random effects models is the package `lme4`. For the dataset in Table 4.4, this can be done as follows

```
library(lme4)
lmer_bw_rep <- lmer(`Body Weight` ~ (1 | Animal), data = tbl_rep_obs_no_breed)
summary(lmer_bw_rep)
```

```
## Linear mixed model fit by REML ['lmerMod']
## Formula: `Body Weight` ~ (1 | Animal)
##    Data: tbl_rep_obs_no_breed
##
## REML criterion at convergence: 82.7
##
## Scaled residuals:
##      Min       1Q   Median       3Q      Max
## -1.36360 -0.50301  0.06086  0.26850  1.43838
##
## Random effects:
##  Groups   Name        Variance Std.Dev.
##  Animal   (Intercept) 954.34   30.892
##  Residual              22.98    4.794
## Number of obs: 12, groups:  Animal, 4
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   503.39      15.51   32.46
```

4.1.4 Mixed Linear Effects Models

A fixed general mean μ or a fixed intercept term b_0 and random residual term e occur in almost all models that were presented so far. Apart from these, all other effects were either all fixed or random[1]. We now consider models where some effects (other than μ and e) are fixed and some are random. Such models are called **mixed linear effects models**[2].

An example dataset which could be analysed with a mixed linear effects model would be, if we would add to each animal in our reference dataset on body weight, breast circumference and breed also the sire of each animal. If some of these animals would share the same sire and hence would be half sibs, the dataset would again as already seen in the repeated observations data, a specific variance structure. This is due to the fact that body weights from half sibs would be expected to be more similar than observations from unrelated animals.

Table 4.5: Body Weight, Breast Circumference, Breed and Sire of Beef Cattle Animals

Animal	Body Weight	Breast Circumference	Breed	Sire
1	471	176	Angus	S1
2	463	177	Angus	S1
3	481	178	Simmental	S3
4	470	179	Angus	S2
5	496	179	Simmental	S3
6	491	180	Simmental	S4
7	518	181	Limousin	S5
8	511	182	Limousin	S5
9	510	183	Limousin	S6
10	541	184	Limousin	S6

When fitting a mixed linear effects model to a dataset as shown in Table 4.5, the question is which effects should be taken as fixed and which should be considered to be random. As already mentioned in this case, `Breast Circumference` and `Breed` would be modelled as fixed effects and `Sire` would be modelled as a random effect. In general, there are not strict rules that would tell us which effects should be modelled as fixed effects an which ones should be considered as random. In our dataset we can certainly say that for `Breast Circumference` and `Breed` we are interested in the effect sizes of the values that are observed in the given datasets. In contrasts to that, we can say that the included sires

[1] Except for a small introduction into repeated measures models, we have not really look at random models in great detail. But they are not of great importance to the treatment of mixed models.

[2] Sometimes these models are just called mixed models. We are using these terms interchangably

are a random sample of a larger population of sires. Furthermore, the primary interest in the sire effects are in the imposed covariance structure of the data due to the sire effects. In the case where the primary interest is in the variance imposed by a certain effect, then the respective effect has to be modelled as a random effect.

The general mixed effects model can be written as

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{Z}\mathbf{u} + \mathbf{e} \tag{4.9}$$

where $\mathbf{y}$ is the vector of observations, $\mathbf{b}$ is the vector of fixed effects, $\mathbf{u}$ is the vector of random effects, $\mathbf{X}$ and $\mathbf{Z}$ are incidence matrices and $\mathbf{e}$ is the vector of random residuals. The random effects are assumed to have expected values of zero and given specific variance-covariance matrices. Hence we can write

$$E\left[\begin{array}{c}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{array}\right] = \left[\begin{array}{c}\mathbf{Xb}\\ \mathbf{0}\\ \mathbf{0}\end{array}\right] \tag{4.10}$$

The variance-covariance matrices are specified as

$$var\left[\begin{array}{c}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{array}\right] = \left[\begin{array}{ccc}\mathbf{ZDZ^T} + \mathbf{R} & \mathbf{ZD} & \mathbf{R}\\ \mathbf{DZ^T} & \mathbf{D} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{array}\right] \tag{4.11}$$

with $var(\mathbf{u}) = E(\mathbf{uu^T}) = \mathbf{D}$ and $var(\mathbf{e}) = E(\mathbf{ee^T}) = R$.

Assuming $\mathbf{V}$ is not singular, the normal equations stemming from the generalized least squares are

$$\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\mathbf{b}^0 = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{4.12}$$

with a solution

$$\mathbf{b}^0 = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} \tag{4.13}$$

From that solution, we can get estimates of estimable functions for the fixed effects as previously discussed for fixed models.

For the random effects $\mathbf{u}$, the conditional expectation of $\mathbf{u}$ given the observations $\mathbf{y}$ are of particular interest as estimators. Assuming multivariate normality for $\mathbf{u}$ and $\mathbf{e}$, we can write

$$\begin{aligned}\hat{\mathbf{u}} = E(\mathbf{u}|\mathbf{y}) &= E(\mathbf{u}) + cov(\mathbf{u}, \mathbf{y}^T)(var(\mathbf{y}))^{-1}(\mathbf{y} - E(\mathbf{y}))\\ &= \mathbf{DZ}^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{Xb}) \qquad (4.14)\end{aligned}$$

Both terms, the solution for $\mathbf{b}^0$ and the estimate $\hat{\mathbf{u}}$ depend on the inverse matrix $\mathbf{V}^{-1}$ which can be extremely large and difficult to compute. In different publications, the research group of Charles Henderson has shown that solving the following system of equations leads to the same estimates for both the fixed and the random effects. This system of equations is called **Mixed Model Equations** and is shown below.

$$\left[\begin{array}{cc} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{D}^{-1} \end{array}\right] \left[\begin{array}{c} \hat{\mathbf{b}} \\ \hat{\mathbf{u}} \end{array}\right] = \left[\begin{array}{c} \mathbf{X}^T\mathbf{R}^{-1}\mathbf{y} \\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{y} \end{array}\right] \tag{4.15}$$

4.2 Pedigree BLUP

The linear mixed effects models as shown above can be applied to datasets in livestock breeding. In such a model, the response variable y corresponds to measurements or observations of phenotypic traits. The vector of fixed effects b contains all information about the known environment such as `Breed`, `Herd`, `Season`, `Age` and possibly other predictors that have an influence on the response. The random effects u contain the breeding values of animals of interest in our livestock breeding population. Once all the informations of the data are collected, it can be transfered into model components. The model components are then used to construct the mixed model equations. Solutions to these equations provide estimates of fixed effects and predictions of breeding values. Properties of the predicted breeding values can be summarized as

- Best: the predicted breeding values have minimum prediction error variance
- Linear: the predicted breeding values are linear functions of the data
- Unbiased: the expected value of the predicted breeding values is equal to the expected value of the true breeding value
- Prediction: because breeding values cannot be observed, the results are called predictions.

The above listed properties are often abbreviated as BLUP.

The application of linear mixed effects models to livestock breeding datasets can be done in two different ways.

1. Sire model: only sires in the dataset get breeding values
2. Animal model: all animals in a datasets (also parents without observations) get breeding values

4.2.1 Sire Model

In a sire model the vector $\mathbf{u}$ of random effects contains all sires in the dataset. For the example data shown in Table 4.5, this corresponds to

$$\mathbf{u} = \begin{bmatrix} S1 \\ S1 \\ S3 \\ S2 \\ S3 \\ S4 \\ S5 \\ S5 \\ S6 \\ S6 \end{bmatrix}$$

Because the sire breeding values ($\mathbf{u}$) are random effects, we also have to specify the expected value and the variance-covariance matrix of $\mathbf{u}$. Because breeding values are defined as deviations, the expected values of the sire breeding values are zero. Hence

$$E(\mathbf{u}) = \mathbf{0} \tag{4.16}$$

$$var(\mathbf{u}) = \mathbf{D} \tag{4.17}$$

with $\mathbf{D}$ beeing the variance-covariance matrix between the sire breeding values. If the sires are not related, then $\mathbf{D} = \sigma_s^2\ I$ where σ_s^2 is a sire variance component. If the sires are related then $\mathbf{D} = \sigma_s^2\ \mathbf{A}_s$ where $\mathbf{A}_s$ is the sire relationship matrix containing elements of probabilities of sharing allels based on identity by descent between related sires as off-diagonal elements. The diagonal elements of $\mathbf{A}_s$ are all one.

For the moment, we assume that the variance component such as σ_s^2 are all given. In reality, such components would also need to be estimated from the data. The discussion on how to estimate variance components from the data is deferred to a later chapter.

4.2.2 Animal Model

The major difference between the sire model and the animal model is that in the animal model all animals in the dataset receive breeding values. Hence in the dataset shown in Table 4.5, we would need to add the dams.

Table 4.6: Body Weight, Breast Circumference, Breed, Sire and Dam of Beef Cattle Animals

Animal	Body Weight	Breast Circumference	Breed	Sire	Dam
1	471	176	Angus	S1	D1
2	463	177	Angus	S1	D2
3	481	178	Simmental	S3	D3
4	470	179	Angus	S2	D2
5	496	179	Simmental	S3	D3
6	491	180	Simmental	S4	D4
7	518	181	Limousin	S5	D5
8	511	182	Limousin	S5	D5
9	510	183	Limousin	S6	D6
10	541	184	Limousin	S6	D7

The vector **u** contains breeding values for all animals in the dataset, also from parents that do not have observations. Hence

$$\mathbf{u} = \begin{bmatrix} S1 \\ S2 \\ ... \\ D1 \\ D2 \\ ... \\ 1 \\ 2 \\ ... \\ 10 \end{bmatrix}$$

The expected value and the variance-covariance matrix of **u** are defined as

$$E(\mathbf{u}) = \mathbf{0} \tag{4.18}$$
$$var(\mathbf{u}) = \mathbf{D} = \mathbf{A}\sigma_u^2 \tag{4.19}$$

where the matrix **A** corresponds to the numerator relationship matrix. This matrix contains the probabilities of two animals sharing alleles identical by descent on the off-diagonal elements. The diagonal elements of **A** are computed as one plus the inbreeding coefficient of an animal. The inbreeding coefficient of an animal is given by half of the relationship coefficient of the parents.

4.3 Genomic BLUP

The term **genomic BLUP** is used for the use of genomic information together with pedigree data and phenotypic observations to predict breeding values. Hence the goal is the same as with the pedigree-based BLUP animal model. The main difference is just in the information that goes into the model. But otherwise, the internal modelling mechanisms are the same as before.

The prediction of genomic breeding values which consists of objective of genomic BLUP can be done in two ways. The two ways are

1. marker effect model
2. breeding value based model

4.3.1 Marker Effect Model

When using marker effect models to predict genomic breeding values, this is done in two steps. In a first step marker effects are estimated from a reference population. In that reference population all animals have a complete set of marker genotypes as well as phenotypic observations of the trait of intertest. In a second step the estimated marker effects ($\hat{\mathbf{q}}^T = \left[\begin{array}{cccc} \hat{q}_1 & \hat{q}_2 & ... & \hat{q}_k \end{array}\right]$) are used to predict genomic breeding values for any animal that has genomic information in the form of marker genotypes available.

In Figure 4.3 the principle of the two step procedure to predict genomic breeding values is shown. A possible linear model to estimate SNP-marker-effects based on the data from the reference population can be defined as follows

$$y = Xb + Mq + e \tag{4.20}$$

where	m	number of SNP markers
	y	vector of observations
	b	vector of fixed effects
	X	design matrix linking fixed effects to observations
	q	random genetic effect of SNP-marker-genotypes
	M	design matrix linking SNP-genotype effects to observations
	e	vector of random residuals

The mixed-model equations resulting from models given in (4.20) have the following structure

$$\left[\begin{array}{cc} X^TR^{-1}X & X^TR^{-1}M \\ M^TR^{-1}X & M^TR^{-1}M + I * \lambda \end{array}\right] \left[\begin{array}{c} \hat{b} \\ \hat{q} \end{array}\right] = \left[\begin{array}{c} X^TR^{-1}y \\ M^TR^{-1}y \end{array}\right] \tag{4.21}$$

where

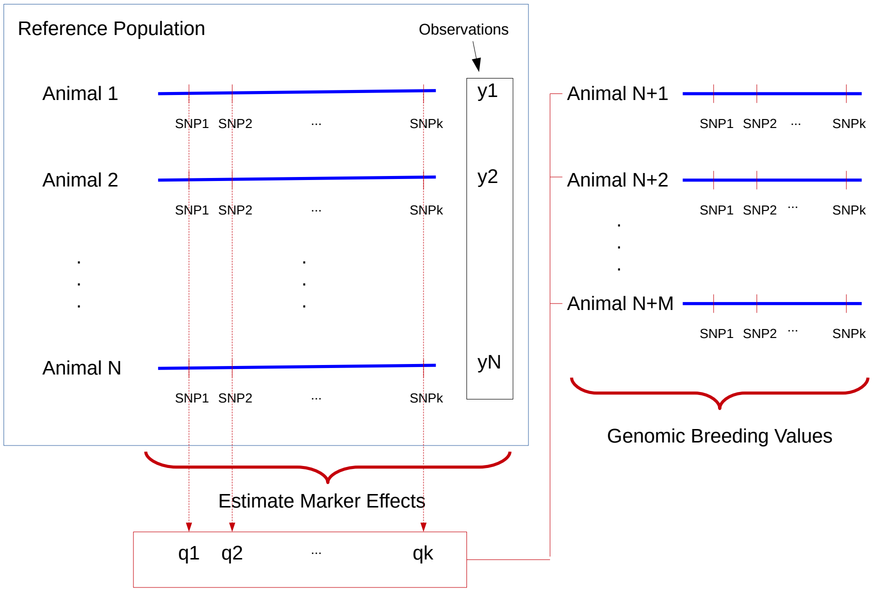

Figure 4.3: Principle of Two-Step Genomic Prediction of Breeding Values

$$\lambda = \frac{\sigma_e^2}{\sigma_q^2} \tag{4.22}$$

In (4.22) σ_q^2 is the total genetic variance explained by the given markers in the dataset

The solutions for $\hat{q}$ from (4.21) correspond to the SNP-genotype effects. The predicted breeding value $\hat{u}$ for any selection candidate i with genomic information is then computed as

$$\hat{u}_i = M_i \cdot \hat{q} \tag{4.23}$$

where M_i corresponds to the vector of SNP-genotypes of selection candidate i.

4.3.2 Breeding Value Based Model

The use of breeding value based models to predict genomic breeding values is also known as **single-step** prediction of genomic breeding values. As the term single-step already alludes to, with this method genomic breeding values are predicted directly from the data. This is done by directly integrating genomic breeding values into the mixed linear effects model where the random effects in the model are the genomic breeding values.

When only looking at the model for predicting genomic breeding values, it looks similar to the pedigree-based animal model as shown below.

$$y = Xb + Zu + e \tag{4.24}$$

The mixed model equations to get solutions used for estimates of fixed effects and predicted genomic breeding values can be written as

$$\left[\begin{array}{cc} X^TR^{-1}X & X^TR^{-1}Z \\ Z^TR^{-1}X & Z^TR^{-1}Z + D^{-1} \end{array}\right] \left[\begin{array}{c} \hat{b} \\ \hat{u} \end{array}\right] = \left[\begin{array}{c} X^TR^{-1}y \\ Z^TR^{-1}y \end{array}\right] \tag{4.25}$$

The difference to the pedigree-based animal model is the matrix D which depended on the numerator realtionship matrix A for the animal model. In single-step genomic BlUP, the matrix D corresponds to

$$D = G * \sigma_u^2$$

where G corresponds to the genomic relationship matrix and σ_u^2 is taken to be the genetic-additive variance. How the matrix G is constructed is shown in the next section.

4.4 Genomic Relationship Matrix

The variance-covariance matrix between the genetic effects u in model (4.24) 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.

4.4.1 Derivation of G

Because the traditional pedigree-based BLUP animal model is very well respected in animal breeding and the defined model (4.24) produces an analogy of the genomic evaluation model to the already known animal model the following properties of u and the genomic relationship matrix G are essential.

1. The genomic breeding values u should correspond to a linear combination of the single SNP-effects q
2. The genomic breeding values u should be defined as deviations from a common mean, leading to the expected value $E\left[u\right] = 0$.

3. The variance-covariance matrix of the vector u corresponds to the product of G times a common variance component σ_u^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.

4.4.2 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 u should be a linear combination of the effects in q means that there exists a matrix U for which we can write

$$u = U \cdot q \tag{4.26}$$

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

4.4.3 Deviation

The genomic breeding values u should be defined as deviation from a common basis. Due to this definition the expected value of the genetic effect is determined by $E\left[u\right] = 0$. 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 M in the marker effect model. Let us further assume that the SNP loci are in Hardy-Weinberg equilibrium. Then w can take the following values

$$w = \left\{\begin{array}{lll} -1 & \text{with probability} & (1-p)^2 \\ 0 & \text{with probability} & 2p(1-p) \\ 1 & \text{with probability} & p^2 \end{array}\right. \tag{4.27}$$

The expected value of w corresponds to

$$E\left[w\right] = (-1) * (1-p)^2 + 0 * 2p(1-p) + 1 * p^2 = -1 + 2p - p^2 + p^2 = 2p - 1 \tag{4.28}$$

The matrix U is computed as the difference between the matrix M 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

$$\begin{aligned} E\left[u\right]_j &= \left[(1-p_j)^2 * (-2p_j) + 2p_j(1-p_j)(1-2p_j) + p_j^2(2-2p_j)\right] q_j \\ &= 0 \end{aligned} \tag{4.29}$$

4.4.4 Variance of Genomic Breeding Values

As already postulated the variance-covariance matrix of the genomic breeding values should be proportional to the genomic relationship matrix G.

$$var(u) = G * \sigma_u^2 \tag{4.30}$$

Computing the same variance-covariance matrix based on equation (4.26)

$$var(u) = U \cdot var(q) \cdot U^T \tag{4.31}$$

The variance-covariance matrix of the SNP effects is $var(q) = I * \sigma_q^2$. Inserting this into (4.31) we get $var(u) = UU^T\sigma_q^2$.

In [Gianola et al., 2009] the variance component σ_u^2 was derived from σ_q^2 leading to

$$\sigma_u^2 = 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 \tag{4.32}$$

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

$$var(u) = G * \sigma_u^2 = UU^T\sigma_q^2 \tag{4.33}$$

In (4.33), σ_u^2 is replaced by the result of (4.32).

$$G * 2\sum_{j=1}^{m} p_j(1-p_j)\sigma_q^2 = UU^T\sigma_q^2 \tag{4.34}$$

Dividing both sides of (4.34) by σ_q^2 and solving for G gives us a formula for the genomic relationship matrix G

$$G = \frac{UU^T}{2\sum_{j=1}^{m} p_j(1-p_j)} \tag{4.35}$$

4.5 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 (4.24) is a mixed linear effects model. The solution for the unknown parameters can be obtained by solving the mixed model equations shown in (4.36). In this form the Inverse G^{-1} of G and the vector $\hat{u}$ 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.

$$\begin{bmatrix} X^TX & X^TZ & 0 \\ Z^TX & Z^TZ + G^{(11)} & G^{(12)} \\ 0 & G^{(21)} & G^{(22)} \end{bmatrix} \begin{bmatrix} \hat{b} \\ \hat{u}_1 \\ \hat{u}_2 \end{bmatrix} = \begin{bmatrix} X^Ty \\ Z^Ty \\ 0 \end{bmatrix} \tag{4.36}$$

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 $\hat{u}_1$ contains the predicted breeding values for the animals with observations and the vector $\hat{u}_2$ contains the predicted breeding values of all animals without phenotypic observations.

Based on the last line of (4.36) the predicted breeding values $\hat{u}_2$ of all animals without phenotypic observations can be computed from the predicted breeding values $\hat{u}_1$ from the animals with observations.

$$\hat{u}_2 = -\left(G^{22}\right)^{-1} G^{21}\hat{u}_1 \tag{4.37}$$

Equation (4.37) is referred to as genomic regression of predicted breeding values of animals without observation on the predicted genomic breeding values of animals with observations.

4.6 Single Step Genomic BLUP With Real-World Datasets

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

$$y = Xb + Zu + e \tag{4.38}$$

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

$$u = \left[\begin{array}{c} u_1 \\ u_2 \end{array}\right]$$

where u_1 is the vector of breeding values for non-genotyped animals and u_2 is the vector of genotyped animals.

$$\left[\begin{array}{cc} X^TX & X^TZ \\ Z^TX & Z^TZ + \lambda * H^{-1} \end{array}\right] \left[\begin{array}{c} \hat{b} \\ \hat{u} \end{array}\right] = \left[\begin{array}{c} X^Ty \\ Z^Ty \end{array}\right] \tag{4.39}$$

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

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

$$H^{-1} = A^{-1} + \left(\begin{array}{cc} 0 & 0 \\ 0 & G^{-1} - A_{22}^{-1} \end{array}\right)$$

where A^{-1} is the inverse numerator relationship matrix and A_22 corresponds to the part of the numerator relationship matrix containing all genotyped animals.